Theor Chem Acc (2001)
DOI 10.1007/s002140100290

Theoretical
Chemistry Accounts

*Regular article*

# The performance of the rapid estimation of basis set error and correlation energy from partial charges method on new molecules of the G3/99 test set

**Sándor Kristyán**[1], **Adrienn Ruzsinszky**[2], **Gábor I. Csonka**[2]

[1] Chemical Research Center, Institute of Chemistry, Hungarian Academy of Sciences, Pusztaszeri út 59–67, 1025 Budapest, Hungary
[2] Department of Inorganic Chemistry, Budapest University of Technology, 1521 Budapest, Hungary

Received: 10 March 2001 / Accepted: 5 July 2001 / Published online: 11 October 2001


**Abstract.** Experimental enthalpies of formation have been approximated using single-point Hartree–Fock (HF)–self-consistent-field (SCF) total energies plus the rapid estimation of basis set error and correlation energy from partial charges (REBECEP) energy corrections. The energy corrections are calculated from the HF–SCF partial atomic charges and optimized atomic energy parameters. The performance of the method was tested on 51 closed-shell neutral molecules (50 molecules from the G3/99 thermochemistry database plus urea, composed of H, C, N, O, and F atoms). The predictive force of the method is demonstrated, because these larger molecules were not used for the optimization of the atomic parameters. We used the earlier RECEP-3 [HF/6-311+G(2d,p)] and REBECEP [HF/6-31G(d)] atomic parameter sets obtained from the G2/97 thermochemistry database (containing small molecules) together with natural population analysis and Mulliken partial charges. The best results were obtained using the natural population analysis charges, although the Mulliken charges also provide useful results. The root-mean-square deviations from the experimental enthalpies of formation for the selected 51 molecules are 1.15, 3.96, and 2.92 kcal/mol for Gaussian-3, B3LYP/6-11+G(3df,2p), and REBECEP (natural population analysis) enthalpies of formation, respectively (the corresponding average absolute deviations are 0.94, 7.09, and 2.27 kcal/mol, respectively). The REBECEP method performs considerably better for the 51 test molecules with a moderate 6-31G(d) basis set than the B3LYP method with a large 6-311+G(3df,2p) basis set.

**Key words:** Enthalpies of formation – Energy parameters – Partial charges – G3 method – B3LYP method

*Correspondence to*: kristyans@chemres.hu
csonka@web.inc.bme.hu

## 1 Introduction

We present a continuation of our previous work on the development and application of the rapid estimation of basis set error and correlation energy from partial charges (REBECEP) method [1–4]. According to this method an approximation to the basis set error and correlation energy is estimated from the Hartree–Fock (HF)–self-consistent-field (SCF) partial charges using atomic energy parameters. A method- and basis-set-dependent energy correction was defined for our purposes:

$$E_{\mathrm{corr}}(\text{method, basis set}) = E_{\mathrm{T}}(\text{method}) - E_{\mathrm{T}}(\text{HF–SCF/basis set}) \ , \quad (1)$$

where $E_{\mathrm{T}}$(method) is a total energy obtained from a good-quality (usually composite) method (e.g. Gaussian-2, [2,5] Gaussian-3, [3,6], or experiment, [4] vide infra), and $E_{\mathrm{T}}$(HF–SCF/basis set) is a HF–SCF total energy calculated with a given basis set [6-311+G(2d,p) and 6-31G(d) basis sets were tested by us previously]. If ''method'' is a complete configuration interaction and the ''basis set'' is infinite, one obtains the exact correlation energy at the given molecular geometry. The geometries were calculated at the second order Møller–Plesset level with the 6-31G(d) basis set in order to compare our results to the earlier results; however, any good-quality molecular geometry in the vicinity of the true molecular geometry could be used.

The advantage of the REBECEP method is that the very expensive $E_{\mathrm{T}}$(method) calculation in Eq. (1) is not necessary because $E_{\mathrm{corr}}$(method, basis set) is estimated from atomic charges for closed-shell ground-state covalent molecules in the vicinity of stationary points (the geometry optimization is not a part of the current method) by an inexpensive atom-by-atom method:

$$E_{\mathrm{corr}}(\text{REBECEP, method, basis set, charge def.}) \equiv \sum_{A \in M} E_{\mathrm{corr}}(N_{\mathrm{A}}, Z_{\mathrm{A}}, \text{method, basis set, charge def.}) \ , \quad (2)$$

where $E_{corr}$(REBECEP, method, basis set, charge def.) approximates $E_{corr}$(method, basis set) in Eq. (1) using one of the atomic charge definitions. The REBECEP energy is the sum of the $M$ atomic correlation energies $E_{corr}(N_A, Z_A$, method, basis set, charge def.). $N_A$ is the "electron content" on atom A, generally noninteger and defined as ($Z_A$–partial charge), where $Z_A$ is the nuclear charge of atom A. Four partial charge definitions were tested (natural population analysis [7], ChelpG [8], Merz–Kollman [9], and Mulliken) in our previous work.

The $E_{corr}(N_A, Z_A$, method, basis set, charge def.) atomic energy terms of Eq. (2) are interpolated:

$$\begin{aligned} &E_{corr}(N_A, Z_A, \text{method, basis set, charge def.}) \\ &\quad = (N_A - N_1) E_{par}(N_2, Z_A, \text{method, basis set, charge def.}) \\ &\qquad + (N_2 - N_A) E_{par}(N_1, Z_A, \text{method, basis set, charge def.}) \ , \end{aligned} \tag{3}$$

where $N_1$ and $N_2$ are integer numbers of electrons, with $N_1 \le N_A \le N_2 = N_1 + 1$, and $N_A$ is the electron content around atom A. $E_{par}(N_2, Z_A$, method, basis set, charge def.) and $E_{par}(N_1, Z_A$, method, basis set, charge def.) in Eq. (3) are the so-called REBECEP atomic parameters that transform the partial charge into energy correction. For hydrogen atoms, we suggest using a single parameter, $E_{corr}(N_A$, 1, method, basis set, charge def.) = $N_A E_{par}$(2, 1, method, basis set, charge def.)/2.

Then the total energy of a method is approximated as follows:

$$\begin{aligned} &E_T(\text{method}) \\ &\quad \approx E_{corr}(\text{REBECEP, method, basis set, charge def.}) \\ &\qquad + E_T(\text{HF–SCF/basis set}) \ . \end{aligned} \tag{4}$$

In this way one can predict the G2 or G3 total energy and experimental thermochemistry results from a single-point HF–SCF/basis set total energy using the partial charges and the corresponding REBECEP parameters in terms of Eq. (3). The most important question in this respect is how to find the suitable atomic parameters. The REBECEP method relies on two basic physical properties: the total energy of a molecule in a ground state can be exactly partitioned into atomic parts and the correlation energy depends on the number of electrons (the electron spin plays a special role in this respect; however, the present investigation is limited to singlet closed-shell systems). In the actual calculations two mathematical problems arise: how to calculate the atomic electron number in a molecular environment and how to deal with the basis set errors. The rapid estimation of correlation energy from partial charges (RECEP) and REBECEP methods have been described in detail elsewhere [2–4]. We briefly review our earlier results here.

In Ref. [1] we tested a priori correlation energy parameters derived from the solution of the Schrödinger equation or from large basis set B3LYP [10–12] calculations for atoms. The results showed that there is no straightforward way to obtain enthalpies of formation within chemical accuracy (±1 kcal/mol) using such a priori parameters [2]. One of the reasons for this is that atomic correlation energies derived from high-spin states are not readily transferable into a low-spin molecular environment. Also various charge definitions would provide quantitatively different results.

Next, fitted correlation energy parameters were used to obtain the best approximation (in the least-squares sense) to $E_{corr}$(G2) for 41 closed-shell neutral molecules of the G2/97 thermochemistry database composed of H, C, N, O, and F atoms [13]. In the fitting procedure we used the energy corrections obtained from the G2 [5] and HF–SCF/6-311+G(2d,p) total energies [Eq. 1, method = G2 and basis set = 6-311+G(2d,p)]. The four partial charge definitions mentioned previously were tested and we obtained four slightly different atomic parameter sets [2]. The G2 total energies were approximated with 1.9 kcal/mol average absolute deviation from a single-point HF–SCF/6-311+G(2d,p) calculation [2]. The basis set error of the HF–SCF/6-311+G(2d,p) energy was compared to the HF energy extrapolated to infinite basis set. It was observed that the relatively small basis set error was effectively compensated by the RECEP-2 parameter set [2].

The more recent G3 method shows a considerably better performance [6] than the G2 method; thus, we reparameterized our method in order to obtain the RECEP-3 parameter set and estimated $E_{corr}$(G3) for 65 closed-shell neutral molecules (again composed of H, C, N, O, and F atoms) of the G2/97 thermochemistry database form single-point HF–SCF/6-311+G(2d,p) calculations [3]. Recently, we tested the possibility to approximate the G3 results from single-point HF–SCF/6-31G(d) calculations for the same set of 65 molecules and obtained a new RECEP-4 parameter set [4]. (Note that considerable basis set error is incorporated into these atomic parameters and this parameter set is also called REBECEP.) In this way the speed of the method was increased considerably (e.g. by more than 1 order of magnitude in the case of aniline) without loss of precision. The RECEP-3 and REBECEP parameter sets yielded rather good approximations to the G3 total energies (Eq. 4). The enthalpies of formation were obtained after the usual procedure (zero-point vibration, and thermal corrections, using atomic energies and enthalpies). The details are given elsewhere [4]. The root-mean-square deviation of the RECEP-3 total energies and enthalpies of formation from the G3 total energies and from the experimental enthalpies of formation for the 65 molecular energies were 1.76 and 2.17 kcal/mol, respectively (the corresponding average absolute deviations were 1.43 and 1.75 kcal/mol, respectively). In order to eliminate the error of the G3 method, we fitted the REBECEP parameter set to the experimental enthalpies of formation as well. The average absolute deviation of the best REBECEP enthalpies of formation from the experimental enthalpies of formation (derived from natural population analysis charges) was 1.39 kcal/mol for the test set of 65 neutral enthalpies of formation.

Recently, the G2/97 thermochemistry database was extended and relatively large molecules were included; thus, currently $E_T$(G3) is available and was tested against 376 test energies (222 neutral enthalpies of formation, 88 ionization potentials, 58 electron affinities, and eight proton affinities) in the full G3/99 test set [14]. This provides a good opportunity to test the

performance of our RECEP-3 and REBECEP parameter sets (reported in Refs. [3, 4]) especially because these molecules were not used during the parameterization of our RECEP method.

## 2 Selection of the test molecules and molecular geometries

The criterion for choosing the molecules for the G3-3 subset of the G3/99 test set was that their experimental enthalpies of formation at 298 K have a quoted uncertainty of $\pm 1$ kcal or less [14], although this is not necessarily a guarantee for the accuracy of the experimental data. All the new molecules contain three or more non-hydrogen atoms. The largest contains ten non-hydrogen atoms. As many as 75 new molecules define the G3-3 subset containing 13 molecules without hydrogen, 16 hydrocarbons, 44 substituted hydrocarbons, and two radicals. We selected 50 molecules from the G3-3 subset of the G3/99 test set listed in Table 1. The corresponding experimental enthalpies of formation and the quoted experimental errors are shown in Table 2. All the selected molecules are closed-shell neutrals composed of H, C, N, O, and F atoms. We excluded the two radicals, because the correlation energy for radicals is rather different from the correlation energy of closed-shell molecules; thus, it requires different considerations [1]. The molecules of the G3-3 subset containing second-row atoms were excluded because of the problems with basis sets, atomic enthalpies of formation, relativistic effects, and the relatively poor MP2 molecular geometries. To illustrate the problem, we mention that accurate results for second-row compounds can be achieved only if high-exponent d and f functions are added to the basis set [15–17]. These so-called ''inner-shell polarization functions'' cause a HF–SCF-level effect, and it has little to do with inner-shell correlation. It dwarfs the latter in importance because contributions as high as 10 kcal/mol were reported [17]. We also note that for Na–Ar the 6-311G(d) basis was defined in a nonuniform manner across the row. To avoid these difficulties the uniformly defined 6-31G(d) basis was used in the G3 method. Furthermore, the calculation of the molecular enthalpy of formation requires precise atomic enthalpies of formation; however, several of the atomic enthalpies of formation (B, Al, Si) have large uncertainties (for further details see Ref. [14]). It is also known that the relativistic effects are important and nonnegligible for second-row elements; however, if spin-orbit corrections can be sufficiently treated at the atomic level – as was done in G3 theory [6] – it would fit well (i.e. it could be incorporated) into our REBECEP method. This possibility will be tested in future work. We added urea (no. 25 in Table 1) to the 50 molecules selected and used these 51 molecules to test the RECEP-3 and REBECEP parameters (Table 2). The experimental value for urea quoted in the Webbook [18] differs from the value of Pedley et al. (–58.7 kcal/mol) [19]; however, our detailed investigation and the value of Frenkel et al. (–56.67 kcal/mol) [20] supports the value in the Webbook database (Table 2, although the experimental error is probably larger than the quoted 0.3 kcal/mol). All the selected molecules contain at least 30 electrons and the largest contains 68 electrons or ten non-hydrogen atoms (Table 1).

Finally, we mention that our previous RECEP-3 and REBECEP parameter sets were optimized for atomic partial charges that were calculated at the MP2(full)/6-31G(d) equilibrium geometries. For the sake of consistency, we must use the same type of geometries in the current test study. Using different geometries would certainly influence the HF–SCF total energies and the partial charges calculated, leading to incomparable results; however, we admit that such expensive geometries are certainly not the optimal choice for future REBECEP studies. Finding a geometry optimization method that is fast enough to use with the REBECEP method falls outside the scope of this work.

## 3 Results and discussion

The G3 energy values and the approximate G3 energy corrections of the 51 molecules selected are shown in Table 1 [Eq. 1 with ''method'' is G3 and ''basis set'' is 6-31G(d) or 6-311+G(2d,p)]. The Gaussian 98 [21] program was used for all the calculations. Inspection of the approximate energy correction values defined in Eq. (1) clearly shows that the HF/6-31G(d) energies require considerably larger corrections in absolute value than the HF/6-311+G(2d,p) energies. The differences between these energies (whose origin is obviously the basis set error) are compensated by the differences in the RECEP-3 and REBECEP correlation energy parameters (the latter being more negative [3, 4]); thus, Eq. (3) provides an estimation of the same energy.

The experimental enthalpies of formation at 298 K, the experimental errors, and the deviations between experimental and calculated G3 and B3LYP/6-311+G(3df,2p) enthalpies of formation are shown in Table 2 for comparison. The inspection of the experimental errors quoted shows that these errors are rather small for the 51 molecules selected (usually less than $\pm 0.4$ kcal/mol), and only three compounds have a quoted error of $\pm 0.8$ kcal/mol (azulene, benzoqinone, and $C_2F_6$). The G3 values approximate the experimental values quite well; however, the B3LYP values show considerable error, as shown in Tables 2 and 3. The distribution of the errors is shown in Fig. 1. The B3LYP method provides considerably more positive enthalpies of formation than experiment. The average deviation between the experimental and the B3LYP values is $-6.82$ kcal/mol (Table 3). The value of the average deviation for the G3 method is 0.04 kcal/mol.

The deviations of our RECEP enthalpies of formation (Eq. 4) from the experimental enthalpies of formation are also shown in Table 2. First, we show the results calculated from HF–SCF/6-311+G(2d,p) total energies, natural population analysis partial charges, and the corresponding RECEP-3 parameter set optimized to reproduce the experimental total energies [4] (Eq. 4, method = experiment., charge definition = natural

**Table 1.** Number of atoms and electrons in 51 test molecules, their G3 total energies (hartree) and the energy corrections (hartree) of Eq. (1) calculated with two different basis sets

| | Species | Number of atoms | Number of electrons | E(G3) | $E_{corr}$ 6-311+G(2d,p) | $E_{corr}$ 6-31G(d) |
|---|---|---|---|---|---|---|
| 1 | $C_4H_6$ (methylallene) | 10 | 30 | −155.9081 | −0.9629 | −1.0098 |
| 2 | $C_5H_8$ (isoprene) | 13 | 38 | −195.2311 | −1.2166 | −1.2755 |
| 3 | $C_5H_{10}$ (cyclopentane) | 15 | 40 | −196.4769 | −1.2611 | −1.3140 |
| 4 | $C_5H_{12}$ (*n*-pentane) | 17 | 42 | −197.6891 | −1.2990 | −1.3566 |
| 5 | $C_5H_{12}$ (neopentane) | 17 | 42 | −197.6963 | −1.3062 | −1.3632 |
| 6 | $C_6H_8$ (1,3-cyclohexadiene) | 14 | 44 | −233.3246 | −1.4309 | −1.4950 |
| 7 | $C_6H_8$ (1,4-cyclohexadiene) | 14 | 44 | −233.3242 | −1.4271 | −1.4923 |
| 8 | $C_6H_{12}$ (cyclohexane) | 18 | 48 | −235.7858 | −1.5153 | −1.5786 |
| 9 | $C_6H_{14}$ (*n*-hexane) | 20 | 50 | −236.9874 | −1.5525 | −1.6203 |
| 10 | $C_6H_{14}$ (3-methylpentane) | 20 | 50 | −236.9884 | −1.5573 | −1.6250 |
| 11 | $C_6H_5CH_3$ (toluene) | 15 | 50 | −271.4507 | −1.6403 | −1.7117 |
| 12 | $C_7H_{16}$ (*n*-heptane) | 23 | 58 | −276.2857 | −1.8057 | −1.8841 |
| 13 | $C_8H_8$ (cyclooctatetraene) | 16 | 56 | −309.4598 | −1.8553 | −1.9397 |
| 14 | $C_8H_{18}$ (*n*-octane) | 26 | 66 | −315.5840 | −2.0589 | −2.1478 |
| 15 | $C_{10}H_8$ (naphthalene) | 18 | 68 | −385.7282 | −2.2793 | −2.3757 |
| 16 | $C_{10}H_8$ (azulene) | 18 | 68 | −385.6708 | −2.2914 | −2.3901 |
| 17 | $CH_3COOCH_3$ (acetic acid methyl ester) | 11 | 40 | −268.2831 | −1.3658 | −1.4499 |
| 18 | $(CH_3)_3COH$ (2-methyl-2-propanol) | 15 | 42 | −233.5910 | −1.3609 | −1.4392 |
| 19 | $C_6H_5NH_2$ (aniline) | 14 | 50 | −287.4877 | −1.6751 | −1.7585 |
| 20 | $C_6H_5OH$ (phenol) | 13 | 50 | −307.3446 | −1.6975 | −1.7887 |
| 21 | $C_4H_6O$ (divinyl ether) | 11 | 38 | −231.1038 | −1.2715 | −1.3453 |
| 22 | $C_4H_8O$ (tetrahydrofuran) | 20 | 40 | −232.3577 | −1.3145 | −1.3830 |
| 23 | $C_5H_8O$ (cyclopentanone) | 14 | 46 | −270.4653 | −1.5241 | −1.6014 |
| 24 | $C_6H_4O_2$ (benzoquinone) | 12 | 56 | −381.2975 | −1.9604 | −2.0698 |
| 25 | $CH_4ON_2$ (urea) | 8 | 32 | −225.1869 | −1.1229 | −1.2052 |
| 26 | $C_4H_4N_2$ (pyrimidine) | 10 | 42 | −264.2111 | −1.4496 | −1.5206 |
| 27 | $N{\equiv}C{-}CH_2{-}CH_2{-}C{\equiv}N$ (butanedinitrile) | 10 | 42 | −264.2017 | −1.4513 | −1.5222 |
| 28 | $C_4H_4N_2$ (pyrazine) | 10 | 42 | −264.2037 | −1.4534 | −1.5241 |
| 29 | $CH_3{-}C(=O){-}C{\equiv}CH$ (acetyl acetylene) | 9 | 36 | −229.8838 | −1.2310 | −1.3001 |
| 30 | $CH_3{-}CH{=}CH{-}CHO$ (crotonaldehyde) | 11 | 38 | −231.1390 | −1.2672 | −1.3383 |
| 31 | $CH_3{-}C(=O){-}O{-}C(=O){-}CH_3$ (acetic anhydride) | 13 | 54 | −381.5785 | −1.8852 | −2.0016 |
| 32 | $(CH_3)_2CH{-}CN$ (2-methylpropane nitrile) | 12 | 38 | −211.3013 | −1.2497 | −1.3088 |
| 33 | $CH_3{-}CO{-}CH_2{-}CH_3$ (methyl ethyl ketone) | 13 | 40 | −232.3764 | −1.3100 | −1.3809 |
| 34 | $(CH_3)_2CH{-}CHO$ (2-methylpropanal) | 13 | 40 | −232.3663 | −1.3123 | −1.3833 |
| 35 | $C_4H_8O_2$ (1,4-dioxane) | 14 | 48 | −307.5394 | −1.6233 | −1.7169 |
| 36 | $C_4H_8NH$ (tetrahydropyrrole) | 14 | 40 | −212.5006 | −1.2939 | −1.3565 |
| 37 | $CH_3{-}CH_2{-}CH(CH_3){-}NO_2$ (nitro-*s*-butane) | 16 | 56 | −362.8203 | −1.9421 | −2.0547 |
| 38 | $CH_3{-}CH_2{-}O{-}CH_2{-}CH_3$ (diethyl ether) | 15 | 42 | −233.5692 | −1.3533 | −1.4260 |
| 39 | $CH_3{-}CH(OCH_3)_2$ (acetaldehyde dimethyl acetal) | 16 | 50 | −308.7526 | −1.6655 | −1.7617 |
| 40 | $(CH_3)_3C{-}NH_2$ (*t*-butylamine) | 16 | 42 | −213.7294 | −1.3396 | −1.4088 |
| 41 | $-CH{=}CH{-}N(CH_3){-}CH{=}CH$ (*N*-methylpyrrole) | 13 | 44 | −249.3768 | −1.4706 | −1.5399 |
| 42 | $C_5H_{10}O$ (tetrahydropyran) | 16 | 48 | −271.6639 | −1.5695 | −1.6478 |
| 43 | $CH_3{-}CH_2{-}CO{-}CH_2{-}CH_3$ (diethyl ketone) | 16 | 48 | −271.6753 | −1.5635 | −1.6443 |
| 44 | $C_5H_{10}O_2$ $CH_3{-}C(=O){-}O{-}CH(CH_3)_2$ (isopropyl acetate) | 17 | 56 | −346.8901 | −1.8740 | −1.9790 |
| 45 | $C_5H_{10}NH$ (piperidine) | 17 | 48 | −251.8073 | −1.5487 | −1.6211 |
| 46 | $(CH_3)_3C{-}O{-}CH_3$ (*t*-butyl methyl ether) | 18 | 50 | −272.8712 | −1.6146 | −1.6974 |
| 47 | $C_6H_4F_2$ (1,3-difluorobenzene) | 12 | 52 | −430.5462 | −2.0171 | −2.1432 |
| 48 | $C_6H_4F_2$ (1,4-difluorobenzene) | 12 | 52 | −430.5452 | −2.0175 | −2.1436 |
| 49 | $C_6H_5F$ (fluorobenzene) | 12 | 44 | −331.3479 | −1.7012 | −1.7949 |
| 50 | $(CH_3)_2CH{-}O{-}CH(CH_3)_2$ (diisopropyl ether) | 21 | 52 | −312.1751 | −1.8651 | −1.9583 |
| 51 | $C_2F_6$ | 8 | 66 | −675.0219 | −2.4306 | −2.6426 |

population analysis) [3]. A slightly different RECEP-3 parameter set, which was optimized to reproduce the G3 total energies [Eq. 4, method = G3, charge definition = natural population analysis, and basis set = 6-311+G(2d,p)] is also available; however, the results obtained with this parameter set are not shown for brevity. We note that owing to linear dependency problems, the natural population analysis analysis with the 6-311 + G(2d,p) basis set was not feasible with the Gaussian98 program package [21] for several molecules, such as naphthalene, azulene, aniline, benzoquinone, and difluorobenzene (see note in Table 2). The results obtained from HF–SCF/6-31G(d) total energies, partial charges, and the corresponding REBECEP parameter set [4] are also shown in Table 2.

The REBECEP enthalpy of formation of azulene shows the largest deviation from the experimental value (−10.7 kcal/mol, Table 2). This molecule is quite difficult

**Table 2.** Experimental enthalpies of formation, deviations of G3 and B3LYP methods from experimental enthalpies of formation, deviations between various RECEP-3 and REBECEP results from experimental results for the 51 enthalpies of formation selected from the G3/99 test set (kcal/mol)

| | Species | Expt. | | Deviation (expt.–theory) | | Deviation (expt.–REBECEP) Natural population analysis, expt.[a] | |
|---|---|---|---|---|---|---|---|
| | | $\Delta H_f^0$ (298 K) | Error (±) | G3 | B3LYP | 6-311+G(2d,p) | 6-31G(d) |
| 1 | $C_4H_6$ (methylallene) | 38.8 | 0.1 | 0.2 | 0.0 | 1.7 | 0.1 |
| 2 | $C_5H_8$ (isoprene) | 18.0 | 0.3 | −0.2 | −4.9 | −0.1 | −2.9 |
| 3 | $C_5H_{10}$ (cyclopentane) | −18.3 | 0.2 | −0.5 | −10.0 | −3.7 | −0.5 |
| 4 | $C_5H_{12}$ (*n*-pentane) | −35.1 | 0.2 | 0.3 | −7.1 | −1.5 | −1.1 |
| 5 | $C_5H_{12}$ (neopentane) | −40.2 | 0.2 | 0.5 | −10.5 | −5.8 | −4.4 |
| 6 | $C_6H_8$ (1,3-cyclohexadiene) | 25.4 | 0.2 | −0.9 | −9.3 | −2.5 | −2.0 |
| 7 | $C_6H_8$ (1,4-cyclohexadiene) | 25.0 | 0.1 | −1.4 | −9.7 | −0.6 | −0.7 |
| 8 | $C_6H_{12}$ (cyclohexane) | −29.5 | 0.2 | −0.2 | −13.4 | −5.6 | −1.9 |
| 9 | $C_6H_{14}$ (*n*-hexane) | −39.9 | 0.2 | 0.6 | −9.3 | −3.3 | −1.6 |
| 10 | $C_6H_{14}$ (3-methylpentane) | −41.1 | 0.2 | 0.2 | −11.7 | −6.4 | −4.5 |
| 11 | $C_6H_5CH_3$ (toluene) | 12.0 | 0.1 | −0.9 | −7.6 | | 0.4 |
| 12 | $C_7H_{16}$ (*n*-heptane) | −44.9 | 0.3 | 0.8 | −11.7 | −4.4 | −2.3 |
| 13 | $C_8H_8$ (cyclooctatetraene) | 70.7 | 0.4 | −1.4 | −11.7 | −3.9 | −5.2 |
| 14 | $C_8H_{18}$ (*n*-octane) | −49.9 | 0.3 | 0.9 | −14.0 | −5.7 | −3.0 |
| 15 | $C_{10}H_8$ (naphthalene) | 35.9 | 0.4 | 0.5 | −11.7[b] | | 0.2 |
| 16 | $C_{10}H_8$ (azulene) | 69.1 | 0.8 | −1.6 | −11.7[b] | | −10.7 |
| 17 | $CH_3COOCH_3$ (acetic acid methyl ester) | −98.4 | 0.4 | 0.7 | −2.9 | −1.5 | 1.1 |
| 18 | $(CH_3)_3COH$ (2-methyl-2-propanol) | −74.7 | 0.2 | 0.8 | −9.0 | −4.6 | −3.4 |
| 19 | $C_6H_5NH_2$ (aniline) | 20.8 | 0.2 | −1.3 | −2.7 | | −1.5 |
| 20 | $C_6H_5OH$ (phenol) | −23.0 | 0.2 | −1.6 | −7.1[b] | | −1.1 |
| 21 | $C_4H_6O$ (divinyl ether) | −3.3 | 0.2 | −0.2 | −1.2 | −1.0 | −3.6 |
| 22 | $C_4H_8O$ (tetrahydrofuran) | −44.0 | 0.2 | −0.2 | −7.6 | −2.1 | 1.9 |
| 23 | $C_5H_8O$ (cyclopentanone) | −45.9 | 0.4 | 0.7 | −8.2 | −1.9 | 2.0 |
| 24 | $C_6H_4O_2$ (benzoquinone) | −29.4 | 0.8 | −1.1 | −8.6[b] | | −0.9 |
| 25 | $CH_4ON_2$ (urea) | −56.3 | 0.3 | −1.3 | | −1.8 | 1.0 |
| 26 | $C_4H_4N_2$ (pyrimidine) | 46.8 | 0.3 | 1.7 | 5.3 | 2.7 | 4.7 |
| 27 | N≡C−$CH_2$−$CH_2$−C≡N (butanedinitrile) | 50.1 | 0.2 | −0.2 | −2.1 | 1.9 | 6.5 |
| 28 | $C_4H_4N_2$ (pyrazine) | 46.9 | 0.3 | −2.7 | 1.4 | 0.6 | 3.5 |
| 29 | $CH_3$−C(=O)−C≡CH (acetyl acetylene) | 15.6 | 0.2 | −2.5 | −5.9 | −2.8 | −3.6 |
| 30 | $CH_3$−CH=CH−CHO (crotonaldehyde) | −24.0 | 0.3 | 0.8 | −1.0 | 2.3 | 1.3 |
| 31 | $CH_3$−C(=O)−O−C(=O)−$CH_3$ (acetic anhydride) | −136.8 | 0.4 | 2.1 | −4.0 | −4.1 | 0.3 |
| 32 | $(CH_3)_2$CH−C≡N (2-methylpropane nitrile) | 5.6 | 0.3 | −1.1 | −5.4 | −0.4 | 1.2 |
| 33 | $CH_3$−CO−$CH_2$−$CH_3$ (methyl ethyl ketone) | −57.1 | 0.2 | 0.3 | −4.5 | −0.5 | 1.1 |
| 34 | $(CH_3)_2$CH−CHO (2-methylpropanal)) | −51.6 | 0.2 | −0.6 | −6.4 | −2.7 | −2.1 |
| 35 | $C_4H_8O_2$ (1,4-dioxane) | −75.5 | 0.2 | 0.9 | −7.8 | −1.3 | 3.7 |
| 36 | $C_4H_8NH$ (tetrahydropyrrole) | −0.8 | 0.2 | −0.7 | −5.3 | −3.2 | −0.8 |
| 37 | $CH_3$−$CH_2$−CH($CH_3$)−$NO_2$ (nitro-*s*-butane) | −39.1 | 0.4 | 1.1 | −5.2 | −4.8 | −1.2 |
| 38 | $CH_3$−$CH_2$−O−$CH_2$−$CH_3$ (diethyl ether) | −60.3 | 0.2 | 0.8 | −4.4 | −0.8 | 1.3 |
| 39 | $CH_3$−CH($OCH_3$)$_2$ (acetaldehyde dimethyl acetal) | −93.1 | 0.2 | 1.7 | −6.5 | −4.4 | −1.4 |
| 40 | $(CH_3)_3$C−$NH_2$ (*t*-butylamine) | −28.9 | 0.2 | −0.1 | −6.5 | −6.4 | −5.4 |
| 41 | −CH=CH−N($CH_3$)−CH=CH (*N*-methylpyrrole) | 24.6 | 0.1 | −0.8 | −2.8 | −7.7 | −4.4 |
| 42 | $C_5H_{10}O$ (tetrahydropyran) | −15.2 | 0.2 | 0.3 | −10.8 | −4.0 | 0.9 |
| 43 | $CH_3$−$CH_2$−CO−$CH_2$−$CH_3$ (diethyl ketone) | −61.6 | 0.2 | 1.1 | −6.4 | −1.3 | 1.3 |
| 44 | $CH_3$−C(=O)−O−CH($CH_3$)$_2$ (isopropyl acetate) | −115.1 | 0.2 | 1.3 | −8.3 | −5.2 | −0.5 |
| 45 | $C_5H_{10}NH$ (piperidine) | −11.3 | 0.1 | −0.9 | −9.2 | −5.7 | −2.0 |
| 46 | $(CH_3)_3$C−O−$CH_3$ (*t*-butyl methyl ether) | −67.8 | 0.3 | 1.4 | −10.2 | −6.9 | −5.0 |
| 47 | $C_6H_4F_2$ (1,3-difluorobenzene) | −73.9 | 0.2 | 0.4 | −4.7[b] | | −1.0 |
| 48 | $C_6H_4F_2$ (1,4-difluorobenzene) | −73.3 | 0.2 | 0.4 | −4.8[b] | | −0.4 |
| 49 | $C_6H_5F$ (fluorobenzene) | −27.7 | 0.3 | −0.4 | −5.1[b] | | 0.7 |
| 50 | $(CH_3)_2$CH−O−CH($CH_3$)$_2$ (diisopropyl ether) | −76.3 | 0.4 | 1.6 | −11.6 | −6.7 | −3.2 |
| 51 | $C_2F_6$ | −321.3 | 0.8 | 2.8 | −7.4 | 0.5 | 2.3 |

[a] Natural population analysis or Mulliken charges are used in the REBECEP calculation, as indicated. G3 and expt. denote the method used to obtain the total energy. The REBECEP parameters were fitted to reproduce the G3 or the experimental total energy
[b] Natural bond orbital cannot handle linearly dependent basis sets

test even for the G3 method, and the B3LYP method fails for this molecule (−11.7 kcal/mol deviation, Table 2). As it was not possible to calculate the HF–SCF/ 6-311+G(2d,p) natural population analysis charges for azulene the results for the larger basis set are missing. Comparison of the correlation energy values in Table 1 indicates that the G3 energy correction (Eq. 1) for the HF–SCF/6-31G(d) energy is considerably larger

**Table 3.** Statistical analysis of the results obtained for the 51 enthalpies of formation selected from the G3/99 test set (kcal/mol)

| | Deviation (expt.−theory) | | Deviation (expt.−REBECEP) | | | | |
|---|---|---|---|---|---|---|---|
| | G3 | B3LYP | 6-311+G(2d,p) | | 6-31G(d) | | |
| | | | Natural population analysis, G3 | Natural population analysis, expt. | Natural population analysis, G3 | Natural population analysis, expt. | Mulliken, G3 |
| Number of molecules | 51 | 50 | 42 | 42 | 51 | 51 | 47 |
| Root-mean-square deviation | 1.15 | 3.96 | 2.67 | 2.71 | 3.15 | 2.92 | 3.79 |
| Average deviation | 0.04 | −6.82 | −2.69 | −2.75 | −1.00 | −0.92 | −1.58 |
| Average absolute deviation from the average | 0.93 | 3.11 | 2.25 | 2.26 | 2.38 | 2.18 | 2.95 |
| Average absolute deviation from the reference | 0.94 | 7.09 | 3.09 | 3.21 | 2.52 | 2.31 | 3.26 |

for azulene (no. 16) than for naphthalene (no. 15), −2.3901 versus −2.3757 a.u., respectively. Detailed analysis of the HF/6-31G(d) natural population analysis partial charge distribution shows that only slightly larger partial charges occur on the ten carbon atoms of azulene than on the ten carbon atoms of naphthalene, leading to slightly different REBECEP energy correction (−2.3755 versus −2.3752 a.u., respectively). Thus, the agreement for naphthalene is good, and it is rather poor for azulene (Table 2: 0.2 and −11.7 kcal/mol). Further investigations show that azulene remains a serious outlier for the RECEP method calculated with any basis sets and charge definitions. Considerable (larger than 5 kcal/mol) basis set independent deviation was observed for neopentane, 3-methylpentane, cyclooctatetraene, *n*-octane, butanedinitrile, *t*-butylamine, *N*-methylpyrrole, and *t*-butyl methyl ether. The charge distribution of the two CN groups in butanedinitrile and the crowded *t*-butyl group in *t*-butyl methyl ether are assumed to be responsible for the large deviation in the RECEP method. Indeed, inspection of the values in Table 2 illustrates that the molecules containing the *t*-butyl group usually show relatively large deviations, for example, −4.4, −3.4, and −5.4 kcal/mol for neopentane (no. 5), *t*-butanol (no. 18), and *t*-butylamine (no. 40), respectively. Using Mulliken charges instead of natural population analysis charges leads to worse results (Table 3).

In order to characterize the G3, B3LYP, and various REBECEP deviations from experiment, the root-mean-square and the average deviations are shown in Table 3. The results show that the G3 method yields the smallest and the B3LYP method yields the largest root-mean-square deviation from experiment. The REBECEP results are consistently better than the B3LYP/6-311+G(3df,2p) results. The average deviations (the center of the distribution) also exhibit similar behavior. The average deviation of the G3 method provides a nearly perfect agreement (0.04 kcal/mol) with experiment, and the average deviation of the B3LYP method is rather large (−6.82 kcal/mol). The B3LYP enthalpies of formation are too positive on average. The smallest average deviation from the experimental enthalpies of formation (−0.92 kcal/mol) was obtained with natural population analysis charges and the 6-31G(d) basis set. The results obtained with the larger 6-311+G(2d,p) basis set show considerable average deviation (−2.75 kcal/mol); however it should be noted that this basis set provides a more focused distribution (smaller root-mean-square and average absolute deviation from the average, Table 3). In order to make comparison easier

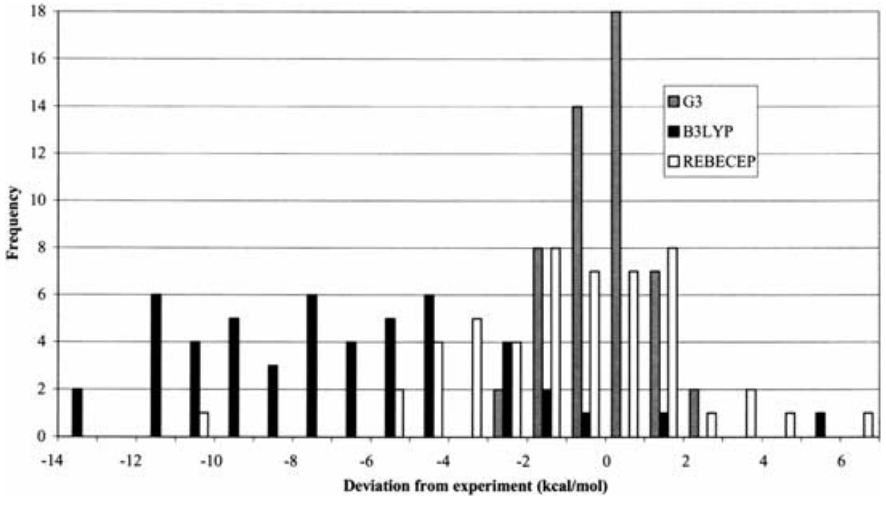


**Fig. 1.** Histogram of G3, B3LYP, and rapid estimation of basis set error and correlation energy from partial charges (*REBECEP*) deviations (experiment minus theory) for the test set of 51 molecules. Each *vertical bar* represents the frequency of a given deviation in a 1-kcal/mol range

with the earlier G3 studies we also show the average absolute deviation from the reference in Table 3 (called average absolute deviation in the G3 studies). We also show the average absolute deviation from the average (not used in the G3 studies). The average deviation and the root-mean-square deviation characterize better the statistical distribution of the errors than the average absolute deviation from experiment (the reference). Using this latter value alone does not show the structure of the distribution. The problem can be illustrated by the distribution of the B3LYP deviations. The main source of errors of the B3LYP results is the too positive average enthalpy of formation. The distribution of the data is considerably better than is suggested by the value of 7.09 kcal/mol average absolute deviation from experiment (3.96 kcal/mol for the root-mean-square deviation and 3.11 kcal/mol average absolute deviation from the average in Table 3). In agreement with the observation in Ref. [6], our preliminary analysis suggests that the density functional theory errors accumulate in large molecules. We analyzed the REBECEP errors with respect to the number of electron pairs and found no correlation in general; however, homologous series show quasilinear accumulation of errors. Further details will be presented elsewhere.

Inspection of the results of the statistical analysis in Table 3 shows that the quality of the HF–SCF/6-31G(d) REBECEP results is comparable to the quality of the HF–SCF/6-311+G(2d,p) RECEP-3 results. Although the latter provides a more focused distribution, however, the center of the distribution is shifted considerably (hint for systematic error). Partly the consistent behavior of the 6-31G(d) basis set is advantageous for REBECEP-type energy calculations. However, one should not forget that the Gaussian98 package failed to provide natural population analysis charges for several molecules (Table 2) with the 6-311+G(2d,p) basis set, while all natural population analysis charges were available for the smaller basis set. Consequently the two test sets are not equal. It should be noted that azulene, the most problematic molecule, is missing from the former distribution. Exclusion of azulene from the distribution would decrease considerably the root-mean-square deviation. These results suggest that the HF–SCF basis set extension energy error can be effectively approximated from partial charges. Further thinking along this line suggests that extrapolations to the infinite basis set could be done using atomic parameters (atomic partition of the basis set error). If this idea prevails, infinite basis set quality results can be reached very effectively. Further investigations are necessary along this line.

The distribution of errors of the three methods is shown as a histogram in Fig. 1. The frequency of the deviations clearly shows the problem with the B3LYP method. It can be seen that the distribution of errors is not a Gaussian-like distribution. The deviations of the B3LYP results are nearly evenly distributed between −4 and −11 kcal/mol. The distribution of the REBECEP deviations is considerably more focused than the distribution of the B3LYP deviations, with the former being closer to the ideal Gaussian distribution; however, it shows a composed structure (e.g. several overlapping Gaussians). The G3 method shows a relatively narrow and nearly ideal Gaussian distribution in Fig. 1.

## 4 Conclusions

We selected 50 large molecules from the new G3-3 subset of the G3/99 test set plus urea for testing the RECEP-3 [HF/6-311+G(2d,p)] and REBECEP [HF/6-31G(d)] atomic energy parameter sets. The REBECEP total energies of these molecules were estimated from atomic charges using our earlier parameters, obtained by fitting to reliable total energies of smaller molecules in the G2/97 database. All the molecules selected were closed-shell neutrals composed of H, C, N, O, and F atoms. We excluded radicals because the correlation energy for radicals is rather different from the correlation energy in closed-shell molecules.

We compared the calculated REBECEP, G3, and B3LYP/6-311+G(3df,2p) enthalpies of formation to the experimental enthalpies of formation of 51 molecules. In the REBECEP calculations the same MP2(full)/6-31G(d) geometries were used as in the G3 calculations. The best REBECEP results were produced with the use of natural population analysis charges, although the results obtained from Mulliken charges can approach chemical accuracy as well. The statistical analysis of the reliability of the G3, B3LYP, and REBECEP methods for the 51 molecules in this study shows that the root-mean-square deviations from the experimental enthalpies of formation for the 51 molecules selected are 1.15, 3.96, and 2.92 kcal/mol for Gaussian-3, B3LYP/6-311+G(3df,2p), and REBECEP (natural population analysis) enthalpies of formation, respectively. The corresponding average deviations from the experimental enthalpies of formation are 0.04, −6.82, and −0.92 kcal/mol, respectively. The corresponding average absolute deviations from the experimental enthalpies of formation are 0.94, 7.09, and 2.27 kcal/mol, respectively. The REBECEP results for the 51 large molecules of the G3/99 test set are worse than those obtained for the 65 molecules of the G2/97 test set (1.4–1.7 kcal/mol); however, it can be observed that the REBECEP enthalpies of formation deviate only for two molecules by more than 6 kcal/mol from the experimental values (azulene by −10.7 kcal/mol and butanedinitrile by +6.5 kcal/mol). These molecules probably require special treatment. These two molecules alone contribute more than 0.35 kcal/mol to the average absolute deviation. We also note that molecules containing the *t*-butyl group and that cyclooctatetraene are sources of the large deviation from the experimental results.

It was observed that a REBECEP energy calculation with the 6-31G(d) basis set can provide results of comparable quality to the results of the RECEP-3 calculations performed with the considerably larger 6-311+G(2d,p) basis set. This feature was attributed to the consistent behavior of the 6-31G(d) basis set. The results suggest that the HF–SCF basis set extension energy error can be effectively approximated from partial charges. The implicit atomic partitioning of the basis set error used in our method could be done explicitly and

the infinite basis set energies could be predicted from atomic parameters and charges. Further investigations along this line seems to be promising.

*Acknowledgements*. S.K. is thankful to the Domus Hungarica Scientarium et Artium 2000 summer grant. This work was supported by the OTKA grant (T 031767, Hungary).

## References

1. Kristyan S, Csonka GI (1999) Chem Phys Lett 307: 469
2. Kristyán S, Csonka GI (2001) J Comput Chem 22: 241. DOI 10.1002/1096-987X(20010130)22:2 < 241::AID-JCC11 > 3.0.CO; 2-C
3. Kristyán S, Ruzsinszky A, Csonka GI (2001) J Phys Chem A 105:1926, http://pubs.acs.org/reprint-request?jp0018192/V6K6
4. Kristyan S, Ruzsinszky A, Csonka GI (2001) Theor Chem Acc (in press)
5. Curtiss LA, Raghavachari K, Trucks GW, Pople JA (1991) J Chem Phys 94: 7221
6. Curtiss LA, Raghavachari K, Redfern PC, Rassolov V, Pople JA (1998) J Chem Phys 109: 7764
7. Reed AE, Weinstock RB, Weinhold FJ (1985) J Chem Phys 83: 735
8. Breneman CM, Wiberg KB (1990) J Comput Chem 11: 361
9. Besler BH, Merz KM Jr, Kollman PA (1990) J Comput Chem 11: 431
10. Lee C, Yang W, Parr RG (1988) Phys Rev B 37: 785
11. Becke AD (1993) J Chem Phys 98: 5648
12. Frisch MJ, Trucks GW, Head-Gordon M, Gill PMW, Wong MW, Foresman JB, Johnson BG, Schlegel HB, Robb MA, Replogle ES, Gomperts R, Andres JL, Raghavachari K, Binkley JS, Gonzalez C, Martin RL, Fox DJ, DeFrees DJ, Baker J, Stewart JJP, Pople JA (1993) Gaussian 92/DFT. Gaussian, Pittsburgh, Pa
13. (a) Curtiss LA, Raghavachari K, Redfern PC, Pople JA (1997) J Chem Phys 106: 1063; (b) Curtiss LA, Redfern PC, Raghavachari K, Pople JA (1998) J Chem Phys 109: 42
14. Curtiss LA, Raghavachari K, Redfern PC, Pople JA (2000) J Chem Phys 112: 7374
15. Bauschlicher CW Jr, Ricca A (1998) J Phys Chem A 102: 8044
16. Feller D, Peterson KA (1998) J Chem Phys 108: 154
17. Martin JML (1998) J Chem Phys 108: 2791
18. Chemistry webbook: http://webbook.nist.gov/chemistry
19. Pedley JB, Naylor RD, Kirby SP (1986) Thermochemical data of organic compounds, 2nd edn. Chapman and Hall, New York
20. Frenkel M, Marsh KN, Wilhoit RC, Kabo GJ, Roganov GN (1994) Thermodynamics of organic compounds in the gas state. Thermodynamics Research Center, College Station, Tex
21. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson BG, Chen W, Wong MW, Andres JL, Head-Gordon M, Replogle ES, Pople JA (1998) Gaussian 98. Gaussian, Pittsburgh, Pa