Theor Chem Acc (2006) 115: 298–307
DOI 10.1007/s00214-005-0039-3

REGULAR ARTICLE

Sandor Kristyan

Rapid estimation of basis set error and correlation energy based on Mulliken charges and Mulliken matrix with the small 6-31g* basis set

Received: 17 August 2005 / Accepted: 31 August 2005 / Published online: 8 December 2005

Abstract Good, density functional quality (B3LYP/6-31G*) ground state total electronic energies have been approximated using single point Hartree–Fock-self consistent field (HF-SCF/6-31G*) total energies and Mulliken partial charges versus. Mulliken matrix (electrons assigned to atoms and atoms pairs from Mulliken population analysis). This is a development of our rapid estimation of basis set error and correlation energy from partial charges (REBECEP) method, published earlier (see references [21,22,30]). The development is as follows: (1) A larger set of atoms (H, C, N, O, F, Si, P and S) are considered as building blocks for closed shell, neutral, ground state molecules at their equlibrium geometry; (2) geometries near equilibrium geometry are also considered; (3) A larger set, containing 115 molecules, was used to fit REBECEP parameters; (4) most importantly, electrons belonging to chemical bonds (between atom pairs) are also considered (Mulliken matrix) in addition to the atoms (Mulliken charges), using more REBECEP parameters to fit and yielding a more flexible algorithm. With these parameters a rather accurate closed shell ground state electronic total energy can be obtained from a small basis set HF-SCF calculation in the vicinity of optimal geometry. The 3.3 kcal/mol root mean square deviation of REBECEP improves to 1.5 kcal/mol when using Mulliken matrix instead of Mulliken charges.

Keywords Ab initio total ground state electronic energy · Basis set error · Correlation energy · Mulliken partial charge · Density functional theory

1 Introduction

Accurate techniques for prediction of the thermochemistry of molecules using ab initio calculations are emerging at a rapid pace. The primary goal is to reach the so-called chemical accuracy (±1 kcal/mol) reliably. One frequently used composite method is the Gaussian-3 (G3) theory [1–4] yielding good results at the expense of larger disc space and CPU time. Several similar techniques were developed recently [5–7]. Common to these techniques is that the correlationenergy is approximated in the classical way by very expensive methods [e.g. CCSD(T) or QCISD(T)], and empirical corrections are used to reach the chemical accuracy. Alternatively, the density functional theory (DFT) methods use considerably faster algorithms [8–11] for the estimation of the correlation energy, for example the B3LYP [12–17]. However, these DFT methods also increase the computation time by a factor of about two, even in the case of smaller molecules, and the increase is more drastic for larger ones; furthermore, computational convergence problems may arise by the numerical integration involved if the number of electrons is high.

In our earlier works [18–23] we analyzed the applicability of a radically different approach to calculate the dynamic correlation energy for closed shell, ground state, and near-equilibrium geometry systems (composed of H, C, N, O, and F atoms) very rapidly and effectively. This procedure is called rapid estimation of basis set error and correlation energy from partial charges (REBECEP). This method is also a DFT procedure, because via certain partial charges from the electron density it partitionates the correlation energy among the atoms using a multilinear functional [see Eqs. (3), (4) below]. The basic idea of the method works even for a separate zero point energy calculation as well [23,24]. In general, the correlation energy (E_{corr}) for the electronic ground state is defined [25] as the difference between the exact non-relativistic complete-CI (configurations interactions) electronic ground state basis set limit total electronic energy (E_{T}(CI)) and the single determinant electronic ground state Hartree-Fock-self consistent field (HF-SCF) basis set limit total electronic energy (E_{T}(HF-SCF)) of a system:

$$E_{\mathrm{corr}} = E_{\mathrm{T}}(\mathrm{CI}) - E_{\mathrm{T}}(\mathrm{HF\text{-}SCF}). \quad (1)$$

Because the calculation of E_{T}(CI) is currently not feasible for most of the molecules (i.e. the number of electrons must be less than, let say, about 20), in our earlier work [21]

S. Kristyan
Chemical ResearchCenter, Institute of Chemistry, Hungarian Academy of Sciences, Pusztaszeri út 59-67, H-1025 Budapest, Hungary
E-mail: kristyans@chemres.hu

we defined the total molecular correlation energy to the level of a reliable theory ("method" in the argument). We redifine this as the difference of a chosen method and the HF-SCF/6-31G* total energy (where the basis in the latter is a moderately good set [17,26,27]):

$$E_{\mathrm{corr}}(\mathrm{method}) = E_{\mathrm{T}}(\mathrm{method}) - E_{\mathrm{T}}(\mathrm{HF\text{-}SCF/6-31G*}). \quad (2)$$

In this work, we have chosen the B3LYP/6-31G* quality total electronic energy [14,16,26] for the "method" in the argument. Although this latter method uses the same basis set now, but in general it is not a restriction in Eq. (2). Equation (2) is a practical definition, because it contains the basis set error as well.

The REBECEP method requires only a single point HF-SCF/6-31G* calculation in the vicinity of equilibrium geometry. A molecular set of 115 molecules (see below) was geometry optimized on MP2(FU)/6-31G* level [17,26,28], as in the G3 set. 89 of them were chosen from that data set [1–4], the other 26 were created to contain certain atoms (starting with Spartan [29] molecular mechanics geometry optimization) to have a balanced set for fitting REBECEP atomic parameters.

In the present paper the use of a smaller basis set, namely the 6-31G* basis set, is analyzed as already considered in ref. [30]. This smaller basis set increases considerably the speed of HF-SCF calculations, but introduces considerable basis set error into the total energy. This increase of speed is critical for larger molecules and the use of this basis set extends considerably the use of REBECEP type methods. We used the Gaussian 98 [26] program package in this work for all ab initio calculations as well as to measure the necessary CPU time and disc space on a 2 GHz personal computer (PC) with LINUX environment. The above mentioned additional molecules were created by a Spartan program package [29] on a Silicon Graphics (Octan 8) machine with Irix 6.5 (Unix) operation system. Thus, after the brief summary of the REBECEP method we present the results obtained with the new parameter set belonging to Mulliken charges and Mulliken matrix. It will show how the correlation energy can be partitioned among atoms or atom-atom pairs in molecules. The choice of Mulliken charges comes from the fact that it is instant after a HF-SCF calculation as well as easily available in commercial packages [17,26]. In fact the Mulliken charge is the "most obvious" partial charge definition [28,31] in the HF-SCF formalism.

2 The rapid estimation of basis set error and correlation energy from partial charges method

The REBECEP formula [18–21,30] for ground state covalent neutral molecules in the vicinity of stationary points (here we deal only with geometry minimums) is the following:

$$E_{\mathrm{corr}}(\mathrm{REBECEP, method, charge\ def., basis\ set}) \equiv \sum_{A=1}^{M} E_{\mathrm{corr}}(N_A, \mathrm{Z_A, method,\ charge\ def., basis\ set}) \quad (3)$$

where E_{corr}(REBECEP, method, charge def., basis set) is the REBECEP molecular correlation energy and basis set error that approximates E_{corr}(method) in Eq. (2) using one of the various available partial charges calculated at HF-SCF/(basis set) level of theory. Again, we have chosen the B3LYP/6-31G* for "method", i.e. Eq. (3) approximates this level of calculation. From our previous reports on REBECEP (see Kristyan et al. [21,22,30]), we can say the following: (1) Different "method, charge def., basis set" cases have different REBECEP parameter sets for the right hand side of Eq. (3) (that is the reason it is indicated in the argument here as well), which have similar magnitudes, but most importantly, yield very close values for the left hand side. By this reason, the most educated choice focuses on the fastest HF-SCF calculation including partial charge values. As a consequence, we have chosen the Mulliken partial charges for "charge def." and 6-31G* for "basis set", the latter is probably the smallest basis set which yields reliable results. (From our unpublished investigations, a 3-21G basis set, is too poor for REBECEP method although it can yield good results, e.g. for partial charges, etc). (2) The G3, B3LYP or other accepted reliable methods for total electronic energies differ from each other by up to 1–2 hartree for systems containing, for example, about 50 electrons, however in computational chemistry we always need energy differences, which can correct a large part of the errors. For example the HF-SCF/6-31G*, B3LYP/6-31G* and G2 methods for equilibrium gas phase benzene (C_6H_6, 42 electrons) give −230.702511, −232.248439 and −231.876360 hartree, respectively. (3) The restriction of neutrality in molecular charge comes from the fact that the range of REBECEP parameters may not be wide enough in Eq. (3), otherwise it can be extended by involving more charged (e.g. protonated) molecules in the molecular set for the fit. (The open shell molecules probably need different REBECEP parameter sets, however open shell ab initio calculations have too many complexities anyway). In the right hand side of Eq. (3), N_A is the "electron content" on atom A, generally non-integer and defined as "Z_A - partial charge", where Z_A is the nuclear charge of atom A. The summation in Eq. (3) runs for all M atoms in the molecule. The two basic assumptions of Eq. (3) are that: (1) the correlation and basis set error energy is the sum of the REBECEP atomic correlation and basis set error energies. (2) The value of these atomic correlation and basis set error energies can be estimated from the atomic electron contents (N_A) in the molecule in the vicinity of stationary points.

The $E_{corr}(N_A,N_A$ method, charge def., basis set) atomic energy terms in Eq. (3) are interpolated linearly as follows:

$$\begin{aligned} &E_{corr}(N_A, Z_A, \mathrm{method, charge\ def., basis\ set}) \\ &= (N_A - N1)E_{\mathrm{fitpar}}(N2, Z_A, \mathrm{method,\ charge\ def.,basis\ set}) \\ &\quad +(N2 - N_A)E_{\mathrm{fitpar}}(N1, Z_A, \mathrm{method,\ charge\ def., basis\ set}), \end{aligned} \quad (4)$$

where $N1$ and $N2$ are integer numbers of electrons, with $N1 \le N_A \le N2 = N1 + 1$, and N_A is the electron content around atom A. $E_{\mathrm{fitpar}}(N1$ or $N2, Z_A$, method, charge

def., basis set) in Eq. (4) is the so called REBECEP atomic parameters that transform the partial charge into energy correction (correlation and basis set). The use of Eq. (4) is obvious, however for hydrogen atoms we suggest using the $E_{\mathrm{corr}}(N_A, Z_A = 1, \mathrm{method, charge\ def., basis\ set}) = N_A E_{\mathrm{fitpar}}$ (2, 1, method, charge def., basis set)/2. The reason for this choice is that although one electron has no correlation effect, a hydrogen atom in a molecule with $0 < N_A \leq 1$, a frequent case, still has correlation contribution. For example in H_2 molecule N_A= 1 and the correlation energy is well defined. These effects are discussed in refs. [19–21]. A posteriori parameters can be obtained from a least square fitting procedure that finds the minimum of $\mathrm{Y} = \Sigma_{(i=1,\mathrm{L})}[E_{\mathrm{corr}}(method)_{\mathrm{i}} - E_{\mathrm{corr}}(\mathrm{REBECEP, method, charge\ def., basis\ set})_i]^2$ in a set of L molecules [20,21]. Here $E_{\mathrm{corr}}(\mathrm{B3LYP/6\text{-}31G^*})_i$ was calculated according to Eq. (2) and $E_{\mathrm{corr}}(\mathrm{REBECEP, B3LYP/6\text{-}31G^*, Mulliken\ charge, 6\text{-}31G^*})_i$ is calculated according to Eqs. (3) and (4). (Again, the two basis sets indicated in the argument do not have to be the same. The latter one refers to the HF-SCF/6-31G* level calculation with the Mulliken charge values.) Since Eqs. (3) and (4) are linear, it is a multilinear fit, and obtaining the REBECEP parameters is strigthforward [20,21,30]. The solution of this system of linear equations yields the desired $E_{\mathrm{fitpar}}(N1, Z_A,$ method, charge def., basis set) fitted values (listed in tables below for Mulliken partial charges). FORTRAN programs are available to the reader from the author, see "Supplementary material" at the end. The chemical accuracy can be reached if these fitted REBECEP atomic parameters are used. One must keep in mind that the set of L molecules used for the fit are in the vicinity of their stationary points, here the geometry minimums. Consequently the REBECEP method [Eq. (3)] is recommended only for similar geometries. The set of L molecules also defines the (N, Z) range of the resulting parameters in tables below. At this point we recall that partial charges are essentially mathematical constructions that serve to represent the electron content around the selected atom of the molecule in certain definitions. Partial charges are not physically measurable quantities, however, in an ideal case they have a relation to the electron distribution in a molecule, which is fundamental in DFT calculations.

Above we have mentioned some energy values for benzene. We would like to provide one further numerical example to justify why we have chosen B3LYP/6-31G* for "method" in Eq. (3). The structure of equilibrium D_{6h} symmetry benzene was modified as follows. We kept the C–H bond distances as 1.0867 Å, however the 1.3948 Å value for C–C bond distances was changed by multiples of ±0.1 Å while keeping the D_{6h} symmetry. This is a certain slice on the potential surface and close to the motion that vibration of benzene has at 621 and 2,415 cm^{-1} in a valley similar to this. With respect to a = 1.39 Å C–C bond distance the (a, Δ E(B3LYP/6-31G*), Δ E(G2)) values are: (1.3 Å, 38.5 kcal/mol, 38.05 kcal/mol) and (1.48 Å, 20.7 kcal/mol, 19.02 kcal/mol), and similar results for other "a" values and even between those "a" values tested; (Δ E is the energy difference with respect to a = 1.39 Å). In other words, the G2 and B3LYP/6-31G* energy curves are very close to each other with respect to energy differences. However, the CPU time and disc space usage of the two methods differs strongly. For B3LYP/6-31G*, CPU is about 12 s and the disc space necessary was 14 Mb on a 2 GHz PC with Gaussian 98. The G2 calculation, which is a composite method of subsequent three MP4, one QCISD(T) and five MP2 calculations with larger basises than 6-31G*, the CPU time was about 272 min, and the largest disc space usage was 1,694 Mb (the QCISD(T) step) under the same conditions. For the sake of brevity no more examples are provided here. Based on these accuracy comparisons, we have chosen the B3LYP/6-31G* for "method" in Eq. (3).

3 The REBECEP/6-31G* algorithm for the two cases: Mulliken charge and Mulliken matrix

The easiest way to avoid too much notation to explain the equations above is to introduce the method via an example. An extract from Gaussian 98 output of HF-SCF single point energy and Mulliken partial charge analysis for the four atom molecule formaldehyde is

```
> # hf/6-31G*
> Formaldehyde (H2C=O)
> SCF Done: E(RHF) = -113.863712881  A.U. after 6 cycles
>              Convg  = 0.8513D-04    -V/T = 2.0037
>              S**2   = 0.0000
>           Condensed to atoms (all electrons):
>              1        2        3        4
>  1  O    8.013325   0.522645  -0.049815  -0.049815
>  2  C    0.522645   4.600151   0.371281   0.371281
>  3  H   -0.049815   0.371281   0.596456  -0.068770
>  4  H   -0.049815   0.371281  -0.068770   0.596456
> Total atomic charges:
>               1
>  1  O   -0.436341
>  2  C    0.134643
>  3  H    0.150849
>  4  H    0.150849
> Sum of Mulliken charges=   0.00000
> Normal termination of Gaussian 98.
```

For example the −0.436341 Mulliken charge value for oxygen listed under "`Total atomic charges`" comes from the sum of the values of oxygen in the first line of the matrix above it as (8.013325 + 0.522645 − 0.049815 − 0.049815) electrons −Z (= 8 protons in oxygen) = −0.436341. The term "`Condensed to atoms (all electrons)`" in this print out is referred to in this work as "Mulliken matrix", and the term "`Total atomic charges`" is referred to as "Mulliken charge". The Mulliken matrix is symmetric, which means that the off-diagonal elements are listed twice as information, i.e., for example, the (1,2) and (2,1) elements are both 0.522645, and this means that the electron content between C and O atoms is represented by this number from "both sides" of the bond in this context. We emphasize again that there are a few different partial charge and bond order definitions in the literature (for example the bond order,

what organic chemists use for example, between the C and O atom in formaldehyde is around 2). The (4,1) or (1,4) elements of the matrix belong to the oxygen and hydrogen atom, and these are low values, meaning that there is practically no bond between O and H in formaldehyde. We use the above charge values in the REBECEP method because these values come from certain integrations on part of the molecular electronic structure [28,31]. It is obvious that the Mulliken matrix tells more about the finer electron distribution than the Mulliken charges. As a consequence the REBECEP parameters based on the first one can be more accurate than using the latter one; at the price of having more REBECEP parameters. Also, the values above come from standard definitions, so these values do not come from specific internal definitions of the Gaussian 98 package. In other words, any ab initio package which calculates these quantities should be fine for REBECEP.

Let us rearrange the data as follows for this neutral molecule at equilibrium geometry:

```
>     0  Molecular charge
> 16  # of atoms in the molecule, listed below
>    8     -0.01332500 Z atomic charge, partial charge
>    6      1.39984900
>    1      0.40354400
>    1      0.40354400
>  608     -0.52264500 dummy Z between  O   C
>  108      0.04981500 dummy Z between  O   H
>  108      0.04981500 dummy Z between  O   H
>  608     -0.52264500 dummy Z between  C   O
>  106     -0.37128100 dummy Z between  C   H
>  106     -0.37128100 dummy Z between  C   H
>  108      0.04981500 dummy Z between  H   O
>  106     -0.37128100 dummy Z between  H   C
>  101      0.06877000 dummy Z between  H   H
>  108      0.04981500 dummy Z between  H   O
>  106     -0.37128100 dummy Z between  H   C
>  101      0.06877000 dummy Z between  H   H
>  -0.63645253 = (-114.50016541)-(-113.86371288)
   an accurate corr. energy in hartree (B3LYP-HF)/6-31G*
> Formaldehyde (H2C=O)
```

The first line is technical, giving the molecular charge to check the consistency under it. The second line tells that this molecule contains 16 atoms. Technically, in our program, it will pick up to 16 partial charge values below it. However, formaldehyde has only four atoms, the 12 others are fictitious atoms, defined as follows. The first four lines are the four real atomic number Z and the converted Mulliken matrix values of the individual atoms. For example 8.013325 electrons on oxygen ($Z = 8$) means -0.01332500 partial charge on oxygen only, and so for the C and two H atoms. (In this way the -0.01332500 charge on O from Mulliken matrix in the converted chart is lower in absolute value than the -0.436341 Mulliken charge value in the previous chart, because in Mulliken matrix a part of electron distribution belongs to atom-atom pairs rather than to the individual atom O). The 12 off-diagonal terms were simply copied from the Mulliken matrix, and attributed to fictitious atomic numbers, because these belong to atom pairs. All these terms are listed twice. For example in the fifth line, the first fictitious Z can be found as "608, -0.52264500, dummy Z between O C". Only the first two numbers are picked up by the algorithm, and it means that between O and C atoms there are 0.52264500 electrons along the bond. We will say that this belongs to a "C and O" atom pair. The atom pair REBECEP parameters are also transferable like the atomic REBECEP parameters, i.e. good for any molecule (near to equilibrium geometry, etc.) and these two atoms can be anywhere in the molecule. The latter means that there is no necessary chemical bond between them. For example, if the Mulliken matrix value is low, there is no chemical bond between, if it is large, there is. However, in programming one should number this relationship for easier algebraic treatment. Definition: the "dummy or fictitious Z" means that its value is $100*Z1 + Z2$ where $Z1$ less or equal to $Z2$. For example, 608 means $Z1$=6 (C atom) and $Z2$=8 (O atom), 101 means that $Z1 = Z2 = 1$, i.e. it concerns two different H atoms in the molecule, and so on, see the 12 examples above. In this work we deal with atoms with $Z < 18$ (Argon), on the other hand, the smallest fictitious Z defined is 101, so there will not be unwanted overlap in atomic number (Z) values. After the four atomic and 12 atom-atom partial charges from Mulliken matrix, there is a B3LYP/6-31G* level correlation energy and basis set error value (only the first number in the line is picked up by the program, and the last line tells the name of the molecule as text). This is the input for fitting REBECEP/6-31G* parameters using Mulliken matrix indicated in the title of this section, and/or for calculating the energy correction by Eq. (3) if the parameters are on hand. If we have the REBECEP parameters on hand, one can estimate the correlation energy and basis set error for any molecule in or outside of the set. The other case of REBECEP/6-31G* uses the Mulliken charge as follows (as in our previous works). The input file has similar structure, now the Mulliken charges are simply copied from the Gaussian 98 output:

```
>    0  Molecular charge
> 4  # of atoms in the molecule, listed below
>    8   -0.43634100 Z atomic charge, partial charge
>    6    0.13464300
>    1    0.15084900
>    1    0.15084900
>      -0.63645253 = (-114.50016541)-(-113.86371288)
an accurate corr. energy in hartree (B3LYP-HF)/6-31G*
> Formaldehyde (H2C=O)
```

Using these inputs we have fitted the REBECEP/6-31G* parameters in both cases (Mulliken matrix and Mulliken charge) for the set of selected 115 molecules. In our previous works only the Mulliken partial charges were used (as well as other partial charge definitions). Now we include more atom types and also introduce the Mulliken matrix, a more detailed charge set than the Mulliken charge, and compare the two cases. As in the Mulliken charge case, one yields parameters for atoms with its ionized states, e.g. for C atom, there will be parameters for C^{+3}, C^{+2}, C^{+1}, C, C^{-1} ions, etc. The (N, Z) range depends on the partial charges coming up in molecules. In the next section we report the REBECEP/6-31G* parameter set (Tables 1, 2) yielded by the fit

Table 1 The case of Mulliken charge. A set of 115 molecules (Table 3) with their Mulliken partial charges has generated the range of these 23 parameters

Atom	N	Z	$E_{\text{fitpar}}(N, Z)$
H	2	1	−0.04172540
C	4	6	−0.17422431
C	5	6	−0.20981117
C	6	6	−0.23745760
C	7	6	−0.25917593
N	6	7	−0.33448715
N	7	7	−0.30484084
N	8	7	−0.32310273
O	7	8	−0.46192945
O	8	8	−0.36921711
O	9	8	−0.37905434
F	9	9	−0.40151786
F	10	9	−0.42644947
Si	12	14	−0.50175714
Si	13	14	−0.55615027
Si	14	14	−0.57734232
P	13	15	−0.55840024
P	14	15	−0.61222298
P	15	15	−0.63072847
S	14	16	−0.63446938
S	15	16	−0.67763142
S	16	16	−0.67971304
S	17	16	−0.68886287

Fitted $E_{\text{fitpar}}(N, Z)$ REBECEP/6-31G* atomic parameters in hartree to use in Eq. (4) for calculating correlation energy and basis set error via Eq. (3) for closed shell neutral molecules in the vicinity of optimum geometry from HF-SCF/6-31G* Mulliken charges containing atoms listed to correct HF-SCF/6-31G* total ground state electronic energy via Eq. (2) to achieve e.g. B3LYP quality. N is the number of electrons and Z is the atomic charge

for the Mulliken matrix and charge cases. Before discussing Tables 1, 2, we show the algorithm with which one can calculate the REBECEP/6-31G* level correlation and basis set error estimation. That is simply the use of Eqs. (3) and (4). As one can see, it can even be done even on a pocket calculator, so after the HF-SCF procedure, it is instant.

For Mulliken matrix case, the calculation is as follows:

```
>Formaldehyde (H2C=O)
>     Zparc.chrg.  N1       NA   N2  E1fitpar  E2fitpar Eweighted
>[a.u.]  [a.u.]                    [hartree] [hartree] [hartree]
>  8   -0.0133250   8  8.0133250  9 -0.3744126-0.4279456-0.3751259
>  6    1.3998490   4  4.6001510  5 -0.1636632-0.1988572-0.1847849
>  1    0.4035440   0  0.5964560  2  0.0000000-0.0662246-0.0197500
>  1    0.4035440   0  0.5964560  2  0.0000000-0.0662246-0.0197500
>608   -0.5226450 608608.5226450609-0.0014455-0.0256796-0.0141113
>108    0.0498150 107107.9501850108 0.0495712 0.0006167 0.0030554
>108    0.0498150 107107.9501850108 0.0495712 0.0006167 0.0030554
>608   -0.5226450 608608.5226450609-0.0014455-0.0256796-0.0141113
>106   -0.3712810 106106.3712810107 0.0001045-0.0189142-0.0069568
>106   -0.3712810 106106.3712810107 0.0001045-0.0189142-0.0069568
>108    0.0498150 107107.9501850108 0.0495712 0.0006167 0.0030554
>106   -0.3712810 106106.3712810107 0.0001045-0.0189142-0.0069568
>101    0.0687700 100100.9312300101 0.0246618-0.0000247 0.0016729
>108    0.0498150 107107.9501850108 0.0495712 0.0006167 0.0030554
>106   -0.3712810 106106.3712810107 0.0001045-0.0189142-0.0069568
>101    0.0687700 100100.9312300101 0.0246618-0.00002470.0016729>
>an accurate Ecorr, Ecorr(REBECEP) [hartree]= -0.63645253,
                                    -0.63989331
>an accurate Ecorr- Ecorr(REBECEP) [hartree]= 0.00344078
```

Table 2 The case of Mulliken matrix. The $Z > 16$ values are fictitious values describing atom-atom pairs, see the text, so for the N. E.g. $Z = 106 = 100 \times Z1 + Z2 \text{with} Z1 \leq Z2 \Rightarrow Z1 = 1 \text{and} Z2 = 6 \Rightarrow$ it is a H–C atom pair, etc. A set of 115 molecules (Table 3) with their Mulliken matrices for partial charge has generated the range of these 106 parameters

Atom	N	Z	$E_{\text{fitpar}}(N, Z)$
H	2	1	–0.06622464
C	3	6	–0.23074779
C	4	6	–0.16366323
C	5	6	–0.19885719
C	6	6	–0.23999147
N	5	7	–0.33387528
N	6	7	–0.30265901
N	7	7	–0.32320268
N	8	7	–0.31547675
O	7	8	–0.27245124
O	8	8	–0.37441258
O	9	8	–0.42794560
F	8	9	–0.56146588
F	9	9	–0.45917123
F	10	9	–0.40105467
Si	10	14	–1.37396128
Si	11	14	–0.24089557
Si	12	14	–0.19384219
P	11	15	–0.36525803
P	12	15	–0.51376891
P	13	15	–0.54378488
P	14	15	–0.58356695
S	12	16	–0.47375909
S	13	16	–0.61628751
S	14	16	–0.61866633
S	15	16	–0.65943013
S	16	16	–0.70051863
H H	100	101	0.02466176
H H	101	101	–0.00002474
H C	105	106	0.02908445
H C	106	106	0.00010452
H C	107	106	–0.01891420
H N	106	107	–0.00785187
H N	107	107	0.00064296
H N	108	107	–0.00441181
H O	107	108	0.04957125
H O	108	108	0.00061670
H O	109	108	0.01110888
H F	108	109	–0.23221507
H F	109	109	0.00317488
H F	110	109	0.06218120
HSi	113	114	–0.36008531
HSi	114	114	–0.17045459
HSi	115	114	0.15404294
H P	114	115	0.04703979
H P	115	115	–0.00996426
H P	116	115	–0.00196052
H S	115	116	0.00484868
H S	116	116	0.00099813
H S	117	116	0.00550390
C C	605	606	0.03838783
C C	606	606	–0.00024197
C C	607	606	–0.03750534
C C	608	606	–0.04927484
C N	606	607	–0.01010163
C N	607	607	–0.00009908

Table 2 (Contd.)

Atom	N	Z	$E_{\mathrm{fitpar}}(N, Z)$
C N	608	607	–0.01659326
C O	607	608	0.05062429
C O	608	608	–0.00144555
C O	609	608	–0.02567958
C F	608	609	–0.05229206
C F	609	609	–0.01489223
C F	610	609	0.15527481
CSi	613	614	0.17134387
CSi	614	614	0.34342506
CSi	615	614	0.69324456
C P	614	615	0.03070217
C P	615	615	0.02223860
C S	615	616	–0.04885000
C S	616	616	0.00237198
C S	617	616	–0.01438090
N N	706	707	–0.07637497
N N	707	707	0.00162535
N N	708	707	0.01900106
N O	707	708	0.02826211
N O	708	708	–0.00020507
N F	708	709	0.31644475
N F	709	709	0.01304137
N F	710	709	0.12844456
O O	807	808	0.04144183
O O	808	808	0.00205831
O O	809	808	–0.06846057
O F	808	809	–0.33836067
O F	809	809	0.01181880
OSi	814	814	–0.28877492
OSi	815	814	–0.23619244
O P	814	815	–0.06232572
O P	815	815	–0.00480354
O P	816	815	–0.01175658
O S	815	816	–0.00606231
O S	816	816	–0.00386373
O S	817	816	–0.01576410
F F	908	909	–0.44396153
F F	909	909	0.00270366
FSi	914	914	–0.12826953
FSi	915	914	0.29276534
F P	915	915	0.00146196
F P	916	915	0.04812473
SiSi	1413	1414	0.58203406
SiSi	1414	1414	0.66704466
SiSi	1415	1414	1.36228331
P P	1514	1515	–0.08698784
P P	1515	1515	0.02160504
P P	1516	1515	0.07410262
S S	1615	1616	–0.35180004
S S	1616	1616	0.02064828

Fitted $E_{\mathrm{fitpar}}(N, Z)$ REBECEP/6-31G* atomic parameters in hartree to use in Eq. (4) for calculating correlation energy and basis set error via Eq. (3) for closed shell neutral molecules in the vicinity of optimum geometry from HF-SCF/6-31G* Mulliken charges containing atoms listed to correct HF-SCF/6-31G* total ground state electronic energy via Eq. (2) to achieve e.g. B3LYP quality. N is the number of electrons and Z is the atomic charge

The first three lines are just heading. After that the atomic contribution of the correlation energy and basis set error of four individual atoms is listed, followed by the 12 atom-atom pair contributions. Notice that in Eq. (3) the M is not 4 but 16, four atoms plus the 12 atom-atom pairs. Let us look in detail at the first atom-atom contribution, all the other lines are analogues. The first number is the fictitious atomic number 608, i.e. it is a C–O atom pair (again any C and O in the molecule, not necessarily chemically bounded). The next value is the Mulliken matrix type partial charge −0.5226450. Because it is negative, it means there are extra electrons on it. Now, as indicated above, one needs the Z=608 fictitious atom with $N1 = Z$ and $N2 = Z + 1$ electron content, i.e. $N1$ = 608 and $N2$=609 electrons. It is obvious, that these fictitious 608 and 609 electrons DO NOT appear in the electronic wave function (which is now a 8+6+1+1=16 electron wave function in the case of formaldehyde). It is necessary only because in an algorithm we must attribute a value to the C–O atom pair. The $E1_{\mathrm{fitpar}}$ and $E2_{\mathrm{fitpar}}$ are the two values in the right hand side of Eq. (4) for linear interpolation taking from Table 2 for Z = 608 with $N1$ = 608 and $N2$ = 609. The N_A value is $Z-$(part.charge) = 608−(−0.5226450) = 608.5226450 from the previous columns. Eweighted = −0.0141113 hartree is the value in the left hand side of Eq. (4), the REBECEP contribution from this C–O atom pair. The four individual atom contributions are analogue, even easier to comprehend, since there is no fictitious atomic number. After completing all the 16 contributions, one simply adds the Eweighted values according to Eq. (3), to obtain a value of −0.63989331 hartree, which compares to the B3LYP/6-31G* value of −0.63645253 hartree. The difference is 0.00344078 hartree (2.2 kcal/mol) see molecule# 13 in Table 3.

The Mulliken charge case is similar:

```
>Formaldehyde (H2C=O)
>   Z   parc.chrg. N1     NA    N2 E1fitpar  E2fitpar Eweighted
>[a.u.]  [a.u.]                   [hartree] [hartree] [hartree]
>   8   -0.4363410  8  8.4363410 9 -0.3692171-0.3790543-0.3735095
>   6    0.1346430  5  5.8653570 6 -0.2098112-0.2374576-0.2337352
>   1    0.1508490  0  0.8491510 2  0.0000000 -0.0417254-0.0177156
>   1    0.1508490  0  0.8491510 2  0.0000000 -0.0417254-0.0177156
>
>an accurate Ecorr, Ecorr(REBECEP) [hartree]= -0.63645253,
                                      -0.64267586
>an accurate Ecorr- Ecorr(REBECEP) [hartree]= 0.00622333
```

Here the difference is 0.00622333 hartree (3.9 kcal/mol) see again molecule# 13 in Table 3. One should notice that the −0.63645253 hartree correlation energy and basis set error on B3LYP/6-31G* level, about 400 kcal/mol value was estimated now. This final REBECEP/6-31G* value must be added to the −113.863712881 hartree HF-SCF/6-31G* value (see the extract of the Gaussian 98 output at beginning of this section) to get an estimation for the total electronic energy in ground state in case of neutral, ground state, covalent, closed shell molecules containing H, C, N, O, F, Si, P, S atoms in the vicinity of their equilibrium geometry. We have two choices now, the "Mulliken matrix" and "Mulliken charge" cases – these will be discussed and compared in the next section (Table 3). (Charged, e.g. protonated closed shell molecules can also be calculated in this way if (N, Z) in Tables 1 and 2 is wide enough, otherwise the REBECEP parameter set must be extended using charged molecules in the set).

Table 3 List of molecule set used in the linear fit to get the REBECEP/6-31G* parameter set (Tables 1, 2). The column B3LYP ≡ (B3LYP/6-31G*)−(HF-SCF/6-31G*) is the B3LYP level correlation and basis set error in hartree. The last two columns show the deviation from B3LYP total ground state electronic energies in kcal/mol belonging to Mulliken charge [CHARGE ≡ (B3LYP/6-31G*)− (REBECEP/6-31G*/Mulliken charge)] and Mulliken matrix [MATRIX ≡ (B3LYP/6-31G*) − (REBECEP/6-31G*/Mulliken matrix)] methods. These two methods are supposed to reproduce the B3LYP/6-31G* correlation energy and basis set error correction via Eqs. (3) and (4) and the total ground state electronic energy via Eq. (2). In the "CHARGE" case the root mean square deviation is 3.3 kcal/mol, which is improved in "MATRIX" method to 1.5 kcal/mol. (one hartree is about 627.5 kcal/mol)

Molecule	B3LYP [hartree]	CHARGE [kcal/mol]	MATRIX [kcal/mol]
1 Methane (CH4)	–0.3233	–1.1	0.2
2 Ammonia (NH3)	–0.3641	0.5	0.4
3 Water (H2O)	–0.3991	1.4	0.3
4 Hydrogen fluoride (HF)	–0.4179	4.1	0.3
5 Silane (SiH4)	–0.6586	1.3	0.0
6 Phosphine (PH3)	–0.6924	0.7	–0.1
7 Hydrogen sulfide (H2S)	–0.7182	0.4	0.2
8 Acetylene (C2H2)	–0.5098	4.6	2.4
9 Ethylene (C2H4)	–0.5563	1.6	1.1
10 Ethane (C2H6)	–0.6018	–0.5	0.7
11 Hydrogen cyanide (HCN)	–0.5516	6.3	2.2
12 Carbon monoxide (CO)	–0.5746	16.6	2.9
13 Formaldehyde (H2CO)	–0.6365	3.9	2.2
14 Methanol (CH3OH)	–0.6802	1.3	1.1
15 Hydrazine (H2N–NH2)	–0.6884	0.7	0.6
16 Hydrogen peroxide (H2O2)	–0.7730	–2.0	0.5
17 Carbon dioxide (CO2)	–0.9522	4.6	1.7
18 Silicon monoxide (SiO)	–0.9446	–3.6	0.0
19 Carbon monosulfide (CS)	–0.9057	7.5	0.1
20 Disilane (H3Si-SiH3)	–1.2777	1.2	0.0
21 Methanethiol (H3CSH, staggered)	–0.9981	1.5	1.1
22 Sulfur dioxide (SO2)	–1.4291	–4.8	1.3
23 CF4	–1.8348	1.3	–0.6
24 OCS (linear)	–1.2793	–0.6	0.1
25 CS2 (linear)	–1.6053	–4.9	0.0
26 COF2	–1.3974	1.1	0.8
27 SiF4	–2.1605	0.7	0.0
28 N2O	–0.9944	5.5	2.3
29 NF3	–1.5393	10.2	0.4
30 PF3	–1.8291	1.5	0.0
31 F2O	–1.2048	0.6	0.6
32 C2F4	–2.0878	–6.5	–0.2
33 CF3CN	–1.9734	1.4	0.1
34 Propyne (C3H4)	–0.7911	3.1	1.9
35 Allene (C3H4)	–0.7974	–1.4	0.8
36 Cyclopropene (C3H4)	–0.7972	–0.4	0.1
37 Propene (C3H6)	–0.8368	1.0	1.3
38 Cyclopropane (C3H6)	–0.8366	1.2	1.5
39 Propane (C3H8)	–0.8808	–0.3	0.8
40 trans-1,3-Butadiene (C4H6)	–1.0739	1.3	0.4
41 Dimethylacetylene (2-butyne)	–1.0720	2.5	–1.1
42 Methylenecyclopropane (C4H6)	–1.0760	–0.5	0.9
43 Bicyclo[1.1.0]butane (C4H6)	–1.0771	–0.7	0.4
44 Cyclobutene (C4H6)	–1.0746	0.8	1.4
45 Cyclobutane (C4H8)	–1.1163	1.0	1.2
46 Isobutene (C4H8)	–1.1173	–0.1	1.0
47 trans-Butane (C4H10)	–1.1599	0.0	0.7
48 Isobutane (C4H10)	–1.1601	–0.2	0.5
49 Spiropentane (C5H8)	–1.3541	0.6	1.0
50 Benzene (C6H6)	–1.5466	2.7	2.6
51 Difluoromethane (H2CF2)	–1.0783	2.5	–0.9
52 Trifluoromethane (HCF3)	–1.4567	2.7	1.1
53 Methylamine (CH3NH2)	–0.6440	0.4	1.4
54 Acetonitrile (CH3CN)	–0.8317	4.8	1.2
55 Nitromethane (CH3NO2)	–1.3550	3.2	1.4
56 Methylnitrite (CH3-O-N=O)	–1.3478	5.7	2.3
57 Methylsilane (CH3SiH3)	–0.9385	1.2	0.0
58 Formic Acid (HCOOH)	–0.9967	2.5	–0.2
59 Methyl formate (HCOOCH3)	–1.2771	2.3	2.2
60 Acetamide (CH3CONH2)	–1.2384	0.9	1.1
61 Aziridine (cyclic)	–0.8820	0.5	0.8
62 Cyanogen (NCCN)	–1.0748	1.5	0.4
63 Dimethylamine ((CH3)2NH)	–0.9248	–0.1	1.3
64 trans-Ethylamine (C2H5-NH2)	–0.9238	0.2	1.6
65 Ketene (H2C=C=O)	–0.8766	0.4	–2.5
66 Oxirane (cyclic -CH2-O-CH2-)	–0.9207	0.5	5.1
67 Acetaldehyde (CH3CHO)	–0.9164	2.5	2.7
68 trans-Glyoxal (O=CH-CH=O)	–1.2316	6.4	4.9
69 trans-Ethanol (CH3CH2OH)	–0.9593	1.4	2.0
70 Dimethyl-ether (CH3-O-CH3)	–0.9616	0.9	0.2
71 Thiooxirane (cyclic -CH2-S-CH2-)	–1.2366	1.3	1.2
72 Dimethylsulfoxide ((CH3)2SO)	–1.6503	5.7	0.3
73 Thio-ethanol (C2H5-SH)	–1.2773	1.6	1.1
74 Dimethyl-thio-ether (CH3-S-CH3)	–1.2785	2.9	0.8
75 Vinyl fluoride (H2C=CHF)	–0.9388	0.3	0.5
76 Cyano-ethylene (H2C=CHCN)	–1.0688	5.2	1.0
77 Acetone (CH3-CO-CH3)	–1.1955	1.7	3.4
78 Acetic acid (CH3COOH)	–1.2746	2.3	2.3
79 Acetyl fluoride (CH3COF)	–1.2971	1.4	1.9
80 Isopropyl alcohol ((CH3)2CH-OH)	–1.2392	0.5	2.2
81 Methyl,ethyl-ether (C2H5-O-CH3)	–1.2407	0.9	1.1
82 Trimethyl amine ((CH3)3N)	–1.2058	–0.9	–0.2
83 Furan (C4H4O, cyclic)	–1.3980	–1.6	–2.7
84 Thiophene (C4H4S, cyclic)	–1.7140	3.6	0.2
85 Pyrrole (C4H4NH, cyclic)	–1.3599	–1.6	–0.2
86 Pyridine (C5H5N, cyclic)	–1.5912	2.3	0.8
87 H3PO4	–2.1154	3.8	–0.4
88 H2CH3PO4	–2.3959	2.5	0.0
89 HO-CH3O-PO-O-PO-OCH3-OH	–4.3942	–0.8	0.8
90 C2H5SO3H	–2.3690	0.6	0.1
91 C6H5SO3H	–3.3170	1.6	0.3
92 (CH3)3PO	–1.8882	0.1	1.3
93 (CH3)2POH	–1.6112	0.5	0.7
94 CH3CONH2	–1.2377	1.2	1.0
95 CH3POH2	–1.3321	2.1	0.5
96 CH3SO3H	–2.0894	1.2	0.2
97 Methyl-ciclopentan	–1.6726	2.6	1.2
98 Methyl-cyclobuthan	–1.3963	0.6	0.6
99 Methyl-cyclohexan	–1.9521	2.5	1.2
100 Alanine	–1.8763	2.5	2.5
101 Allenyl-CH3	–1.0763	–0.9	1.5
102 Glycine	–1.5965	2.9	1.3
103 m-Methyl-Ethyl-Benzene	–2.3877	0.6	0.4
104 o-Methyl-Ethyl-Benzene	–2.3885	0.4	0.0
105 p-Methy- Ethyl-Benzene	–2.3879	0.6	0.3
106 Trinitro-toluol (TNT)	–4.9438	–7.2	1.3
107 Valine	–2.4362	1.8	–1.0
108 CH3-NH-CH2-NH-CH3	–1.5259	0.9	2.4
109 CH3-NH-NH2	–0.9696	–0.1	1.2
110 CH3-POH-CH2-POH-CH3	–2.9003	–0.9	0.5
111 CH3-POH-POH-CH3	–2.6305	0.3	–0.8
112 CH3-SiH2-CH2-SiH3	–1.8345	2.9	0.0
113 CH3-SiH2-SiH2-CH3	–1.8388	0.3	0.0
114 NH2-CH2-NO2	–1.6782	2.6	0.8
115 Quinuclidine N(CH2CH2)3CH	–2.2332	3.1	0.8

4 Results and discussion

Here we show the results obtained for REBECEP atomic parameters. Tables 1 and 2 shows the fitted values to reproduce B3LYP/6-31G* quality total energies by Eqs. (3) and (4) as described above. The above procedure also shows how the correlation and basis set error can be partitioned to atoms (Mulliken charge case) and atom and atom-atom pairs (Mulliken matrix case). These parameters were obtained by the linear fit indicated above using the ground state energies and partial charges of 115 molecules composed of H, C, N, O, F, Si, P, S atoms listed in Table 3. (As indicated in our previous works, this method cannot be applied for individual atoms and homonuclear diatoms, but this is not a serious restriction. The reason is that in these systems the partial charges are always zero). With these atom types, proteins, DNS, RNS and a large range of organic molecules can be calculated, and the restriction is only that which the HF-SCF/6-31G* calculation has. As a consequence, the CPU time and disc space needed is the same as HF-SCF/6-31G* needs, because the Mulliken analysis and the REBECEP/6-31G* method are instant.

As analyzed below, the Mulliken matrix case is more accurate by the fact that it has more parameters. Additional information must be added here. Because the Mulliken matrix REBECEP/6-31G* parameter set contains so many parameters, the 115 molecule set was not big enough to get a stable linear fit. For this reason, all the first heavy atoms (i.e. not hydrogen x coordinate) in the molecules were changed to +/−0.19 Å to get more molecular geometries in the vicinity of optimum geometry. In case of formaldehyde above, particularly it was the oxygen atom. Generally, this change shifts the total ground state energy by about 5–30 kcal/mol for molecules listed in Table 3. All together, $3\times115=345$ molecular geometries were considered in the fitting procedure to obtain the REBECEP/6-31G* parameters for Mulliken charge and Mulliken matrix cases in Tables 1, 2, and 3. (More exactly, we excluded some identical cases coming from the symmetry. For example in case of#85 (pyrrole), the first heavy atom listed was the N atom. Being a planar molecule and placed in the (y, z) plain, the change in x coordinate by +/−0.19 Å created the same geometry, i.e. not two new cases but only one new geometry was created. In this way its B3LYP value in Table 3 changed from −1.3599 hartree to −1.3604 hartree, and the CHARGE value from −1.6 to −1.9 kcal/mol, as also the MATRIX value from −0.2 to −0.4 kcal/mol). The root mean square deviation values in Table 3 (3.3 and 1.5 kcal/mol) belong to these increased number of cases, not only to the 115 equilibrium geometries listed.

Finally, we have obtained transferable parameters, good for any other molecules outside of the set. In other words, it is expected that good quality total energy can be obtained for any other molecules provided that the parameter set in Tables 1 and 2 is applicable for it (i.e. it consists of the atoms listed and the necessary HF-SCF/6-31G* results are available). Again, in the Mulliken matrix case, the atom-atom pair parameters do not require necessarily bounded neighbor atoms in the molecule, but can be in any position. In both cases, we have obtained stable REBECEP parameters from the linear fit, and they change systematically in the table. In the case of Mulliken matrix the atomic parameters are larger in absolute value than the atom -atom pair parameters and always negative (correlation energy is always negative). Furthermore, these lower absolute value atom-atom parameters have alternate signs.

In Table 3 we show the B3LYP/6-31G* energy corrections to the HF-SCF/6-31G* total ground state electronic energies, and the deviation of REBECEP Mulliken charge case (CHARGE) and matrix case (MATRIX) from it (see the exact definitions in the table head). All of these energies are listed for the set of 115 closed shell molecules previously mentioned. The B3LYP/6-31G* type of energy corrections in Table 3 and the corresponding Mulliken matrix ("Condensed to atoms (all electrons)") and charges ("Total atomic charges") from a HF-SCF/6-31G* calculation (not listed) were the input for the fitting procedure that resulted in the fitted REBECEP atomic energy parameters in Tables 1 and 2 (more exactly the extended set via the manipulation of the first heavy atom x coordinate). The CHARGE and MATRIX values in Table 3 constitute the outcome which represents the quality of the fit. We do not report here the calculated partial charges in order to save space, these charges can be easily calculated e.g. by the GAUSSIAN program package [17, 26]. We note that the speed increase by REBECEP method compared to the B3LYP method in CPU time is a factor of 2. The required disc space is the same. (In the Kohn–Sham equations/method, what the B3LYP method uses, a HF-SCF procedure is developed in the programming using some DFT functionals which correct the energy outcome). REBECEP scales as the HF-SCF/6-31G* method scales with the increase of the size of the molecule. More importantly, the two REBECEP procedures above show how the energy correction can be partitioned among atomic parameters. For molecules not included in the database here, we suggest using the experimental geometries if available. If experimental geometries are not available, probably the best choice would be using the B3LYP/6-31G* equilibrium geometries, because it was previously found that such geometries are useful alternatives to the considerably more expensive MP2 geometries [32].

These results show that the fitted REBECEP parameter set is capable of providing ground state total electronic energies approaching chemical accuracy for the selected 115 molecules (closed shell, neutral and covalent in near equilibrium geometry). The root mean square deviation from the B3LYP total energies of the Mulliken charge method (CHARGE in Table 3) is 3.3 kcal/mol for the test set of 115 molecules, calculated using the parameters in Table 1, while for Mulliken matrix method it is 1.5 kcal/mol (MATRIX in Table 3), calculated using the parameters in Table 2. Thus, the REBECEP method with Mulliken matrix approximates closer, very likely because it uses more parameters.

Analysis of the results in Table 3 reveals that the REBECEP Mulliken matrix method strongly improves the calculation compared to Mulliken charge case. See e.g. molecules

like#12 (CO),#29(NF_3),#32(C_2F_4) or#106(TNT). It is fundamental to recognize that the parameters in Tables 1 and 2 are applicable only for closed shell molecules in the vicinity of their stationary points, because all types of REBECEP parameters should converge toward the corresponding $E_{corr}(N, Z)$ high spin atomic correlation energies in free space (now including the 6-31G* basis set error) as the molecule separates into atoms in free space and the partial charges converge toward zero (or Z-N). This is a basic property of the physics of REBECEP parameters [19–21]. If the basis goes to the infinite basis in the HF-SCF algorithm, the REBECEP parameters converge to a parameter set which contains only correlation correction (if the reference calculation ("method") is accurate). The restrictions, atom types and vicinity of stationary points and closed shells, come from the training set of 115 molecules used to obtain the parameters. Further extensions are straightforward. These are: inclusion of new atom types, charged molecules (for the effects of protonation and deprotonation, etc.), and radicals. (This latter case probably requires different parameters from the closed shell parameters).

We make a short note about the number of REBECEP parameters. One generally needs the ionic states of the eight atoms listed in Table 1 for the method of "CHARGE". Except hydrogen, 2–4 ionic states of atom types appear, basically it is determined by the Mulliken partial charge as they arise in molecules. (The Mulliken partial charge values, in fact, originate from the electronegativity). This makes 23 parameters necessary in Table 1. The method of "MATRIX" in Table 2 is a bit more complex. The number of individual atomic parameters is similar to Table 1, although between 3 to 5 ionic states arise for non-hydrogen atoms, instead of 2–4, as before. For atom-atom pairs, the eight types of atoms make $8\times7/2=28$ possible couples with different atoms (e.g. H–C) as well as eight couples containing the same atoms (e.g. H–H, C–C, etc.). And these couples are listed 2–4 times as their "ionized" states appear (see the fictitious (N, Z) values). Some couples are missing, since certain atoms pair together rarely. We excluded the chlorine atom from our consideration because of the relativistic effects involved. All together, Table 2 ended up with 106 parameters. One can see that, e.g. for sulfur, the range from S^{2+} to S^{-} (i.e. Z = 16 with N = 14,15,16 and 17) in Table 1 changes to S^{4+} to S (i.e. Z = 16 with N = 12,13,14,15 and 16) in Table 2. Its origin lies in the difference between what we have defined with CHARGE and MATRIX cases above (see Sect. 3).

As a last example, we mention the case of the cinchonidine molecule. It is a 44 atom alkaloid which is important in the enantioselective hydrogenation of pyruvates [33]. This molecule was not in the example set (Table 3). Its structure is a threefold substituted methane with OH group, quinoline and quinuclidine (# 115 in Table 3). The HF-SCF/6-31G* calculation on a 2 GHz PC machine with Gaussian 98 requires about 14 min and 69 Mb disc space, while the B3LYP/6-31G* needs 21 min. The HF-SCF/6-31G* energy of the "closed" [33] form is −916.006160 hartree and the B3LYP/6-31G* energy is lower by −5.933366 hartree. The REBECEP/6-31G* with Mulliken charge approximates the latter one within 0.003 hartree instantly.

As a last note, we mention, that while the DFT functionals in the Kohn–Sham formalism (like the B3LYP) make the corrections inside the algorithm using the (spin) electron density, the REBECEP makes the correction outside of the HF-SCF routine, based on some kind of weighted summation of Mulliken partial charges, however, the partial charges are always some measure (integrals) of the electron density. Other authors have used the Mulliken matrix (or Mulliken population analysis) for drawing chemical conclusions, by interpreting the numerical quantities as atomic charges and bond populations (see first paragraph of Sect. 3), has little direct relevance to the present (REBECEP) use of these quantities, hence those criticism do not apply here.

5 Conclusions

With our new REBECEP atomic parameter sets for Mulliken charge and matrix, it is possible to reproduce good quality total ground state electronic energy from single point small basis set ab initio HF-SCF/6-31G* total energy calculations including Mulliken analysis. The only time consuming step in the REBECEP method is the HF-SCF/6-31G* calculation, which determines the CPU time requirement and limits the size of molecule. The two REBECEP atomic parameter sets were optimized for 115 molecules composed of the H, C, N, O, F, Si, P and S atoms. The optimization was done by a simple multi-linear least square fit to approximate the B3LYP total electronic energies for this set of molecules using the Mulliken partial charge analysis available in commercial program packages. The instant REBECEP calculation can be used with this relatively small 6-31G* basis set. We thus conclude that the basis set errors can be absorbed by the parameter set (i.e. it can be treated at atomic level). The REBECEP total energies calculated with the above-mentioned parameters and with Mulliken matrix and charges can approach the required accuracy (cf. Tables 1, 2, 3). The root mean square deviation of REBECEP total energies from the B3LYP/6-31G* total energies is 1.5 kcal/mol in Mulliken matrix case of REBECEP. The REBECEP atomic parameters are listed in Tables 1 and 2 and recommended for calculating molecular ground state total electronic energies of neutral closed shell molecules in the vicinity of their equilibrium, containing the atoms listed. We have also shown that use of Mulliken matrix contra Mulliken charge in REBECEP method improves, as one uses more parameters (these are the atom-atom pair values beside the atomic values); but both calculations are instant.

Supplementary material

The molecular geometries and the HF-SCF/6-31G* partial charges of the 115 molecules are available from the author.

The FORTRAN program that calculates the REBECEP parameters and/or the REBECEP correlation and basis set error energies is also available to the reader via email or from the web site web.inc.bme.hu/∼kristyan.

References

1. Curtiss LA, Raghavachari K, Redfern PC, Rassolov V, Pople JA (1998) J Chem Phys 109:7764
2. Curtiss LA, Raghavachari K, Redfern PC, Pople JA (1997) J Chem Phys 106:1063
3. Curtiss LA, Redfern PC, Raghavachari K, Pople JA (1998) J Chem Phys 109:42
4. Curtiss LA, Raghavachari K, Redfern PC, Pople JA (2000) J Chem Phys 112:7374
5. Ochtersky JW, Petersson GA, Montgomery JrJA (1996) J Chem Phys 104:2598
6. Fast PL, Sanchez ML, Corchado JC, Truhlar DG (1999) J Chem Phys 110:11679
7. Martin JML, de Oliveira GJ (1999) Chem Phys 111:1843
8. Kristyan S, Pulay P (1994) Chemical Physics Letters 229:175
9. Kristyan S (1995) J Chem Phys 102:278
10. Kristyan S (1995) Chem Phys Lett 247:101
11. Kristyan S (1996) Chem Phys Lett 256:229
12. Parr RG, Yang W (1989) Density functional theory of atoms and molecules. Oxford University Press, New York
13. Dreizler RM, Gross EKV (1990) Density functional theory. Springer, Berlin Heidelberg New York
14. Becke AD (1993) J Chem Phys 98:5648
15. Perdew JP (1991) In: Ziesche P, Eschrig H, (eds) Electronic structure of solids. Akademie Verlag, Berlin Heidelberg New York, p 11
16. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785
17. 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 D, DeFrees DJ, Baker J, Stewart JJP, Pople JA (1993) Gaussian 92/DFT. Gaussian, Pittsburgh PA
18. Kristyan S (1997) Chem Phys 224:33
19. Kristyan S, Csonka GI (1999) Chem Phys Lett 307:469
20. Kristyan S, Csonka GI (2001) J Comput Chem 22:241
21. Kristyan S, Ruzsinszky A, Csonka GI (2001) J Phys Chem A 105:1926
22. Kristyan S, Ruzsinszky A, Csonka GI (2001) Theoret Chem Account 106:404
23. Kristyan S, Mol J (2004) Struct Theochem 712:153
24. Ruzsinszky A, Kristyan S, Margitfalvi JL, Csonka GI (2003) J Phys Chem A 107:1833
25. Löwdin PO (1959) Adv Chem Phys 2:207
26. 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, Inc., Pittsburgh, PA
27. Hehre WJ, Radom L, Schleyer PR, Pople JA (1986) Ab initio molecular orbital theory. Wiley, New York
28. Szabo A, Ostlund NS (1982) Modern quantum chemistry: introduction to advanced electronic structure theory. McMillan, New York
29. Hehre WJ, Huang WW, Klunzinger PE, Deppmeier BJ, Driessen AJ (1997) *Spartan Manual*, Wavefunction, Inc., 18401 Von Karman Ave., Suite 370, Irvine, CA 92612
30. Kristyan S, Ruzsinszky A, Csonka GI (2001) Theoret Chem Account, 106:319
31. Mulliken RS (1962) J Chem Phys 36:3428
32. Baboul AG, Curtiss LA, Redfern PC, Raghavachari KJ (1999) Chem Phys 110:7650
33. Margitfalvi J, Tfirst E (1999) J Mol Cat A Chem 139:81–95