



Viewpoint

Organizing atomic partial charges into a database

Gábor I. Csonka^{a,*}, Adrienn Ruzsinszky^{a,1}, Julianna Oláh^{a,b,1}, Christian Van Alsenoy^b^aDepartment of Inorganic Chemistry, Budapest University of Technology and Economics, H-1521 Budapest, Hungary^bDepartment of Chemistry, University of Antwerp UIA, Universiteitsplein 1, B-2610 Antwerpen, Belgium

Received 13 March 2002; accepted 24 April 2002

Abstract

A partial charge database was constructed for 117 molecules. The actual database contains Mulliken, natural population analysis, and stockholder atomic charges calculated from the HF/6-31G(d) or HF/6-311+G(2d,p) wavefunctions. The MP2/6-31G(d), B3LYP/6-31G(d), or HF/6-31G(d) equilibrium geometries were used for these calculations. The database also contains the corresponding total energies calculated at HF/6-31G(d) or HF/6-311+G(2d,p) levels of theory. Gaussian 3 or G3SX [Curtiss L.A.; Redfern, P.C.; Raghavachari, K.; Pople, J.A. *J.Chem. Phys.* 114 (2001) 108.] composite energies with experimental, zero point vibration energy, or thermal corrections are included. The database is used to obtain atomic correction parameters of Rapid Estimation of Basis set Error and Correlation Energy from Partial charges method [S. Kristián, A. Ruzsinszky, G.I. Csonka, *Theoret. Chem. Acc.* 106 (2001) 319 and 404]. An XML database structure is also suggested. © 2002 Published by Elsevier Science B.V.

Keywords: Mulliken atomic charges; Natural population analysis atomic charges; Stockholder atomic charges

Atomic partial charges in the molecules play an important role in understanding the molecular electron distribution and aspects of molecular reactivity. Partial charges are essentially mathematical constructions that serve to reflect the electron content around the selected atom of the molecule. These partial charges are not physically measurable quantities, as they depend on the definition of the partition scheme of the electron density. The partitioning of the molecular electron density can be done either in the Hilbert space spanned by the atomic orbital basis of the molecular wave function or in real the three-dimensional space. The general problem of any

population analysis method consists in the partitioning of the molecular electron density and its assignment to the atoms that compose the molecule. The primary use of partial charges is to help chemists to establish empirical rules, e.g. for describing chemical bonds. However, partial charges can be defined to reproduce the measurable dipole moment and electrostatic potential of the molecule, for example. It should be noted that atomic multipole moments are also required for correct dipole moment predictions (due to nonspherical electron distribution about an atom in the molecule). The partial charges were successfully applied to identify the electron rich (nucleophil) and electron poor (electrophil) functional groups of molecules. Thus partial charges in ideal case are able to represent (in a simplified manner) the electron distribution in a molecule.

* Corresponding author.

E-mail address: csonka@web.inc.bme.hu (G.I. Csonka).

¹ <http://web.inc.bme.hu/csonka>

```

<?xml version="1.0" ?>
<!DOCTYPE rebecca
[
<!ELEMENT rebecca (version, comment, molecule+)>
  <!ELEMENT version (#PCDATA)>
  <!ELEMENT comment (#PCDATA)>
  <!ELEMENT molecule (name, geometry+, tenergy+)>
    <!ATTLIST molecule id CDATA #IMPLIED>
    <!ATTLIST molecule natoms CDATA #IMPLIED>
  <!ELEMENT name (#PCDATA)>
  <!ELEMENT geometry (atoms, energy+)>
    <!ATTLIST geometry opt CDATA #IMPLIED>
  <!ELEMENT atoms (atom+)>
    <!ATTLIST atoms xyzunit CDATA #IMPLIED>
    <!ELEMENT atom (atsym, atnum, x, y, z, id, charge+)>
      <!ELEMENT atsym (#PCDATA)>
      <!ELEMENT atnum (#PCDATA)>
      <!ELEMENT x (#PCDATA)>
      <!ELEMENT y (#PCDATA)>
      <!ELEMENT z (#PCDATA)>
      <!ELEMENT id (#PCDATA)>
      <!ELEMENT charge (#PCDATA)>
        <!ATTLIST charge wavefn CDATA #IMPLIED>
        <!ATTLIST charge method CDATA #IMPLIED>
    <!ELEMENT energy (#PCDATA)>
      <!ATTLIST energy wavefn CDATA #IMPLIED>
      <!ATTLIST energy unit CDATA #IMPLIED>
  <!ELEMENT tenergy (#PCDATA)>
    <!ATTLIST tenergy method CDATA #IMPLIED>
    <!ATTLIST tenergy correction CDATA #IMPLIED>
    <!ATTLIST tenergy fit CDATA #IMPLIED>
    <!ATTLIST tenergy unit CDATA #IMPLIED>
]
>

```

Fig. 1. Suggested document type definition for atomic charge database.

Our efforts to use the atomic charges for rapid estimation of correlation charges [1] require a reliable partial charge database. In the current database we collect the Mulliken [2] natural population analysis (NPA) [3], and stockholder [4] atomic partial charges of 117 molecules. These atomic charge calculation methods will not be described here in full detail; the interested reader should refer to the literature. The Mulliken population analysis is done in the atomic orbital basis and it evenly splits the electrons that arise ‘between’ two atoms via overlap population. This method can be programmed readily. However, nonphysical results might occur (e.g. negative orbital population or orbital populations greater than 2) [2]. One of the best orbital-based population analyses is the NPA, in which the atomic orbitals are defined by the density distribution. Natural atomic orbitals (NAO) are eigen-

functions of atomic sub-blocks of the density matrix. The NAOs are orthonormal atomic orbitals of maximum occupancy and usually localized about a single atom. The summation of the occupancies of the NAOs gives the atomic population for a specific atom [3]. The Mulliken and NPA methods give atomic populations by summing the occupancies of atomic orbitals.

Fundamentally different approaches are counting the electrons in a given region of space about an atom. The stockholder recipe requires a superposition of the free-atom electron densities placed on the positions the atoms occupy in the molecule (promolecule density). The electron density in each point in space is divided among the atoms of the molecule in proportion to their respective contributions to the promolecule density at that point. The charge of an atom is obtained by integrating the charge density assigned to a given atom. This method is advantageous in several respects as Hirshfeld stated [4]: “the stockholder recipe is completely general, requires no arbitrary placing of boundary surfaces, and produces well-defined atomic fragments that differ from the free atoms only to the degree that the molecule itself differs from a superposition of free-atom densities”.

In the current version of the database the HF/6-31G(d) and HF/6-311 + G(2d,p) charges were calculated at MP2/6-31G(d),² B3LYP/6-31G(d),² and HF/6-31G(d) equilibrium geometries. The MP2 and B3LYP geometries were taken from the input files published on the internet,² the HF geometries, the charges, and the total energies were calculated using GAUSSIAN 98 program [5].

In order to obtain the stockholder charges [6] electron densities of atoms in a specific spectroscopic state were calculated at the HF–SCF level of theory with ATOMSCF [7]. In all cases the atomic electron densities were calculated with the same basis set as was used to calculate the molecular electron density and, unless otherwise indicated, the atomic densities were calculated for atoms in their spectroscopic ground state. The numerical integration procedure used in this database is based on the procedure proposed by Becke [8] making use of Euler–Maclaurin [9] and Lebedev [10] grids for the

² <http://chemistry.anl.gov/compmat/compterm.htm>

```

<?xml version="1.0" ?>
<!DOCTYPE rebecca (View Source for full doctype...)>
<rebecca>
<version>1.0</version>
<comment>REBECCA Program: The applied REBECEP parameters References:
http://web.inc.bmc.hu/csonka/p96.html</comment>
<molecule id="CH4" natoms="5">
  <name>Methane</name>
  <geometry opt="MP2/6-31G(d)">
    <energy wavefn="HF/6-31G(d)" unit="hartree">-40.19507</energy>
    <energy wavefn="HF/6-311+G(2d,p)" unit="hartree">-40.21020</energy>
    <atoms xyzunit="angstrom">
      <atom>
        <atsym>C</atsym>
        <atnum>6</atnum>
        <x>0.00000</x>
        <y>0.00000</y>
        <z>0.00000</z>
        <id>1</id>
        <charge wavefn="HF/6-31G(d)" method="Mulliken">-0.658545</charge>
        <charge wavefn="HF/6-31G(d)" method="NPA">-0.86663</charge>
        <charge wavefn="HF/6-31G(d)" method="Stockholder">-0.102618</charge>
        <charge wavefn="HF/6-311+G(2d,p)" method="NPA">-0.7110</charge>
      </atom>
      <atom>
        <atsym>H</atsym>
        <atnum>1</atnum>
        <x>-0.629118</x>
        <y>0.629118</y>
        <z>0.629118</z>
        <id>2</id>
        <charge wavefn="HF/6-31G(d)" method="Mulliken">0.164636</charge>
        <charge wavefn="HF/6-31G(d)" method="NPA">0.21666</charge>
        <charge wavefn="HF/6-31G(d)" method="Stockholder">0.025655</charge>
        <charge wavefn="HF/6-311+G(2d,p)" method="NPA">0.17775</charge>
      </atom>
      + <atom>
      + <atom>
      + <atom>
    </atoms>
  </geometry>
  <geometry opt="B3LYP/6-31G(d)">
    <energy wavefn="HF/6-31G(d)" unit="hartree">-40.19491374</energy>
    <energy wavefn="HF/6-311+G(2d,p)" unit="hartree">-40.21007133</energy>
    + <atoms xyzunit="angstrom">
  </geometry>
  <geometry opt="HF/6-31G(d)">
    <energy wavefn="HF/6-31G(d)" unit="hartree">-40.1951709</energy>
    + <atoms xyzunit="angstrom">
  </geometry>
  <tenergy method="G3" fit="Expt" correction="">-40.49969</tenergy>
  <tenergy method="G3" fit="Expt" correction="0K">-40.45703</tenergy>
  <tenergy method="G3" fit="Expt" correction="298K">-40.46006</tenergy>
  <tenergy method="G3SX" fit="Expt" correction="">-40.50439</tenergy>
  <tenergy method="G3SX" fit="Expt" correction="0K">-40.46028</tenergy>
  <tenergy method="G3SX" fit="Expt" correction="298K">-40.46331</tenergy>
</molecule>
+ <molecule id="NH3" natoms="4">
.
.
.
</rebecca>

```

Fig. 2. The XML layout of the database.

radial and angular parts of the atomic grids, respectively.

The simple text database extracted from the output files is currently being transformed into an extensible markup language (XML) document.³ An XML document is valid if it has an associated document type definition (DTD) and the document complies with the constraints expressed in DTD. The DTD of this database is shown in Fig. 1. The root element (*rebecca*) contains another three **elements**, **version**, **comment** and **molecule** (child elements). If an element name is followed by + the element may occur several times, so **molecule** + means that the XML document contains several molecules. The **molecule** (**name**, **geometry** + , **tenergy** +) in Fig. 1 means that the molecule element contains a name element and several **geometry** and **tenergy** (total energy) elements. The molecule element also has two attributes the **id** and **natoms**. The latter shows the number of atoms in the molecule. The **opt** attribute of the **geometry** element shows the geometry optimization method (as it was mentioned currently we use MP2/6-31G(d), B3LYP/6-31G(d), and HF/6-31G(d) geometries). The **geometry** element contains **atoms** (attribute: xyzunit). The **atoms** element contains one or more **atom** elements. The **atom** element contains atomic symbol (**atsym**), atomic number (**atnum**), **x**, **y**, **z** coordinates, identifier (**id**), and **charge** element. Depending on the wave function and charge calculating method one atom can have several different partial charges (attributes: **wavefn** and **method**). The single point energy (element: **energy**) is calculated by a given wave function. Various, very precise composite energies are also given for the molecule (element **tenergy** with **method**, **correction**, **fit**, and **unit** attributes). The database contains GAUSSIAN 3 [11] or G3SX [12] composite energies with corrections (experimental, zero point vibration energy, or thermal corrections). These energies are required for REBECEP energy calculation and for obtaining atomic correction parameters [1].

The following list shows in a simplified manner the structure of the DTD shown in Fig. 1 (the elements are shown in bold and the attributes of one element are shown in the same line, explanations are given in parentheses):

```

rebecca (root)
  version
  comment
  molecule—id natoms
    name
    geometry—opt (geometry optimization method)
      atoms
        atom—xyzunit
          atsym (atomic symbol)
          atnum (atomic number)
          x (x coordinate)
          y
          z
          id
          charge—wavefn method (partial
            charge calculated from a wavefn with
            a given method)
        energy—wavefn unit
      tenergy—method correction fit unit
  
```

Fig. 2 shows a detail of the actual XML document as shown in an actual browser. As it was already mentioned the currently used programs provide a simple text output. The necessary data from these output files were first transformed into a text database that contains the various molecules and the related data (coordinates, charges and energies) in a labeled form similar to the original output. The current version of the database contains 117 molecules. This database can be used for charge analysis and REBECEP type parameter search. The current size of the text database is about 600 printed pages and it will be continuously expanded. It is available through the internet.⁴ This database will be used for a series of papers for charge analysis, REBECEP parameter fitting, and zero-point vibration energy analysis. The final version will be transformed into an XML document described in the current paper.

Acknowledgments

This work was supported by the OTKA Grant (T 031767, and 029976 Hungary). A part of this work

³ <http://www.w3.org/TR/REC-xml>

⁴ <http://web.inc.bme.hu/csonka/p96.html>

was realized within the framework of a Bilateral Cooperation Agreement between the Governments of Hungary and the Flemish community of Belgium under Project BIL98/03.

References

- [1] S. Kristyán, A. Ruzsinszky, G.I. Csonka, *Theor. Chem. Acc.* 106 (2001) 319 see also page 404.
- [2] R.S. Mulliken, W.C. Emler, *Diatomic Molecules: Results of ab initio Calculations*, Academic Press, New York, 1977.
- [3] A.E. Reed, R.B. Weinstock, F.J. Weinhold, *J. Chem. Phys.* 83 (1985) 735.
- [4] F.L. Hirshfeld, *Theor. Chim. Acta Berl.* 44 (1977) 129.
- [5] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P.M.W. Gill, B.G. Johnson, W. Chen, M.W. Wong, J.L. Andres, M. Head-Gordon, E.S. Replogle, J.A. Pople, *GAUSSIAN 98*, Gaussian Inc., Pittsburgh, PA, 1998.
- [6] B. Rousseau, A. Peeters, C. Van Alsenoy, *Chem. Phys. Lett.* 324 (2000) 189.
- [7] B. Roos, C. Salez, A. Veillard, E. Clementi, *Tech. Rep.*, IBM Res. RJ 518 (1968) 1.
- [8] A.D. Becke, *J. Chem. Phys.* 88 (1988) 2547.
- [9] C.W. Murray, N.C. Handy, G.J. Laming, *Mol. Phys.* 78 (1993) 997.
- [10] V.I. Lebedev, D.N. Laikov, *Doklady Math.* 59 (1999) 477.
- [11] L.A. Curtiss, K. Raghavachari, P.C. Redfern, V. Rassolov, J.A. Pople, *J. Chem. Phys.* 109 (1998) 7764.
- [12] L.A. Curtiss, P.C. Redfern, K. Raghavachari, J.A. Pople, *J. Chem. Phys.* 114 (2001) 108.