QUANTUM CHEMICAL BENCHMARK ENERGY AND GEOMETRY DATABASE FOR MOLECULAR CLUSTERS AND COMPLEX MOLECULAR SYSTEMS (www.begdb.com): A USERS MANUAL AND EXAMPLES

Jan Řezáč[a], Petr Jurečka[b], Kevin E. Riley[a], Jiří Černý[a], Haydee Valdes[a], Kristýna Pluháčková[a], Karel Berka[a], Tomáš Řezáč[a], Michal Pitoňák[a], Jiří Vondrášek[a] and Pavel Hobza[a1,b,*]

[a] *Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v.v.i. and Center for Biomolecules and Complex Systems, 166 10 Prague 6, Czech Republic; e-mail:* [1] *hobza@uochb.cas.cz*

[b] *Department of Physical Chemistry, Palacký University, třída Svobody 26, 771 46 Olomouc, Czech Republic*

Received October 1, 2008
Accepted October 6, 2008
Published online November 27, 2008

Dedicated to Professor Rudolf Zahradník on the occasion of his 80th birthday.

Our previous benchmark CCSD(T)/ complete basis set limit calculations were collected into a database named begdb – Benchmark Energy and Geometry DataBase. Web-based interface to this database was prepared and is available at www.begdb.com. Users can browse, search and plot the data online or download structures and energy tables.

Keywords: *Ab initio* calculations; CCSD(T); Coupled cluster method; Internet database; Molecular clusters.

The coupled cluster method covering single and double electron excitations iteratively and triple electron excitations non-iteratively (CCSD(T)) provides highly accurate energies, geometries and various properties for molecular clusters and complex molecular systems. In our laboratory, we consistently use the CCSD(T) method extrapolated to complete basis set limit (CBS) and during the past several years we have collected data for several hundreds of molecular complexes and complex molecular systems (mostly peptides). These calculations are, however, time consuming and for systems with more than about 50 atoms are still impractical. This is due to the fact that the method scales as N^7 where N is the number of basis functions. To attain similar accuracy for extended systems requires the use of new techniques,

Collect. Czech. Chem. Commun. 2008, Vol. 73, No. 10, pp. 1261–1270
© 2008 Institute of Organic Chemistry and Biochemistry
doi:10.1135/cccc20081261

which should be parametrized and are thus, strictly speaking, not from the ab initio family.

Recently many new promising methods with improved efficiency and/or accuracy have been introduced, however, in order for these methods to be used correctly it is necessary to properly characterize their strengths and limitations. Each method should thus be parameterized and/or tested against reliable benchmarks, either from experiment or from accurate high-level *ab initio* calculations. In this paper, we shall consider the latter case. Although experiments can potentially be extremely accurate and guarantee an unquestionable description of reality, the use of such data is still limited. On the other hand comparing one type of calculation to another has many advantages. Firstly, we can directly compare the variables our method yields – namely the energy of a molecule or molecular complex – not a derived property that can be measured experimentally. The comparison can be based on the same molecular geometry, which makes the comparison simple and well-defined, or we can compare energies of the geometric minima obtained by the respective method. Finally, we can prepare a customized set of structures that represent a more complex problem we intend to study with the tested method to prepare a more accurate, but specialized method.

The CCSD(T)/CBS interaction energies, which represent benchmark data, were obtained as MP2/CBS interaction energies corrected with the difference between MP2 and CCSD(T) interaction energies (so called CCSD(T) correction term) calculated using a smaller basis set[1]. Although this strategy is not as computationally demanding as the original CCSD(T)/CBS method, the calculations involved are still very expensive and we realized that they can also be of use for other scientists working in the development and parametrization of new fast computational procedures.

Based on our experience with high-level calculations involving biomacromolecules and their building blocks, we have prepared several sets of benchmark data. The first of these datasets, dubbed S22 [1], consists of 22 molecular complexes covering both hydrogen bonds, dispersion interaction and their combinations in a balanced way. The study of complexes was further extended to more biomolecules contained in the JSCH-2005 dataset[1]. Recently, the S22 set was extended as S26 to emphasize the hydrogen bonded complexes[2]. We have also covered the phenomena of halogen bonding[3]. Aside from noncovalent complexes, we have studied conformers of small peptides[4–6] at the same level of theory. This list is not final; we are working on more systematic high-level studies. All these results were published along with the molecular structures.

The reference data published in the above mentioned publications have since been extensively used as benchmarks. In particular, they have been used in methodology development by us[7] and many others[8–12] and to validate or assess the newly developed methods[13–33]. It is of special interest that a very wide variety of computational methods have been tested on a single set (S22), which allows for their direct comparison. The methods tested range from the pure DFT methods[9,13,14] through the combination of the DFT theory with empirical dispersion (DFT-D)[15–17], hybrid and double hybrid DFT[18–20], DFT with fully non-local correlation[21,22], spin component scaled methods in MP2[23–25] and CCSD[26], r12 methods[27], semiempirical methods[8] to quantum Monte Carlo[28], symmetry adapted perturbation theory (SAPT)[29] and others[30–33]. Accurate interaction energies and geometries have also been useful as references in various applications[34] from biomolecules to nano-chemistry[34k,34l] and, in the case of the peptides database, both for experimentalists[35] and theoreticians[36].

To make the work with the benchmark energies and geometries easier, we decided to collect all our data in a database and make it accessible on the internet. We have launched a website www.begdb.com (Benchmark Energy and Geometry DataBase) for easy access to the database, where anyone can browse, search and download the data, including the high-quality geometries used for their calculation (for screenshots, see Figs 1–3). Currently, the database contains only results from our laboratory, but we plan to open it to the other authors in the near future.

FEATURES

For convenience, the data are organized in datasets originating from the respective publications. This grouping is logical, putting together only directly comparable results. The basic view on the data is a table featuring all the methods evaluated in the study (Fig. 1). Users can sort the table in different ways and remove columns or download the whole table for later processing in a spreadsheet application. For each structure, all the necessary details are provided and the geometry can be downloaded in .xyz format. In addition, the structure can be viewed online using the Jmol[37] java applet embedded into the webpage.

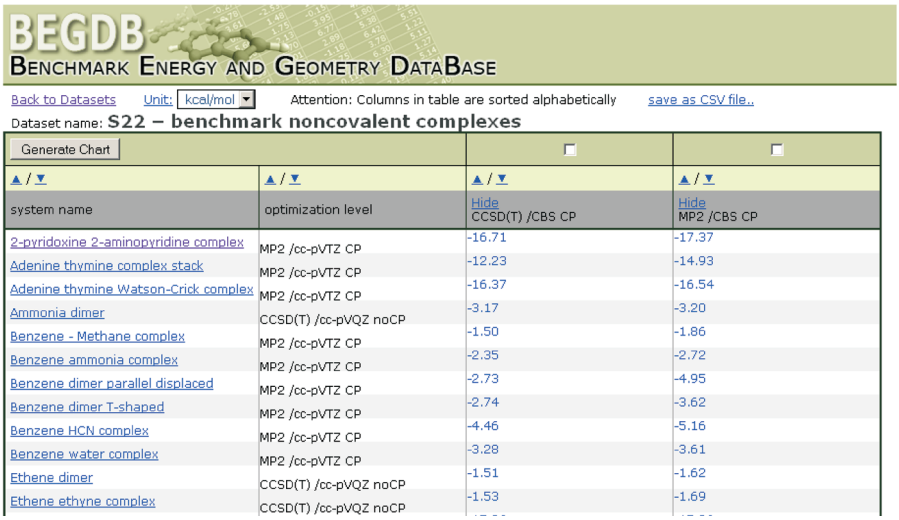
Selected methods can also be compared visually in a graph, plotting energies for all the structure in the dataset (Fig. 2).

If the user is interested in a particular compound, the quick search feature can be used to search the database for its name.

The advanced search function allows one to enter more complicated search expressions to obtain specific information across the datasets (Fig. 3). Search results are browsed in the same manner as the datasets, including download of the table and plotting. To make the search function more powerful, the structures are tagged with keywords. The autocomplete function in the search form provides list of all possible keywords. Using these tags, it is possible to look for specific structures, for example list all peptides containing H-bonds, complexes containing nucleic acid bases etc. Apart from the keywords, the molecules can be searched by the methods used for calculation, the energy value and more. The advanced search function can be used to do very specific analysis of the results presented in the database; an example is given in the following text.

IMPLEMENTATION

The web-based interface to the database can be found at www.begdb.com. It is written in PHP programming language[38] and uses Javascript and AJAX to implement advanced components of the user interface. The data themselves are stored in a MySQL database[39]. Both the website and the database run on a dedicated server.


BEGDB
BENCHMARK ENERGY AND GEOMETRY DATABASE

Back to Datasets Unit: kcal/mol Attention: Columns in table are sorted alphabetically save as CSV file..

Dataset name: **S22 – benchmark noncovalent complexes**

Generate Chart

system name	optimization level	Hide CCSD(T) /CBS CP	Hide MP2 /CBS CP
2-pyridoxine 2-aminopyridine complex	MP2 /cc-pVTZ CP	-16.71	-17.37
Adenine thymine complex stack	MP2 /cc-pVTZ CP	-12.23	-14.93
Adenine thymine Watson-Crick complex	MP2 /cc-pVTZ CP	-16.37	-16.54
Ammonia dimer	CCSD(T) /cc-pVQZ noCP	-3.17	-3.20
Benzene - Methane complex	MP2 /cc-pVTZ CP	-1.50	-1.86
Benzene ammonia complex	MP2 /cc-pVTZ CP	-2.35	-2.72
Benzene dimer parallel displaced	MP2 /cc-pVTZ CP	-2.73	-4.95
Benzene dimer T-shaped	MP2 /cc-pVTZ CP	-2.74	-3.62
Benzene HCN complex	MP2 /cc-pVTZ CP	-4.46	-5.16
Benzene water complex	MP2 /cc-pVTZ CP	-3.28	-3.61
Ethene dimer	CCSD(T) /cc-pVQZ noCP	-1.51	-1.62
Ethene ethyne complex	CCSD(T) /cc-pVQZ noCP	-1.53	-1.69

FIG. 1
Database entries listed in a table – S22 dataset (only a part of the database is shown)

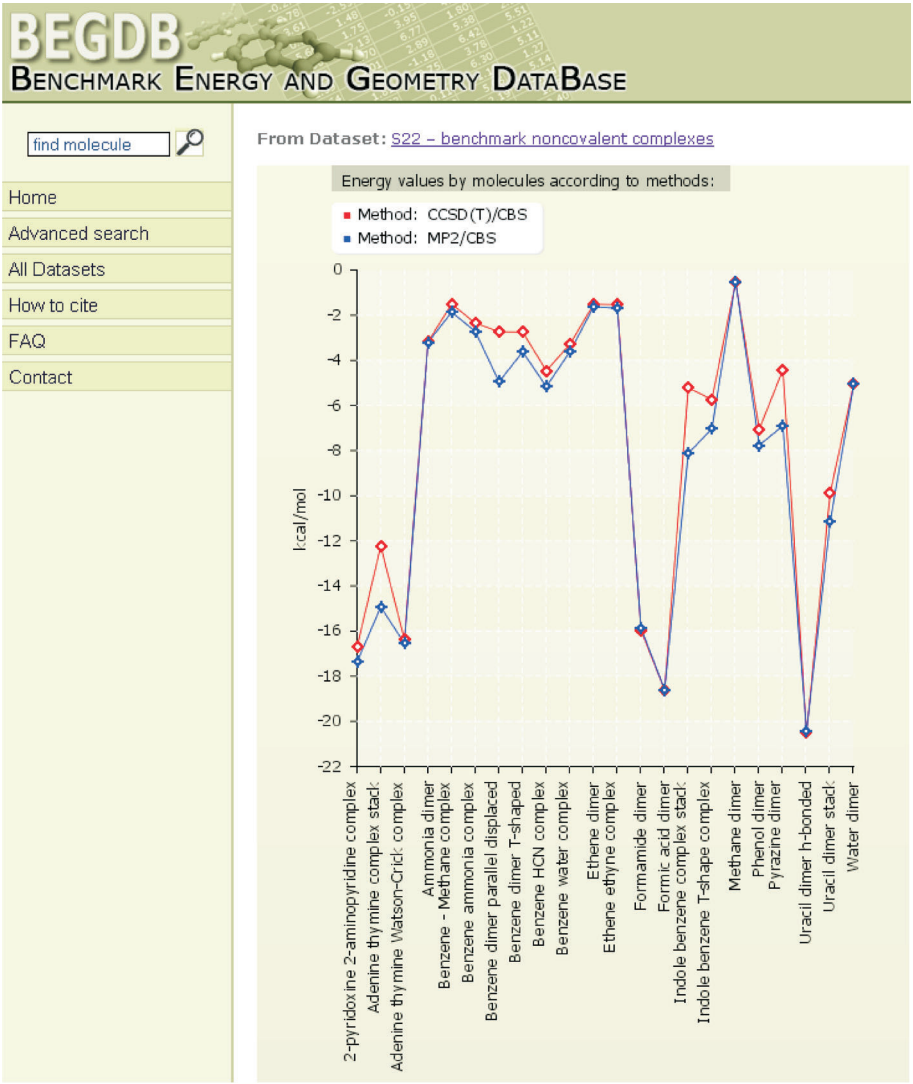


FIG. 2
Results of different computational methods can be compared in plots

Example 1

The first example covers usage of the database for assessment of the performance of a newly developed method for calculation of interaction energies. Our method is rather expensive, so the testing set should be reasonably small, but well balanced. The S22 database[1] is a good starting point. We would also like to include the four extra complexes featuring single hydrogen bonds[2] that complement S22 set forming the set named S26. In the list of datasets at www.begdb.com, we see that these structures are divided into two sets: S22 and S26 extras. To work with S26 more conveniently, we can use the Advanced search to join the datasets, using OR operator in the search expression (also in Fig. 3):

```
Dataset     LIKE     S22
OR
Dataset     LIKE     S26
```

As a result, we obtain the combined S26 dataset in one table, and we can download this table in a .csv format that can be opened in any spreadsheet editor. The other file we are going to download is the archive of geometries of the complexes.

On these geometries, we then run calculations using the tested method, and then compare the results against the CCSD(T)/CBS benchmark values from the database.

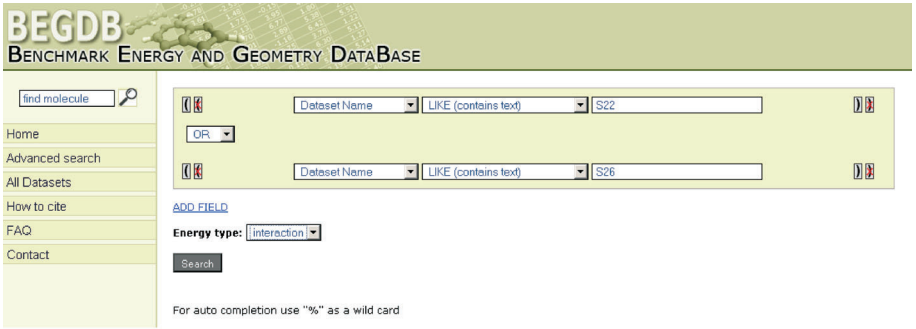


FIG. 3
Advanced search form allows to enter complex queries

Example 2

In the second example, we would like to show the advanced capabilities of the database. We will use the Advanced search to evaluate the performance of the SCC-DFTB-D [40] (self-consistent charges density functional tight binding improved by empirical dispersion correction) method when it is applied to study of small peptides. The CCSD(T)/CBS data will serve as a benchmark. Energies of conformers of the peptides are made relative to the average energy in each method and molecule.

We know that the SCC-DFTB-D method often underestimates hydrogen bonds that can be important for stabilization of the structure of a peptide (expressed as a relative energy of the conformer). We will operate on the Peptides dataset (which lists relative energies – what must be specified in the search), and we will specifically select structures of GFA tripeptide with zero, one and two H-bonds. It is possible because information on the H-bonds is provided as tags for each structure. The results can be viewed in a form of plots or downloaded for further processing.

To list all peptides without hydrogen bonds, we will use following expression in Advanced search:

Dataset	LIKE	Peptides
AND		
Molecular name	LIKE	GFA
AND		
Tag	NOT LIKE	H-bond

The operator “LIKE” is used for loose and case-insensitive comparison of text, looking for tags containing the desired substring. Analogically, we can search for structures containing one (see example) or more H-bonds:

Dataset	LIKE	Peptides
AND		
Molecular name	LIKE	GFA
AND		
Tag	LIKE	1 H-bond

For a first impression, we can select the respective columns and view the results in graphs. What is clearly visible is that structures without hydrogen bonds are more stable and structures with more hydrogen bonds less stable compared to the benchmark. For more rigorous analysis, we can download the tables, open them in a spreadsheet and calculate average difference between SCC-DFTB-D and CCSD(T) for each count of H-bonds. As a result, we see the H-bond underestimated by about 0.75 kcal/mol per hydrogen bond in the SCC-DFTB-D method.

CITATION

When using the BEGDB database it is recommended to cite the original paper (referred in the database) as well as the present one.

CONCLUSIONS

• The BEGDB database allows easy access to high-quality molecular geometries and benchmark CCSD(T)/CBS calculations on them. Other methods are also added for comparison.

• The results are organized in logical datasets according to their nature and source.

• The database will be open to other authors in future.

• The interface allows simple browsing of the results as well as advanced search functions applicable across the datasets.

• The Advanced search can be used to combine and analyze the results in new ways.

This work was a part of the research project No. Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v.v.i. and it was supported by Grants No. LC512 and MSM 6198959216 from the Ministry of Education, Youth and Sports of the Czech Republic. The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P. Hobza in 2007 is also acknowledged.

REFERENCES AND NOTES

1. Jurečka P., Šponer J., Černý J., Hobza P.: *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
2. Riley K. E., Hobza P.: *J. Phys. Chem. A* **2007**, *111*, 8257.
3. Riley K. E., Hobza P.: *J. Chem. Theory Comput.* **2008**, *4*, 232.
4. Valdes H., Spiwok V., Řezáč J., Řeha D., Abo-Riziq A. G., de Vries M. S., Hobza P.: *Chem. Eur. J.* **2008**, *14*, 4886.
5. Valdes H., Pluháčková K., Pitoňák M., Řezáč J., Hobza P.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 2747.

6. Řeha D., Valdes H., Vondrášek J., Hobza P., Abu-Riziq A., Crews B., de Vries M. S.: *Chem. Eur. J.* **2005**, *11*, 6803.
7. Jurečka P., Černý J., Hobza P., Salahub D. R.: *J. Comput. Chem.* **2007**, *28*, 555.
8. Tuttle T., Thiel W.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 2159.
9. a) Zhang Y., Vela A., Salahub D. R.: *Theor. Chem. Acc.* **2007**, *118*, 693; b) Zhang Y., Salahub D. R.: *Chem. Phys. Lett.* **2007**, *436*, 394.
10. Riley K. E., Vondrasek J., Hobza P.: *Phys. Chem. Chem. Phys.* **2007**, *9*, 5555.
11. a) Hill J. G., Platts J. A.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 2785; b) Hill J. G., Platts J. A.: *J. Chem. Theory Comput.* **2007**, *3*, 80.
12. a) McNamara J. P., Sharma R., Vincent M. A., Hillier I. H., Morgado C. A.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 128; b) Morgado C. A., McNamara J. P., Hillier I. H., Burton N. A.: *J. Chem. Theory Comput.* **2007**, *3*, 1656; c) McNamara J. P., Hillier I. H.: *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.
13. a) Zhao Y., Truhlar D. G.: *Theor. Chem. Acc.* **2008**, *120*, 215; b) Zhao Y., Truhlar D. G.: *J. Phys. Chem. C* **2008**, *112*, 4061; c) Zhao Y., Truhlar D. G.: *Acc. Chem. Res.* **2008**, *41*, 157; d) Zhao Y., Truhlar D. G.: *J. Chem. Theory Comput.* **2007**, *3*, 289; e) Zhao Y., Truhlar D. G.: *J. Chem. Phys.* **2006**, *125*, 194101.
14. van der Wijst T., Guerra C. F., Swart M., Bickelhaupt F. M.: *Chem. Phys. Lett.* **2006**, *426*, 415.
15. a) Grimme S., Antony J., Schwabe T., Muck-Lichtenfeld C.: *Org. Biomol. Chem.* **2007**, *5*, 741; b) Antony J., Grimme S.: *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287.
16. Morgado C., Vincent M. A., Hillier I. H., Shan X.: *Phys. Chem. Chem. Phys.* **2007**, *9*, 448.
17. Ducere J. M., Cavallo L.: *J. Phys. Chem. B* **2007**, *111*, 13124.
18. a) Muck-Lichtenfeld C., Grimme S.: *Mol. Phys.* **2007**, *105*, 2793; b) Schwabe T., Grimme S.: *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397.
19. Goll E., Leininger T., Manby F. R., Mitrushchenkov A., Werner H. J., Stoll H.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 3353.
20. Benighaus T., DiStasio R. A., Lochan R. C., Chai J. D., Head-Gordon M.: *J. Phys. Chem. A* **2008**, *112*, 2702.
21. Cooper V. R., Thonhauser T., Langreth D. C.: *J. Chem. Phys.* **2008**, *128*, 204102.
22. Sato T., Tsuneda T., Hirao K.: *J. Chem. Phys.* **2007**, *126*, 234114.
23. Antony J., Grimme S.: *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287.
24. Hill J. G., Platts J. A.: *J. Chem. Theory Comput.* **2007**, *3*, 80.
25. Distasio R. A., Head-Gordon M.: *Mol. Phys.* **2007**, *105*, 1073.
26. Takatani T., Hohenstein E. G., Sherrill C. D.: *J. Chem. Phys.* **2008**, *128*, 124111.
27. Marchetti O., Werner H. J.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 3400.
28. Korth M., Luchow A., Grimme S.: *J. Phys. Chem. A* **2008**, *112*, 2104.
29. a) Hesselmann A., Jansen G., Schutz M.: *J. Am. Chem. Soc.* **2006**, *128*, 11730; b) Sedlák R., Jurečka P., Hobza P.: *J. Chem. Phys.* **2007**, *127*, 075104.
30. a) Rubeš M., Bludský O., Nachtigall P.: *ChemPhysChem* **2008**, *9*, 1702; b) Bludský O., Rubeš M., Soldán P., Nachtigall P.: *J. Chem. Phys.* **2008**, *128*, 114102.
31. a) Lin I. C., Rothlisberger U.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 2730; b) Lin I. C., von Lilienfeld O. A., Coutinho-Neto M. D., Tavernelli I., Rothlisberger U.: *J. Phys. Chem. B* **2007**, *111*, 14346.
32. Kubař T., Jurečka P., Černý J., Řezáč J., Otyepka M., Valdes H., Hobza P.: *J. Phys. Chem. A* **2007**, *111*, 5642.
33. Johnson E. R., McKay D. J. J., DiLabio G. A.: *Chem. Phys. Lett.* **2007**, *435*, 201.

34. a) Barone V., Biczysko M., Pavone M.: *Chem. Phys.* **2008**, *346*, 247; b) Cysewski P., Czyznikowska Z., Zalesny R., Czelen P.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 2665; c) Rutledge L. R., Durst H. F., Wetmore S. D.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 2801; d) Csontos J., Palermo N. Y., Murphy R. F., Lovas S.: *J. Comput. Chem.* **2008**, *29*, 1344; e) Kysel O., Budzak S., Medved M., Mach P.: *Int. J. Quantum Chem.* **2008**, *108*, 1533; f) Lin I. C., von Lilienfeld O. A., Coutinho-Neto M. D., Tavernelli I., Rothlisberger U.: *J. Phys. Chem. B* **2007**, *111*, 14346; g) Schreiber M., Gonzalez L.: *J. Comput. Chem.* **2007**, *28*, 2299; h) Langner K. M., Sokalski W. A., Leszczynski J.: *J. Chem. Phys.* **2007**, 127; i) Kolar M., Hobza P.: *J. Phys. Chem. A* **2007**, *111*, 5851; j) Mamdouh W., Kelly R. E. A., Dong M. D., Kantorovich L. N., Besenbacher F.: *J. Am. Chem. Soc.* **2008**, *130*, 695; k) McNamara J. P., Sharma R., Vincent M. A., Hillier I. H., Morgado C. A.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 128; l) Piquemal J. P., Chevreau H., Gresh N.: *J. Chem. Theory Comput.* **2007**, *3*, 824.
35. a) Vaden T. D., de Boer T., Simons J. P., Snoek L. C.: *Phys. Chem. Chem. Phys.* **2008**, *10*, 1443; b) Plusquellic D. F., Siegrist K., Heilweil E. J., Esenturk M.: *ChemPhysChem* **2007**, *8*, 2412.
36. a) Schwabe T., Grimme S.: *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397; b) Holroyd L. F., van Mourik T.: *Chem. Phys. Lett.* **2007**, *442*, 42; c) Riley K. E., Op't Holt B. T., Merz K. M.: *J. Chem. Theory Comput.* **2007**, *3*, 407; d) van Mourik T., Karamertzanis P. G., Price S. L.: *J. Phys. Chem. A* **2006**, *110*, 8.
37. *Jmol: An Open-Source Java Viewer for Chemical Structures in 3D*. http://www.jmol.org/
38. http://www.php.net/
39. http://www.mysql.com/
40. a) Elstner M., Pozerag D., Jungnickel G., Elsner J., Haugk M., Frauenheim T., Suhai S., Seifert G.: *Phys. Rev. B* **1998**, *58*, 7260; b) Elstner M., Hobza P., Frauenheim T., Suhai S., Kaxiras E.: *J. Chem. Phys.* **2001**, *114*, 5149.