

Doctoral (Ph.D.) thesis

**Benchmarking experimental and
computational chemistry data**

Dóra Barna

Supervisor:

Dr. Gyula Tasi

Associate Professor

Doctoral School of Chemistry

University of Szeged

Department of Applied and Environmental Chemistry

2012

Szeged

1. Introduction

The progress of chemistry and chemical technology is highly dependent on benchmark thermochemical, spectroscopic, and kinetic parameters and on their constantly improving accuracy. In order to present new reaction mechanisms, to prepare models with improved predictive ability, or to design technological procedures more accurately and, accordingly, more economically, it is essential to know the corresponding high-quality experimental results. Progress in this area is also encouraged by the strong dependence of computational chemistry on highly accurate experimental parameters, partly as the latter serves as benchmark for more and more powerful theoretical chemistry methods. On the other hand, experimental values can also be used as empirical corrections to contribute to the quality of computational chemistry data.

During the determination of high-quality experimental results, evaluation of measured data sets counts as a neglected area. According to general practice, the ordinary linear least-squares method is used most frequently. However, this widespread technique may lead to incorrect results, especially, if the measured data sets are corrupted with a serious amount of noise, or in case of an originally nonlinear model being transformed into linear form, in order to make parameter estimation less complicated.

In our days, it has become possible for a – at present yet small, but constantly growing – subset of molecules to carry out quantum chemical computations with an accuracy, which can be competitive with the quality of the most precise experiments. To bring this goal to affect, one of the composite methods of quantum chemistry is usually applied, which are designed in a way which endeavors to take into consideration every kind of important contribution. Following this way, the error bar of a quantum chemical calculation can be reduced even to the level of a few cal mol⁻¹.

In regard to the fact that such computations are as yet limited by the size of the system under study, there is a serious need for theoretical chemistry methods, which are less demanding computationally, and as a consequence, can be used for a wide variety of molecular systems. In principle, this role can be filled in by one of the already existing theoretical model chemistries. To confirm this regarding one of these models, they have to be subjected to a detailed analysis, and it is also worth studying whether they can be improved by some minor modifications in their algorithms.

Quantum chemical basis sets have crucial impact on the accuracy of calculated chemical values. In practice, basis sets with the same parameter set are used for every system at every theoretical level. However, if we aim to determine benchmark chemical data with the highest possible accuracy, it is worth to analyse the suitability of the basis set to the studied system and the chosen computational level.

2. Research objectives

Our main objective was to provide a clear picture of the points, which should be taken into consideration every time one wants to determine benchmark chemical values. Since nowadays an aim like this can be approached both from experimental and computational side, we put a high emphasis on considering both approaches in equal measure. To demonstrate our findings we needed an example which has been standing in the center of interest of theoretical as well as experimental chemists for decades. For this purpose we chose the determination of the conformational enthalpy difference of butane, a fundamental parameter of chemistry, and of molecular structure in particular. Although this question has already inspired several scientific studies, the accurate, unquestionable answer, result of either measurements or calculations, is still lacking.

Since we found that experimental determination of chemically meaningful parameters does not get the attention it is entitled to, we first wanted to explore how the various regression methods influence the precision of the parameters. However, since the most popular software products can only manage the simplest form of the least-squares method, first and foremost we had to create a parameter estimation method, which is able to accomplish any kind of regression.

Following exhaustive tests our next goal was to ascertain whether the so far most accurate conformational enthalpy differences of butane can be improved using a parameter estimation method more fitting to the current problem.

In the next step we also wanted to determine the conformational enthalpy difference using quantum chemical computations. In order to do this we constructed a new *ab initio* model chemistry, which enabled us to determine relative thermochemical quantities of hydrocarbons with higher precision than ever.

Our aim was to follow the same procedure for pentane conformers, but the increased size of the molecule did not allow us to carry out calculations on such a high level. As a consequence, we began to look for a less demanding quantum chemical model, which – either in its original form or after minor modifications – could prove an acceptable alternative regarding the balance between the required time and the quality of the final results. For the sake of finding such a model, we analysed and evaluated several model chemistries in fine detail.

Since only finite basis sets can be applied in quantum chemical calculations in practice, it is important to use basis functions of the best possible quality for constructing the wave function of a system. We intended to introduce a method, which uses the extended virial theorem for evaluating existing quantum chemical basis sets and designing new, better ones.

3. Applied methods

Our parameter estimation method is based on the simplex routine, which has primarily established as a tool for determining extrema of functions. The simplex procedure is a robust, direct-search method: it needs smooth function's values only. Originally the simplex method did not possess any algorithm for error analysis. We recovered this deficiency by integrating a numerical error estimating procedure into the original simplex routine. As a result we created a method, which is able to handle any kind of regression or model function. In other words, it can prove useful in solving practically every parameter estimation problem.

The FORTRAN program, written on the grounds of the above mentioned principles, was applied in the re-evaluation of experimental conformational enthalpy differences in the cases of butane and pentane. Three regression methods were used in our investigations: ordinary and weighted linear least-squares and ordinary nonlinear least-squares method.

The improved *ab initio* model chemistry, primarily prepared for the case of butane, follows the principles of the Weizmann-*n* and HEAT models: the main building blocks are coupled cluster calculations with correlation consistent Dunning basis sets. The equilibrium structures were determined at CCSD(T)/cc-pVQZ level, to the highest level single point computation the CCSDT(Q)/cc-pVDZ method was used. Furthermore diagonal Born-Oppenheimer correction and scalar relativistic contribution are also included in the model. The CCSDT(Q) computations were carried out with the MRCC suite of quantum chemical programs interfaced to the CFOUR package. For MP2 and scalar relativistic computations, the MOLPRO package was used; all other results were obtained with CFOUR. In all calculations restricted Hartree-Fock orbitals were used.

During the detailed analysis of already existing theoretical model chemistries the models integrated into the Gaussian 09 program suite were studied. Among them we find the Gaussian-*n* models (*Gn*, *n*: 1-4), the G3B3, and the Weizmann-1RO protocols, along with three Complete Basis Set models: CBS-4M, CBS-QB3, and CBS-APNO. Our intention was to improve these model chemistries by modifying the steps which determine the equilibrium geometry and the harmonic vibrational frequencies. These computations were also carried out using the Gaussian 09 package.

The evaluation of the quantum chemical basis sets was carried out using the extended virial theorem. Within the Born-Oppenheimer approximation and for general wave functions the extended virial theorem can be described with the following formula:

$$2T + V + \sum_{\alpha} \bar{R}_{\alpha} \frac{\partial E}{\partial \bar{R}_{\alpha}} + \sum_q \bar{P}_q \frac{\partial E}{\partial \bar{P}_q} - \sum_j \zeta_j \frac{\partial E}{\partial \zeta_j} = 2T + V + \Delta R + \Delta P + \Delta Z = 0 \quad (1)$$

The first two terms in the above equation contain the kinetic and the potential energy, respectively. In the last three terms, the derivatives of the total energy can be found with respect to the coordinates of the atomic nuclei, the coordinates of the centers of basis functions, and the exponents of the basis functions, respectively. The necessary computations were carried out using a FORTRAN program, which can perform Hartree-Fock and MP2 level calculations. Besides, the program supports the use of atom-centered, bond-centered, floating-centered, and floating-exponent Gaussian-type basis functions with arbitrary angular momentum.

4. Summary of new research findings

4.1. The simplex method extended with error analysis[1, 2]

With the help of the numerical algorithm for calculating error propagation, every kind of parameter estimation problems can be handled, the value of the parameters along with their estimated standard errors can be determined with any arbitrary accuracy. It should be noted that the simplex method was only a choice of convenience from our part. If the number of parameters in the actual problem exceeds twenty (an approximate number which limits the applicability of the simplex routine), the simplex procedure can be replaced *e.g.* by Newton or by quasi-Newton algorithms.

Based on the tests we have accomplished, this procedure meets the condition regarding that the result should be invariant against transposing the roles of dependent and independent variables. We could also ascertain that choosing the least-squares method, best fitting to the current problem, the distortion of parameters and standard errors could be avoided. This latter phenomenon could be the result of either the linearization of our model or the not sufficient treatment of the noise in the measured data sets.

4.2. Conformational enthalpy difference of butane based on the re-evaluation of experimental data[3]

During the last seven decades there were many attempts to determine the enthalpy difference between the two conformers of butane using experimental techniques. Nevertheless, we cannot forget that even after the most precise measurements the same error has been committed every time: the ordinary linearized least-squares method has been used in a case where the original model function was nonlinear, and both the dependent and the independent variable could be measured with non-negligible error. The conformational enthalpy differences, considered as benchmark values so far, all carry the error resulting from this oversimplification. In order to determine the most accurate chemical parameters we repeated the evaluation of the original experimental data sets using the weighted linearized and the ordinary nonlinear least-squares methods, too. As a consequence, the benchmark value of the conformational enthalpy difference of butane is $668 \pm 20 \text{ cal mol}^{-1}$, referring to the 133 - 196 K temperature range. The respective value

in the 223-297 K temperature range is $653 \pm 125 \text{ cal mol}^{-1}$, although, as we can see from the estimated standard error, its uncertainty is much more significant. These values also reveal that the frequently fully neglected temperature dependence of the enthalpy can only be accepted as a first order approximation even in this 100 K wide temperature range.

4.3. Conformational enthalpy difference of butane using quantum chemical computations[3]

Similarly to experimental values, there was not an agreement over the most accurate theoretical value of the conformational enthalpy difference of butane. We intended to close this dispute by creating an improved *ab initio* model chemistry, and using it to determine the value, which can be accepted as the true benchmark result of the field. The primary features of our model is that the equilibrium geometry is computed at the CCSD(T)/cc-pVQZ level, for the harmonic frequencies the CCSD(T)/cc-pVTZ method was used, and the anharmonic correction was calculated at the MP2/6-31G* level. In order to calculate the various contributions to the electronic energy we utilized Hartree-Fock, CCSD, CCSD(T), and CCSDT(Q) computations, along with the diagonal Born-Oppenheimer correction and the scalar relativistic contribution. The guiding principle behind this quantum chemical model is that the level of computation for every contribution must be increased systematically until convergence is reached. As a result of these calculations, the energy difference between the butane conformers has proved to be $595.5 \pm 2.0 \text{ cal mol}^{-1}$, and finally we could state that the theoretical value of the conformational enthalpy difference is $647.4 \pm 7.0 \text{ cal mol}^{-1}$ at 298.15 K.

Further evidence beside the improved quantum chemical model is that after determining the computational equivalents of the experimental data sets in both temperature ranges mentioned above, and performing regression analysis using the three methods mentioned in the previous point, we found that the three results were virtually the same in both of the temperature ranges. In addition, we could also make certain that the error bar was especially low in every case, a proof that the noise in the theoretical data sets is practically non-existent. As a consequence of the two different approaches,

experimental and quantum chemical, it could be asserted that their results mutually support each other, providing further evidence beside the accuracy of these values.

4.4. Conformational enthalpy differences in case of pentane[3]

The conformers of pentane, the next member of the alkane homolog series after butane, have also been studied extensively. In this case there are four non-equivalent conformers, three of which allow accurate spectroscopic measurements due to their more significant conformational ratios. Similarly to butane, there was not a consensus on the accurate values of these enthalpy differences. Accordingly, we re-evaluated the most precise experimental data sets for pentane using the same methods as in the case of butane. These regression analyses resulted in an enthalpy difference of $618 \pm 5 \text{ cal mol}^{-1}$ for the two most stable conformers. The enthalpy of the least stable conformer proved to be $954 \pm 15 \text{ cal mol}^{-1}$ relative to the most stable all-*trans* isomer.

The larger molecule size of pentane along with its lack of symmetry relative to butane do not make it possible to carry out such highly accurate computations within a rational time frame as we have done with the smaller homolog. This explains why we cannot present benchmark-quality theoretical enthalpy differences for the pentane conformers yet.

4.5. Comparing different *ab initio* model chemistries[4]

The molecules like pentane require a quantum chemical model, which is able to supply results with acceptable accuracy but still within a reasonable time. In order to find such a model, we studied some of the less demanding models through the example of the butane conformers. After comparing the results of standard quantum chemical models (Gaussian-*n*, Complete Basis Set, and Weizmann-*n* models) with the theoretical benchmark enthalpy difference, mentioned at point 4.3., we have come to the conclusion that, if we have to find a substitute for the reference *ab initio* protocol, either the G3 or the CBS-APNO model could be our best choice for investigating alkanes.

In the hope that we could improve our computational results by minor modifications in the algorithms of these models, we replaced the steps determining the equilibrium geometries and harmonic vibrational frequencies in the original *Gn* protocols

with those of the improved *ab initio* model, which we accepted as reference. An exchange like this should, by all means, improve the accuracy of the results, since both the new geometry and the new frequencies are calculated at a significantly higher level. Nevertheless, the actual results all reflect the exact opposite of this expectation: the modifications resulted in the increased inaccuracy of the newly calculated total energies relative to the benchmark values. This statement indicates that the modified models provide conformational enthalpy differences which are also further away from their benchmark equivalents than were the results of the original protocols. Considering that the applied changes did not lead to enhanced precision in the final results, we can state that minor modifications only are not able to improve the model chemistries studied by us. In order to reach real achievements on this field, it is advisable to conduct a thorough and detailed analysis to understand how these models were constructed and the reason behind it, what is the role of the individual energy contributions and what is the connection between them.

4.6. Evaluation of quantum chemical basis sets based on the extended virial theorem[5]

In the majority of cases, quantum chemical basis sets have decisive influence on the results of chemical computations. Since, in practice, only finite basis sets can be applied, it is crucial to select basis functions of the highest quality possible. The extended virial theorem provides a tool for quantitatively evaluating the parameterization of each and every basis function. This method can help us to decide whether the chosen basis set is able to describe the studied system on the given theoretical level with sufficient accuracy.

In the stationary points of the potential energy surface only the last one of the three virials in **equation 1** of **Section 3** has to be considered, if we use atom-centered basis functions:

$$2T + V - \sum_j Z_j \frac{\partial E}{\partial Z_j} = 2T + V + \Delta Z = 0 \quad (2)$$

The third term in **equation 2**, the virial resulting from the exponents of the basis functions, is the sum of atomic contributions, which in turn can be calculated by

summing up of the individual virials stemming from basis functions centered on the given atom:

$$\Delta Z = \sum_A \Delta Z(A) \quad (3)$$

$$\Delta Z(A) = - \sum_{j \in A} \zeta_j \frac{\partial E}{\partial \zeta_j} \quad (4)$$

If the exponents of the basis functions are treated as variational parameters, the last term of **equation 2** becomes zero, too. This equation is the “ideal” form of the quantum mechanical virial theorem, applicable for fully variational wave functions.

Evaluation of individual atomic basis functions can be carried out using **equation 4**. The larger is the virial of a basis function j , the lower is the quality of the exponent of the given Gaussian-type function. Computations carried out at RHF/6-31G** and RMP2(full)/6-31G** level for carbon monoxide revealed that parameterization of the 6-31G** basis set had been done less adequately for the oxygen atom than for carbon. Furthermore, it has also been stated that this basis set is less sufficient for electron correlation calculations than for calculations done at Hartree-Fock level.

The extended virial theorem can also be used for designing quantum chemical basis sets of a better quality. Determining parameters for the basis functions while minimizing the total energy of the system leads to a basis set, which is more suitable to the chosen theoretical level and to the studied system.

5. List of scientific publications

Publications on which the dissertation is based

1. G. Tasi, **D. Barna**:
Analytical and numerical computation of error propagation of model parameters
Journal of Mathematical Chemistry, 49 (2011) 1322-1329.
IF = 1.303
2. **D. Barna**, L. Gyevi-Nagy, G. Tasi:
To linearize or not to linearize: analytical and numerical computation of error propagation of parameters (in Hungarian)
Magyar Kémikusok Lapja, 66 (2011) 383-388.
3. **D. Barna**, B. Nagy, J. Csontos, A. G. Császár, G. Tasi:
Benchmarking Experimental and Computational Thermochemical Data: A Case Study of the Butane Conformers
Journal of Chemical Theory and Computation, 8 (2012) 479-486.
IF₂₀₁₁ = 5.215
4. **D. Barna**, G. Samu, G. Tasi:
Evaluation of ab initio model chemistries through the example of butane conformers
In preparation
5. G. Tasi, **D. Barna**:
Energy decomposition based on the extended virial theorem: Hartree-Fock and second-order Møller-Plesset results
International Journal of Quantum Chemistry, 109 (2009) 2599-2605.
IF = 1.315

$\Sigma_{\text{IF}} = 7.833$

Independent citations: 0

Conference attendance

1. **D. Barna:**

Energy decomposition of molecules and molecular complexes

(Molekulák és molekulakomplexek energiadekompozíciós analízise)

XXXII. Kémiai Előadói Napok

Szeged, 2009.

Lecture in Hungarian

2. **D. Barna, G. Tasi:**

Energy decomposition of alkyl-substituted furan molecules

9th Central European Symposium on Theoretical Chemistry

Novy Smokovec, 2010.

Poster