J Mol Model (2004) 10:6–12
DOI 10.1007/s00894-003-0157-6

ORIGINAL PAPER

James J. P. Stewart

Comparison of the accuracy of semiempirical and some DFT methods for predicting heats of formation

Received: 22 April 2003 / Accepted: 17 July 2003 / Published online: 4 December 2003

Abstract A comparison is made of the relative accuracy of some NDDO semiempirical methods and the DFT functionals LYP and PW91 using both double and triple zeta basis sets. The comparison is between the calculated heat of formation and that reported in the NIST database.

Keywords Semiempirical methods · MNDO · AM1 · PM3 · PM5 · Density functional theory

Electronic Supplementary Material Parameters for the tailored method are available in the supplementary material; these are suitable for use with MOPAC2002. All raw data (experimental heats of formation, geometries, total energies and heats of formation for the various methods, etc.) are also provided in CAChe format. Supplementary material is available for this article if you access the article at http://dx.doi.org/10.1007/s00894-003-0157-6. A link in the frame on the left on that page takes you directly to the supplementary material.

Introduction

Computational methods are increasingly being used for predicting properties of compounds, of which the heat of formation of the gas-phase system at 298 K is one of the more important.

There are three main types of computational methods: a classical method, molecular mechanics, and two quantum chemical methods: ab initio and semiempirical. In turn, ab initio methods exist as two large families of methods: Hartree–Fock and density functional. For the purposes of this work, only semiempirical and density functional methods will be used.

J. J. P. Stewart (✉)
Stewart Computational Chemistry,
15210 Paddington Circle, Colorado Springs, CO 80921, USA
e-mail: Jstewart@fujitsu.com

Computational method

All calculations were performed using the CAChe ProjectLeader software, [1] with the starting geometries being generated using CAChe WorkSpace. In some instances it was not immediately obvious which of two or more conformers had the lower energy. Whenever that occurred, all likely conformers were calculated, and the conformer with the lowest energy identified and used in all subsequent calculations.

Density functional methods

For the purpose of this work, all density functional calculations were performed using Dgauss [2] within CAChe. This program optimizes the electronic structure and geometry of molecules by solving the Kohn-Sham [3, 4] equations in a SCF molecular orbital approximation. As such, the DFT methods are first principles ab initio methods. Two exchange-correlation energy functionals were used, one developed by Lee, Yang, and Parr, B88-LYP [5] and the other by Perdew and Wang, B88-PW91. [6, 7] Both correlation energy functionals use the same exchange functional. [8] Two basis sets that were developed specifically for density functional Gaussian-type orbital calculations were also used: [9] a double zeta and triple zeta, DZVP and TZVP.

Because of the large variety of compounds used, all calculations were run with the default setting changed so as to maximize the probability of achieving a self-consistent field. Thus, the default number of iterations allowed in DFT calculations was increased from 50 to 200, and the density matrix-mixing ratio, MIXD, was decreased from 1.0 to 0.2.

Calculation of heat of formation at 298 K

First-principles calculations yield total energies in atomic units: that is, the energy released when atomic nuclei and electrons coalesce to form a molecular system in its lowest energy state. Conventionally, heats of formation are defined as “the energy in kcal mol^{-1} required to form one mole of a compound, in the gas phase, at standard temperature and pressure, from its elements in their natural state.”

Converting the total energy, E_{tot}, into heat of formation, ΔH_{f}, is straightforward, [10] Eq. (1).

$$\Delta H_{\mathrm{f}}(0) = 627.51E_{\mathrm{tot}} - \sum_{A} C'_A N_A \tag{1}$$

In this expression, the sum A is over all elements in the molecule, N_A is the number of atoms of element A, and C'_A is a constant for each element. The effects of zero point energy can be approximated by atom additive constants, so an explicit term for zero point energy is not necessary. Thus for two isomers of $C_{14}H_{30}$, tetradecane and octamethylhexane, the zero point energies are 255.7 and 254.7 kcal mol^{-1}, a difference of 1.0 kcal mol^{-1}. The resulting heat of formation corresponds to the $\Delta H_{\mathrm{f}}(0)$; internal energy terms, such as rotational and vibrational energies, need to be included in order to reproduce $\Delta H_{\mathrm{f}}(298)$. However, the effects of such terms can be approximated by constants for each atom in the system, and can thus be incorporated into the above expression, Eq. (2).

$$\Delta H_{\mathrm{f}}(298) = 627.51E_{\mathrm{tot}} - \sum\nolimits_{A} C_A N_A \tag{2}$$

An estimate was made of the magnitude of the errors introduced by using this approximation. For systems of about 50 to 60 atoms, the error was less than 2 kcal mol^{-1}, which indicated that the use of the approximation was justified.

Determination of the constants, C_A

The quantities C_A in the above expression are parameters whose values were defined to give the smallest error in the predicted heat of formation. The standard method of least squares fitting was used, with two modifications:

In order to minimize error, the first modification was the replacement of the simple expression above by the more general form, Eq. (3).

$$\Delta H_f(298) = C_{\mathrm{E}}E_{\mathrm{tot}} - \sum_{A} C_A N_A + C \tag{3}$$

This does not involve any loss of rigor, because the number of reference data used (more that 1,000 data) still vastly exceeded the number of parameters to be fitted: eleven (C_{E}, C, and one atomic C_A for each of the elements H, C, N, O, S, F, Cl, Br, I) in the case of double zeta calculations and seven (C_{E}, C, and one atomic C_A for each of the elements H, C, N, O, F) for the triple zeta.

The second modification was made necessary by the presence of a few large errors. By its very nature, least-squares fitting minimizes the root-mean-square differences between calculated and observed quantities. However, the quantity most used in computational chemistry is the average unsigned error. The modification made was to exclude from the fitting all data for which the calculated and predicted errors were large. This involved a two step process. In the first step, all data were used in determining the constants. The resulting errors were then sequenced in order of increasing unsigned error. Data for the lowest 50% were then used in a second least-squares fitting and the resulting parameters used for the subsequent analysis. Preliminary tests showed that the average unsigned and median errors were fairly insensitive to the fraction of data used, when the fraction was in range of 40% to 60%.

The main effect of this second modification was to minimize the unsigned error for systems for which the error was already relatively small. A second effect was to increase the error for those systems for which the underlying DFT method was not accurate. Since theoretical methods are not normally applied to systems where they are known to be inaccurate, concentrating any errors into such systems, with a resulting increase in accuracy of the remaining systems, was considered justified.

Semiempirical methods

Only NDDO methods of the type used in MOPAC were considered. In order of original development, the methods used were MNDO, [11, 12] AM1, [13] PM3, [14, 15] and PM5. [16] These methods all yield gas phase heats of formation at 298 K. All calculations were performed using MOPAC 2002. [17] The effect of using a linear regression to correct the heats of formation was minimal, so the results of semiempirical calculations were used without further modification.

Tailored semiempirical method

All previous NDDO methods were designed to be general, to be able to model a wide range of phenomena for a wide range of compounds. Because of this, the resulting methods were not optimal for predicting heats of formation. For this work, the parameters used in NDDO work were re-optimized so as to minimize the errors in heats of formation for a subset of the compounds used in this work. This involved the simultaneous optimization of a total of 167 parameters to fit 1,611 reference data for the eight elements H, C, N, O, F, S, Cl, Br, and I. The method used for optimizing parameters has been described elsewhere. [14] As with the subset used in defining the parameters in the DFT calculations, the set used in defining the values of the semiempirical parameters consisted only of compounds for which the errors were known to be small. No attempt was made to produce a method for general use; because of this the tailored method was not given a name.

Table 1 Average errors in heats of formation for various methods. All data (kcal mol^{-1})

Method	No. in set	Median unsigned error	Average unsigned error	Root mean square error	Largest error
MNDO	1,276	6.91	15.38	31.41	+223.0
AM1	1,276	6.69	10.31	16.05	+150.5
PM3	1,276	4.74	6.54	10.74	+153.8
PM5	1,276	3.75	5.57	9.79	+155.1
Tailored method	1,276	2.73	4.48	9.51	+163.9
B88-LYP(D)[a]	1,276	4.31	6.69	11.34	+155.6
B88-PW91(D)[a]	1,276	4.08	6.77	11.98	+155.8

[a] Heats of formation estimated using a simple atomic additivity expression

Table 2 Experimental data with possible errors detected using experimental data

Compound	Formula	Expt. ΔH_f	Calc. ΔH_f [a]	Diff
Perfluorobut-2-ene	C_4F_8	−226	−389.9	−163.9
2-*tert*-Butyl-*p*-cresol	$C_{11}H_{16}O$	49.5	−52.0	−101.5
Octafluorotoluene	C_7F_8	−303.4+/−1.9	−356.8	−53.4
Undecylcyclohexane	$C_{17}H_{34}$	−35.4	−85.3	−49.9
2,6,6-Trimethyl-2-cyclohexen-1-one	$C_9H_{14}O$	−10.6+/−0.53	−55.6	−45.0
1,2,3,6-Tetrahydropyridine	C_5H_9N	33.6	7.1	−26.5
Pentacyclo hexacosa nonane	$C_{26}H_{26}$	97.9+/−1.3	78.0	−19.9
3-Chloro-4-methylaniline	C_7H_8NCl	18+/−1.7	4.0	−14.0
6-Methyl-5-hepten-2-one	$C_8H_{14}O$	−72.6+/−0.38	−60.1	12.4
4-Hydroxy-4-methylpentan-2-one	$C_6H_{12}O_2$	−129.2	−116.2	13.0
2,2’-Biquinoline	$C_{18}H_{12}N_2$	83.2+/−1.9	101.8	18.6
Teraphthalamide	$C_8H_8N_2O_2$	−89.9+/−1.0	−69.7	20.2
Isophthalamide	$C_8H_8N_2O_2$	−91.4+/−0.1	−70.3	21.1
2,4,6-Trimethoxy-s-triazine	$C_6H_9N_3O_3$	−92.7+/−0.36	−70.1	22.6
Fluorodinitrophenylmethane	$C_7H_5N_2O_4F$	−44.2+/−0.6	−16.9	27.3
Diethyl malonate	$C_7H_{12}O_4$	−220.3+/−1.0	−190.1	30.2
1,3,5-Tricyanobenzene	$C_9H_3N_3$	74.9+/−1.1	121.8	46.9

[a] Calculated using the tailored method

Reference data

A subset of the set of experimental reference heats of formation given in the NIST WebBook [18] formed the set used for this study. Results for the various theoretical methods are presented in Table 1.

Four measures of accuracy are shown. The first is the median unsigned error: that is, half of all unsigned errors are larger than the value indicated and half are smaller. While not normally used as a measure of accuracy, it is a useful quantity to be aware of. All computational methods have limitations, cases where the methods do not work well, and errors for such systems can be quite large. Measures normally used, such as the average error and, more importantly, the root-mean-square error, are influenced to a larger degree by the few instances where the method has been shown to be inaccurate. However, as soon as the weaknesses of a particular method are known, the use of that method can be limited to those cases where it is known to be reliable. The median error is then a good indicator of the relative accuracy of the method when applied to systems other than those where the method is known to be of limited accuracy.

The second measure in Table 1 is the average unsigned error. This is the normal quantity used in reporting the accuracy of a method. Although commonly used in computational chemistry work, it is difficult to justify in terms of statistical significance, so, for completeness, the term most frequently used in statistical analyses, the root-mean-square error, is also given. Finally, the value of the largest error for any compound in the set used is given.

An examination of the entries in the NIST database indicated that several reported heats of formation were of questionable accuracy. Therefore, before investigation of the accuracy of the computational methods was performed, reference data that was suspect was removed. This was done in two stages.

Removal of reference data inconsistent with other reference data

The assumption was made that, where the results of calculation and experiment agreed, the experimental data were accurate. A logical consequence of this assumption was that, where the results disagreed, the possibility existed that the reference data were inaccurate. Of course, the possibility also existed that the calculations were inaccurate. To resolve this issue, in each case where a reference datum was of questionable accuracy, an attempt was made to compare the datum with reference data for closely related species. This was possible in several instances (a list of compounds of this type is given in Table 2), details of which are presented elsewhere. [19]

In six other instances, the molecular structure described in the NIST WebBook corresponded to the lowest

Table 3 Experimental data with possible errors detected using theoretical methods

Compound	Formula	Expt. ΔH_f	Tailored method		B88-LYP	
			Work			
			Calc. ΔH_f	Diff	Calc. ΔH_f	Diff
Pentafluoroiodobenzene	C_6F_5I	−131.1+/−3.0	−176.6	−45.5	−178.1	−47.0
Perfluorobutadiene	C_4F_6	−225.2	−253.4	−28.2	−241.6	−16.4
Bis-(*n*-perfluoropropyl ether)	C_6OF_{14}	−742.1+/−0.8	−769.3	−27.2	−757.3	−15.2
Dodecafluorocyclohexane	C_6F_{12}	−566.5+/−2.0	−590.5	−24.0	−587.4	−20.9
Bromopentafluorobenzene	C_6F_5Br	−170.2+/−1.3	−191.3	−21.1	−191.8	−21.6
Perfluoroacetone	C_3OF_6	−325.2	−342.6	−17.4	−337.7	−12.5
Hexafluorobenzene	C_6F_6	−228.5+/−0.29	−242.5	−14.0	−241.6	−13.1
Thietane	C_3H_6S	14.6+/−0.3	4.4	−10.2	8.0	−6.6
DL-3,4-Di-1-cyclohexen-1-yl-2,2,5,5-tetramethyl hexane	$C_{22}H_{38}$	−62.1+/−1.5	−50.4	11.7	−17.5	44.6
Dioxybismethanol	$C_2H_6O_4$	−136.6+/−1.6	−124.5	12.1	−119.2	17.4
2,5,8-Trioxanonane	$C_6H_{14}O_3$	−138.9+/−0.25	−124.6	14.3	−124.2	14.7
2,4,6-Trimethylphenyl isocyanide	$C_{10}H_{11}N$	40.0	56.6	16.6	48.8	8.8
6-(1,1-dimethylethyl)-2,3-dihydro-1,1-dimethyl-1H-Indene	$C_{15}H_{22}$	−41.7	−24.9	16.7	−19.2	22.4
n-Perfluorobutane	C_4F_{10}	−533.9	−515.3	18.6	−515.2	18.7
Pyrazine-1,4-dioxide	$C_4H_4N_2O_2$	44.6	36.3	−8.3	33.5	−11.1
para-Hydroxybiphenyl	$C_{12}H_{10}O$	8.5	0.0	−8.5	0.7	−7.8
5,6-Dibutyl-5,6-bis(4-*tert*-butylphenyl)decane	$C_{38}H_{62}$	−83.8+/−0.8	−58.6	25.2	−40.1	43.7

Table 4 Average errors in heats of formation for various methods excluding possible experiment errors detected by calculation (kcal mol^{-1})

Method	No. in set	Median unsigned error	Average unsigned error	Root mean square error	Largest error
MNDO	1,238	6.66	14.51	29.69	178.8
AM1	1,238	6.45	9.60	13.80	86.1
PM3	1,238	4.57	5.84	7.82	38.1
PM5	1,238	3.58	4.87	6.65	33.8
Tailored method	1,238	2.63	3.64	5.35	36.2
B88-LYP(D)[a]	1,238	4.10	5.86	8.50	40.3
B88-PW91(D)[a]	1,238	3.95	5.82	8.51	39.8

[a] Heats of formation estimated using a simple atomic additivity expression

energy structure in aqueous solution, not to the lowest energy structure in the gas phase. Deleting these two sets of compounds resulted in a lowering of the average error.

Removal of reference data inconsistent with theoretical predictions

After the inconsistent reference data were removed, several other data were identified as being of questionable accuracy, but no ready comparison with reference data for closely related compounds was possible. The possibility clearly existed that the origin of the discrepancy lay in the theoretical method used, so, in an attempt to resolve the origin of the error, a comparison was made between different theoretical methods, specifically the tailored semiempirical method described above and the density functional method used in DGauss. Since these two methods are very different, errors in one method are not, in general, present in the other. In addition, any systematic errors in each method can be determined with good confidence. Thus, for example, DFT functionals using a double zeta basis set were found to consistently predict heats of formation of sulfate compounds to be too positive, by about 20 kcal mol^{-1}. In those instances where both the semiempirical and the DFT methods predict a heat of formation significantly different from that observed experimentally, and no significant systematic error was present in either theoretical method, then the experimental value was considered inaccurate, and removed from the set. The set of such compounds is shown in Table 3.

Results

A comparison of the accuracies of the various methods is presented in Table 4.

Within the version of DGauss used, triple zeta basis sets were defined for elements up to fluorine only, limiting triple zeta calculations to compounds of H, C, N, O, and F. This reduced the number of compounds to 1,001; errors for these systems are also shown in Table 5.

Table 5 Average errors in heats of formation for various methods including triple zeta (kcal mol^{-1})

Method	No. in set	Median unsigned error	Average unsigned error	Root mean square error	Largest error
MNDO	1,001	6.64	12.01	21.29	140.0
AM1	1,001	6.92	9.95	14.25	86.1
PM3	1,001	4.35	5.36	7.10	37.7
PM5	1,001	3.78	5.03	6.84	33.8
Tailored method	1,001	2.76	3.70	5.32	36.2
B88-LYP(DZVP)[a]	1,001	3.95	5.25	7.30	42.1
B88-PW91(DZVP)[a]	1,001	3.90	5.22	7.26	37.2
B88-LYP(TZVP)[a]	1,001	3.30	4.94	7.23	46.3
B88-PW91(TZVP)[a]	1,001	3.08	4.71	6.78	37.0

[a] Heats of formation estimated using a simple atomic additivity expression

Table 6 Structural elements of compounds badly predicted by various methods

Method	Structural element
MNDO	S(VI), e.g. sulfates Nitro and nitrate groups Polyfluorinated compounds
AM1	Polymethylene, e.g. *n*-decane Polyethers
PM3	Sulfones
PM5	Adamantane and substituted adamantanes
Tailored method	Diazines Acetylenes
DFT	S(VI), e.g. sulfates Cyanide group

Identification of systematic errors in methods

Of their nature, computational methods are entirely systematic, resulting in errors in prediction that are also systematic. For any given method, if large errors can be associated with specific structural elements, then limiting the applicability of that method to compounds that do not contain such structural elements produces an increase in accuracy. That is, if methods are used only for systems for which they are suitable, more confidence can be placed in the resulting predictions.

Large systematic errors were found in all the methods used here. A list of the more important relevant structural elements is shown in Table 6. When compounds involving these structural elements are removed, the accuracy increases significantly, as shown in Table 7.

Semiempirical heats of formation are normally obtained from a program such as MOPAC and used without further modification. To allow a comparison of equal quantities, the effect of a linear regression of the type used in Eq. (3) on the calculated heats of formation was evaluated. This is also presented in Table 7. For MNDO and the tailored method, the regression equations are presented in Eqs. (4) and (5). As expected, the effect of linear regression is small, and decreases in the order of increasing accuracy.

$$\begin{aligned}&\Delta H_{\mathrm{f}}(\mathrm{MNDO}) =\\&-1.73573N_{\mathrm{H}} + 0.952956N_{\mathrm{O}} - 8.83055N_{\mathrm{O}} + 2.26152N_{\mathrm{N}}\\&+ 1.22043N_{\mathrm{S}} - 9.30977N_{\mathrm{F}} - 4.46947N_{\mathrm{Cl^-}} - 2.21406N_{\mathrm{Br}}\\&+ 1.79588N_{\mathrm{I}} + 0.832201\Delta H_{\mathrm{f}}(\mathrm{MNDO}) + 10.3669\end{aligned} \tag{4}$$

$$\begin{aligned}&\Delta H_{\mathrm{f}}(\mathrm{tailored}) =\\&- 0.114127N_{\mathrm{H}} + 0.0526197N_{\mathrm{C}} - 0.807535N_{\mathrm{O}}\\&- 0.504739N_{\mathrm{N}} - 1.14932N_{\mathrm{S}} - 0.537479N_{\mathrm{F}}\\&- 0.61802N_{\mathrm{Cl^-}} - 0.649609N_{\mathrm{Br}} - 0.45479N_{\mathrm{I}}\\&+ 0.988847\Delta H_{\mathrm{f}}(\mathrm{tailored}) + 1.65975\end{aligned} \tag{5}$$

Table 7 Average errors in heats of formation for various methods excluding systematic errors in methods (kcal mol^{-1})

Method	No. in set	Median unsigned error	Average unsigned error	Root mean square error	Largest error
MNDO	1,175	6.23	9.74	15.30	130.2
AM1	1,193	6.28	8.56	11.73	63.5
PM3	1,225	4.52	5.67	7.51	38.1
PM5	1,228	3.52	4.70	6.25	28.1
Tailored method	1,217	2.60	3.45	4.92	36.2
MNDO[a]	1,175	5.40	8.27	12.08	79.9
AM1[a]	1,193	5.23	7.20	9.88	54.8
PM3[a]	1,225	4.14	5.19	6.92	35.1
PM5[a]	1,228	3.53	4.62	6.13	27.2
Tailored method[a]	1,217	2.59	3.39	4.78	37.7
B88-LYP(DZVP)[b]	1,179	3.66	4.79	6.92	55.7
B88-PW91(DZVP)[b]	1,179	3.74	4.93	6.55	34.5

[a] Heats of formation obtained by linear regression from NDDO results
[b] Heats of formation estimated using a simple atomic additivity expression

Table 8 Heats of formation for isomers of $C_{14}H_{30}$

Method	*n*-Tetradecane ΔH_{f}	Error	Octamethylhexane ΔH_{f}	Error	Diff.	Error
Expt.	−79.4±0.4		−59.4±0.6		+20.0±0.7	
MNDO	−77.0	+2.4	+46.7	+106.1	+123.7	+103.7
AM1	−99.8	−20.4	−33.5	+25.9	+66.3	+46.3
PM3	−83.5	−4.1	−54.5	+4.8	+29.0	+9.0
PM5	−82.5	−3.1	−47.6	+11.7	+34.9	+14.9
Tailored method	−79.4	0.0	−62.2	−2.8	+17.2	−2.8
B88-LYP (DZVP)[a]	−91.7	−12.4	−38.9	+20.5	+52.8	+32.8
B88-PW91 (DZVP)[a]	−92.0	−12.6	−40.1	+19.2	+51.9	+31.9
B88-LYP (TZVP)[a]	−92.4	−13.0	−40.1	+19.3	+52.3	+32.3
B88-PW91 (TZVP)[a]	−90.8	−11.4	−40.1	+19.2	+50.7	+30.7

[a] Heats of formation estimated using a simple atomic additivity expression

In some instances, smaller but still important errors were found in some structural elements. In the density functional methods, the most important of these errors involved the differences between straight chain and branched hydrocarbons. This is best illustrated using the compounds tetradecane and octamethylhexane. Because these are isomers of $C_{14}H_{30}$, any effects due to the use of atomic additivity terms are eliminated. Errors in heats of formation for these two compounds are shown in Table 8.

Both compounds are simple hydrocarbons; there was no obvious reason to doubt the experimental results. Early semiempirical methods did not reproduce the difference in heats of formation of the two isomers, which is 20 kcal mol^{-1}, but later methods were more accurate. In the parameterization used here, the calculated difference of 17.2 kcal mol^{-1} was quite close to that observed. With density functional methods, the difference was approximately 52 kcal mol^{-1}, no matter which functional or basis sets were used.

To determine whether this discrepancy between calculated and experimentally observed energies could be explained by differences in internal energies, the validity of the assumption that the effect of internal energies could be approximated by atomic additivities was determined by an explicit semiempirical calculation of the internal energies at 298 K. These were 11.6 kcal mol^{-1} for octamethylhexane and 13.3 kcal mol^{-1} for tetradecane. The effect of correcting for internal energy contributions would be to lower the difference in isomer energies from about 52 to about 50 kcal mol^{-1}, which is still considerably above that observed.

Discussion

Unlike semiempirical methods, density functional methods do not yield heats of formation directly. By making the assumption that atomic contributions were constant, an expression was derived for relating heats of forma-tion to total energies. For the B88-LYP method in DGauss, the expressions used are presented in Eqs. (6) and (7).

$$\begin{aligned}\Delta H_{\mathrm{f}} = {}& \\ & 384.170N_{\mathrm{H}} + 24979.741N_{\mathrm{C}} + 49270.383N_{\mathrm{N}} \\ & + 35898.264N_{\mathrm{O}} + 260960.163N_{\mathrm{S}} + 65395.349N_{\mathrm{F}} \\ & + 301576.518N_{\mathrm{Cl}^-} + 1686830.458N_{\mathrm{Br}} + 4535361.678N_{\mathrm{I}} \\ & + 655.410E_{\mathrm{tot}}(DZVP) + 3.802 \end{aligned} \tag{6}$$

$$\begin{aligned}\Delta H_{\mathrm{f}} = {}& \\ & 383.274N_{\mathrm{H}} + 24936.811N_{\mathrm{C}} + 49183.777N_{\mathrm{N}} \\ & + 35835.298N_{\mathrm{O}} + 65279.766N_{\mathrm{F}} + 654.168E_{\mathrm{tot}}(\mathrm{TZVP}) \\ & + 5.365 \end{aligned} \tag{7}$$

For the B88-PW91 functional, the corresponding expressions are presented in Eqs. (8) and (9).

$$\begin{aligned}\Delta H_{\mathrm{f}} = {}& \\ & 383.370N_{\mathrm{H}} + 24726.839N_{\mathrm{C}} + 48760.180N_{\mathrm{O}} \\ & + 35530.597N_{\mathrm{N}} + 258262.044N_{\mathrm{S}} + 64715.724N_{\mathrm{F}} \\ & + 298467.440N_{\mathrm{Cl}^-} + 1669482.073N_{\mathrm{Br}} + 4488702.800N_{\mathrm{I}} \\ & + 648.630E_{\mathrm{tot}}(\mathrm{DZVP}) + 1.203 \end{aligned} \tag{8}$$

$$\begin{aligned}\Delta H_{\mathrm{f}} = {}& \\ & 391.198N_{\mathrm{H}} + 25237.145N_{\mathrm{C}} + 49766.581N_{\mathrm{N}} \\ & + 35835.298N_{\mathrm{O}} + 66050.780N_{\mathrm{F}} + 661.912E_{\mathrm{tot}}(\mathrm{TZVP}) \\ & + 2.269 \end{aligned} \tag{9}$$

The development and parameterization of semiempirical methods has been based on the assumption that the reference data, that is, the experimental data, were accurate. Using very different theoretical methods, evidence has been presented (Table 4) of potential inaccuracies in experimental heats of formation. Because theoretical methods were used, the heats of formation generated by such methods were only predictions, and should not be regarded as being as reliable as most experimental values. Nevertheless, the computational results presented here indicate potential problems with the experimental values.

All examples presented involve large differences between experimental and computed values. This was deliberate, in that large errors are easier to detect than small ones. Admittedly, in those cases where the accuracies of experimental data used here were reported in the NIST WebBook, they were still considerably lower than for most other compounds.

Because of the limited accuracy of theoretical methods, only when large deviations were detected were they reported. The existence of these large errors suggests the possibility that other errors exist, but the methods used here are not yet of sufficient reliability to allow such errors to be identified with confidence. However, to make the assumption that such errors do not exist appears to be unwarranted.

When experimental data of dubious accuracy are removed from the set of reference data used in determining the accuracy of theoretical methods, the accuracy of such methods rises significantly, with the root-mean-square error decreasing the most, and the median unsigned error decreasing only by about 5%.

A measure of the true accuracy of semiempirical and DFT methods for predicting standard heats of formation was made by comparison with a reference base of experimental values. However, an examination of the reference data set revealed the existence of many potentially inaccurate values. Such values would have distorted the measures used for determining accuracy. In particular, the root-mean-square error would have been significantly increased. Therefore, before the comparison was made, the reference data set used was purged of those data considered to be of dubious accuracy.

Once this was done, the resulting accuracies of the various methods were determined. In the semiempirical methods cited here, all averages (mean unsigned error, average unsigned error, and root-mean-square error) decreased monotonically in the order in which they were developed. All these methods use the same basic NDDO set of approximations, with only minor variations, such as the modification of the core repulsion function in AM1 and the introduction of diatomic parameters in PM5. The most important differences between the methods were the degree of parameter optimization and the composition of the data sets used in the optimization. Early parameterizations were severely limited by the amount of computer power available; in the original MNDO optimization, data on only 32 molecules were used in the parameter optimization and required years of effort. A consequence of the considerable increase in computer power over the past few decades has been the ability to use more data. The parameterization described here used a reference data set of about 1,000 molecules, and required only a few CPU days of work on a 2.2-GHz desktop PC.

When the set of molecules used in determining accuracy was limited to well-behaved ground-state systems, the average error in heats of formation obtained using recent semiempirical methods dropped below that of DFT methods, even when the triple zeta basis set was used.

This observation should not be construed as disparaging first-principles methods. Semiempirical methods are parameterized to reproduce certain phenomena. In the most recent parameterizations reported here, the parameterization was designed to reproduce heats of formation with increased accuracy, at the expense of other phenomena, such as electronic structure. Additionally, the reference data set used in the parameterization was predominantly composed of the data used in the statistical analyses presented here. Conversely, DFT methods are not parametric in the sense that semiempirical methods are, and consequently can be applied to a wider range of types of system.

Conclusion

A comparison has been made between heats of formation predicted by various theoretical methods and the values determined experimentally. When the application of the methods was limited to those systems where large systematic errors in the methods were known to be absent, average unsigned errors of less than 5 kcal mol^{-1} were obtained using DFT methods, and less than 4 kcal mol^{-1} using a semiempirical method. Heats of formation were obtained directly from the semiempirical methods and estimated, using a simple atomic additivity expression, from the total energies from DFT calculations.

Acknowledgment This work was funded by grant 1 R43 GM067327-01 from the National Institutes of Health.

References

1. Purvis G (2002) CAChe Worksystem Pro 5.0. CAChe Group, Fujitsu America, Inc, USA
2. Andzelm J, Wimmer E (1992) J Chem Phys 96:1280–1303
3. Kohn W, Sham LJ (1965) Phys Rev B 140:A1133-A1138
4. Hohenberg P, Kohn W (1964) Phys Rev B 136:B864-B871
5. Lee C, Parr RG, Yang W (1988) Phys Rev B 37:785–789
6. Perdew JP (1991) Electronic properties of solids. Akademie Verlag, Berlin
7. Perdew JP, Chevary JA, Vosko SH, Jackson KA, Pederson MR, Singh DJ, Fiolhais C (1992) Phys Rev B 46:6671–6687
8. Becke A (1988) Phys Rev A 38:3098–3100
9. Godbout N, Salahub DR, Andzelm J, Wimmer E (1992) Can J Chem 70:560-571
10. Dewar MJS, Storch DM (1985) J Am Chem Soc 107:3898–3902
11. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:4907–4917
12. Dewar MJS, McKee ML (1977) J Am Chem Soc 99:5231–5241
13. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909
14. Stewart JJP (1989) J Comput Chem 10:209–220
15. Stewart JJP (1989) J Comput Chem 10:221–264
16. Stewart JJP (2002) MOPAC 2002. CAChe Group, Portland
17. Stewart JJP (2002) MOPAC2002 1.0. Fujitsu, Ltd, Tokyo, Japan
18. Linstrom PJ, Mallard WG (2003) NIST Chemistry WebBook, NIST Standard Reference Number 69 (http://webbook.nist.gov/chemistry)
19. Stewart JJP (2003) J Phys Chem Ref Data (accepted)