JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Benchmarking Semiempirical Methods for Thermochemistry, Kinetics, and Noncovalent Interactions: OMx Methods Are Almost As Accurate and Robust As DFT-GGA Methods for Organic Molecules

Martin Korth* and Walter Thiel*

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany

ABSTRACT: Semiempirical quantum mechanical (SQM) methods offer a fast approximate treatment of the electronic structure and the properties of large molecules. Careful benchmarks are required to establish their accuracy. Here, we report a validation of standard SQM methods using a subset of the comprehensive GMTKN24 database for general main group thermochemistry, kinetics, and noncovalent interactions, which has recently been introduced to evaluate density functional theory (DFT) methods (*J. Chem. Theory Comput.* **2010**, *6*, 107). For all SQM methods considered presently, parameters are available for the elements H, C, N, and O, and consequently, we have extracted from the GMTKN24 database all species containing only these four elements (excluding multireference cases). The resulting GMTKN24-hcno database has 370 entries (derived from 593 energies) compared with 715 entries (derived from 1033 energies) in the original GMTKN24 database. The current benchmark covers established standard SQM methods (AM1, PM6), more recent approaches with orthogonalization corrections (OM1, OM2, OM3), and the self-consistent-charge density functional tight binding method (SCC-DFTB). The results are compared against each other and against DFT results using standard functionals. We find that the OMx methods outperform AM1, PM6, and SCC-DFTB by a significant margin, with a substantial gain in accuracy especially for OM2 and OM3. These latter methods are quite accurate even in comparison with DFT, with an overall mean absolute deviation of 6.6 kcal/mol for PBE and 7.9 kcal/mol for OM3. The OMx methods are also remarkably robust with regard to the unusual bonding situations encountered in the "mindless" MB08−165 test set, for which all other SQM methods fail badly.

1. INTRODUCTION

Semiempirical quantum mechanical (SQM) methods are based on self-consistent-field molecular orbital theory. They employ a minimal valence basis set, integral approximations, and parametrized matrix elements that are normally fitted against experimental data. The most popular SQM methods make use of the NDDO (neglect of diatomic differential overlap) integral approximation as implemented in the MNDO (modified neglect of differential overlap) model,[1] for example, AM1,[2] PM3,[3] and PM6.[4] More recent NDDO-based approaches go beyond the MNDO model by including orthogonalization corrections into the Fock matrix. OM1[5] incorporates these corrections only in the one-center part of the core Hamiltonian matrix, while OM2[6,7] and OM3[8] include them also in the two-center part. OM3 differs formally from OM2 by neglecting some of the smaller correction terms, which results in a speedup without loss of accuracy.[8,9]

SQM methods are widely applied as an efficient tool in computational studies of large molecules, and there are several reviews that describe the underlying theory and typical applications.[10−15] In their formalism, they retain the essential physics of molecular systems by variationally optimizing the electronic wave function (thereby taking into account, for example, polarization and charge transfer effects). However, the severe approximations adopted in these methods must cause errors which can only partially be compensated for by the parametrization against (mostly experimental) reference data. Careful validation of SQM methods is thus essential to establishing their accuracy and robustness, and corresponding evaluations are available for all popular SQM methods. These evaluations normally utilize benchmark data sets assembled in the SQM community[2−4,8,9,16−19] as well as data commonly used in the ab initio and DFT communities like the G2 and G3 sets.[20−22]

Validations of existing SQM methods need to be updated when more comprehensive and/or more accurate benchmark data sets become available. In this article, we report the results of such an evaluation against the recently proposed GMTKN24 database[23] for two of the established MNDO-type methods (AM1, PM6) and three NDDO-based methods with orthogonalization corrections (OM1, OM2, OM3). AM1 is still the most widely used SQM method, PM6 is the latest and most refined parametrization of MNDO-type models, and the methods of the OMx family have been most promising in previous SQM validations.[9] In our comparisons, we include the self-consistent-charge density functional tight binding (SCC-DFTB) method,[24] which is derived within a simplified DFT formalism but shares many features with standard SQM methods.[9] Furthermore, we also address the performance of standard DFT methods (PBE,[25] B3LYP[26,27]) to put the SQM results into perspective.

2. THE GMTKN24 BENCHMARK DATABASE

Recently, Goerigk and Grimme introduced a comprehensive quantum chemistry benchmark database for general main group thermochemistry, kinetics, and noncovalent interactions named GMTKN24.[23] It consists of 24 different, chemically relevant subsets that are either taken from the literature or compiled

Received: June 22, 2011
Published: August 05, 2011

ACS Publications © 2011 American Chemical Society

Table 1. Description of the Subsets within the GMTKN24-hcno Database[a]

set	description	no. entries (orig.[b])	av. energy	reference data
ACONF	relative energies of alkane conformers	15(15)	1.8	W1h-val
BH76RC	reaction energies of the BH76 set	17(30)	21.5	W1 and theor. est.
BH76	barriers of substitution and association reactions	38(76)	18.5	W1 and theor. est.
BHPERI	barriers of pericyclic reactions	22(26)	19.4	W1 and CBS-QB3
DARC	reaction energies of Diels−Alder reactions	14(14)	32.2	est. CCSDT/CBS
DC9	nine difficult cases for DFT	6(9)	35.7	theor. and expt.
G21EA	adiabatic electron affinities	11(25)	33.6	exptl.
G21IP	adiabatic ionization potentials	13(36)	250.8	exptl.
G2RC	reaction energies of selected G2/97 systems	12(25)	50.6	exptl.
IDISP	intramolecular dispersion interactions	6(6)	14.1	theor. and exptl.
ISO34	isomerization energies of organic molecules	34(34)	14.3	exptl.
MB08−165	decomposition energies of artificial molecules	21(165)	117.2	est. CCSD(T)/CBS
O3ADD6	energies and barriers for ozone reactions	6(6)	22.7	est. CCSD(T)/CBS
PA	adiabatic proton affinities	8(12)	174.9	est. CCSD(T)/CBS and W1
PCONF	relative energies of tripeptide conformers	10(10)	1.5	est. CCSD(T)/CBS
RSE43	radical stabilization energies	28(43)	7.5	est. CCSD(T)/CBS
S22	binding energies of noncovalently bound dimers	22(22)	7.4	est. CCSD(T)/CBS
SCONF	relative energies of sugar conformers	17(17)	4.9	est. CCSD(T)/CBS
SIE11	self-interaction error related problems	4(11)	34.0	est. CCSD(T)/CBS
W4−08woMR	atomization energies of small molecules	39(83)	237.5	W4
WATER27	binding energies of water/H^+/OH^- clusters	27(27)	82.0	est. CCSD(T)/CBS; MP2/CBS

[a] Based on a similar table in ref 23. [b] Number of entries in the full GMTKN24 database.

specifically for the purpose of benchmarking. It includes both theoretical or experimental reference values. When running the full benchmark (including multireference cases), 1049 single-point calculations are needed to determine 731 relative energies, which can then be compared with the accurate reference data. Extensive tests have shown the chemical relevance of the GMTKN24 database and its usefulness for evaluating the overall performance of theoretical methods. Goerigk and Grimme recommend validation against their benchmark data for the evaluation of the “true” performance of new quantum mechanical methods.[23]

Table 1 lists, in alphabetical order, the 21 subsets of the GMTKN24 database that contain molecules consisting only of the elements H, C, N, and O (excluding three presently irrelevant subsets without such molecules). Several subsets focus on noncovalent interactions and conformational preferences (ACONF, IDISP, PCONF, S22, SCONF, WATER27). Others address reaction, isomerization, and atomization energies (BH76RC, DARC, G2RC, ISO34, W4−08); barrier heights (BH76, BHPERI); electron affinities (G21EA); ionization potentials (G21IP); proton affinities (PA); and radical stabilization energies (RSE43), and there are two sets that collect difficult cases for DFT methods (DC9, SIE11). The MB08−165[28] subset is special in that it is based on a diversity-oriented approach. It consists of randomly generated artificial molecules which are constructed by applying systematic constraints (rather than any chemical bias) to open the narrow structural space of chemical intuition and to produce “electronically difficult” species with unusual and diverse geometries. The artificial molecules in the MB08−165 subset contain eight main-group atoms and are of single-reference character. The MB08−165 reference data are reaction energies for decomposition into small hydrides and diatomics obtained from coupled cluster [CCSD(T)] calculations with complete basis set (CBS) extrapolation.[28] The performance for the MB08−165 subset is considered to be a good indicator for general robustness in diverse chemical applications; this is supported by the finding that the performance ranking for DFT functionals is similar for the MB08−165 subset and for the whole GMTKN24 database, indicating the usefulness of this “mindless” benchmark for a quick performance assessment.[23,28]

In the original GMTKN24 publication, the authors used mean absolute deviations (MADs) in their comparisons for individual subsets and weighted total MADs (WTMADs) in their overall statistical analysis.[23] WTMADs take into account the number of entries in the test set (like a simple overall MAD would) but also subset-specific factors defined as the ratio between the corresponding MADs for the BLYP and B2PLY-D methods, in order to capture the “difficulty” of a certain subset and the importance of crucial dispersion interactions. Because of the large difference in the accuracy of BLYP and B2PLYP-D for dispersion effects, the subsets focusing on noncovalent interactions acquire particularly large weights and thus contribute prominently to the WTMAD values: the IDISP, PCONF, and S22 sets account for 38 out of 740 entries in the database and typically for almost one-fifth of the WTMAD value.[23] As a result, the inclusion of dispersion corrections in DFT methods leads to a general and rather large improvement with regard to WTMAD but to a less pronounced and less systematic improvement with regard to the overall MAD (OVMAD) for the whole database (which is available from the Supporting Information of the original GMTKN24 publication[23]). Furthermore, for DFT methods without dispersion corrections, the ranking of the different functionals is similar with respect to the WTMAD and OVMAD values. In this article, we shall present both WTMAD and OVMAD values but focus on the latter in the statistical analysis since they seem better suited for a general-purpose evaluation

Table 2. Mean Absolute Deviations (MADs) in kcal/mol for the GMTKN24-hnco and Full GMTKN24 Sets with PBE

method	PBE/(aug-)def2-QZVP[a]	PBE/TZVP[b]	PBE/TZVP[b]
sets	GMTKN24	GMTKN24	GMTKN24-hcno
ACONF	0.6	0.65	0.65
BH76	9.2	9.92	8.94
BH76RC	4.3	3.63	3.54
BHPERI	2.9	2.34	2.84
DARC	6.8	5.48	5.48
DC9	10.8	12.03	9.30
G21EA	3.4	6.43	7.17
G21IP	3.9	3.71	4.83
G2RC	6.2	9.07	7.71
IDISP	12.3	11.50	11.50
ISO34	1.8	1.68	1.68
MB08−165	9.0	9.26	9.87
O3ADD6	4.4	4.71	4.71
PA	2.1	2.29	2.47
PCONF	3.9	3.54	3.54
RSE43	3.4	3.48	3.62
S22	2.6	2.31	2.31
SCONF	0.4	0.78	0.78
SIE11	12.0	10.66	12.78
W4−08	11.0	8.65	10.41
WATER27	3.2	20.87	20.87

[a] From the Supporting Information of ref 23. Diffuse functions added to aug-cc-pVQZ for G21EA and WATER27. [b] This work.

of SQM methods (avoiding special emphasis on dispersion interactions). Regardless of this choice, it is clear that dispersion corrections are essential when noncovalent interactions play an important role, and we thus evaluate the performance of SQM methods without and with dispersion corrections.[29−33]

An extended version of the GMTKN24 database, named GMTKN30,[34] was published shortly after completion of our study. The new database contains six additional benchmark sets, three of which are made up of molecules with elements other than H, C, N, and O (ALK6, RG6, HEAVY28) and can thus not be applied here. The remaining three sets address further noncovalent interactions (ADIM), further isomerization reactions (ISOL22) for large molecules with less accurate SCS-MP3/CBS reference data, and hydrocarbon bond separation reactions (BSR36) with significant differences between the theoretical and experimental reference values. We consider these additional data in the GMTKN30 database[34] to be less crucial in the present context and thus decided to disregard them and to focus on the comparison between GMTKN24 and GMTKN24-hcno results.

3. THE GMTKN24-HCNO BENCHMARK DATABASE

Starting from the original GMTKN24 database, we have compiled the GMTKN24-hcno benchmark for SQM methods by stripping the 715 relative energies (1033 single-point energies) of the original set (excluding multireference cases) from all entries that contain elements other than H, C, N, and O, thus arriving at 370 relative energies (593 single-point energies) that are suited for all common SQM methods considered presently.

Table 1 shows the resulting number of entries for all subsets. From the original GMTKN24 database, the three subsets focusing on aluminum (AL2X), boron (NBRC), and sulfur (CYCONF) chemistry were completely skipped, as all entries contain either aluminum, boron, or sulfur. Owing to its pronounced diversity, only 21 of 165 entries of the MB08−165 benchmark could be kept. A large number of entries had to be skipped also for the G21EA, G21IP, G2RC, SIE11, and W4−08 sets (up to two-thirds) and for the BH76RC, BH76, BHPERI, DC9, PA, and RSE43 sets (up to one-half). All entries could be retained for ACONF, DARC, IDISP, ISO34, O3ADD6, PCONF, S22, SCONF, and WATER27.

Is the difficulty of the GMTKN24 benchmark greatly diminished by leaving out the entries specified above? This question can be addressed by comparing MADs from DFT calculations for the full GMTKN24 database and the reduced GMTKN24-hcno database. We have thus performed PBE/TZVP calculations for the reduced and full sets and compared the results with the published PBE/(aug-)def2-QZVP data for the full set.[23] The MADs for most subsets are obviously quite similar (see Table 2), indicating that the reduced GMTKN24-hcno subsets indeed retain the characteristic features of the full GMTKN24 subsets. Large discrepancies are found only in subsets that focus on electron affinities (G21EA) or negatively charged species (WATER27) and thus exhibit a strong basis set dependence in DFT calculations, as has been documented previously.[23]

One may also ask whether the overall performance of SQM methods can be assessed from benchmarking systems containing only H, C, N, and O. This can be checked for SQM methods that have also been parametrized for other elements, by adding to GMTKN24-hcno the corresponding reference data from the full GMTKN24 database. We tested this with PM6 for the full MB08−165 subset (165 rather than 21 entries) and found only small changes in performance (MAD 119.0 rather than 128.4 kcal/mol). In the case of the orthogonalization-corrected SQM methods (OM1, OM2, OM3), we extended the GMTKN24-hcno database by including all molecules also containing fluorine (413 instead of 370 entries), which led to only minor changes of 0.1−0.4 kcal/mol in the OVMAD values, with similar trends in performance for the fluorine-containing and other molecules (for further details see section 5.3).

In summary, these tests suggest that the GMTKN24-hcno benchmark database is well suited to serve for the purpose of evaluating SQM methods.

4. COMPUTATIONAL DETAILS

PBE[25] and B3LYP[26,27] DFT calculations with and without dispersion corrections of DFT-D2 type[35] were done using the Turbomole 5.9 software,[36] TZVP Gaussian basis sets,[37] and (in the case of PBE) the resolution-of-identity approximation[38,39] for two-electron integrals. SQM calculations were carried out with MOPAC2009[40] for AM1 and PM6; with DFTBplus[41] for SCC-DFTB; and with MNDO99[42,43] for OM1, OM2, and OM3, as well as PM3[3] and PM3-PDDG.[18] The SQM methods were enhanced with standard D2 dispersion corrections using the published parameters for AM1-D,[33] PM6-D,[33] OMx-D,[30] and SCC-DFTB-D.[31] These corrections do not involve any changes in the standard SQM parameters, unlike an alternative AM1-based approach.[29] The SQM calculations for open-shell molecules employed a restricted ROHF treatment in the case of MNDO99 and an unrestricted UHF scheme in the case of MOPAC2009. Entries involving triplets or quartets were skipped for SCC-DFTB, because the available software did not allow black-box benchmarking of such species.

5. RESULTS AND DISCUSSION

We begin with two introductory remarks. First, when calculating proton affinities with SQM methods, we follow the convention to use the experimental heat of formation for the proton since all investigated SQM methods are known to be off by several tens of kilocalories per mole for this quantity. Second, the cage/bowl isomerization of C_{20} in the DC9 subset—which is known to be problematic even for high-level *ab initio* methods because of partial multireference effects—is not described adequately by any of the investigated SQM methods, with errors exceeding 100 kcal/mol (see Table 3), which are thus much larger than the estimated uncertainty of about 10 kcal/mol in the *ab initio* reference value.[23] We therefore decided to exclude this item from the statistical analysis, thus reducing the number of entries in our GMTKN24-hcno database to 370. Removing this outlier decreases the OVMAD values for the SQM methods by 0.3−0.8 kcal/mol but does not influence our conclusions on their relative merits.

Table 3. Errors (kcal/mol) with PBE and SQM Methods for the C_{20} Cage/Bowl Isomerization, Relative to the CCSD(T)/CBS Estimate

	PBE/TZVP	PM6	AM1	OM3	OM2	OM1
C_{20} cage/bowl	−5.7	102.6	203.8	206.9	193.4	325.4

The results of our benchmarks are presented as follows: Tables 4 and 5 show SQM and DFT results for the GMTKN24-hcno benchmark database without and with empirical dispersion corrections. Table 6 summarizes the OVMAD values for several SQM and DFT methods. Table 7 shows the effect of including entries with F on the OMx MADs. Figures 1 and 2 compare the MAD values of OM3 with those of PM6 and PBE, respectively. Figure 3 shows the element-wise error of PM6 for the MB08−165 subset. We use Tables 4 and 5 and Figure 1 to compare the SQM methods with each other (subsection 5.1), Tables 4−6 and Figure 2 to compare the SQM methods with DFT (subsection 5.2), and Table 7 and Figure 3 to discuss the effect of taking other elements into account (subsection 5.3).

5.1. Comparison of SQM Methods. Perusing Table 4, the following observations regarding the different SQM methods can be made:

For most subsets, all OMx methods perform roughly similarly well. Exceptions are the PA and O3ADD6 sets where OM1 is best, whereas OM2 and OM3 are better than OM1 for the G21IP, G21EA, and WATER27 sets as well as in the description of noncovalent interactions (see for instance the IDISP set; the effect of including empirical dispersion corrections is discussed below separately). Consequently, the overall deviation (OVMAD) is substantially lower for OM2 and OM3 (8.3 and 7.9 kcal/mol, respectively) than for OM1 (10.9 kcal/mol). The differences between OM2 and the slightly faster OM3 method are small, but

Table 4. Mean Absolute Deviations (MADs) in kcal/mol for the GMTKN24-hnco Sets: PBE/TZVP, B3LYP/TZVP, and SQM Methods

set	PBE	B3LYP	OM3	OM2	OM1	PM6	AM1	SCC-DFTB
ACONF	0.65	0.77	0.86	0.63	0.52	0.56	0.44	0.23
BH76	8.94	4.82	8.69	7.58	10.42	13.81	10.88	14.82
BH76RC	3.54	2.42	6.18	4.09	5.13	17.64	12.46	12.64
BHPERI	2.84	4.42	8.82	8.79	11.31	10.36	10.59	6.98
DARC	5.48	13.65	4.91	7.25	4.10	3.91	4.65	3.55
DC9	9.30	11.36	13.20	13.60	11.40	5.18	15.68	15.22
G21EA	7.17	8.50	9.91	11.70	24.45	22.06	23.03	7.77
G21IP	4.83	4.82	12.72	12.53	22.07	40.14	24.31	15.96
G2RC	7.71	2.69	4.53	8.58	8.68	30.87	12.43	27.97
IDISP	11.50	17.04	6.67	8.19	14.17	14.27	14.01	13.13
ISO34	1.68	2.39	4.37	4.44	4.45	3.46	6.45	4.66
MB08−165	9.87	5.85	21.32	22.00	18.82	128.43	44.44	100.20
O3ADD6	4.71	1.83	10.97	12.24	4.01	2.03	10.57	7.51
PA	2.47	3.06	11.85	14.69	4.90	18.41	12.82	18.58
PCONF	3.54	3.84	1.32	1.28	3.60	2.27	5.35	1.68
RSE43	3.62	2.50	5.24	4.28	3.95	5.20	2.46	9.56
S22	2.31	3.49	3.58	3.07	5.14	3.41	6.83	3.55
SCONF	0.78	0.33	1.32	1.66	5.87	2.62	2.39	2.08
SIE11	12.78	6.53	5.00	9.38	5.15	3.29	10.65	20.83
W4−08woMR	10.41	3.46	11.82	12.79	12.08	15.57	14.35	13.90
WATER27	20.87	11.42	9.19	12.11	36.09	17.81	48.60	22.87
OVMAD	6.60	4.82	7.86	8.33	10.93	18.19	14.52	
OVMAD[a]	5.89	4.73	6.76	7.06	9.68	14.60	13.54	13.87
OVMAD*[b]	6.40	4.76	7.05	7.51	10.46	11.56	12.72	
OVMAD*[a,b]	5.78	4.80	6.21	6.68	9.51	10.04	12.17	10.26
WTMAD	5.7	5.1	6.4	6.7	9.1	13.3	11.6	

[a] Without entries involving triplets or quartets, which reduces the size of the GMTKN24-hcno benchmark database from 370 to 299 entries. [b] Overall MAD without contributions from the MB08−165 subset.

Table 5. Mean Absolute Deviations (MADs) in kcal/mol for the GMTKN24-hnco Sets: PBE/TZVP, B3LYP/TZVP, and SQM Methods with Empirical Dispersion Corrections (-D)

set	PBE-D	B3LYP-D	OM3-D	OM2-D	OM1-D	PM6-D	AM1-D	SCC-DFTB-D
ACONF	0.22	0.21	0.32	0.31	0.41	0.69	1.83	0.53
BH76	9.06	3.90	8.98	7.51	10.01	13.78	10.62	14.88
BH76RC	3.56	2.50	6.30	4.00	5.22	17.65	12.47	12.65
BHPERI	3.55	3.07	7.51	7.14	9.47	9.78	8.50	7.27
DARC	3.67	9.61	8.30	9.99	3.98	4.50	7.65	4.62
DC9	8.04	8.44	12.31	15.07	9.75	4.68	12.47	15.27
G21EA	8.85	8.45	9.92	11.71	24.46	22.06	23.03	7.77
G21IP	4.83	4.83	12.73	12.55	22.09	40.14	24.31	15.96
G2RC	7.87	2.68	4.02	8.10	8.20	30.85	12.04	28.00
IDISP	6.11	7.38	8.42	12.43	13.24	15.51	15.29	12.24
ISO34	1.60	2.08	4.43	4.44	4.37	3.42	6.54	4.60
MB08−165	9.92	5.57	21.92	21.88	18.78	128.51	45.85	100.02
O3ADD6	4.93	1.86	11.13	12.67	3.57	1.81	9.48	7.46
PA	2.62	3.05	11.66	14.76	4.98	18.40	12.67	18.54
PCONF	1.36	0.58	2.07	2.01	4.01	3.02	5.13	0.67
RSE43	3.43	2.27	4.98	4.03	3.84	5.19	2.38	9.53
S22	0.86	0.76	1.10	1.06	2.41	1.65	2.95	1.86
SCONF	0.87	0.57	1.40	1.64	5.18	2.68	1.87	2.12
SIE11	13.51	7.32	5.51	9.96	5.08	3.11	8.77	21.01
W4−08woMR	10.41	3.67	11.65	12.61	11.90	15.56	14.20	13.90
WATER27	26.19	19.43	7.80	4.24	27.58	14.92	36.78	22.21
OVMAD	6.76	4.58	7.70	7.69	9.87	17.89	13.40	
OVMAD[a]	6.05	4.60	6.57	6.31	8.42	14.23	12.14	13.72
OVMAD*[b]	6.57	4.52	6.84	6.83	9.34	11.24	11.45	
OVMAD*[a,b]	5.92	4.64	5.97	5.89	8.19	9.65	10.65	10.11
WTMAD	5.2	3.9	6.2	6.3	8.8	12.9	10.5	

[a] Without entries involving triplets or quartets, which reduces the size of the GMTKN24-hcno benchmark database from 370 to 299 entries. [b] Overall MAD without contributions from the MB08−165 subset.

Table 6. Overall Mean Absolute Deviations (OVMADs) in kcal/mol for the GMTKN24-hnco Set: DFT/TZVP and SQM Methods

method	without -D	with -D
PM6	18.2	17.9
AM1	14.5	13.4
OM3	7.9	7.7
BLYP	6.6	6.3
PBE	6.6	6.8
BP86	6.0	6.2
TPSS	5.5	5.6
B3LYP	4.8	4.6

Table 7. Mean Absolute Deviations (MADs) in kcal/mol for the GMTKN24-hnco Set and the GMTKN24-hcnof Set (Including Entries with F): OM1, OM2, and OM3

	OM1		OM2		OM3	
set	hcno	hcnof	hcno	hcnof	hcno	hcnof
BH76	10.39	10.42	7.58	9.72	8.69	10.66
BH76RC	5.28	5.13	4.09	4.29	6.18	5.37
G21EA	24.81	24.45	11.70	11.39	9.91	9.31
G21IP	22.45	22.07	12.53	12.00	12.72	11.45
G2RC	9.07	8.68	8.58	8.23	4.53	4.16
MB08−165	19.47	18.82	22.00	22.47	21.32	19.46
RSE43	3.86	3.95	4.28	4.02	5.24	4.96
SIE11	4.40	5.15	9.38	9.38	5.00	5.00
W4−08woMR	11.49	12.08	11.82	12.28	11.82	11.38
all	11.01	10.93	8.33	8.68	7.86	8.01

OM3 generally outperforms OM2 when larger differences occur (G2RC, SIE11), making OM3 overall the best OMx model in the GMTKN24-hcno benchmark.

PM6 improves upon AM1 for a number of demanding subsets (DC9, O3ADD6, SIE11) and in the treatment of noncovalent interactions (PCONF, S22, WATER27) but is less convincing than AM1 for a number of other sets with electronically complicated species (BH76RC, G21IP, G21EA, G2RC, PA, RSE43). AM1 has large problems with the MB08−165 subset (MAD 44.4 kcal/mol), but PM6 performs even worse (MAD 128.4 kcal/mol). Mostly for this reason, the overall deviation (OVMAD) is larger for PM6 (18.2 kcal/mol) than for AM1 (14.5 kcal/mol). OM3 clearly outperforms both PM6 and AM1 on a number of subsets (G21IP, G21EA, G2RC, IDISP, WATER27, and most importantly MB08−165) but fails to reach the outstanding accuracy of PM6 for DC9 and O3ADD6.

SCC-DFTB improves upon PM6 and AM1 for the G21EA, G21IP, and G2RC sets but shows large errors for SIE11 and

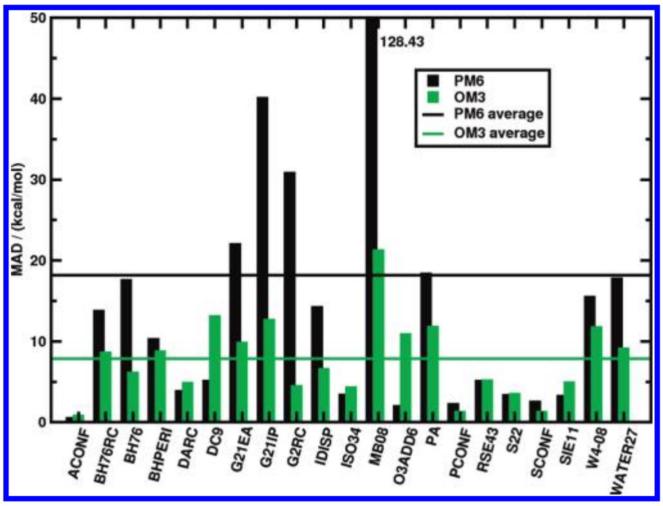

Figure 1. Comparison of OM3 and PM6 results for the GMTKN24-hcno set.

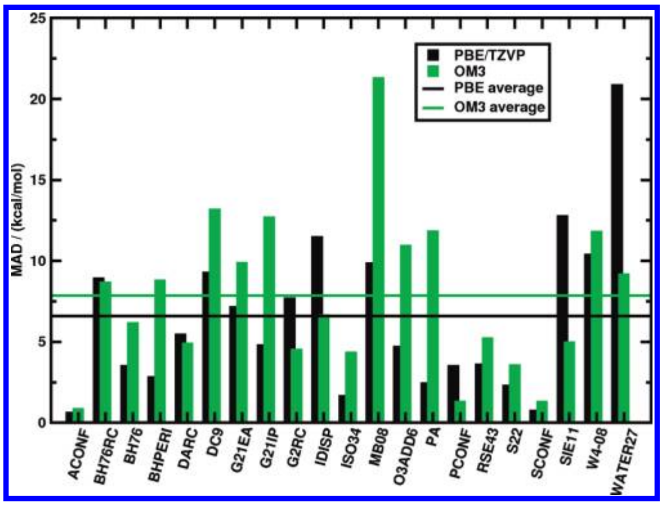

Figure 2. Comparison of OM3 and PBE/TZVP results for the GMTKN24-hcno set.

especially for the MB08−165 set (like PM6 and AM1). SCC-DFTB is inferior to OM3 for the G2RC, PA, RSE43, and WATER27 sets and even more so for SIE11 and MB08−165. As noted above (section 4), triplet and quartet species were excluded from the SCC-DFTB benchmark runs for technical reasons, and hence the overall deviation (OVMAD 13.9 kcal/mol) refers to a smaller sample. For the sake of comparison, OVMAD values for this smaller sample are given in Table 4 also for the other methods.

In addition to the methods shown in Table 4, we have also looked at the performance of the pairwise distance directed Gaussian (PDDG)[18] approach in combination with PM3[3] as implemented in MNDO99. The OVMAD value of PDDG-PM3 is 17.1 kcal/mol (12.5 kcal/mol without the MB08−165 set), about 2 kcal/mol (1 kcal/mol) higher than the value for PM3 itself (14.7 and 11.4 kcal/mol), which performs similarly to AM1 (14.5 and 12.7 kcal/mol) for our database.

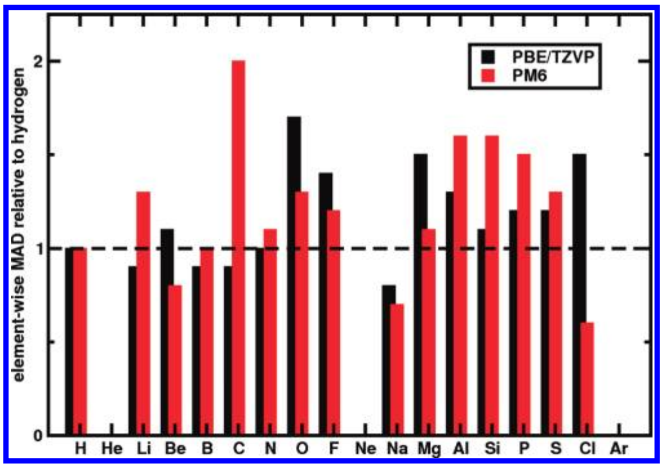

Figure 3. Element-wise MAD values relative to hydrogen for the MB08−165 set (see text for details).

Since the “mindless” MB08−165 benchmark set with its artifical molecules is particularly demanding, we also provide in Table 4 overall deviations without the MB08−165 contributions (OVMAD*). These deviations decrease in the following order:

AM1 > *SCC-DFTB*≈PM6 > OM1 ≫ OM2≈OM3 > PBE > B3LYP

The OVMAD* values for AM1, PM6, SCC-DFTB, and OM1 are rather similar (ranging between 12.7 and 10.3 kcal/mol) although the performance for a specific subset may be quite different between these methods. OM2 and OM3 are distinctly more accurate (OVMAD* 7.5 and 7.1 kcal/mol, respectively) and actually approach the overall accuracy of PBE (OVMAD* 6.4 kcal/mol) while B3LYP performs best (OVMAD* 4.8 kcal/mol).

Inclusion of the artificial molecules from the MB08−165 set generally deteriorates the error statistics, but to different extents. The overall deviations (OVMAD vs OVMAD*) become much worse for PM6 and SCC-DFTB (increase by 6.6 and 3.6 kcal/mol), somewhat worse for AM1 (by 1.8 kcal/mol), and only slightly worse for the OMx methods (by 0.5−0.8 kcal/mol). The OMx methods are thus remarkably robust in this regard, again reminiscent of the performance of PBE (increase by 0.2 kcal/mol). The superior performance of the OMx methods (compared with the other SQM methods) may be viewed as support for the OMx approach of going beyond the MNDO model: including more physics in the model appears to be a better strategy for coping with the electronically demanding MB08−165 species than going for a better parametrization. On the basis of the OVMAD values, the overall deviations decrease in the following sequence:

PM6 ≫ *SCC-DFTB*≈AM1 ≫ OM1 ≫ OM2≈OM3 > PBE > B3LYP

Figure 1 illustrates the relative merits of OM3 and PM6 by plots of the overall deviations (OVMAD) and of the deviations (MAD) of the individual subsets of the GMTKN24-hcno benchmark.

Table 5 shows that the inclusion of empirical dispersion corrections into the SQM methods leads to a clear improvement for all subsets in which intermolecular noncovalent interactions play an important role (most notably S22 and WATER27), but there is also some minor deterioration for other subsets. Overall, there is a general small gain in accuracy for all SQM methods considered, with a decrease of 0.2−1.3 kcal/mol in the OVMAD and OVMAD* values. The accuracy ranking of the SQM methods in our benchmark is not affected by the inclusion of empirical dispersion corrections, since their effect on the statistics is smaller than the underlying intrinsic differences and since all SQM methods benefit to a similar extent.

As noted above (see section 2), the use of weighted total MADs (WTMADs) emphasizes the importance of noncovalent interactions in the benchmark. These WTMADs are listed in Tables 4 and 5. They clearly reflect the gain from adding dispersion corrections to the DFT methods: Upon inclusion of these corrections, the WTMADs decrease by 0.5 kcal/mol for PBE and by 1.2 kcal/mol for B3LYP, while the OVMADs change by only 0.2 kcal/mol (increasing for PBE and decreasing for B3LYP). In the case of the SQM methods, both the WTMADs and OVMADs are generally lowered by the dispersion corrections, typically by 0.2 to 0.6 kcal/mol (for OM1 and AM1 by up to 1.1 kcal/mol). Focusing on the dispersion-corrected methods (Table 5) the WTMADs are always smaller than the OVMADs (PBE-D by 1.6 kcal/mol, B3LYP-D by 0.7 kcal/mol, OMx-D by 1.1 to 1.5 kcal/mol, other SQM-D by 2.9 to 5.0 kcal/mol), with OM2-D and OM3-D again showing the best performance among all SQM methods and approaching PBE-D accuracy. Since the emphasis of this study is not on noncovalent interactions, we will disregard dispersion corrections from now on and again employ OVMAD instead of WTMAD values in the analysis.

5.2. Comparison of SQM and DFT Methods.

OM3 is the most accurate and robust of the SQM methods considered presently and has therefore been chosen for a comparison with DFT/TZVP methods. Among the available DFT functionals, we focus mainly on PBE and B3LYP, which are commonly used representatives of GGA (generalized gradient approximation) and hybrid-GGA functionals. The data in Tables 4 and 6 indicate that B3LYP (OVMAD 4.8 kcal/mol) is on average more accurate than PBE (OVMAD 6.6 kcal/mol) for organic molecules. Other common GGA functionals perform similarly to PBE (OVMAD 6.6 kcal/mol for BLYP, 6.0 kcal/mol for BP86), while at the meta-GGA level, TPSS shows an intermediate performance (OVMAD 5.5 kcal/mol). OM3 (OVMAD 7.9 kcal/mol) is surprisingly close in overall accuracy to standard DFT-GGA methods. This is visualized in Figure 2 showing the MADs for the subsets of the GMTKN24-hcno benchmark for OM3 and the PBE functional. It is obvious that PBE outperforms OM3 for several sets (BHPERI, G21IP, MB08−165, O3ADD4, PA), but there are also sets where the opposite is true (IDISP, SIE11, WATER27). These latter cases merit further comments.

The IDISP set contains molecules in which intramolecular noncovalent interactions are of crucial importance. Such intramolecular effects are partially taken into account by the OM3 parametrization (unlike intermolecular effects that are not covered), and it is therefore not surprising that the inclusion of empirical dispersion corrections actually deteriorates the OM3 results for IDISP (in contrast to the improvements for S22 where intermolecular dispersion effects are dominant). On the other hand, dispersion is generally missing at the PBE level, and the dispersion corrections in PBE-D thus yield substantial improvements both for IDISP and S22. Consequently, PBE-D has slightly lower MADs than OM3-D for both sets (IDISP, S22). This also suggests that a reparameterization of SQM methods with dispersion corrections included in the fit process is likely to be worthwhile, offering the chance for a more balanced treatment of intra- and intermolecular dispersion interactions.

The WATER27 set contains several negatively charged species. It is well-known[23] that accurate PBE calculations on these species require basis sets that are larger than the TZVP basis used presently, and extending the basis from TZVP to (aug-)def2-QZVP lowers the MAD of PBE for the WATER27 set from 20.9 to 3.2 kcal/mol (Table 2). When using a sufficiently large basis, PBE thus outperforms OM3 also for the WATER27 set (OM3 MAD 9.2 kcal/mol). On the other hand, the PBE problems related to self-interaction errors (SIE11) seem genuine since they are not alleviated by basis set extension (Table 2).

In summary, the PBE/TZVP results for the GMTKN24-hcno benchmark are somewhat more accurate than the OM3 results (OVMAD 6.6 vs 7.9 kcal/mol), which remains true after including empirical dispersion corrections (OVMAD 6.8 vs 7.7 kcal/mol). It should also be noted, that "high-end" functionals like the M0n family or double hybrids are substantially more accurate than the commonly used PBE and B3LYP methods for the benchmark sets featured in the GMTKN24 and GMTKN30 databases: The GMTKN24 WTMAD and OVMAD values for M06−2X[44] are both 2.2 kcal/mol, compared to 6.2 and 7.0 kcal/mol for PBE. The M06−2X OVMAD value for our reduced "hcno" set drops from 2.2 to 1.8 kcal/mol, well in line with the change for PBE from 7.0 to 6.6 kcal/mol, which provides further support to the transferability of our hcno results. An extensive collection of DFT data and a detailed analysis of the relative performance of a wide range of DFT functionals can be found in the original GMTKN24 and GMTKN30 publications by Grimme and Goerigk.[23,34]

5.3. Element-Specific Error Analysis.

PM6 has parameters for all elements appearing in the full GMTKN24 benchmark. We can therefore use PM6 to check whether and how the errors depend on the elements that are present in the reference molecules. For this purpose, we have analyzed the performance of PM6 for the full MB08−165 benchmark set (i.e., our most demanding subset containing artificial molecules with complicated electronic structure). Figure 3 shows the element-specific errors obtained as follows: For a given entry in the MB08−165 set, each element is assigned a fraction of the error in the computed absolute reaction energy according to its occurrence (number of atoms present divided by the total number of atoms). Then, an average is taken over the resulting values for all entries and weighted with the elemental occurrence in the full set, and finally this average value is divided by the corresponding value for hydrogen for the purpose of normalization.[28] The resulting error distribution in Figure 3 looks fairly balanced both for PM6 and for the PBE functional, indicating that the quality of the PM6 results is reasonably uniform for different elements. This implies that the present GMTKN24-hcno benchmark is expected to be of general relevance and that the conclusions derived from molecules containing only H, C, N, and O may be valid in general.

In a second test, we have extended the GMTKN24-hcno database by including all species from the full database that also include fluorine atoms. The resulting GMTKN24-hcnof database contains 413 entries derived from 654 single-point calculations (compared with 371 entries and 595 single-point calculations in GMTKN24-hcno). The corresponding OMx results are shown in Table 7 for all subsets that differ in the two databases. It is obvious that the MAD values remain essentially unchanged upon the addition of fluorine-containing molecules, with variations in the overall deviations (OVMAD) for the complete benchmark database of 0.1−0.4 kcal/mol.

The outcome of both tests suggests that our GMTKN24-hcno benchmark database is indeed well suited to serve the purpose of evaluating SQM methods.

5.4. Computational Costs.

The overall performance of OM3 seems satisfactory especially in view of the fact that the OM3 calculations are about 3 orders of magnitude faster than the PBE/TZVP calculations: Computation times are 4.5 s for OM3, 8296 s for PBE/TZVP, and 11865 s for B3LYP/TZVP (ratio 1:1844:2637) on one core of an Intel Xeon 5670 processor for the whole GMTKN24-hcno benchmark database. To compare the computational costs of the different SQM methods with each other, we use averages over 100 calculations of the complete database. For the systems investigated here, substantial differences are found mainly between programs and less so between methods. MNDO99 needs on average about 4 s for the complete database (AM1 3.9s, PM3 4.0s, OM2 4.2s, OM1 4.3s, OM3 4.5s); this value is roughly doubled for MOPAC2009 (AM1 8.0s, PM6 8.1s) and larger by a factor of about 9 for DFTB+ (SCC-DFTB 35.4s)

6. SUMMARY

We have presented a thorough evaluation of semiempirical QM methods based on a reduced version of the recently introduced GMTKN24 benchmark database. We find that the OMx family of methods outperforms the established SQM methods AM1, PM6, and SCC-DFTB by a significant margin. The overall differences between AM1, PM6, and SCC-DFTB are rather small, while OM2 and OM3 are substantially more accurate (by about 3 kcal/mol on average). Furthermore, the OMx family of methods is remarkably robust with regard to the unusual bonding situations in the artificial molecules from the "mindless" MB08−165 benchmark, where all other SQM methods fail badly. This provides further support to the OMx strategy of improving the adopted semiempirical model (instead of further parameter refinement). In the present GMTKN24-hcno benchmark, the OM2 and OM3 results are reasonably accurate and robust even in comparison to DFT(PBE) calculations.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: korth@mpi-muelheim.mpg.de, thiel@mpi-muelheim.mpg.de.

■ ACKNOWLEDGMENT

M.K. was supported by Grants LPDS-2009-19 and LPDR-2009-3 from the German National Academy of Sciences Leopoldina. The authors would like to thank Lars Goerigk and Stefan Grimme for kindly supplying additional unpublished DFT data for their benchmark databases.

■ REFERENCES

(1) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4799.

(2) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(3) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221.

(4) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.

(5) Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993**, *14*, 775.

(6) Weber, W. Ph.D. thesis, University of Zürich, Zürich, Switzerland, 1996.

(7) Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495.

(8) Scholten, M. Ph.D. thesis, University of Düsseldorf, Düsseldorf, Germany, 2003.

(9) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751.

(10) Stewart, J. J. P. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Vol. 1, pp 45−81.

(11) Thiel, W. *Adv. Chem. Phys.* **1996**, *93*, 703.

(12) Clark, T. J. *THEOCHEM* **2000**, *530*, 1.

(13) Thiel, W. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; John von Neumann Institut fur Computing: Jülich, 2000; NIC Series, Vol. 3, p 261.

(14) Bredow, T.; Jug, K. *Theor. Chem. Acc.* **2005**, *113*, 1.

(15) Thiel, W. In *Theory and Applications of Computational Chemistry*; Dykstra, C. E., Kim, K. S., Frenking, G., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 559−580.

(16) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4907.

(17) Thiel, W.; Voityuk, A. A. *J. Phys. Chem. Soc.* **1996**, *100*, 616.

(18) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.

(19) Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2008**, *4*, 297.

(20) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063.

(21) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.

(22) Redfern, P. C.; Zapol, P.; Curtiss, L. A.; Raghavachari, K. *J. Phys. Chem. A* **2000**, *104*, 5850.

(23) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2010**, *6*, 107.

(24) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.

(25) Perdew, J. P.; Burke, K.; Enzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(26) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(27) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(28) Korth, M.; Grimme, S. *J. Chem. Theory Comput.* **2009**, *5*, 993.

(29) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.

(30) Tuttle, T.; Thiel, W. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2159.

(31) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.

(32) Rezac, J.; Fanfrlik, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749.

(33) Korth, M.; Pitonak, M.; Rezac, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344.

(34) Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2011**, *7*, 291.

(35) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.

(36) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

(37) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(38) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652.

(39) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119.

(40) MOPAC2009. See http://openmopac.net/MOPAC2009.html (accessed Mar 8, 2011).

(41) DFTBplus. See http://www.dftb-plus.info (accessed Mar 8, 2011).

(42) Thiel, W. *MNDO99*, version 6.1; Max-Planck-Institut für Kohlenforschung: Mülheim, Germany, 2007.

(43) The OMx methods are implemented in the MNDO99 code, which is distributed by Scienomics (http://www.scienomics.com, accessed Aug 4, 2011) and is also available from one of the authors (W.T.).

(44) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.