J Mol Model (2003) 9:408–414
DOI 10.1007/s00894-003-0156-7

ORIGINAL PAPER

Paul Winget · Anselm H. C. Horn · Cenk Selçuki · Bodo Martin · Timothy Clark

AM1* parameters for phosphorus, sulfur and chlorine

Received: 11 June 2003 / Accepted: 16 July 2003 / Published online: 4 September 2003
© Springer-Verlag 2003

Abstract An extension of the AM1 semiempirical molecular orbital technique, AM1*, is introduced. AM1* uses AM1 parameters and theory unchanged for the elements H, C, N, O and F. The elements P, S and Cl have been reparameterized using an additional set of *d* orbitals in the basis set and with two-center core–core parameters, rather than the Gaussian functions used to modify the core–core potential in AM1. Voityuk and Rösch's AM1(d) parameters have been adopted unchanged for AM1* with the exception that new core–core parameters are defined for Mo–P, Mo–S and Mo–Cl interactions. Thus, AM1* gives identical results to AM1 for compounds with only H, C, N, O, and F, AM1(d) for compounds containing Mo, H, C, N, O and F only, but differs for molybdenum compounds containing P, S or Cl. The performance and typical errors of AM1* are discussed.

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00894-003-0156-7. Tables 2 and 4–7 and a full list (Tables S1, S2) of geometrical parameters and barrier heights are given in the supplementary material.

Keywords Semiempirical parameterization · AM1* · NDDO

Introduction

Despite many predictions of its demise, semiempirical molecular orbital (MO) theory remains important in modern computational chemistry. [1] The speed and excellent scaling characteristics of current semiempirical methods make them ideal computational tools for fast preliminary scans before using more expensive methods, for cheminformatics [2] and even for calculating the complete wavefunctions of proteins. [3, 4, 5] Perhaps the most important characteristic of semiempirical methods is that they allow far more complete and comprehensive studies of large systems than methods such as density functional theory. This is especially true when the system becomes large enough that adequate conformational sampling becomes an issue. Gregersen, Lopez and York recently impressively demonstrated the power of semiempirical methods in this respect. [6] The proliferation of commercial semiempirical MO software underlines the importance of current NDDO-based [7] techniques. Thiel and Voityuk's formulation of the integral approximations for *d* orbitals in MNDO/d [8, 9, 10, 11, 12, 13] opened the way for several newer methods based on AM1 [14, 15] or PM3. [16, 17, 18] Unfortunately, with the laudable exception Voityuk and Rösch's parameterization of an AM1(d) technique for molybdenum, [19] these parameterizations have not been published, presumably because the software companies expect some commercial advantage. Thus, the parameters for PM3(tm), [20] have not been published although those for the extended versions of AM1 and PM3 should soon become available. [21] Additionally, a completely new parameterization of an AM1-like method, PM5 is commercially available, but has not been published. [21] We thus have the unhappy situation that some published AM1 and PM3 parameters for sodium [22] are not those available under the same name in commercial software. [21].

The situation that the parameterizations, or even the details of the computational method, are not publicly available for almost all parameter sets for transition metals and that in some cases even the parameterization data are not available, is clearly incompatible with the guidelines to publishing semiempirical results [23] and with good scientific practice. We have therefore set out to develop, parameterize and publish an extension of AM1 [14, 15] that uses Thiel and Voityuk's *d* orbital formulation [8, 9, 10, 11, 12, 13] for an extended series of elements including transition metals. We have based our

P. Winget · A. H. C. Horn · C. Selçuki · B. Martin · T. Clark (✉)
Computer-Chemie-Centrum,
Friedrich-Alexander-Universität Erlangen-Nürnberg,
Nägelsbachstraße 25, 91052 Erlangen, Germany
e-mail: clark@chemie.uni-erlangen.de

method on AM1, rather than MNDO [24, 25] or PM3 [16, 17, 18] because AM1 reproduces the energies of hydrogen bonds (but not their geometries) relatively well and generally performs better for rotation barriers of partial double bonds (such as the C–N bond in amides) than the other two methods. We have named the new method AM1* to emphasize its relationship to AM1 and to distinguish it from Voityuk and Rösch's AM1(d), [19] which does not use *d* orbitals for elements of the second long period, and from the AM1(d) method available in MOPAC. [21] We have, however, used Voityuk and Rösch's AM1(d) parameters for molybdenum [19] with some slight changes for AM1*. In accord with the practice followed for MNDO/d, [8, 9, 10, 11, 12, 13] we have used the original AM1 parameters for H, C, N, O and F unchanged. [14, 15] We now report AM1* parameters for the main group elements P, S and Cl. Of these elements, phosphorus is probably the most difficult to treat with an *sp* basis, as demonstrated by the relatively high errors given by the newly parameterized PM5 [21] for phosphorus compounds. MNDO/d, on the other hand, uses *d* orbitals for phosphorus with considerable success. [8, 9, 10, 11, 12, 13] Very recently, Lopez and York published an AM1(d) parameter set for phosphorus designed to be used specifically for nucleophilic attack on biological phosphates. [26] Such specialized parameterizations should prove extremely useful in studying specific problems. We emphasize here that AM1* is intended only as an extension and improvement of the existing AM1 method. It does not represent a new stage in the development of NDDO-based theories, but rather is intended to provide a freely available fully documented semiempirical MO technique that performs well for elements of the second long period and that can be extended to transition metals easily.

Theory

AM1* uses standard MNDO [24, 25] approximations for all integrals involving *s* and *p* orbitals and those from MNDO/d [8, 9, 10, 11, 12, 13] for *d* orbitals. For H, C, N, O and F, AM1* is identical to the published AM1 method. [14, 15] The only major deviation from standard AM1 theory for the elements that have been newly parameterized is the use of element-pair specific parameters, α_{ij} and δ_{ij} to describe the core–core interactions between elements *i* and *j*. Element-pair specific core–core parameters are used in MNDO/d [8, 9, 10, 11, 12, 13] and by Voityuk and Rösch for their AM1(d) parameterization for molybdenum. [19] For the "pure AM1" elements, H–F outlined above, AM1* uses the additional Gaussian functions common for AM1, PM3 and, apparently, PM5 to modify the core–core repulsion. However, Voityuk and Rösch found an alternative formalism with fewer parameters using α_{ij} and δ_{ij} to be more effective. This formalism has the disadvantage that it requires specific parameters for every pair of elements. However, it does not lead to spurious minima as the Gaussian modification can. [27] For the newly parameterized elements, the core–core repulsion energy $E^{\text{core}}(i–j)$ between elements of elements *i* and *j* is given by

$$E^{\text{core}}(i-j) = Z_i Z_j \rho_{\text{ss}}^0 \left[1 + \delta_{ij} \exp\left(-\alpha_{ij} r_{ij}\right)\right] \quad (1)$$

where Z_i and Z_j are the effective (valence only) core charges of elements *i* and *j*, $\rho_{\text{ss}}{}^0$ is defined as for the original MNDO/d method [8, 9, 10, 11, 12, 13] and r_{ij} is the distance between the atoms *i* and *j*. Note that the δ_{ij} parameters given by Voityuk and Rösch are half those used here (i.e. the preexponential factor used in [19] for molybdenum in AM1(d) is actually $2\delta_{ij}$). [28]

The validating density functional calculations were performed with Gaussian 98. [29] Geometries were optimized using the Becke three-parameter functional [30, 31] in conjunction with the Lee–Yang–Parr correlation functional [32, 33] using the 6-31+G(d) basis set. [34, 35] The standard B3LYP implementation in Gaussian 98 was used. Harmonic vibrational frequencies were also calculated at this level and used for extrapolation to 298 K. Single-point B3LYP/6-311+G(2d,f)//B3LYP/6-31+G(d) [36, 37, 38, 39, 40, 41, 42, 43] calculations were used to refine the energies and dipole moments. Heats of formation were calculated using these Born–Oppenheimer energies and the thermodynamic corrections derived from the smaller basis set.

Parameterization data

Parameterization data was taken largely from the MNDO/d parameterization dataset, [8, 11, 12] but were extended with data that were included in the PM3 and AM1 datasets. We also added bond lengths and angles for a series of sulfamides and sulfonamides; methane sulfamide, 1,1,1-trifluoromethanesulfamide, ethane sulfonamide, and *N*,*N*-dimethylsulfonamide, where target values were taken from B3LYP/6-31+G(d) calculations. In most cases heats of formation, dipole moments and geometries were checked using the DFT-based scheme outlined above. The individual datasets and the values used for parameterization are outlined in the tables and supplementary material. The set contains 300 heats of formation, 156 ionization potentials, 90 dipole moments, 315 bond lengths, 209 bond angles and eight dihedral angles. We will draw special attention to some molecules for which the experimental situation is unclear in the following sections.

We have modified the target values for the heats of formation for 13 compounds where the tabulated values were significantly different than those from DFT or G2 calculations. The value for $HOSO_2$ was changed from −98.0 in the MNDO/d data set to −59.3. The values for SOF_3, SCl_6, SO_2Cl, OPF_2, SCl_4, $OPCl_2$, and $PSCl_3$ were changed from the values in the PM3 data set to 21.9, −45.9, −171.1, −55.9, −70.9, and −69.5 kcal mol^{-1} respectively. In the cases of SO_2F, SOF_4 and phosphorus pentoxide, there were multiple experimental values used

for the parameterizaton, none of which were within 15 kcal mol^{-1} of the calculated values, which were then substituted as target values. In the case of $(OCN)_3PO$ the DFT value of −64.3 was used. While certainly there are cases where calculated heats of formation by both DFT and G2 are significantly in error, this is balanced by the consistency that using these calculated values provides.

For parameters involving Mo, we used a subset of the compounds containing S and Cl taken from the paper of Voityuk and Rösch [19] which contains ten heats of formation, 26 bond lengths and 28 bond angles. Since phosphorous was not included in the original parameterization we extended the set to contain the calculated energetic barrier heights for two reactions, the topomeric isomerization of $MoP_6C_6Me_6$ and $Mo(P_2C_2Me_2)(P_2C_2H_2)_2$ recently reported at B3LYP/LanL2dz using polarization functions [44, 45] were used with a weighting factor of 10 mol $kcal^{-1}$. In addition, 33 calculated Mo–P bond lengths and one P–Mo–P bond angle in these compounds were used.

Parameterization

Our goal in the parameterization was not to find a set of parameters with the lowest error for the parameterization data at all costs, but rather to produce a chemically reasonable set of parameters that perform moderately well for as many applications as possible. Because of the relative paucity of experimental data, we cannot use an independent validation dataset, as would be desirable for such a parameterization, and so we have concentrated on avoiding overtraining at the cost of a more general method. In addition, because we have used the original AM1 parameters for the elements H–F, the quality of our results for elements containing these compounds is limited by the existing parameters. Ideally, all elements should be parameterized at once, the philosophy behind PM3, [16, 17, 18] but this is clearly not possible without losing backwards compatibility with AM1 for the first row elements. After initial explorations of the behavior of the parameters and setting up a set of initial guess parameters based on the results, the parameters were optimized using a Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimizer [46, 47, 48, 49, 50] with numerical gradients. We initially optimized the parameters to give low atomic forces at the B3LYP geometry, a technique introduced by Stewart for the PM3 parameterization (although with experimental geometries), [16, 17, 18] but performed the final parameterizations with full geometry optimization and an error function based on the internal coordinates. This strategy requires significantly more computational effort, but avoids the weighting of geometrical parameters by the associated force constants, a potential problem with using the gradients alone. Using this scheme a small geometrical error in a strong bond has a larger effect on a gradient-based parameterization than the same error in a "softer" geometrical coordinate. We have retained the error function and weighting factors of MNDO. However, utilizing the current training set, the emphasis placed on heats of formation is significantly higher than on the other components in the error function. Note that we have used the root mean square deviation (RMSD), rather than the mean unsigned error (MUE) as the error function. This discriminates against very large deviations for individual compounds, which we hope will result in a more robust parameterization.

Results

The optimized AM1* parameters are shown in Table 1. Geometry optimizations using MNDO/d, AM1, PM3, and the new AM1* parameterization were done using VAMP 8.1, [51] while the PM5 calculations used LINMOPAC2002. [52] Results for MNDO/d, AM1, and

Table 1 Optimized AM1* parameters

Parameter	P	S	Cl
U_{ss} [eV]	−45.6707151	−58.6147064	−90.5010619
U_{pp} [eV]	−35.2098162	−47.1543086	−74.9323907
U_{dd} [eV]	−23.6885421	−27.1670804	−44.3751518
ζ_s [$bohr^{-1}$]	2.0894704	2.3827146	4.5899310
ζ_p [$bohr^{-1}$]	1.9476331	1.6189739	2.3382401
ζ_d [$bohr^{-1}$]	1.2697580	1.2888468	0.9886585
β_s [eV]	−10.3868963	−3.8082753	−22.6208745
β_p [eV]	−10.7694019	−7.4192147	−15.4461422
β_d [eV]	−4.9129999	−1.9225157	−2.3720965
α [Å^{-1}]	1.8232300	1.9717900	2.9456800
g_{ss} [eV]	10.9221093	12.3977366	13.7252953
g_{pp} [eV]	8.5031975	9.8433852	13.7139758
g_{sp} [eV]	5.6174929	8.8485884	12.4487535
g_{p2} [eV]	7.8119356	7.2121425	10.2681036
h_{sp} [eV]	0.7461127	3.0283882	3.2315813
z_{sn} [$bohr^{-1}$]	1.6351391	1.6251800	2.1424382
z_{pn} [$bohr^{-1}$]	0.9773978	1.1875234	1.4930146
z_{dn} [$bohr^{-1}$]	0.8744020	1.2636009	1.5854230
ρ(core) [bohr]	1.2437106	1.1436458	0.9216254
$\Delta H°_f$(atom) [kcal mol^{-1}]	75.57	66.40	28.99
F^0_{sd} [eV]	11.6055655	12.7223726	3.0250995
G^2_{sd} [eV]	12.9748658	40.2365349	10.1746511
α_{ij}			
H [Å^{-1}]	1.7054944	1.9688663	2.3544058
C [Å^{-1}]	1.7662992	2.1440293	2.0250324
N [Å^{-1}]	2.2875170	2.0308348	1.9463635
O [Å^{-1}]	2.1041690	1.9550810	2.7096282
F [Å^{-1}]	1.9919322	2.1151785	2.8705654
P [Å^{-1}]	1.9690320		
S [Å^{-1}]	1.8731991	1.8750321	
Cl [Å^{-1}]	1.6020362	1.8902611	2.1841135
Mo [Å^{-1}]	1.4120112	2.1707174	2.4269681
δ_{ij}			
H	1.0906700	0.9561224	1.1348265
C	1.0734607	1.3145695	1.0333870
N	2.3031769	1.0000831	0.9393650
O	1.6352693	0.9150090	2.4647496
F	1.3763366	1.2706784	4.6676202
P	3.1258913		
S	1.4768864	0.9659725	
Cl	0.9242251	1.0781803	1.9732923
Mo	1.1152178	2.9826189	4.7956244

Table 3 Error statistics for the heats of formation (experimental, DFT, MNDO/d, PM3, PM5, AM1 and AM1* all in kcal mol^{-1}) for the parameterization compounds

	DFT	MNDO/d	PM3	PM5	AM1	AM1*
All compounds (*N*=300)	(*N*=298)					
MSE	−8.70	1.32	4.38	2.57	−3.27	2.45
MUE	10.38	8.76	11.54	12.20	18.20	13.12
RMSD	14.09	12.98	18.35	24.54	43.93	17.83
Most +ve error	13.56	43.84	135.5	150.26	81.99	71.54
Most −ve Error	−47.94	−61.99	−31.46	−161.3	−465.13	−64.19
Phosphorus compounds (*N*=73)						
MSE	−6.11	9.09	10.35	10.03	1.72	3.09
MUE	9.42	14.23	17.05	16.48	14.89	16.12
RMSD	14.16	18.48	26.29	28.34	19.29	19.60
Most +ve error	11.97	42.18	135.5	150.26	55.71	52.27
Most −ve error	−39.09	−61.99	−31.46	−37.33	−54.00	−50.89
Sulfur compounds (*N*=133)	(*N*=131)					
MSE	−9.98	−1.23	1.62	4.32	7.99	3.52
MUE	11.18	8.39	9.11	9.26	13.69	9.61
RMSD	14.64	12.42	12.55	14.46	20.30	13.28
Most +ve error	13.56	43.84	46.17	60.83	81.99	48.58
Most −ve error	−47.94	−45.26	−30.32	−33.77	−37.23	−35.56
Chlorine compounds (*N*=132)	(N=131)					
MSE	−9.07	0.81	4.88	−0.31	−13.40	1.91
MUE	10.32	6.55	10.53	13.62	24.89	14.28
RMSD	13.88	9.68	16.74	29.10	62.86	19.67
Most +ve error	13.56	39.21	87.56	64.16	81.99	71.54
Most −ve error	−41.52	−23.50	−19.70	−161.30	−465.13	−64.19

PM3 are essentially identical for the two programs. Table 2 (see Electronic Supplementary Material) shows the heats of formation (experimental, DFT, AM1*, AM1, PM3, PM5 and MNDO/d) obtained for the parameterization compounds. A summary of the results for the entire dataset and for compounds containing P, S and Cl is shown in Table 3.

Tables 4 and 5 (see Electronic Supplementary Material) report results for ionization potentials and dipole moments, while Table S1 of the supplementary material gives their geometrical details. We will discuss the results for the three elements chlorine, sulfur and phosphorus separately. We note that the performance of the DFT-based method is only moderate. The mean unsigned error (MUE) for the parameterization dataset is 13.2 kcal mol^{-1}, compared with 10.6 and 11.1 for MNDO/d and PM3, respectively. Only AM1 is significantly less accurate with an MUE of 20 kcal mol^{-1}. This result is perhaps surprising, but indicates the high level of accuracy achieved by modern semiempirical methods.

Chlorine-containing compounds

During the parameterization, we realized that we could not obtain satisfactory results in which four oxygens are bound to a central chlorine. This bonding pattern, however, also proved to be difficult for sulfur and phosphorus as the central atom, so that we attribute it to a weakness in the AM1 parameterization for oxygen, which was used here unchanged. This hypothesis is supported by the large negative errors given by AM1 for these compounds. Negative errors larger than −30 kcal mol^{-1} are found for perchloryl fluoride (−64.2 kcal mol^{-1}), the chloride anion (−38.9), ClF_5 (−38.2) and ClF (−31.1). However, we note that the DFT technique also gives an error of −20.2 kcal mol^{-1} for perchloryl fluoride, so that the experimental value may be in error and an AM1* error of about 40 kcal mol^{-1} is likely. Other compounds with large negative errors are ClO_4^- (−28.3), dichloroacetylene (−26.0), reflecting the general difficulty of AM1 for triple bonds, and the *m*- and *p*-chlorobenzaldehydes (−27.5 and −25.6, respectively). However, there must be some doubt about the experimental values for the chlorobenzaldehydes (−15.1, −36.6 and −34.8 kcal mol^{-1} for the *o*-, *m*- and *p*-isomers, respectively). Taken at face value, these imply an isomerization energy of around −20 kcal mol^{-1} from the *m*- or *p*-isomer to *o*-chlorobenzaldehyde, an unreasonably high value. The DFT errors (−15.5, −32.7 and −30.2 kcal mol^{-1} for *o*-, *m*- and *p*-isomers, respectively) imply that the values for the *m*- and *p*-compounds are in error. Unfortunately, compounds involving Cl_2 (−19.9), Cl^- (−38.9) and HCl (−8.1) also give large negative errors. These include HCl_2^- (−21.2) and $ClHF^-$ (−22.0). Clearly, it would be preferable to reproduce the energies of these small, important molecules correctly. However, we have chosen to accept large errors for these compounds in order to obtain a more robust model. Quite generally, semiempirical techniques tend to give poor results for small compounds, for which the NDDO approximation is most severe. This is the case for AM1*.

The most positive error (71.5 kcal mol^{-1}) is found for FCl_2O^-, as for the other Cl–O compounds ClO, ClO_2 and ClO_3 (21.9, 40.9 and 40.5, respectively). Otherwise, the chlorofluorocarbons CCl_2F_2 (56.9), CF_3Cl (42.3), $CFCl_3$ (49.9) and C_2Cl_6 (46.9) and the five and six-coordinate sulfur halogens also give large positive errors. Many of

these errors are partly caused by the original AM1 parameterization for O and F.

Sulfur-containing compounds

The energies calculated for sulfur compounds show smaller errors than for chlorine or phosphorus compounds for all of the semiempirical methods used, but slightly larger for DFT. Once again, we find different types of oxygen- and fluorine-containing compounds at both extremes of the error scale, supporting the idea that the AM1 parameterization for oxygen may be at fault. $(CH_3O)_3SO$ (−35.6 kcal mol^{-1}) and SO_2 (−27.6) give the most negative errors and $HOSO_2$ (48.3) and SOF_3 (48.6) are among the compounds with large positive errors. The multiply bonded species CS, CS_2 and OCS all give large negative errors (−26.7, −11.7 and −17.9, respectively).

Phosphorus-containing compounds

Phosphorus-containing compounds have proven particularly difficult to parameterize in semiempirical techniques. Our results for AM1* are comparable to those obtained for chlorine-containing compounds, but not as good as for sulfur. The most negative errors are given by trimethyl and triethyl phosphites (−37.8 and −50.9, respectively), triethyl phosphate (−43.1) and PO_3 (−34.7) and HMPTA (−41.2). However, once again, the DFT technique also gives large negative errors (from −14.7 to −35.0) for all of these compounds except PO_3 (−8.1). The most positive errors are given for P_2O_5 (52.3) and hexachloro-1, 3, 5, 2, 4, 6-triazatriphosphorine (34.3). These errors once again reflect problems with the original AM1 parameterization for oxygen.

Molybdenum compounds with sulfur, phosphorus and chlorine ligands

The molybdenum parameters of Voityuk and Rösch [19] were used without change except in combination with the elements P, S and Cl. In these cases, new values of α_{ij} and δ_{ij} were obtained for i=Mo and j=P, S or Cl. The values obtained are shown in Table 1.

Comparison to other semiempirical methods

An error analysis for our dataset of 300 compounds is shown in Table 3. The data show small systematic deviations in the mean signed error for each technique. Thus MNDO/d, PM3, PM5 and AM1* give heats of formation that are on average 1.3–4.4 kcal mol^{-1} to unstable, whereas AM1 slightly (−3.3 kcal mol^{-1}) and DFT significantly (−8.7) overestimate the stability. The mean unsigned errors (MUE) suggest that much of the error in the DFT calculations is caused by the systematic deviation, which can be removed by a suitable parameterization. [53] Of the semiempirical methods, MNDO/d performs the best with an MUE of 8.8 and a root mean square deviation (RMSD) of 13.0 kcal mol^{-1}. AM1, PM3 and PM5 all give significantly lower MUEs (18.2, 11.5 and 12.2 kcal mol^{-1}, respectively) than RMSDs (43.9, 18.4 and 24.5), presumably because they were parameterized using the MUE as the error function. AM1*, which used the RMSD for the parameterization, gives a moderate MUE (13.1), but a respectable RMSD (17.8). This is reflected in the smaller largest absolute error (71.5 for $FCLO_2^-$) for AM1* compared to PM3 (135.5), PM5 (150.3) and AM1 (−465.1 for Cl_2O_7). We prefer to use the RMSD, rather than the MUE, in order to avoid very large outliers. The largest error for both PM3 and PM5 is given by PO_3, a compound for which the DFT technique gives only a small error. The improvement in the performance of AM1 on going to AM1* is significant. The MUE is reduced by 28%, the RMSD by 60% and the largest absolute error by 85%. However, the MUE and RMSD for this dataset are still 4–5 kcal mol^{-1} higher than those given by MNDO/d. We attribute this difference to the fact that we have used the AM1 parameter set for H–F, whereas MNDO/d used the standard MNDO parameters for these elements. We note, however, that we have achieved our goal of making AM1, which gives hydrogen bonds and performs best of the published methods for rotation barriers in conjugated bonds, more reliable for P, S and Cl. One surprise in our data is that AM1 is quite good for phosphorus compounds (in fact, we have not been able to improve the reliability of the energy calculations for phosphorus compounds). A further surprising observation is that the new parameterization PM5 performs slightly worse than its predecessor for this dataset.

Of the individual elements, sulfur gives the best results, as outlined above. Phosphorus and chlorine give very similar errors for AM1*, whereas phosphorus is far better than chlorine with AM1.

Dipole moments

Table 4 (see Electronical Supplementary Material) shows the dipole moments used for the parameterization and the results given by the semiempirical methods. Once again, the performance of AM1* in this respect is influenced strongly by the weighting factors used to calculate the error function. However, AM1* performs in general better than the *s,p* methods and worse than MNDO/d for dipole moments. This result is consistent with the proposal that the AM1 parameters for H–F limit the performance of AM1*.

Ionization potentials

Table 5 (see Electronical Supplementary Material) lists the experimental and Koopmans' theorem ionization

potentials from the parameterization set. There is little to choose between the performance of the various methods, although MNDO/d, AM1 and AM1* are slightly better than PM3 and PM5. Large negative errors are found for SO_3 and large positive ones for thiocyanogen for all methods. AM1* performs worse than the other methods for ClO_2, but is otherwise relatively reliable with no very large outliers. We have retained the practice of using Koopmans' theorem for parameterization although it is necessarily an approximation. However, semiempirical methods have traditionally used this approach, which indirectly places some constraints on the electronic parameters, so that we have retained it here. Future methods may require a more satisfactory treatment of the ionization potentials.

Structures

The geometrical parameters used to parameterize AM1* and the values given by the different calculational methods are shown in Table S1. A summary of the errors appears in Tables 6 and 7 (see Electronic Supplementary Material). Generally, the performance of the semiempirical methods is comparable with MNDO/d once again giving the smallest deviations from experiment. We emphasize, however, that the performance of the method with respect to geometries can be tuned by means of the weighting factors used for determining the error function during the parameterization. We have chosen a compromise to give roughly the same order of accuracy as found for the other methods. AM1* gives significant errors for many O–Cl and F–Cl bonds. We attribute these errors to the use of the AM1 parameters for oxygen and fluorine and note that, at least for Cl–O, AM1* gives smaller errors than AM1. AM1* also consistently calculates the C–Cl bond lengths to be too short, especially those in acid chlorides, which deviate from the experimental values by 0.1 Å or more. Quite generally, AM1* performs worst for chlorine-containing molecules and gives errors in bond lengths that are comparable with the other methods for other elements.

The situation is similar for bond angles. MNDO/d performs slightly better than the other techniques, which are otherwise fairly similar. AM1* gives larger errors than expected for sulfur-containing compounds, but AM1* deviations from experimental bond angles are otherwise comparable with those given by other methods. However, we note that AM1* is no improvement over AM1 for bond angles, a surprising result.

Conclusions

Semiempirical methods, especially ones such as that presented here that are based on existing parameterizations, will never be perfect. However, we have attempted to provide not necessarily the numerically most accurate parameter set for a limited dataset, but rather as robust a parameterization as possible for general use. Our parameterization datasets are therefore more varied than those used for other methods so that we can expect AM1* to behave correctly in many cases where other techniques may not. However, this work demonstrates very clearly the quality of the MNDO/d parameterization, which performs best in almost all respects for our dataset. However, MNDO/d suffers from its very poor performance for hydrogen bonds and for rotation barriers in π systems, so that it cannot, for instance, be used for most biological studies. This work does, however, suggest that the inclusion of *d* orbitals in the basis set does not necessarily improve the performance of the method for, for instance phosphorus compounds. However, AM1* is a considerable improvement over AM1 for our dataset.

Some issues of the reliability of the parameterization data arose during this work. The ideal procedure of using completely independent training and validation datasets is not yet applicable for semiempirical parameterizations because not enough data are available, particularly for ionization potentials and dipole moments. This situation can be improved by liberal use of, for instance, DFT geometries and G2 or G3 heats of formation. Brothers and Merz [22] have expressed reservations about the amount of calculated data used for their AM1 and PM3 parameterizations for sodium. While we share some of their reservations, we note that a semiempirical MO technique that behaves as well as, for instance, B3LYP/6-31G(d) would also be a significant achievement and would approach the spirit of the original work on PRDDO. [54, 55, 56] We will in any case be forced to resort to more and more data calculated at higher levels of theory to parameterize future semiempirical methods. To this end, we are making our parameterization dataset freely available on the web in a form that can be searched, data exported etc. [57] Others are invited to submit their parameterization data for inclusion in the database.

References

1. Clark T (2000) J Mol Struct (Theochem) 530:1–10
2. Beck B, Carpenter JE, Horn A, Clark T (1998) J Chem Inf Comput Sci 38:1214–1217
3. Stewart JJP (1997) Theochem 401:195–205
4. Stewart JJP (1996) Int J Quantum Chem 58:133–146
5. Gogonea V, Suarez D, van der Vaart A, Merz Jr K (2001) Curr Opin Struct Biol 11:217–223
6. Gregersen BA, Lopez X, York DM (2003) J Am Chem Soc 125:7178–7179
7. Pople JA, Santry DP, Segal GA (1965) J Chem Phys 43:S129–S135
8. Thiel W, Voityuk AA (1992) Int J Quantum Chem 44:807–829
9. Thiel W, Voityuk AA (1992) Theor Chim Acta 81:391–404
10. Thiel W, Voityuk AA (1996) Theor Chim Acta 93:315
11. Thiel W, Voityuk AA (1994) Theochem 119:141–154
12. Thiel W, Voityuk AA (1996) J Phys Chem 100:616–626
13. Thiel W (1998) In: Schleyer PvRS, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR (eds) Encyclopedia of computational chemistry, vol 3. Wiley, Chichester, pp 1604–1605
14. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909

15. Holder AJ (1998) In: Schleyer PvRS, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR (eds) Encyclopedia of computational chemistry, vol 1. Wiley, Chichester, pp 8–11
16. Stewart JJP (1989) J Comput Chem 10:209–220
17. Stewart JJP (1989) J Comput Chem 10:221–264
18. Stewart JJP (1998) In: Schleyer PvRS, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR (eds) Encyclopedia of computational chemistry, vol 3. Wiley, Chichester, p 2080
19. Voityuk AA, Rösch N (2000) J Phys Chem A 104:4089–4094
20. Wavefunction, Inc. (http://www.wavefun.com)
21. (a) Stewart JJP (2003) AM1 and PM3: submitted to J Mol Model; (b) Stewart JJP (2002) PM5: CAChe Group, Fujitsu America Inc, Mopac (http://www.cachesoftware.com/Mopac/index.shtml)
22. Brothers EN, Merz Jr K (2002) J Phys Chem B 106:2779–2785
23. Stewart JJP (2000) Pure Appl Chem 72:1449–1452
24. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:4899–4907
25. Thiel W (1998) In: Schleyer PvRS, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR (eds) Encyclopedia of computational chemistry, vol 3. Wiley, Chichester, p 1599
26. Lopez X, York DM (2003) Theor Chem Accounts 109:149–159
27. Klebe G (1990) Struct Chem 1:597–616
28. Voityuk AA (2002) personal communication
29. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratman RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu C, Liashenko A, Piskorz P, Komaromi, I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson BG, Chen W, Wong MW, Andres JL, Gonzales C, Head-Gordon M, Replogle ES, Pople JA (1998) Gaussian 98. Gaussian, Pittsburgh, Pa.
30. Becke AD (1996) J Chem Phys 104:1040–1046
31. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ (1994) J Phys Chem 98:11623–11627
32. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
33. Miehlich B, Savin A, Stoll H, Preuss H (1989) Chem Phys Lett 157:200–206
34. Hehre WJ, Ditchfield R, Pople JA (1972) J Chem Phys 56:2257–2261
35. Clark T, Chandrasekhar J, Spitznagel GW, Schleyer PvRS (1983) J Comput Chem 4:294–301
36. McLean AD, Chandler GS (1980) J Chem Phys 72:5639–5648
37. Krishnan R, Binkley JS, Seeger R, Pople JA (1980) J Chem Phys 72:650–654
38. Wachters AJH (1970) J Chem Phys 52:1033–1036
39. Hay PJ (1977) J Chem Phys 66:4377–4384
40. Raghavachari K, Trucks GW (1989) J Chem Phys 91:2457–2460
41. Binning Jr RC, Curtiss LA (1990) J Comput Chem 11:1206–1216
42. Curtiss LA, McGrath MP, Blaudeau J-P, Davis NE, Binning Jr RC, Radom L (1995) J Chem Phys 103:6104–6113
43. McGrath MP, Radom L (1991) J Chem Phys 94:511–516
44. Topf C, Clark T, Heinemann FW, Hennemann M, Kummer S, Pritzkow H, Zenneck U (2002) Angew Chem, Int Ed Engl 41:4047–4052
45. Topf C, Clark T, Heinemann FW, Hennemann M, Kummer S, Pritzkow H, Zenneck U (2002) Angew Chem 114:4221–4226
46. Broyden CG (1970) J Inst Math Appl 6:222
47. Fletcher R (1970) Comput J 13:317
48. Goldfarb D (1970) Math Comput 24:23
49. Shanno DF (1970) Math Comput 24:647
50. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in fortran. Cambridge University Press, Cambridge
51. Clark T, Alex A, Beck B, Burkhardt F, Chandrasekhar J, Gedeck P, Horn AHC, Hutter M, Martin B, Rauhut G, Sauer W, Schindler T, Steinke T (2003) VAMP 8.1. Accelrys Inc, San Diego
52. Stewart JJP (2002) LinMopac 2002. Fujitsu Ltd, FQS Poland Sp z o o, Krakow
53. Winget P, Clark T (2003) J Comput Chem submitted
54. Marynick DS, Lipscomb WN (1982) Proc Natl Acad Sci U S A 79:1341–1345
55. Halgren TA, Lipscomb WN (1973) J Chem Phys 58:1569–1591
56. Marynick DS (1998) In: Schleyer PvRS, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR (eds) Encyclopedia of computational chemistry, vol 3. Wiley, Chichester, pp 2153–2160
57. Martin B, Winget P, Horn AHC, Selcuki C, Clark T (2003) in preparation