

Extension of the PDDG/PM3 Semiempirical Molecular Orbital Method to Sulfur, Silicon, and Phosphorus

Ivan Tubert-Brohman, Cristiano Ruch Werneck Guimarães, and William L. Jorgensen*

*Department of Chemistry, Yale University, 225 Prospect St.,
New Haven, Connecticut 06520-8107*

Received February 11, 2005

Abstract: The PDDG/PM3 semiempirical molecular orbital method has been parametrized for molecules, ions, and complexes containing sulfur; the mean absolute error (MAE) for heats of formation, ΔH_f , of 6.4 kcal/mol is 35–40% smaller than those for PM3, AM1, and MNDO/d. For completeness, parametrization was also carried out for silicon and phosphorus. For 144 silicon-containing molecules, the ΔH_f MAE for PDDG/PM3, PM3, and AM1 is 11–12 kcal/mol, whereas MNDO/d yields 9.4 kcal/mol. For the limited set of 43 phosphorus-containing molecules, MNDO/d also yields the best results followed by PDDG/PM3, AM1, and PM3. The benefits of the d orbitals in MNDO/d for hypervalent compounds are apparent for silicon and phosphorus, whereas they are masked in the larger dataset for sulfur by large errors for branched compounds. Overall, for 1480 molecules, ions, and complexes containing the elements H, C, N, O, F, Si, P, S, Cl, Br, and I, the MAEs in kcal/mol for ΔH_f are 6.5 (PDDG/PM3), 8.7 (PM3), 10.3 (MNDO/d), 10.8 (AM1), and 19.8 (MNDO).

Introduction

Semiempirical methods based on the neglect of diatomic differential overlap (NDDO)¹ approximation, such as MNDO,² AM1,³ PM3,⁴ and MNDO/d,⁵ occupy an important place in computational chemistry because of their speed and excellent scaling with increasing system size. Even as constant advances in computer resources permit application of ab initio and DFT methods to ever-larger systems of chemical interest, semiempirical methods allow the exploration of new frontiers such as full quantum mechanical calculations for proteins⁶ or long Monte Carlo and molecular dynamics simulations of reactions in solution and in enzymes by means of coupled quantum and molecular mechanics (QM/MM).^{7–16}

In recent articles, a new NDDO-based method, PDDG/PM3, was introduced.^{17,18} It is derived from the PM3 method by the addition of small pairwise distance-directed Gaussians to the core repulsion function. The method was initially parametrized for the basic organic elements C, H, N, and O¹⁷ and later extended to the halogens F, Cl, Br, and I.¹⁸ The use of the PDDG function in conjunction with extensive

parametrization using large datasets resulted in a reduction of about 30% in the mean absolute error (MAE) for heats of formation in comparison to PM3. Several systematic errors such as for homologation and branching were overcome, and improvements for activation barriers for S_N2 reactions involving halogens were also obtained. The PDDG/PM3 method yields heats of formation and isomerization energies that are more accurate than those obtained from B3LYP/6-31G* calculations, and for some classes of compounds, chemical accuracy is approached; for example, the MAE is 1.17 kcal/mol for alkanes. It has also been used successfully in QM/MM studies of nucleophilic aromatic substitution (S_N-Ar) and S_N2 reactions in solution.^{7,8}

In this paper, the PDDG/PM3 method is extended to sulfur, silicon, and phosphorus. With these additions, the complete set of “organic elements” is available for calculations. This will allow the application of the PDDG/PM3 method to a wide variety of molecules of chemical and biological interest. Although it is expected that methods that do not include d orbitals will have difficulties with hypervalent compounds, the question is how far can one go with an sp basis set and the PDDG approach?

* Corresponding author email: william.jorgensen@yale.edu.

PDDG Formalism

The rationale behind the PDDG formalism has been discussed.¹⁷ The key difference between the PDDG/PM3 method and its predecessor is the addition of pairwise distance-directed Gaussian terms (PDDG) to the core repulsion function, as given in eq 1.

$$\text{PDDG}(A,B) = \frac{1}{n_A + n_B} \left\{ \sum_{i=1}^2 \sum_{j=1}^2 (n_A P_{A_i} + n_B P_{B_j}) \exp[-10 \text{ \AA}^{-2} (R_{AB} - D_{A_i} - D_{B_j})^2] \right\} \quad (1)$$

This equation is an empirical correction to the core repulsion between atoms *A* and *B*, which are separated by a distance R_{AB} ; each element requires four parameters, P_{A1} , P_{A2} , D_{A1} , and D_{A2} . The function is weighted using n_A and n_B , which are the number of valence electrons for atoms *A* and *B*, respectively. For each pair of atoms, there are four Gaussians, which depend only on atomic parameters. The Gaussians are small compared to those used in the original AM1 and PM3 core repulsion functions and work by addressing systematic errors associated with bonds and functional groups.^{17,18}

Another difference between the traditional NDDO methods and the PDDG method is the way in which the molecular energy obtained from the self-consistent field calculation is converted to a heat of formation at 298 K. In both cases, the difference between the heat of formation of each atom and its electronic energy (*isol*) is added to the molecular energy. In the traditional method, *isol* is obtained as a derived parameter by calculating the energy of an isolated atom with a restricted single-determinant wave function using the semiempirical formalism and parameter set;¹⁹ for the PDDG methods, *isol* is treated as an optimizable parameter, obtained from a through-origin linear regression so that the conversion from molecular energy to heat of formation gives as small an error as possible. A similar approach has been applied in recent work by Winget and Clark using density functional theory (DFT).²⁰

Parameter Optimization

To extend the PDDG/PM3 method to sulfur, silicon, and phosphorus, a similar procedure to the one used for the halogens was followed: the U_{ss} , U_{pp} , β_s , β_p , ζ_s , ζ_p , and α MNDO parameters, as well as the PM3 and PDDG/PM3 Gaussian pre-exponential and distance terms, were optimized by a combination of gradient-based methods (Fletcher–Powell) and simulated annealing.¹⁸ The available reference data had very few molecules involving more than one of the elements discussed in this paper, so little coupling between different elements' parameters was expected, which allowed each element to be optimized separately. A total of 527 reference values were used in the error function during the optimization of the three elements, as detailed in Table 1. The error function that was minimized was the weighted sum of the square deviation between the calculated and reference values, as discussed previously.¹⁸ Most of the optimizations were done with fixed PM3 geometries, adding gradients to the error function to ensure that the geometric

Table 1. Composition of the Training Sets^a

data type	S	Si	P
heat of formation	81	47	23
ionization potential	28	15	13
dipole moment	16	12	7
bond length	78	96	16
bond angle	54	33	8

^a For detailed data and references, see the Supporting Information.

minimum for each molecule did not stray far from the PM3 minimum. Only the final optimization stage involved fully flexible geometries; in this stage, the gradients were not included in the error function, since they are almost zero.

The prior PDDG/PM3 parametrizations were performed as local optimizations that tried to stay close to the original PM3 parameters.^{17,18} However, preliminary results for sulfur showed that the local approach could provide little improvement over PM3, so a more global optimization was undertaken. As a first step, to perturb the PM3 parameters reasonably far from their original values, 384 (3×2^7) initial parameter sets were generated by crossover between the U_{ss} , U_{pp} , β_s , β_p , ζ_s , ζ_p , and α parameters of PM3, MNDO, and AM1. These parameter sets were optimized in the usual way while using the PM3 Hamiltonian and local optimization for the PM3 Gaussians; the PDDG Gaussians were not added yet. The six best results were chosen, and an exhaustive crossover was performed between them, resulting in 1920 (15×2^7) parameter sets. After optimizing the latter, the best set was chosen, and 256 (4^4) different combinations of PM3 Gaussian parameters were tried in a grid-like fashion. After optimization of these, 256 combinations of PDDG Gaussians were added to the best result. The best parameter set from the last step was then subjected to 3000 steps of stochastic search; at each step, a random “kick” was applied to the parameters, which were then optimized using the same gradient-based algorithm. The best parameter set obtained so far was subjected to a flexible geometry optimization using simulated annealing, resulting in the final parameter set for sulfur. Suffice it to say that this protocol was not planned in advance in its entirety but was the result of much trial and error. However, it resulted in remarkable improvement over the simple local optimization, as discussed below.

Similar procedures were tried for silicon and phosphorus, with a few variations. The MNDO method was no longer included in the crossover part, as it was found to not be helpful; also, the stochastic search was not performed for these elements. For silicon, it was found that the results for the global optimization were not better than those from the simple local optimization, whereas for phosphorus, gains were made over PM3. The buildup procedure of parametrization for C, H, N, and O first; then the halogens; and then S, Si, and P clearly leads to the largest errors for the last elements. Simultaneous optimization for all elements is desirable but logistically taxing.

Results and Discussion

The optimized and dependent parameters for the PDDG/PM3 method are shown in Table 2 along with the original PM3 parameters for comparison. Some PDDG/PM3 parameters

Table 2. Optimized PDDG/PM3 Parameters for S, Si, and P, along with the Standard PM3 Parameters^a

	PDDG/PM3			PM3		
	S	Si	P	S	Si	P
U_{ss}	-43.906 366	-26.332 522	-37.882 113	-49.895 371	-26.763 483	-40.413 096
U_{pp}	-43.461 348	-22.602 540	-30.312 979	-44.392 583	-22.813 635	-29.593 052
β_s	-2.953 912	-3.376 445	-12.676 297	-8.827 465	-2.862 145	-12.615 879
β_p	-8.507 779	-3.620 969	-7.093 318	-8.091 415	-3.933 148	-4.160 040
ζ_s	1.012 002	1.586 389	2.395 882	1.891 185	1.635 075	2.0175 63
ζ_p	1.876 999	1.485 958	1.742 213	1.658 972	1.313 088	1.504 732
a	2.539 751	2.215 157	2.005 294	2.269 706	2.135 809	1.940 534
<i>eisol</i>	-166.336 554	-66.839 000	-117.212 854	-183.453 740	-67.788 214	-117.959 174
<i>DD</i>	1.006 989	1.310 515	0.893 978	1.121 431	1.314 455	1.064 495
<i>QQ</i>	0.891 487	1.126 089	0.960 457	1.008 649	1.274 340	1.112 039
ρ_0^b	1.517 625	2.695 556	1.743 870	1.517 625	2.695 556	1.743 870
ρ_1^b	0.711 672	1.630 757	1.050 851	0.748 602	1.633 605	1.160 242
ρ_2^b	0.754 336	0.949 200	1.208 907	0.814 668	1.025 130	1.339 579
a_1	-0.330 692	-0.071 314	-0.398 055	-0.399 191	-0.390 600	-0.611 421
b_1	6.000 000	6.000 000	1.997 272	6.000 669	6.000 054	1.997 272
c_1	0.823 837	0.237 995	0.950 073	0.962 123	0.632 262	0.794 624
a_2	0.024 171	0.089 451	-0.079 653	-0.054 899	0.057 259	-0.093 935
b_2	6.000 000	6.000 000	1.998 360	6.001 845	6.007 183	1.998 360
c_2	2.017 756	1.897 728	2.336 959	1.579 944	2.019 987	1.910 677
P_{A1}	0.120 434	-0.091 928	0.462 741			
P_{A2}	-0.002 663	-0.040 753	-0.020 444			
D_{A1}	0.672 870	1.163 190	0.714 296			
D_{A2}	2.032 340	2.190 526	2.041 209			

^a Units are (eV) U_{ss} , U_{pp} , β_s , β_p , *eisol*, a_1 , a_2 , P_{A1} , P_{A2} ; (au) ζ_s , ζ_p ; (Bohr) *DD*, *QQ*, ρ_0 , ρ_1 , ρ_2 ; (Å) c_1 , c_2 , D_{A1} , D_{A2} ; (Å⁻¹) α , b_1 , b_2 . ^b For use in MOPAC 6, $\rho_0 = 0.5/AM$, $\rho_1 = 0.5/AD$, $\rho_2 = 0.5/AQ$.

for sulfur, such as U_{ss} , ζ_s , and β_s , are quite different from the PM3 parameters. This is a result of the global optimization method along with the differences in the training set. However, the PDDG/PM3 parameters are still more similar to PM3 than to AM1 values, while the magnitude of the differences between PDDG/PM3 and PM3 parameters is similar to the differences between PM3 and AM1. The differences in the parameters for silicon and phosphorus are generally smaller than those for sulfur.

Sulfur Thermochemistry. The performances of PDDG/PM3 and the other common semiempirical methods for the calculation of standard heats of formation are summarized in Table 3. Full details are provided for all individual compounds in the Supporting Information. The MAE for the entire set of 249 molecules is 6.4 kcal/mol for PDDG/PM3, which is 35–40% smaller than the MAE for the older methods. This large improvement was observed for most classes of compounds, but it could only be achieved after the extensive global optimization discussed above. Initial tests with local optimizations resulted in only ca. 15% improvements with respect to PM3. It is remarkable that the errors for hypervalent molecules with PDDG/PM3 and its sp basis set are not large and that notably improved performance for classes such as sulfones is achieved, while simultaneously improving the performance for normal valence functional groups. The MAEs for thiols, sulfides, and disulfides are decreased by 63%, 13%, and 76%, respectively, when compared to PM3. Compounds with both sulfur and halogens were the only notable class where PDDG/PM3 did not yield improvement over PM3 or MNDO/d. The problem cases here

Table 3. MAEs for Heats of Formation for Sulfur Compounds (kcal/mol)

	N	PDDG	MNDO	AM1	PM3	MNDO/d
all	249	6.4	41.2	10.6	10.5	10.0
training	81	7.1	34.2	9.9	10.6	10.0
test	168	6.0	44.6	10.9	10.4	10.1
halides	15	13.7	56.5	14.5	6.6	5.1
sulfoxides	7	5.4	43.2	3.8	6.1	5.6
sulfones	36	5.9	143.8	18.9	18.1	10.4
sulfates	5	1.7	162.9	10.5	5.1	5.9
sulfites	4	8.9	54.1	22.0	11.2	7.6
thiols	29	1.4	5.5	4.4	3.8	7.3
sulfides	38	3.9	9.5	5.9	4.5	6.4
disulfides	14	1.8	9.7	5.1	7.6	9.5
aromatics	8	1.3	4.7	2.6	4.5	1.8
thioamides	9	7.4	14.6	11.8	22.2	27.5
dithiocarbamates	5	4.6	19.2	9.1	17.8	34.6
thioesters	10	5.4	11.6	6.5	7.0	7.8
thiocarbonates	7	4.6	11.8	5.6	15.6	8.2
anions	7	5.3	5.7	5.3	7.2	3.5
cations	7	11.3	19.2	18.0	23.0	11.0
transition structs.	4	10.6	26.7	26.2	14.9	29.9
complexes	18	13.0	41.6	15.0	17.5	19.6
others	16	10.2	10.2	9.1	8.4	6.6

are predominantly small fluorine-containing species including SF₂, FSSF, the transition state for the SH⁻ + CH₃F S_N2 reaction, and the HFHS⁺ complex. If more data were available for larger molecules, the benefits of PDDG for homologation and branching could be expected to dominate for this class too.

Table 4. Heat of Formation Results (kcal/mol) for Selected Molecules that Highlight Systematic Problems

formula	name	exp	Deviation (calc-exp)				
			PDDG	PM3	AM1	MNDO	MNDO/d
CH ₄ S	methanethiol	-5.4	0.2	-0.1	1.1	-1.9	1.1
C ₄ H ₁₀ S	1-butanethiol	-21.1	2.2	1.6	-3.1	-1.8	1.8
C ₁₀ H ₂₂ S	1-decanethiol	-50.7	-1.3	-1.3	-14.6	-0.5	3.1
C ₄ H ₁₀ S	isobutyl thiol	-23.1	0.3	4.3	1.4	4.2	8.5
C ₅ H ₁₂ S	isopentyl thiol	-27.4	-0.5	2.4	-1.1	4.4	8.0
C ₅ H ₁₂ S	neopentyl thiol	-30.8	-0.7	7.7	7.0	15.3	20.4
C ₄ H ₁₀ S	sec-butyl thiol	-23.0	1.3	4.4	1.0	2.9	7.3
C ₄ H ₁₀ S	tert-butyl thiol	-26.0	1.0	8.1	7.3	10.9	14.9
C ₂ H ₆ S	dimethyl sulfide	-8.9	-5.3	-2.1	-0.4	-8.2	0.4
C ₄ H ₁₀ S	diethyl sulfide	-20.0	-0.7	2.8	-1.8	-9.0	0.5
C ₈ H ₁₈ S	diisobutyl sulfide	-42.9	-2.4	5.7	-0.9	3.1	14.7
C ₆ H ₁₄ S	diisopropyl sulfide	-33.9	-0.8	7.8	3.2	0.6	11.0
C ₈ H ₁₈ S	di-tert-butyl sulfide	-45.1	-6.1	12.0	11.3	21.5	31.6
C ₂ H ₆ S ₂	dimethyl disulfide	-5.8	-3.5	1.0	1.6	-9.0	-0.7
C ₂ H ₆ S ₃	dimethyl trisulfide	-3.0	-2.0	-3.9	-1.7	-10.2	-2.4
C ₈ H ₁₈ S ₂	dibutyl disulfide	-37.9	-1.2	3.5	-6.8	-7.9	3.0
C ₈ H ₁₈ S ₂	di-tert-butyl disulfide	-47.8	0.8	20.1	16.0	21.0	34.5
C ₂ H ₆ O ₂ S	dimethyl sulfone	-89.2	1.7	12.9	18.9	142.9	3.6
C ₄ H ₁₀ O ₂ S	diethyl sulfone	-102.6	9.6	21.7	21.4	143.2	5.6
C ₈ H ₁₈ O ₂ S	dibutyl sulfone	-121.8	7.7	19.5	13.3	143.6	7.1
C ₈ H ₁₈ O ₂ S	diisobutyl sulfone	-128.0	7.9	27.6	24.4	160.2	25.0
C ₈ H ₁₈ O ₂ S	di-tert-butyl sulfone	-130.6	8.0	38.2	41.8	178.9	49.3

The latter point becomes apparent in Table 4, which highlights some of the systematic errors found with the different semiempirical methods. All methods except PDDG/PM3 have notable problems with homologation or branching or both. In particular, AM1 yields errors proportional to the number of methylene groups that results, for example, in a deviation of 14.6 kcal/mol for decanethiol. All of the methods except PDDG/PM3 tend to overestimate the heats of formation of branched compounds. The problem is particularly large for both MNDO and MNDO/d; the latter method yields deviations of 31.6 kcal/mol for di-tert-butyl sulfide and 49.3 kcal/mol for di-tert-butyl sulfone. Although a major part of the branching errors may be attributed to the hydrocarbon parameters, sulfur parameters must also play a role, because MNDO/d, while having the same parameters as MNDO for the first-row elements, tends to have larger branching errors. All of the methods, including PDDG/PM3, overestimate the heat of formation of sulfones to varying degrees. MNDO/d might be expected to be the best method for sulfones because of its inclusion of *d* functions; however, its advantages for small sulfones are minor, and the branching problems for larger sulfones lead to large errors. PDDG/PM3 still overestimates the heat of formation of sulfones, but when all the sulfones in the dataset are considered, the mean signed error is +2.7 kcal/mol versus about +18 kcal/mol for AM1 and PM3.

The 18 complexes in the dataset are considered further in Table 5, which focuses on the computed enthalpies of complexation, that is, the difference between the heats of formation of the complex and the separated molecules. PDDG/PM3 gives the lowest MAEs for neutral as well as positively and negatively charged complexes. The overall

MAE of 3.8 kcal/mol from PDDG/PM3 shows large improvement over the alternative semiempirical methods. The improvements here are clearly not from better treatment of branching because the molecules are all small. Perhaps surprisingly, the original MNDO method gives the second best results for all classes; it does not show the serious overestimate of binding for positively charged complexes as from AM1 or PM3. Thus, its problems with the heats of formation of hypervalent molecules such as SO₂ and dimethyl sulfoxide mostly cancel out when calculating interaction enthalpies.

A new method based on AM1, called AM1*, has been reported recently.²⁶ This method includes *d* orbitals for phosphorus, sulfur, and chlorine as well as a modified core-repulsion function with two-element parameters. The standard AM1 parameters and formalism are used for hydrogen and the first-row elements. The intersection of the dataset used in the AM1* study with the dataset used here has 82 sulfur-containing molecules and results in the following MAEs in kcal/mol: MNDO/d (6.0), PM3 (7.2), PDDG/PM3 (7.3), AM1* (8.0), AM1 (10.1), and B3LYP/6-311+G(2d,p)//LYP/6-31+G(d) (11.8). The dataset contains few large or branched molecules, so MNDO/d and PM3 perform relatively well. There is no clear benefit for the AM1* methodology, whereas the B3LYP results are the poorest. As discussed previously,^{17,18} PDDG/PM3 consistently outperforms the much slower DFT methods for calculating heats of formation owing, presumably, to some systematic errors with the DFT methods.

The lower errors for heats of formation with PDDG/PM3 carry over to lower errors for isomerization enthalpies, as illustrated for several series in Table 6. PDDG/PM3 some-

Table 5. Interaction Enthalpies of Sulfur-Containing Complexes

complex	ref	PDDG	MNDO	AM1	PM3	MNDO/d
NH ₃ ...SO ₂	-4.49 ^a	-4.26	-2.93	-13.37	-16.55	-2.74
MeNH ₂ ...SO ₂	-6.70 ^a	-3.41	-2.48	-10.63	-10.69	-1.94
Me ₂ NH...SO ₂	-7.93 ^a	-3.42	-2.18	-10.00	-1.05	-1.53
NMe ₃ ...SO ₂	-10.65 ^a	-2.98	-1.46	-8.12	-0.89	-0.92
NH ₃ ...SO ₂ ...NH ₃	-3.00 ^a	-7.60	-5.43	-23.34	-2.31	-4.68
MeNH ₂ ...SO ₂ ...MeNH ₂	-4.28 ^a	-6.06	-4.60	-18.25	-1.87	-3.83
CH ₃ SH...F ⁻	-38.04 ^b	-34.73	-60.55	-103.16	-65.37	-61.76
CH ₃ F...HS ⁻	-12.54 ^b	-9.78	-5.35	-7.77	-5.37	-4.30
CH ₃ SH...H ₂ O ^c	-1.16 ^d	-3.56	-0.52	-1.90	-1.02	-0.57
CH ₃ SH...H ₂ O ^e	-2.21 ^d	-3.42	-0.62	-2.37	-2.43	-0.51
CH ₃ SCH ₃ ...H ₂ O	-2.88 ^d	-4.32	-0.69	-2.44	-2.45	-0.25
DMSO...H ₂ O	-6.56 ^d	-8.15	-2.20	-5.61	-6.17	-1.90
EtSH...CH ₃ S ⁻	-5.40 ^f	-12.39	-11.45	-9.90	-14.37	-3.32
NH ₃ -HS ⁺	-111.45 ^g	-104.00	-101.22	-138.30	-141.19	-70.1
H ₂ O-HS ⁺	-72.51 ^g	-64.71	-69.00	-92.11	-89.45	-42.4
H ₂ S-HS ⁺	-93.36 ^g	-93.88	-84.32	-124.81	-122.12	-86.61
HF-HS ⁺	-33.17 ^g	-24.90	-48.45	-102.27	-64.40	-39.39
HCl-HS ⁺	-60.87 ^g	-58.60	-49.50	-84.69	-90.34	-36.89
MAE (negative or neutral)		3.21	5.23	9.88	6.19	5.26
MAE (positive)		5.26	9.89	34.16	27.23	21.68
MAE (all, kcal/mol)		3.78	6.52	16.62	12.03	9.82

^a MP3/6-31+G(2d,p) values at 298 K.²¹ ^b CCSD(T)/aug-cc-pVTZ ZPE corrected with MP2 frequencies.²² ^c Water as hydrogen-bond acceptor.
^d MP2(fc)/6-31++G(2d(X+),p) ZPE corrected.²³ ^e Water as hydrogen-bond donor. ^f MP2/6-31+G** ZPE corrected.²⁴ ^g G2 values at 0 K.²⁵

Table 6. Heat of Formation Results (kcal/mol) for Some Isomeric Series

	exptl.	PDDG	MNDO	AM1	PM3	MNDO/d
CH ₃ CH ₂ SCH ₃	-14.2	-17.4	-23.0	-15.6	-14.1	-14.0
CH ₃ CH ₂ CH ₂ SH	-16.2	-15.3	-18.2	-17.4	-14.1	-14.7
(CH ₃) ₂ CHSH	-18.2	-17.1	-16.2	-16.0	-14.4	-12.5
HSCH ₂ CH ₂ CH ₂ SH	-7.1	-6.2	-11.0	-9.8	-3.8	-4.0
HSCH ₂ CHSHCH ₃	-7.1	-7.1	-8.3	-9.5	-5.0	-0.4
CH ₃ CH ₂ SSCH ₃	-11.8	-13.7	-20.6	-10.7	-8.8	-11.5
CH ₃ CH ₂ SCH ₂ CH ₃	-20.0	-20.7	-29.0	-21.8	-17.2	-19.5
CH ₃ CH ₂ CH ₂ CH ₂ SH	-21.1	-18.9	-22.9	-24.2	-19.5	-19.4
(CH ₃) ₂ CHSCH ₃	-21.4	-24.8	-25.3	-20.1	-18.4	-15.9
CH ₃ CH ₂ CHSHCH ₃	-23.0	-21.7	-20.0	-21.0	-18.6	-15.7
(CH ₃) ₂ CHCH ₂ SH	-23.1	-22.8	-18.9	-21.6	-18.8	-14.6
(CH ₃) ₃ CSH	-26.0	-25.0	-15.1	-18.7	-17.9	-11.1
HSCH ₂ CH ₂ CH ₂ CH ₂ SH	-12.0	-11.5	-16.0	-17.1	-9.7	-8.9
CH ₃ CH ₂ SSCH ₂ CH ₃	-17.9	-18.0	-26.5	-17.3	-12.7	-16.4
(CH ₃) ₂ CHSSCH ₃	-18.9	-19.5	-23.4	-15.2	-11.9	-13.6

what overestimates the stability of sulfides relative to thiols but does particularly well for dithiols relative to disulfides and for branching isomers.

Silicon Thermochemistry. Table 7 summarizes the results for heats of formation for various classes of silicon-containing compounds. Considering all 144 molecules in the dataset, MNDO/d provides the smallest MAE, and the differences between PM3, AM1, MNDO/d, and PDDG/PM3 are small. There are some classes of compounds for which MNDO/d is particularly advantageous, specifically, divalent silicon molecules, ions, and complexes. PDDG/PM3 gives results that are generally similar to those of PM3 or AM1 for most classes, though it does particularly well for silyl ethers. All of the methods have difficulties with the compounds containing halogens and silicon; however, most of these are very small polyhalosilanes.

Phosphorus Thermochemistry. Reliable thermochemical data for organophosphorus compounds has historically been scarce,²⁷ which makes the development of semiempirical methods difficult. The present dataset consists of only 43 molecules, taken mostly from the paper reporting the parametrization of MNDO/d.⁵ The dataset is both too small and poorly representative of key functional groups containing phosphorus. Thus, the parametrizations should be considered tentative and used with caution. The summary of results in Table 8 should be viewed in this light. PDDG/PM3 gives a MAE for heats of formation similar to that of AM1 and shows a 17% smaller error than PM3. The best method for this dataset is MNDO/d, with a MAE of 12.7 kcal/mol. Because the dataset contains mostly small molecules, the advantage of d orbitals apparently outweighs the homologation and branching problems of MNDO. The MAE for

Table 7. MAEs for Heats of Formation for Silicon Compounds (kcal/mol)

	N	PDDG	MNDO	AM1	PM3	MNDO/d
all	144	11.9	16.2	11.7	11.1	9.4
training	56	10.3	13.0	9.1	9.8	7.4
test	88	12.9	18.2	13.3	11.9	10.7
CHSi	40	7.6	8.2	7.5	6.4	8.3
CHONSi	5	4.4	7.2	4.5	13.5	2.7
halides	31	15.2	26.5	13.5	9.8	15.0
ethers	9	6.9	11.1	7.8	8.6	9.6
anions	4	17.0	21.7	14.4	9.2	6.7
cations	10	13.8	21.0	8.5	24.6	7.3
radicals	14	11.9	14.1	16.8	9.5	7.8
complexes	9	21.7	14.0	20.6	28.5	11.9
divalent	11	14.6	19.0	12.1	12.2	5.2
others	9	13.0	14.3	11.8	6.3	6.8

Table 8. MAEs for Heats of Formation for Phosphorus Compounds (kcal/mol)

	N	PDDG	MNDO	AM1	PM3	MNDO/d
all	43	17.9	44.8	18.2	21.5	12.7
training	21	11.3	47.7	14.5	18.8	9.0
test	22	24.1	42.1	21.7	24.0	16.3
low valence	1	9.0	24.8	36.7	27.6	7.9
normal valence	25	11.0	20.9	11.6	16.3	9.4
hypervalent	14	31.0	94.6	28.3	30.9	18.5
halides	8	20.0	56.5	20.8	14.7	14.4
radicals	3	16.8	18.4	19.3	19.1	15.8
phosphines	4	10.9	14.9	8.8	7.0	4.8
phosphates	3	34.6	80.3	6.3	34.4	15.3
phosphites	4	9.4	54.7	15.6	42.5	9.2

phosphorus halides increases for PDDG/PM3 when compared to PM3, as for sulfur and silicon halides. This trend suggests that the problem may be the halogen parameters. Though they give significantly improved results for CHNOX molecules, where X is a halogen, the halogen parameters are not optimal for molecules with second-row atoms.

For comparison with the new AM1* method, the results for the intersection of the AM1* dataset for phosphorus and the dataset used in this work were analyzed. For the 33 common molecules, the MAEs (kcal/mol) are MNDO/d (8.6), AM1 (13.1), PDDG/PM3 (13.8), PM3 (17.4), AM1* (17.6), and B3LYP/6-311+G(2d,p)//B3LYP/6-31+G(d) (12.6). MNDO/d, again, gives the best result; PDDG/PM3, AM1, and the DFT method give similar MAEs near 13 kcal/mol, and AM1* and PM3 yield MAEs of 17–18 kcal/mol.

Another recent development is the optimization by Lopez and York of the AM1/d method to treat nucleophilic attack on biological phosphates.²⁸ The method was optimized using a set of B3LYP reference values for phosphates, phosphoranes, and transition structures. Because of the different scope of the AM1/d and the PDDG/PM3 parametrizations and the lack of overlap between the datasets used, making a fair comparison is difficult; it is to be expected that the AM1/d method will perform significantly better for the types of systems for which it was parametrized, whereas PDDG/PM3 has a wider scope in terms of functional groups included in the parametrization.

Table 9. MAEs for Ionization Potentials (IP), Dipole Moments, and Geometries

	element	N	PDDG	MNDO	AM1	PM3
IP (eV)	S	32	0.66	0.71	0.48	0.35
	Si	15	1.08	0.86	0.63	1.21
	P	13	0.42	1.27	0.66	0.61
dipole (D)	S	28	0.54	0.44	0.43	0.46
	Si	26	0.92	0.84	0.45	0.72
	P	12	0.77	0.83	0.79	0.58
bond length (Å)	S	78	0.06	0.13	0.07	0.08
	Si	96	0.09	0.13	0.06	0.11
	P	16	0.12	0.08	0.08	0.06
angle (deg)	S	54	5.50	9.46	6.74	7.42
	Si	33	3.25	2.47	2.39	2.75
	P	8	3.36	3.39	3.76	1.91

Table 10. MAEs for Heats of Formation (kcal/mol) for All Molecules Studied

	N	standard NDDO methods				PDDG	
		MNDO	AM1	PM3	MNDO/d	MNDO	PM3
CHNO ^a	622	8.4	6.7	4.4	8.4	5.2	3.2
halogens ^b	422	14.0	11.1	8.1	13.4	6.6	5.6
S	249	41.2	10.6	10.5	10.0		6.4
Si	144	16.2	11.7	11.1	9.4		11.9
P	43	44.8	18.2	21.5	12.7		17.9
all	1480	19.8	10.8	8.7	10.3		6.5

^a Reference 17. ^b Reference 18.

Other Observables. The parametrization of the PDDG/PM3 method has focused on improving heats of formation and, therefore, enthalpies of reaction in general while retaining accuracy comparable to prior NDDO-based methods for other observables such as geometries, ionization potentials, and dipole moments. Table 9 shows the MAEs for these properties, while the complete results are, again, reported in the Supporting Information. The performances of the various methods are similar overall, with no striking problems.

Summary

The parametrization of PDDG/PM3 for sulfur showed the benefits of a global optimization over simpler local optimizations and the value of the availability of a large amount of experimental data for training and testing. Through global optimization, a method was obtained that yields MAEs for heats of formation that are 40% smaller than those from the alternatives, including semiempirical methods that employ d orbitals and DFT at the B3LYP/6-311+G(2d,p)//B3LYP/6-31+G(d) level. Extension to silicon and phosphorus was also carried out, though it was hampered by limited experimental data.

PDDG/PM3 can now be used to treat molecules with any combination of the elements C, H, N, O, F, Cl, Br, I, S, Si, and P. Table 10 summarizes the results for all molecules, ions, and complexes that have been treated so far. The PDDG/PM3 method is the most accurate semiempirical method available for calculating heats of formation with a MAE of 6.5 kcal/mol for the 1480 molecules in our full dataset; PM3 is the next best with an MAE of 8.7 kcal/mol.

Particularly striking improvements have been obtained for molecules containing C, H, N, O, S, and the halogens. Moreover, the PDDG/PM3 method can be easily implemented into existing software such as MOPAC 6.²⁹

Acknowledgment. Gratitude is expressed to the National Science Foundation (CHE-013996) and to the National Institute of Health (GM032136) for support of this work.

Supporting Information Available: An Excel table with the experimental and calculated heats of formation, ionization potentials, and dipole moments for the molecules in this study. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Pople, J. A.; Beveridge, D. L. *Approximate Molecular Orbital Theory*; McGraw-Hill: New York, 1970.
- (2) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4917.
- (3) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 7, 3902–3909.
- (4) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221–264.
- (5) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100*, 616–626.
- (6) Gogonea, V.; Suarez, D.; Van der Vaart, A.; Merz, K. M., Jr. *Curr. Opin. Struct. Biol.* **2001**, *11*, 217–223.
- (7) Acevedo, O.; Jorgensen, W. L. *Org. Lett.* **2004**, *6*, 2881–2884.
- (8) Vayner, G.; Houk, K. N.; Jorgensen, W. L.; Brauman, J. I. *J. Am. Chem. Soc.* **2004**, *126*, 9054–9058.
- (9) Guimaraes, C. R. W.; Repasky, M., P.; Chandrasekhar, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2003**, *125*, 6892–6899.
- (10) Guimaraes, C. R. W.; Udier-Blagović, M.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2005**, *127*, 3577–3588.
- (11) Chandrasekhar, J.; Shariffskul, S.; Jorgensen, W. L. *J. Phys. Chem. B* **2002**, *106*, 8078–8085.
- (12) Ruiz-Pernia, J. J.; Silla, E.; Tunon, I.; Marti, S.; Moliner, V. *J. Phys. Chem. B* **2004**, *108*, 8427–8433.
- (13) Ranaghan, K. E.; Mulholland, A. J. *Chem. Commun.* **2004**, 1238–1239.
- (14) Devi-Kesavan, L. S.; Garcia-Viloca, M.; Gao, J. *Theor. Chem. Acc.* **2003**, *109*, 133–139.
- (15) Devi-Kesavan, L. S.; Gao, J. *J. Am. Chem. Soc.* **2003**, *125*, 1532–1540.
- (16) Kaminski, G. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 1787–1796.
- (17) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601–1622.
- (18) Tubert-Brohman, I.; Guimaraes, C. R. W.; Repasky, M. P.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 138–150.
- (19) Stewart, J. J. P. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–105.
- (20) Winget, P.; Clark, T. *J. Comput. Chem.* **2004**, *25*, 725–733.
- (21) Wong, M. W.; Wiberg, K. B. *J. Am. Chem. Soc.* **1992**, *114*, 7527–7535.
- (22) Gonzales, J. M.; Cox, R. S., III; Brown, S. T.; Allen, W. D.; Schaefer, H. F., III. *J. Phys. Chem.* **2001**, *105*, 11327.
- (23) Rablen, P. R.; Lockman, J. W.; Jorgensen, W. L. *J. Phys. Chem. A* **1998**, *102*, 3782–3797.
- (24) Gronert, S.; Lee, J. M. *J. Org. Chem.* **1995**, *60*, 6731–6736.
- (25) Solling, T. I.; Radom, L. *Chem.—Eur. J.* **2001**, *7*, 1516–1524.
- (26) Winget, P.; Horn, A. H. C.; Selcuki, C.; Martin, B.; Clark, T. *J. Mol. Model.* **2003**, *9*, 408–414.
- (27) Pilcher, G. In *The Chemistry of Organophosphorus Compounds*; Hartley, F. R., Ed.; John Wiley & Sons: New York, 1990; Vol. 1, pp 127–136.
- (28) Lopez, X.; York, D. M. *Theor. Chem. Acc.* **2003**, *109*, 149–159.
- (29) Directions, code, and parameter files for implementation of the PDDG methods in MOPAC are freely available for download at <http://www.jorgensenresearch.com>. The PDDG methods are also available in the BOSS program; see <http://www.cemcomco.com>.

CT0500287