Reaction Enthalpies Using the Neural-Network-Based X1 Approach: The Important Choice of Input Descriptors

Matthew D. Wodrich and Clémence Corminboeuf*

Laboratory for Computational Molecular Design, Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

Received: January 8, 2009; Revised Manuscript Received: February 12, 2009

Artificial neural networks represent a simple but efficient way to model and correct known errors existing between commonly used density functional computations and experimental data. The recently proposed X1 approach combines B3LYP energies with a neural-network correction. The latter receives input from a set of physical descriptors, which are primarily based on B3LYP energies. The method shows remarkable improvements for enthalpies of formation and bond energies, for molecules containing first and second row elements, in comparison to B3LYP. Here, reaction enthalpies of organic compounds containing H, C, N, and O are derived using the X1 method, as well as B3LYP, M05-2X, and G3. Despite the seemingly impressive results obtained with X1, our study reveals that underlying problems with B3LYP descriptions of medium and long-range correlation remain. Thus, X1, like B3LYP, breaks down when describing both linear and branched organic molecules. These deficiencies likely arise from the improper or insufficient selection of physical descriptors. To improve the B3LYP energies by means of a neural-network correction, we stress the importance of considering protobranching-dependent descriptors in the input layer of the neural network.

Introduction

Becke's three-parameter hybrid method (B3LYP)[1,2] remains the most commonly used exchange correlation density functional among chemists, despite recognized errors when computing energies of basic organic molecules.[3−17] Known problems, such as predictions of alkane isomerization energies, have been attributed to failures in describing van der Waals interactions,[18] medium-range electron correlation,[8] and many electron self-interaction errors.[19−23] Various approaches designed to lessen these problems have been developed to date (see e.g., refs 24–47). Most are based on the extension of the exchange-correlation potential (ν_{xc}) by a nonlocal van der Waals term,[43,44] the superimposition of a pairwise interatomic potential,[27] the direct fitting of ν_{xc} to data sets of weakly bond compounds,[45] or the empirical calibration of dispersion corrected atomic potentials.[28,46,47]

Another simple yet efficient way to correct such errors is to model the complex relationship between the first-principles computation and the experimental data with the goal of using this relation to eliminate the deviation. Artificial neural networks (NN) are one approach of this kind. Inspired by the primary work of Hu et al.,[48] who developed a neural-network-based approach to improve the B3LYP heats of formation of 180 organic molecules, Wu and Xu[49−51] recently introduced a similar NN-based method called X1. The latter was shown to drastically improve the prediction of B3LYP heats of formation on an extended set of 370 molecules presenting large structural diversity. Both methods combine B3LYP with a three layered neural-network correction but differ in the selection of input descriptors. In the input layer of the X1 method, the network receives inputs from a set of physical descriptors consisting of B3LYP energies, unscaled zero-point vibrational energy (ZPE), the number of electrons, and the number of each constituent element. The output layer provides the corrected values, while the hidden layer serves to adjust the connections so that the error in the output values is minimal.[50] This NN scheme was used to compute the heats of formation of compounds containing all first and second row elements except He, Ne, and Ar, producing results similar to those of the Gaussian composite methods[52] for the G3/99 test set. The error for bond energies of compounds containing first and second row elements was even reduced by 3.81 kcal/mol over B3LYP.[53] Despite this remarkable performance, one may wonder if "molecular" descriptors based primarily on B3LYP potential energy surfaces would not result in deficiencies similar to those of B3LYP for pathological systems. In this Article, we address this question by evaluating

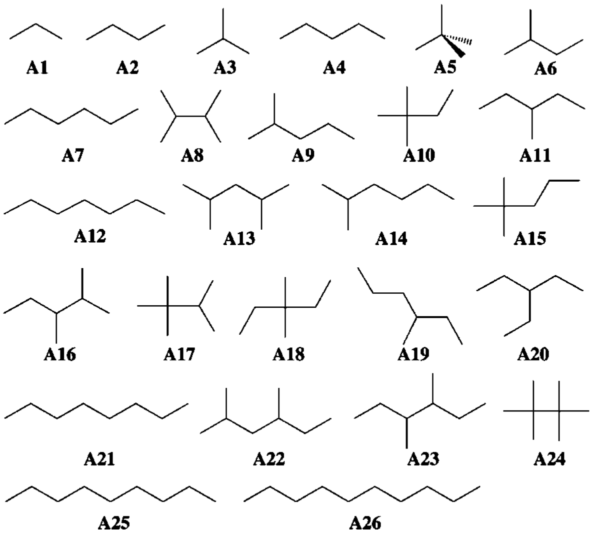

Figure 1. Schematic representation of linear and branched alkane compounds (A set).

* Corresponding author. E-mail: clemence.corminboeuf@epfl.ch.

10.1021/jp9002005 CCC: $40.75
Published on Web 03/10/2009

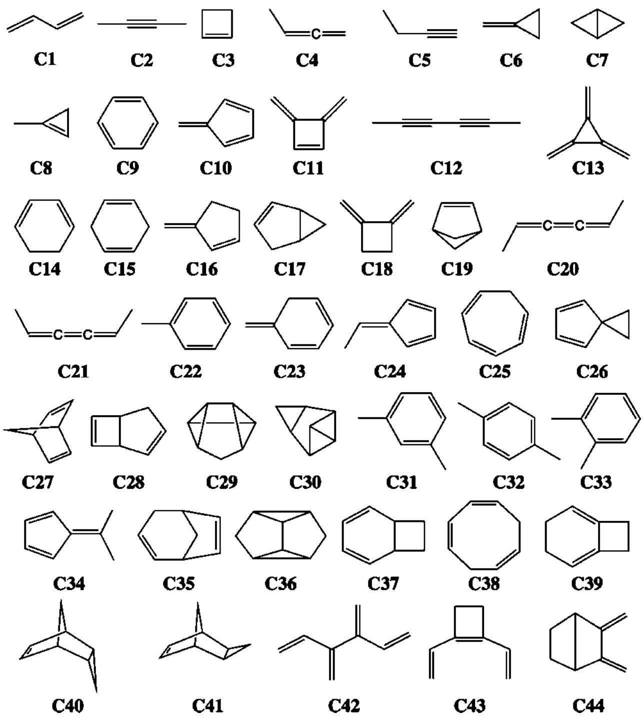

Figure 2. Schematic representation of other hydrocarbon molecules (C set).

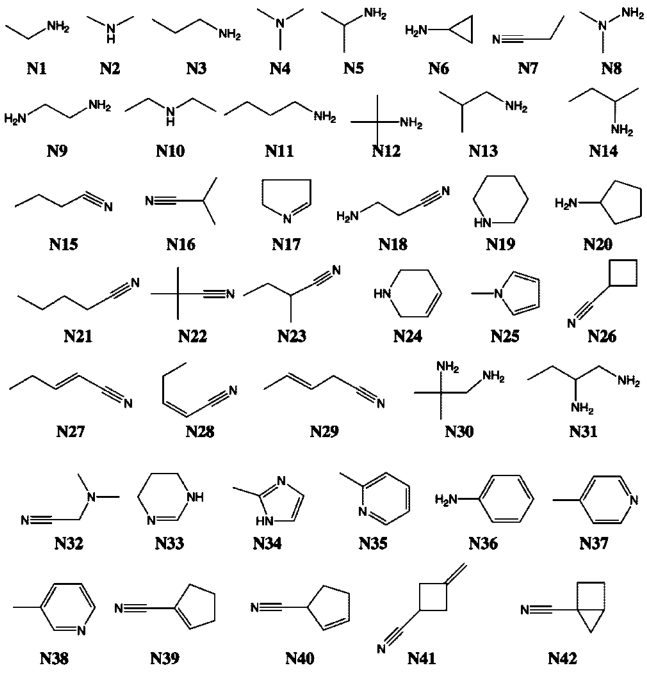

Figure 3. Schematic representation of nitrogen-containing organic molecules (N set).

the performance of such a method in the determination of the enthalpies of reaction for small to medium-sized organic molecules.

Methods

165 H-, C-, N-, and O-containing compounds with diverse structural and electronic features (e.g., branched, conjugated, hyperconjugated, benzene moieties, etc.) having experimental data were selected. Results were compared to the standard B3LYP functional, the M05-2X functional[25] (which was previously shown to be the best performing functional for hydrocarbon thermochemistry[16]), and the G3 composite method.[52] Bond separation reaction[54-57] enthalpies for a series of 26 linear and branched alkanes (**A1**−**A26**, Figure 1), 44 additional hydrocarbons (**C1**−**C44**, Figure 2), 42 nitrogen-containing organic (**N1**−**N42**, Figure 3), and 53 oxygen-containing organic compounds (**O1**−**O53**, Figure 4) were compared to experimental data derived from the NIST thermochemical database.[58] B3LYP and M05-2X computations employed the same basis sets as used for the X1 method [geometry optimization and zero-point/thermal corrections at 6-311+G(d,p) and single point electronic energy using the 6-311+G(3df,2p) basis set]. All computations were done using the Gaussian 03[59] and NWChem[60] suite of programs. Mean absolute deviations (MADs) and mean signed deviations (MSDs) serve to evaluate the performance of the various methods. The MAD gives the average unsigned deviation from experiment, while the MSD gives the average signed deviation.

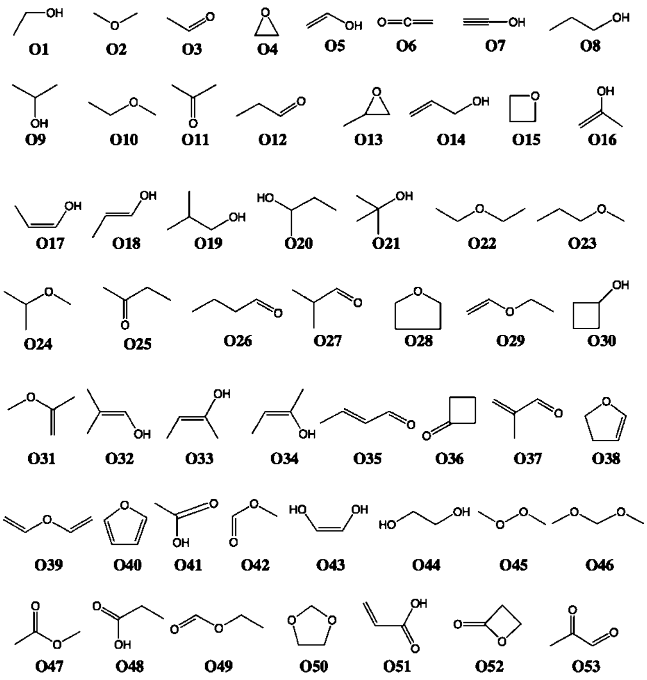

Figure 4. Schematic representation of oxygen-containing organic molecules (O set).

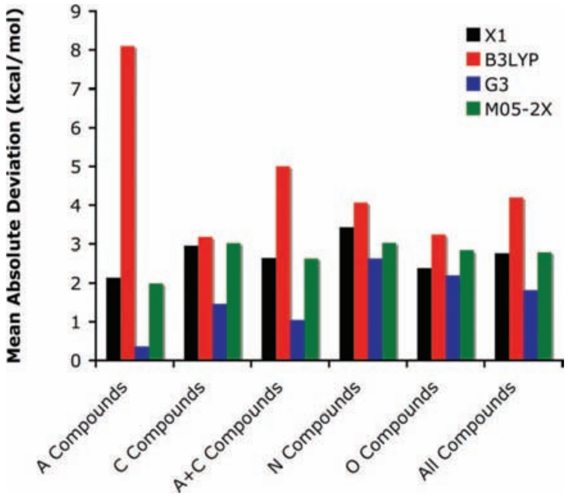

Figure 5. Mean absolute deviations for bond separation reactions for various molecular sets.

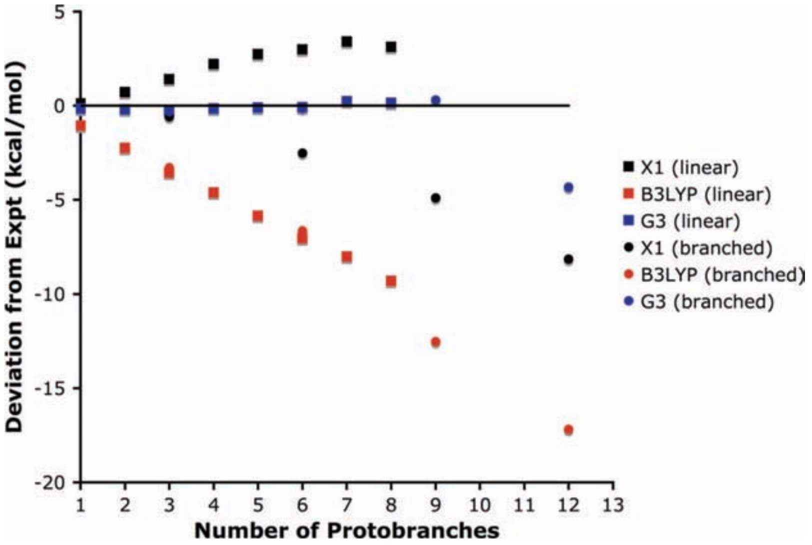

Figure 6. Deviations from experiment for bond separation reactions for linear alkanes containing one (propane) to eight (*n*-decane) protobranches and branched alkanes containing three (A3), six (A5), nine (A17), and twelve (A24) photobranches. In contrast to B3LYP, the X1 method overestimates protobranching stabilization in linear *n*-alkanes but underestimates branched alkanes.

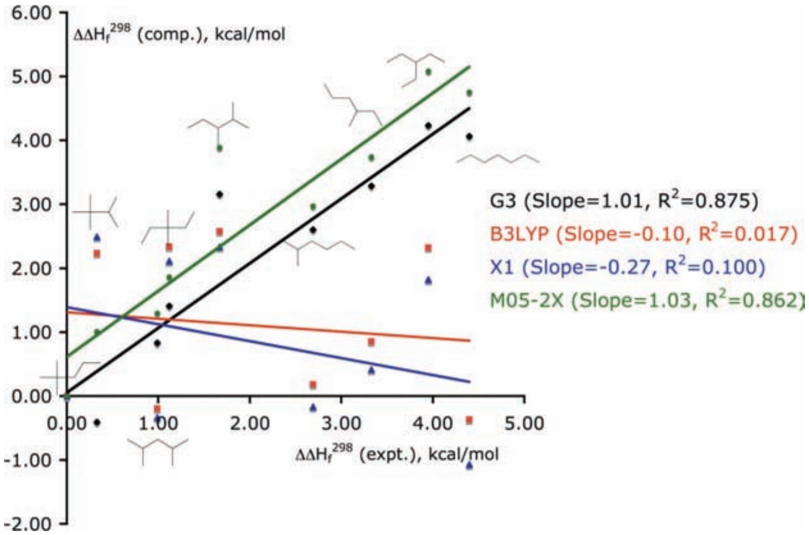

Figure 7. Computed C_7H_{16} alkane isomer energy differences at various computational levels versus experiment.

Results and Discussion

Figure 1 illustrates a test set of 26 linear and branched hydrocarbon molecules (A set) for which DFT and in particular B3LYP has been shown to perform poorly.[12,42] These failures have been attributed to the poor description of protobranching[61] interactions present in linear and branched alkanes, believed to arise from medium and long-range electron correlation.[18] The results carry no surprise (see Table 1); the B3LYP functional is the worst performer among the four methods tested. M05-2X greatly reduces errors over B3LYP, due to its enhanced description of medium-range electron correlation brought about by improved dependence on the spin kinetic energy density. The smallest errors are given by the G3 composite method, which incorporates electron correlation by using perturbation and configuration interaction theory. At first glance, the B3LYP-based X1 method performs impressively well: the MAD (Figure 5 and Table 1) is ~75% less than that of B3LYP, and the MSD (Table 1) indicates that the method is nearly free from systematic errors. While the combination of B3LYP with a neural-network correction seems to compensate the poor description of reaction

TABLE 1: Mean Absolute (MAD) and Mean Signed (MSD) Deviations from Experiment (in kcal/mol) for the Bond Separation Reactions of Relevant Test Sets

	B3LYP		X1		G3		M05-2X	
	MAD	MSD	MAD	MSD	MAD	MSD	MAD	MSD
A set	8.10	−8.10	2.14	−0.83	0.36	−0.32	1.99	−1.99
C set	3.18	−1.63	2.95	0.53	1.46	−0.37	3.02	−2.45
A+C sets	5.01	−4.03	2.65	0.02	1.05	−0.35	2.64	−2.28
N set	4.07	−2.60	3.43	−0.61	2.64	−0.36	3.03	−1.09
O set	3.24	−2.44	2.39	−0.18	2.20	−1.05	2.84	−1.65
all	4.20	−3.16	2.77	−0.21	1.82	−0.58	2.80	−1.77

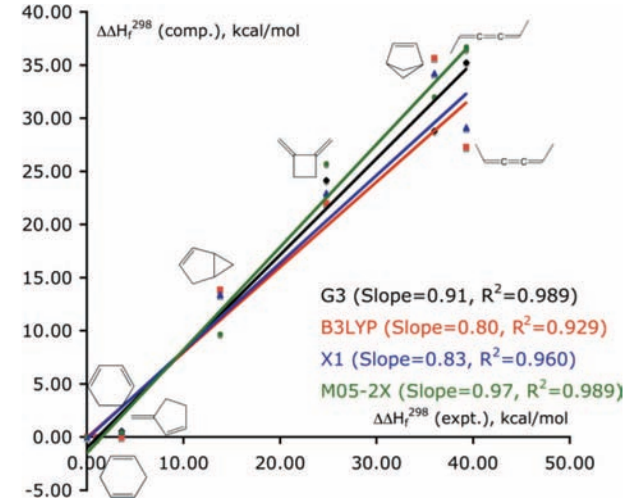

Figure 8. Computed C_6H_8 isomer energy differences at various computational levels versus experiment.

enthalpies by B3LYP, a close inspection indicates that X1 evidently fails to describe (proto)branched compounds. For instance, while X1 overestimates the BSE of all linear alkane chains (e.g., **A25** by 3.40 kcal/mol), that of branched alkanes is significantly underestimated (e.g., **A17** by −4.89 kcal/mol). As further illustrated in Figure 6, the NN-based B3LYP correction suffers from a lack of consistency with respect to the protobranch point topology by making an artificial distinction between branched and linear alkanes.

Even more striking is the poor performance of X1 for predicting alkane isomerization reactions (see Figure 7). In these cases, both B3LYP and X1 show dramatic errors with respect to the experimental (NIST) heats of formation of the medium size C_7H_{16} isomers chosen as an example. In sharp contrast, M05-2X shows a more systematic improvement in the treatment of (proto)branched compounds reducing the mean absolute deviation.

The set of compounds given in Figure 2 (C set) contains fewer examples of alkane branching and more examples of other stereoelectronic effects such as hyperconjugation, conjugation, ring strain, and resonance. The B3LYP description of the BSE reactions of this set is far better than that for the first set of molecules (Table 1). In fact, the MADs for B3LYP, M05-2X, and the X1 method are relatively close, with the G3 composite method again providing a superior description. Similarly, all four methods perform relatively well for describing the isomerization energies such as those illustrated by the C_6H_8 series in Figure 8. A safe conclusion to draw is that the X1 method performs well for hydrocarbons as long as B3LYP captures most of the dominating electronic effects present in the set of compounds being studied.

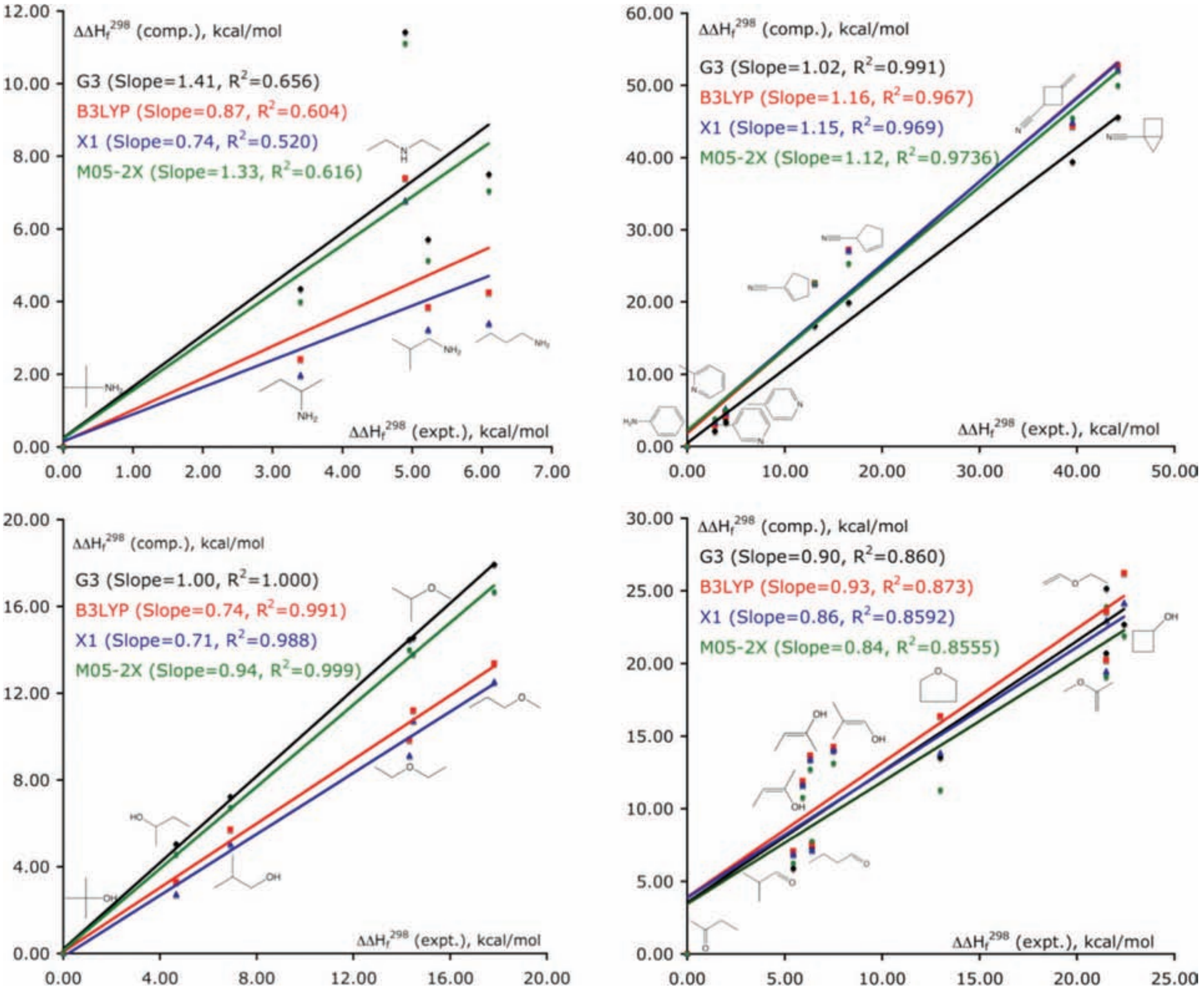

Figure 9. Computed isomer energy differences for $C_4H_{10}N$ (upper left), C_6H_7N (upper right), $C_4H_{10}O$ (lower left), and C_4H_8O (lower left) at various computational levels versus experiment.

These general findings recur for the nitrogen and oxygen derivatives that include the additional stereoelectronic effects arising from the nitrogen and oxygen lone pairs (Figures 3 and 4). BSE reactions of branched compounds such as **N2**, **N8**, **O2**, and **O10** are overestimated by X1 and underestimated by B3LYP. Although to a lesser extent than hydrocarbons, the disagreement with experiment is more apparent when considering the reactions of the branched structural isomers of $C_4H_{11}N$ and $C_4H_{10}O$ as illustrated in Figure 9. Other stereoelectronic effects are generally well described, although aromatic rings incorporating nitrogen atoms (i.e., **N34**, **N35**, **N37**, and **N38**) have large errors with all methods tested.

We now return to the important question regarding the selection of proper physical descriptors in the X1 neural-network-based method. Because this study emphasizes the poor description of protobranching interactions, which arise in B3LYP prediction of alkane isomerization energies, any improvements should focus on describing the branching in a more accurate or explicit manner. However, the descriptors selected for the input layer of the X1 method are either branching-independent (N_e, N_C, etc.) or B3LYP-dependent (ΔH_f^{B3LYP}, ZPE). This inadequate selection or insufficient number of descriptors in the input layer is a likely explanation for the observed collapse of the X1 method for saturated hydrocarbons and their derivatives. There is no doubt that the inclusion of the number of branching points in the input descriptors will lead to more satisfactory results for these sets of compounds.

Conclusions

The recently proposed X1 method based on a neural-network scheme has a significantly lower mean absolute deviation than does B3LYP for the 165 bond separation reactions tested herein (Figure 5 and Table 1). This is noteworthy as the heats of formation computed by the X1 method are derived directly from B3LYP energies, zero-point vibrational energies, and a compound's stochiometry. However, despite this seemingly overall improvement, the NN-based correction (Table 1) suffers from the same inherent defects as B3LYP, which lead to poor BSE reaction energies and disastrous predictions of isomerization energies for linear and branched alkanes. For conjugated hydrocarbons as well as their nitrogen- and oxygen-containing derivatives, the agreement with experiments is, in contrast, generally quite good. Such a deficiency again points to the importance of propane-like branching interactions, whose treatment remains a problem for DFT but also provides insights for improvement. To improve the B3LYP energies by means of a neural-network correction, it is therefore necessary to introduce more physical descriptors in the input layer such as the number of protobranched points or the number of adjacent sp^3 carbons.

Acknowledgment. The Sandoz Family Foundation, the Swiss NSF (grant 200021_121577/1), and the EPFL are acknowledged for supporting this work. We dedicate this paper to Lily E. Wodrich.

Supporting Information Available: Bond separation reactions for compounds in Figures 1−4, and computed energies and heats of formation (for X1) as well as experimental heats of formation used to create Table 1 and Figure 5. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
(2) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
(3) Choi, C. H.; Kertesz, M.; Karpfen, A. *Chem. Phys. Lett.* **1997**, *276*, 266.
(4) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.
(5) Woodcock, H. L.; Schaefer, H. F.; Schreiner, P. R. *J. Phys. Chem. A* **2002**, *106*, 11923.
(6) Check, C. E.; Gilbert, T. M. *J. Org. Chem.* **2005**, *70*, 9828.
(7) Izgorodina, E. I.; Coote, M. L.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 7558.
(8) Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460.
(9) Izgorodina, E. I.; Coote, M. L. *J. Phys. Chem. A* **2006**, *110*, 2486.
(10) Izgorodina, E. I.; Coote, M. L. *Chem. Phys.* **2006**, *324*, 96.
(11) Schreiner, P. R.; Fokin, A. A.; Pascal, R. A., Jr.; de Meijere, A. *Org. Lett.* **2006**, *8*, 3635.
(12) Wodrich, M. D.; Corminboeuf, C.; Schleyer, P. v. R. *Org. Lett.* **2006**, *8*, 3631.
(13) Grimme, S.; Steinmetz, M.; Korth, M. *J. Chem. Theory Comput.* **2007**, *3*, 42.
(14) Izgorodina, E. I.; Brittain, D. R. B.; Hodgson, J. L.; Krenske, E. H.; Lin, C. Y.; Namazian, M.; Coote, M. L. *J. Phys. Chem. A* **2007**, *111*, 10754.
(15) Schreiner, P. R. *Angew. Chem., Int. Ed.* **2007**, *46*, 4217.
(16) Wodrich, M. D.; Corminboeuf, C.; Schreiner, P. R.; Fokin, A. A.; Schleyer, P. v. R *Org. Lett.* **2007**, *9*, 1851.
(17) Brittain, D. R. B.; Lin, C. Y.; Gilbert, A. T. M.; Izgorodina, E. I.; Gill, P. M. W.; Coote, M. L. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1138.
(18) Wodrich, M. D.; Wannere, C. S.; Mo, Y.; Jarowski, P. D.; Houk, K. N.; Schleyer, P. v. R. *Chem.-Eur. J.* **2007**, *13*, 7731.
(19) Mori-Sanchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2006**, *125*, 201102.
(20) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. *Phys. Rev. B* **2008**, *77*, 115123.
(21) Mori-Sanchez, P.; Cohen, A. J.; Yang, W. *Phys. Rev. Lett.* **2008**, *100*, 146401.
(22) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. T. *Science* **2008**, *321*, 792.
(23) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2007**, *126*, 104102.
(24) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397.
(25) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
(26) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
(27) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
(28) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.
(29) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 154108.
(30) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 014104.
(31) Johnson, E. R.; Becke, A. D. *Chem. Phys. Lett.* **2006**, *432*, 600.
(32) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2006**, *124*, 174104.
(33) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.
(34) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.
(35) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.
(36) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2006**, *28*, 555.
(37) Ducere, J.-M.; Cavallo, L. *J. Phys. Chem. B* **2007**, *111*, 13124.
(38) Zimmerli, U.; Parrinello, M.; Koumoutsakos, P. *J. Chem. Phys.* **2004**, *120*, 2693.
(39) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.
(40) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515.
(41) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287.
(42) Wodrich, M. D.; Jana, D. F.; Schleyer, P. v. R.; Corminboeuf, C. *J. Phys. Chem. A* **2008**, *112*, 11495.
(43) Andersson, Y.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **1996**, *76*, 102.
(44) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401.
(45) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.
(46) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. B* **2005**, *71*, 195119.
(47) Tkatchenko, A.; von Lilienfeld, O. A. *Phys. Rev. B* **2007**, *75*, 205131.
(48) Hu, L.; Wang, X.; Wong, L.; Chen, G. *J. Chem. Phys.* **2003**, *119*, 11501.
(49) Wu, J.; Xu, X. *J. Chem. Phys.* **2007**, *127*, 214105.
(50) Wu, J.; Xu, X. *J. Comput. Chem.*, in press.
(51) X1 heats of formation can be computed from B3LYP computations using the following website: http://www.pcoss.org/users/xinxu/X1.php.
(52) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703.
(53) Wu, J.; Xu, X. *J. Chem. Phys.* **2008**, *129*, 164103.
(54) A reaction in which all formal bonds between heavy (nonhydrogen) atoms are separated into the simplest (or parent) molecules with the same type of bond.

(55) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. *J. Am. Chem. Soc.* **1970**, *92*, 4796.

(56) Radom, L.; Hehre, W. J.; Pople, J. A. *J. Am. Chem. Soc.* **1971**, *93*, 289.

(57) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; John Wiley & Sons: New York, 1986.

(58) Afeefy, H. Y.; Liebman, J. F.; Stein, S. E. Neutral Thermochemical Data. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD, 2004; http://webbook.nist-.gov.

(59) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(60) Straatsma, T. P.; Apra, E.; Windus, T. L.; Bylaska, E. J.; de Jong, W.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Harrison, R.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Brown, E.; Cisneros, G.; Fann, G.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dashsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers, Version 5.1*; Pacific Northwest National Laboratory: Richland, WA, 2004.

(61) Protobranching is defined as the stabilization arising from 1,3-alkyl−alkyl interactions present in all linear and branched alkanes, but not in methane or ethane. See ref 18.

JP9002005