JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2018, 58, 12−26

pubs.acs.org/jcim

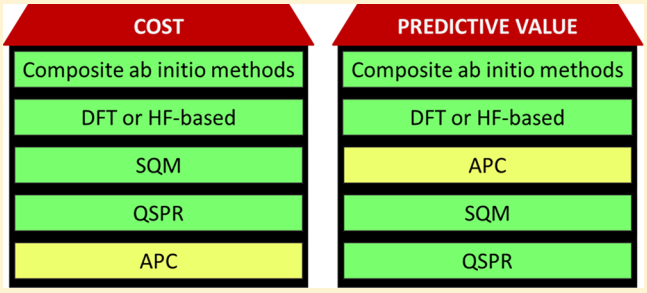
Atom Pair Contribution Method: Fast and General Procedure To Predict Molecular Formation Enthalpies

Didier Mathieu*

CEA, DAM, Le Ripault, 37260 Monts, France

Supporting Information

ABSTRACT: An atom pair contribution (APC) model aimed at predicting the gas-phase formation enthalpy ($\Delta_f H°$) of molecules is reported. In contrast to bond contribution (BC) or group contribution (GC) methods, it relies on increments associated with pairs of bonded and geminal atoms along with 15 structural corrections. Another distinctive feature of the present APC method is the large amount of experimental and high-level theoretical data specially compiled in this work to fit and validate the model (2671 entries). Unlike GC methods, the present APC model has wide applicability with the number of adjustable parameters limited to 68. Although it requires only a structural formula as input and involves only back-of-the-envelope calculations, it is more reliable than state-of-the-art quantitative structure−property relationship methods, popular semiempirical Hamiltonians, and even low-level density functional theory approaches based on gradient-corrected functionals. It is therefore a valuable tool for fast screening applications or whenever chemical accuracy is not necessary.



INTRODUCTION

The standard formation enthalpy ($\Delta_f H°$) of a molecule is one of its most fundamental properties. Accordingly, a wealth of procedures to predict this quantity have been reported, covering a wide spectrum from simple additivity schemes to high-level composite ab initio methods.[1] Between these two ends, other prominent approaches include quantitative structure−property relationship (QSPR) methodologies,[2−10] molecular mechanics (MM) force fields,[11−13] semiempirical quantum-chemical methods (SQM),[14−18] procedures based on Hartree−Fock (HF) or density functional theory (DFT) combined with empirical corrections or isodesmic/isodesmotic reaction schemes,[19,20] and a variety of alternative hybrid procedures.[21−23] As shown in Figure 1, these approaches can be loosely classified on the rungs of a Jacob's ladder according to their computational requirements.

The specific ranking illustrated in this figure is actually not without ambiguities. For instance, the complexity of a QSPR method depends mainly on the descriptors involved. The latter are usually classified according to the level of chemical structure representation required for their determination: 1D descriptors depend only on the molecular formula (such as the oxygen balance), and 2D descriptors depend on the structural formula (i.e., the molecular topology), whereas 3D descriptors are conformation-dependent and therefore include all quantum-chemically derived quantities.[24] Obviously, QSPR methods based on AM1 or PM3 descriptors, like the ones put forward respectively by Bagheri et al.[9] and Vatani et al.,[2] are at least as costly as the straightforward application of the corresponding SQM method. Therefore, the classification of the QSPR technique just above the lowest rung of the ladder in Figure 1 is justified only for models based exclusively on easily accessible descriptors (i.e., 1D and 2D), excluding those involving 3D descriptors.

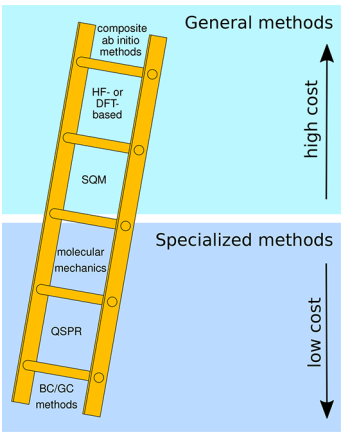


Figure 1. Classification of "black box" procedures to estimate $\Delta_f H°$ on the rungs of a Jacob's ladder reflecting their relative complexity and cost (see the text for details).

The most computer-intensive procedures, namely, ab initio methods based on post-HF calculations, are clearly the most accurate, especially when combined into composite schemes.[25] However, more computationally efficient methods remain of

Received: October 19, 2017
Published: December 22, 2017

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

interest. They are invaluable for the automatic generation of reaction mechanisms,[26] for experimentalists in need of quick enthalpy estimates to analyze measurements,[27] for education purposes,[28] or when it comes to high-throughput screening of $\Delta_f H°$ or related properties with the goal of identifying the most promising compounds with regard to a given application, especially in such fields as energetic materials, fuels, and alternative power sources.[29,30] In such cases, additivity methods are especially attractive because of their outstanding simplicity and extremely low cost. However, simple bond contribution (BC) methods exhibit clear limitations,[31,32] while the application of group contribution (GC) schemes to compounds of practical interest is often hampered by a lack of suitable parameters.[6]

Consequently, SQM methods, especially those based on the neglect of diatomic differential overlap (NDDO), are often used when enthalpies of formation are needed for many molecules with diverse chemical functionalities.[18,30,33−36] For instance, the reaction mechanism generator (RMG) software for detailed kinetic modeling of chemical reaction networks relies on the GC method of Benson for aliphatic compounds but switches to quantum-chemical computations for cyclic systems because of the limited availability of ring-strain corrections.[26] Not surprisingly, in the lack of heavy reliance on empirical fitting (i.e., for procedures on the highest rungs of the ladder), better results go hand in hand with more accurate approximations and higher complexity. However, this is not necessarily the case for procedures lying on the bottom rungs, where the introduction of numerous fitting parameters partially makes up for the crude approximations introduced. The development of accurate and efficient methods through carefully parametrized approximate expressions usually incurs a loss of generality (Figure 1). For instance, GC and MM approaches often provide very accurate $\Delta_f H°$ values, although their scope is usually restricted to specific families of chemicals in view of the many parameters needed.[6] In practice, GC methods have been reported to yield rather fair predictions (although clearly not as good as those obtained using composite ab initio methods) for a wide range of compounds, including rather unusual structures like high-energy compounds.[37] Despite the well-identified limitations of GC methods for ring systems and unusual functional groups, SQM approaches are not necessarily more accurate, as observed in several studies.[38−41] In fact, the popularity of SQM methods primarily stems from their all-terrain character rather than from their predictive value.

Rather surprisingly in view of the encouraging performance of GC methods, it seems that little effort has been made to date to develop simpler additivity methods that are possibly less accurate but have a generality comparable to that of SQM methods. This raises the question of how well such a scheme could perform. In an attempt to provide an answer, the present work describes an extremely simple and generally applicable atom pair contribution (APC) model with two major distinctive features:

1. In contrast to all existing methods, it is based on the assumption that atom−atom interactions between bonded and geminal atom pairs are transferable, while 1,4-interactions are neglected. This approximation has been previously considered only in the context of reactive potentials for molecular dynamics simulations.[42]
2. The present implementation was developed on the basis of 2365 reference enthalpies and validated using 306 additional values, i.e., using a data set much larger than used to fit and validate previous methods. This data set was obtained by merging previously available compilations and additional values taken from recent research papers, as detailed below.

In addition to the present APC approach, this paper describes standard DFT-based atom-equivalent schemes. They are presently used as reference quantum-chemical methods in view of a rigorous comparison of APC with such approaches on the basis of the same data sets. Full details regarding data sets and parametrization procedures are provided. A preliminary model restricted to hydrocarbon compounds is first described. The latter, denoted APC(HC), is shown to exhibit good performance compared with state-of-the-art machine learning techniques. A more general APC model is finally described, demonstrating its superiority over alternative procedures with similar generality and efficiency. This article concludes with a thorough discussion of currently available approaches for the fast evaluation of $\Delta_f H°$.

PRESENT APC MODEL

The standard formation enthalpy ($\Delta_f H°$) of a molecule is defined as the enthalpy change during the formation of the compound from its constituent elements, all species being considered in their standard states. Such a transformation may be defined by the set of covalent bonds broken or formed in the process. Therefore, an especially straightforward approach to describe $\Delta_f H°$ is to assign a fixed binding enthalpy value to each kind of covalent bond.[31,43] Historically, in its simplest version, such an approach was quickly dropped as it proved less accurate than more heavily parametrized group contribution approaches.[44] This is to be expected, as such a bond contribution (BC) method ignores medium-range interactions (in addition to long-range ones), even those between atoms in geminal positions. Such interactions between geminal atoms are implicitly taken into account by GC methods, albeit at the cost of numerous fitting parameters. Traditional BC models imply that energy changes are zero for isodesmic reactions, which is clearly a very crude approximation.

However, in recent investigations aimed at developing a reactive potential for molecular dynamics simulation of chemical reactions, it was observed that explicitly accounting for interactions between atoms in geminal positions through empirical constants already leads to a dramatic improvement over the BC approach, as reflected by a decrease by a factor of 3.5 in the average absolute error (AAE) and root-mean-square error (RMSE).[42] This finding led to a preliminary model based on transferable pairwise interactions. In contrast to BC models, this one includes not only interactions through bonded atoms but interactions between geminal atoms as well. However, since it was designed with the development of a continuous potential in mind, it includes only pairwise atom−atom parameters that would subsequently be expressed as continuous functions of dynamical bond orders. In particular, it does not use any parameter associated with structural elements such as rings or crowded chemical environments, as such bonding patterns are likely to change in the course of chemical reactions. In addition, it is trained against a relatively small amount of experimental data, which might be insufficient in view of possible near linearities.[42]

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

The present work introduces a new variant of this approach that is more specifically aimed at predicting $\Delta_f H°$ for organic molecules. We still approximate the total enthalpy as a sum of 1−2 and 1−3 atom−atom interactions that do not significantly depend on the molecular geometry (because of the constraints on bond lengths and valence angles). However, we now introduce additional corrections that prove necessary to get satisfactory results. In particular, three contributions are introduced for three-, four-, and five-membered rings (hereafter denoted R_3, R_4, and R_5) in order to account for the associated strain. Distinct contributions R_{4a} and R_{5a} are introduced for four- and five-membered aromatic rings (three-membered aromatic rings were not encountered in the present data set). In addition, specific corrections denoted as $\Delta_f H°(444)$ and $\Delta_f H°(666)$ are assigned to every atom simultaneously involved in three four-membered and six-membered nonaromatic rings, respectively, in order to account for the enthalpy change associated with the cage structures of cubane and adamantane derivatives. Similarly, following previous findings regarding the prediction of formation enthalpies for polycyclic aromatic hydrocarbons (PAHs),[45] extra corrections $\Delta_f H°(aa)$ and $\Delta_f H°(3a)$ are introduced for every pair of fused aromatic rings and for every atom involved in three aromatic rings, respectively. A small number of group corrections are introduced as well to account for systematic errors to be expected for some substructures with adjacent multiple bonds. In this work, such substructures are detected using SMARTS strings,[46] as implemented in the RDKit library.[47] In addition to ring corrections, the present model relies on five enthalpy corrections associated with the following substructures with adjacent multiple bonds:

- any carbon atom (=C=) bonded through two cumulative double bonds (in cumulenes);
- any nitro group (NO_2), associated in the present scheme with two N=O bonds;
- any azide group (N_3), associated here with two N=N bonds.

In addition, two ad hoc structural corrections have been introduced a posteriori in view of large and systematic overestimations of $\Delta_f H°$ observed in preliminary fits of the model. These two corrections are denoted as $\Delta_f H°(2/3/NO_2)$ and $\Delta_f H°(CA)$. They apply respectively to any sp^3 carbon atom bearing two or three nitro groups and to any cyclic amide, in other words, an amide group whose C and N atoms belong to a ring. Although we had not anticipated this correction, it is quite natural, as the amide linkage is well-known for conferring structural rigidity.[48] Therefore, it must lead to significant strains when introduced into a ring.

Finally, predicting the formation enthalpy of any molecule simply consists of evaluating the following sum:

$$\Delta_f H^\circ = \sum_{A} \Delta_f H^\circ(A) + \sum_{A-B} \Delta_f H^\circ(A{\sim}B) + \sum_{A..B} \Delta_f H^\circ(A..B) + \sum_{SC} \Delta_f H^\circ(SC) \quad (1)$$

where A, A−B, A..B, and SC denote atoms, covalent bonds, geminal pairs, and structural corrections, respectively, and $\Delta_f H°(X)$ denotes the contribution of any structural pattern X to the formation enthalpy of the molecule. To make the notation less cluttered, we may denote such a term simply as "X" instead of $\Delta_f H°(X)$. In what follows, the symbols "∼" and ":" denote covalent bonds with unspecified and aromatic bond orders, respectively, in line with the SMARTS notation,[49] while "—", "=" and "≡" denote single, double, and triple bonds. For simplicity, any difference between the value of $\Delta_f H°$ calculated from a structural formula and the corresponding experimental value is denoted by δ and expressed in kJ/mol, where the unit is implied.

The formation enthalpies of gaseous atoms, $\Delta_f H°(A)$, are experimentally available,[50] and the other contributions are fitting parameters obtained using the ordinary least-squares routine implemented in the StatsModels package (http://www.statsmodels.org). The fitting ability and predictive value of the model are assessed using various statistical indicators: AAE, RMSE, and minimal (MIN) and maximal (MAX) errors (all in kJ/mol) and the determination coefficient R^2.

REFERENCE QUANTUM-CHEMICAL PROCEDURES

The APC model is compared in this paper to a range of existing procedures, taking advantage of test sets previously introduced to assess their predictive value. In addition, in order to get the most from the large collection of data specially compiled for this work, the corresponding formation enthalpies were evaluated using standard quantum-chemical methods based either on NDDO Hamiltonians implemented in FIREFLY,[51] which is partly based on the GAMESS(US) source code,[52] or DFT functionals available in ORCA.[53] Initial geometries were built with the help of the RDKit library[47] and DG-AMMOS.[54] Because of the size of the present database, no attempt was made to carry out conformational searches using quantum-chemical methods. Therefore, it must be noted that failures to identify the lowest-energy conformations might contribute to the overall error in the quantum-chemically predicted enthalpies.

For DFT-based procedures, the initial geometries were optimized using the efficient BP functional associated with the small def2-SVP basis set (standard ORCA split-valence basis) and dispersion corrections with Becke−Johnson damping (D3BJ keyword). In the first procedure (AE1), the total electronic energy E_{tot} (including nuclear repulsion) was taken directly from the last step of the BP/def2-SVP+D3BJ optimization. In the second procedure (AE2), it was obtained from a subsequent single-point calculation at the B3LYP/DefBas-4 level, keeping D3BJ dispersion.

Finally, the total energy E_{tot} was converted into the formation enthalpy $\Delta_f H°$ through the addition of atom equivalents (AEs), according to a popular approach.[21,55−57] Following the recommendations of Dewar and O'Connor,[55] the AEs depend only on the chemical element under consideration and not on the bonds and groups to which the atom belongs. For a reference set of energetic compounds studied at the B3LYP level, this point was recently shown to get increasingly valid as larger basis sets are considered (see the Supporting Information for ref 58). The present AE values are provided in the Supporting Information for this article (Table S1). They were derived from a fit against the same training set as used to derive the APC parameters using the StatsModels routine as mentioned above.

DATA SETS AND PARAMETRIZATION

For any empirical model, a number of independent experimental enthalpies just equal to the number of unknown parameters is in principle sufficient to unambiguously fix their values. However, in that case any measurement error would

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

directly impact the parameters and thus negatively affect the predictive ability of the model.

In the lack of sufficient amounts of reliable data to confidently fit all 68 APC parameters on the basis of a minimal number of accurate reference values,[59] we assume that quantity can at least partially make up for quality, as suggested by the demonstration that data availability is as important as data quality to develop successful models.[60] In other words, we rely on the expectation that noise in the reference data arising from random errors gets averaged out in a large data set.[61] A further advantage of a large data set is that rampant near-linearities that might plague smaller data sets can be alleviated, thus contributing to reduce the model sloppiness.[62] Systematic errors that might arise for specific chemical families are taken into account through the above-mentioned structural corrections.

The present data were taken from previous compilations, including the NIST Chemistry WebBook,[63] from which most of the values come; recent compilations of data for hydrocarbons,[4] CHO compounds,[64] and refrigerants;[65] and carefully curated data sets for nitroaromatic compounds (NACs),[66] azides,[67] aliphatic nitro compounds, and nitramines.[68] Additional data were also retrieved from original research papers, as detailed in the full database provided in the Supporting Information. Only closed-shell organic molecules with at least four atoms were considered. Whenever several distinct values were listed in the NIST Chemistry WebBook, the most recent one was retained.

For the sake of comparison, this work makes use of various test sets previously introduced in the literature to assess earlier predictive schemes. More specifically, three test sets focused on hydrocarbon compounds (HCs) are considered:

- HC-100: experimental data for 100 general HCs introduced by Teixeira et al.[4] to assess the predictive value of a random forest model;
- PAH-103: experimental data for 103 PAHs taken from Table 13 in ref 45, which are especially useful to test the performance of the structural corrections $\Delta_f H^\circ(aa)$ and $\Delta_f H^\circ(3a)$;
- CUB-20: G3MP2B3 data for 20 cubane derivatives studied at the G3MP2B3 level,[69] used in validating the structural correction $\Delta_f H^\circ(444)$ associated with cubane-like cages.

In addition, two other test sets are used to validate the APC model for general organic compounds with heteroatoms:

- ORG-20: experimental data for 20 general organic compounds previously used to validate a third-order GC model;[70]
- ORG-10: data for an alternative set of 10 organic compounds recently used to assess various procedures based on atomic corrections to DFT or semiempirical calculations.[71]

Finally, in view of assessing APC in the context of the design of new refrigerants, explosives, or propellants, we also consider two specific test sets focused on the corresponding molecular families:

- FRIG-32: experimental data for 32 fluorinated and chlorinated refrigerants recently considered by Demenay et al.;[65]
- HEDM-45: data set of 45 high-energy compounds with carefully checked data.[19,20,41]

The chemical specificity of FRIG-32 and HEDM-45 stems from the high F/Cl content of the molecules in FRIG-32 and the high O/N content of those in HEDM-45, with O atoms preferentially linked to N (rather than C) atoms (typically within nitro groups).

These test sets were previously set up and curated by the original authors, taking advantage of earlier compilations in addition to the NIST Chemistry WebBook as detailed therein. Therefore, the data included in the present paper come from a variety of sources and were obtained using a range of experimental techniques. These differences are irrelevant in the present context, since the deviations from experiment to be expected from the present model are much larger than the uncertainties associated with measurement techniques.

A small number of obvious outliers and compounds associated with dubious data were handled separately, as detailed below, and some errors were identified and corrected. First, the training set entry for tricyclo[4.3.1.0^{8,10}]decane included in the database of Teixeira et al.[4] was deleted, as we could not find any experimental $\Delta_f H^\circ$ data for this compound. The database originally contained the value $\Delta_f H^\circ$ = −85.9 kJ/mol, which is in fact the value measured for protoadamantane. This error stems from the fact that the latter compound is mistakenly associated with the CAS registry number of tricyclo[4.3.1.0^{8,10}]decane in the NIST Chemistry WebBook.[63] On the other hand, a value of +359 kJ/mol measured at high temperature (812 °C) for the compound shown in Figure 2 has been corrected.[63] More specifically, we used the lower value of +249.3 kJ/mol deduced from the data reported by Wodrich et al.[72]

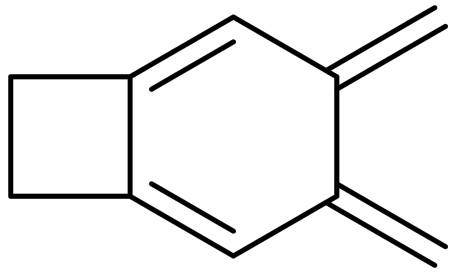


Figure 2. A difficult case for present models: 3,4-dimethylenebicyclo[4.2.0]octa-1,5-diene (CAS registry number 136846-70-3), a compound combining ring strain and diradical character.

As mentioned above, a preliminary procedure denoted APC(HC) and restricted to hydrocarbon compounds is first examined and compared to the random forest (RF) of Teixeira et al.[4] Although many other QSPR methods are available for HCs,[3−5,7] this recent RF model is especially suitable as a reference scheme for comparison with the present approach because of its especially good predictive value. In addition, it takes full advantage of state-of-the-art machine learning techniques for variable selection, model training, and validation. Finally, it was carefully assessed through a cross-validation against the training set and applied to the HC-100 external test set.

The general model put forward in this article, hereafter denoted simply as APC, is applicable to virtually all classes of organic compounds. Therefore, a large and diverse training set was necessary to fit the corresponding parameters. We first considered the set obtained after all of the molecules belonging

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

to one of the above-mentioned test sets were removed, as required to obtain a reliable assessment of the true predictive value of the model. However, a first tentative fit revealed a significant limitation of the present approach, or of any approach based on a representation of molecules as atoms linked together through covalent bonds with either integer or aromatic formal bond orders. The issue concerns any molecule with diradical character (like the one shown in Figure 2) or partial aromatic character (e.g., as a result of a quinoid−aromatic competition), for which formal bond orders cannot be unambiguously assigned. For such compounds, any algorithm determining formal bond orders (including the procedure implemented in RDkit and used in this work) is to some extent arbitrary and prone to yield an unrealistic representation of the electronic structure for some molecules.

Moreover, our initial data set contained some molecules with obviously strained structures (such as fused three-membered rings) for which no specific correction could be implemented because of a lack of reference data to check the transferability of the corresponding parameters. These molecules include 3*H*-diazirine, 2-methylthiirane, and bullvalene. They were removed from the present database. After curation, there were 2365 compounds in the training set aimed at fitting the general APC procedure put forward in this article. It might be worth emphasizing that because it was obtained by first removing all of the test set molecules from the overall data compilation, it does not overlap with the test sets.

All of the training and test sets used in this work are provided in the Supporting Information (in Tables S2 and S3 for APC(HC) and Tables S4 and S5 for APC) along with the observed and presently calculated values of $\Delta_f H°$ that make it easy to carry out further analyses. Corresponding statistical data characterizing the overall performance of the various procedures are provided in Table S6.

APC(HC) MODEL FOR HYDROCARBONS

Optimized Parameters. Before the discussion of the predictions of APC(HC) reported in Tables S2 and S3, the consistency of the APC(HC) fitting parameters is checked in this section. It should be noted that some values like $\Delta_f H°(3a)$ and, to a lesser extent, $\Delta_f H°(C{\equiv}C)$ and $\Delta_f H°(=C=)$, should be considered with caution. Indeed, the Teixeira et al. training set presently used is somewhat unbalanced with regard to the structures of the molecules considered. For instance, cubane is the only compound with a cubic cage structure, associated with the $\Delta_f H°(444)$ correction, while pyrene is the only molecule with carbon atoms belonging to three aromatic cycles, for which the $\Delta_f H°(3a)$ correction is deemed necessary. Similarly, there are only four cumulenic compounds and five alkynes in the training set. Therefore, empirical parameters specific to such structural features might be statistically ill-defined. Nevertheless, in view of the fact that APC(HC) involves a relatively small number of parameters with a straightforward physical interpretation, this issue is probably less problematic than with purely empirical approaches like QSPR procedures.

In fact, a preliminary fit of APC(HC) against the training set revealed that some of the parameters were zero within statistical uncertainties, namely, the cage corrections $\Delta_f H°(444)$ and $\Delta_f H°(666)$ and the geminal interactions C..H and C..C. Therefore, the model was eventually fit with these parameters ignored. Interestingly, this implies that although geminal interactions prove to be crucial to improve oversimple BC models for general organic compounds (as shown in ref 42), they are of little significance when only hydrocarbons are considered. The optimal parameters are listed in Table 1.

Table 1. Values of the 11 Parameters of the APC(HC) Model

k	$\Delta_f H°(k)$ (kJ/mol)	$\sigma(k)$ (kJ/mol)	$N(k)$
Bond Enthalpy Contributions			
C—H	−413.26	0.4	357
C—C	−347.65	0.8	344
C:C	−509.59	0.4	83
C═C	−623.65	1.9	120
C≡C	−829.85	9.6	5
Structural Corrections			
R_3	116.31	4.8	22
R_4	104.86	4.2	12
R_5	18.68	2.8	55
═C═	60.42	14.0	4
aa	141.35	3.6	29
3*a*	−159.12	15.2	1

Throughout this paper, $\Delta_f H°(k)$ denotes the contribution of a structural feature k to $\Delta_f H°$, $\sigma(k)$ the corresponding standard deviation, and $N(k)$ the number of compounds in the training set contributing to $\Delta_f H°(k)$.

A significant asset of APC over empirical QSPR models stems from the clear physical meaning of the parameters, which makes consistency checks easy. For instance, the fact that bond contributions to $\Delta_f H°$ become more negative as the bond order increases is to be expected. On the other hand, the contribution of aromatic C:C bonds exhibits an extra stabilization (by about 15 kJ/mol) with respect to a direct extrapolation from the contributions of other C∼C bonds for integer values of the bond order, as illustrated in Figure 3. Again, this is consistent in view of the availability of several equivalent resonance forms in aromatic systems. On the other hand, the (positive) strain energy associated with small rings (three to five atoms) and cumulenic carbons is unambiguously quantified. Depending on the type of descriptors used, QSPR models may also lend

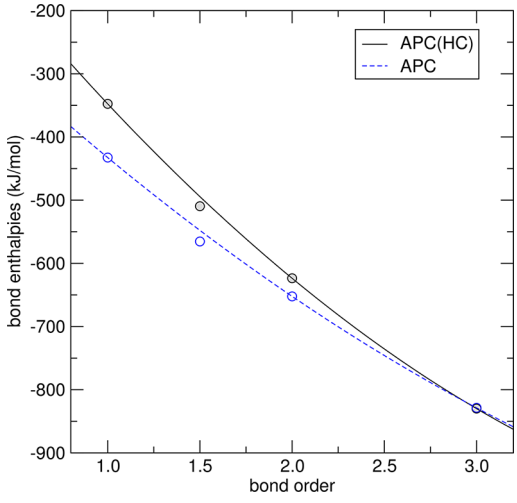


Figure 3. Contribution of C∼C bonds to $\Delta_f H°$ as a function of the bond order, illustrating the extra stabilization due to aromaticity and the increased stability of the bonds in going from APC(HC) (fit against hydrocarbons with no geminal interactions) to APC (fit against all molecules with geminal interactions explicitly included).

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

themselves to easy physical interpretation. However, the latter comes a posteriori and does not in itself constitute a validation of the model.

Results and Comparison with QSPR. Since APC(HC) relies on the same data sets as the RF model of Teixeira et al.,[4] which represents the current state of the art among QSPR models, the two methods may be easily compared. For this purpose, APC(HC) is assessed on the basis of a 10-fold cross-validation and through application of the model to the external test set, as done by those authors.[4] In regard to the cross-validation, APC(HC) yields an RMSE value of 34.5 kJ/mol, which can be compared to the value of 34.1 kJ/mol reported for the RF model. As indicated in Table S6, application of APC(HC) to the external test set yields RMSE = 46.8 kJ/mol, which is very close to the RMSE value of 48.6 kJ/mol reported for the RF model.

By far the most significant error obtained using APC(HC) is an underestimation of $\Delta_fH°$ by −302 kJ/mol for 136846-70-3 (Figure 2), despite the fact that the overestimated value originally reported for this compound (actually measured at 812 °C) was replaced with a more plausible value as reported above. This very large error comes as no surprise in view of the diradical character of this structure, similar to that of quinodimethanes. The second most significant error (δ = −207) is for pyracyclene (187-78-0), shown in Figure 4. Like

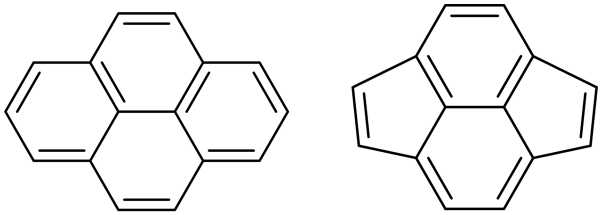


Figure 4. Enhanced strain energy in going from pyrene (129-00-0) to pyracyclene (187-78-0).

136846-70-3, this molecule cannot be represented by a single limiting structure, and the most suitable representation of this compound has been a matter of debate.[73] Another reason why $\Delta_fH°$ is underestimated for this compound is the occurrence of two carbon atoms simultaneously involved in three aromatic rings, one of them being five-membered. These two carbon atoms are associated with the $\Delta_fH°(3a)$ correction. However, as mentioned above, this correction was derived from a single compound from the training set (pyrene), which exhibits only six-membered aromatic rings. As is clear from Figure 4, the fact that two six-membered rings are replaced with five-membered rings in going from pyrene to pyracyclene introduces some constraints due to the fact that the optimal values of the valence angles are different for the two kinds of rings.

All of the remaining deviations are <100 kJ/mol. The third most significant deviation (δ = −98) is for *cis,trans*-1,5-cyclooctadiene (5259-71-2). Again, this is easy to understand, as the presence of two C═C double bonds in this eight-membered cyclic structure induces constraints that are ignored by the present model. Similarly, the fourth most significant deviation (δ = −83) is observed for a compound with obvious strain that is presently ignored, namely, tri-*tert*-butylmethane (35660-96-9). More generally, the main deviations observed from experiment systematically arise from obvious geometric constraints that are ignored by APC(HC), although they could easily be taken into account through additive correction terms if additional reference data were available.

GENERAL APC MODEL FOR ORGANIC MOLECULES

Model Parameters. Like the AE1 and AE2 procedures described above, the general APC model was fit against the present training set of 2671 compounds (Table S4) and applied to the various test sets (Table S5). Unlike the Teixeira et al. training set, the new one introduced in this work (Table S5) enables the determination of statistically significant values for the geminal C..C and C..H parameters and for the structural corrections $\Delta_fH°(444)$ and $\Delta_fH°(666)$ associated with cubane and adamantane cages. Only geminal interactions involving Br/I atoms were found to be zero within statistical uncertainties and thus simply ignored. The values eventually obtained for the parameters associated with bond contributions, geminal contributions, and structural corrections are reported in Table 2, Table 3, and Table 4, respectively, using the same notations as in Table 1.

Table 2. Bond Contributions to $\Delta_fH°$ for the General APC Model[a]

k	$\Delta_fH°(k)$ (kJ/mol)	$\sigma(k)$ (kJ/mol)	$N(k)$
C—H	−415.52	0.3	2268
C—F	−440.55	4.5	116
C—Cl	−340.08	4.7	185
C—Br	−261.01	2.6	58
C—I	−200.87	3.4	44
C—C	−432.62	5.2	1983
C:C	−565.35	3.4	780
C═C	−652.27	2.3	372
C≡C	−828.65	4.1	44
C—N	−376.29	5.5	533
C:N	−493.09	3.9	176
C═N	−614.66	5.9	35
C≡N	−860.71	2.9	119
C—O	−454.60	4.6	805
C:O	−496.24	5.3	31
C═O	−787.39	3.1	627
C—S	−338.28	7.7	167
C:S	−401.01	7.0	14
C═S	−537.67	8.5	23
N—H	−390.78	1.3	230
N—N	−244.96	9.6	33
N:N	−375.91	5.7	30
N═N	−448.48	7.8	54
N—O	−267.91	7.3	18
N→O	−328.47	8.5	38
N:O	−354.91	6.4	21
N═O	−559.08	8.6	192
N—F	−229.69	3.8	13
O—H	−457.68	1.7	398
O—O	−252.11	7.2	28
O—S	−323.48	7.3	12
O═S	−501.63	6.3	62
S—H	−364.51	4.9	35
S—S	−284.38	11.1	15

[a]The arrow in "N→O" denotes a dative bond between nitrogen and oxygen atoms, as encountered in N-oxides.

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

Table 3. Geminal Contributions to $\Delta_f H°$ for the General APC Model

k	$\Delta_f H°(k)$ (kJ/mol)	$\sigma(k)$ (kJ/mol)	$N(k)$
C..H	14.91	0.8	2229
C..C	30.31	1.8	2179
C..N	23.39	2.3	722
C..O	37.31	2.1	1247
C..F	8.21	1.9	111
C..Cl	12.88	2.6	174
C..S	23.29	3.2	181
N..H	17.61	1.6	403
N..N	15.98	4.0	161
N..O	8.12	3.4	147
N..F	15.06	3.1	16
N..Cl	25.41	4.8	7
O..H	31.62	1.5	537
O..O	18.14	3.5	647
O..S	35.34	5.7	11
F..H	−8.54	3.7	24
F..F	−34.45	3.3	89
S..S	14.17	7.0	16
Cl..Cl	15.90	3.8	37
H..S	12.75	2.6	149

Table 4. Structural Corrections to $\Delta_f H°$ for the General APC Model

k	$\Delta_f H°(k)$ (kJ/mol)	$\sigma(k)$ (kJ/mol)	$N(k)$	$N_{test}(k)$
R_3	189.90	5.5	88	13
R_4	94.79	3.4	58	32
R_5	11.30	2.1	181	45
R_{4a}	425.09	23.3	2	–
R_{5a}	53.34	6.8	92	5
=C=	31.50	4.9	23	–
aa	127.11	2.8	118	22
3*a*	−148.55	12.9	2	5
444	−12.49	3.9	4	6
666	−6.99	1.0	31	3
CA	239.14	7.1	15	–
2/3/NO_2	61.23	9.7	14	4
NO_2	212.62	17.0	177	33
N_3	−72.80	14.3	20	8

Let us first consider the bond contributions listed in Table 2. For C∼C bonds, these contributions are compared in Figure 3 to the values previously obtained for APC(HC). As expected, the new values for APC are more negative since they have to make up for the positive geminal interactions explicitly included in this more general procedure. Furthermore, the difference gets larger as the bond order decreases, in line with the fact that lower bond orders are associated with larger numbers of geminal pairs.

The dependence of the APC bond contributions on the atomic number of the bonded atoms and on the bond order is illustrated in Figure 5. As previously noted for C∼C bonds in the APC(HC) scheme, a smooth dependence on the value of the bond order is now observed for C∼C and C∼N bonds. Moreover, bonds linking a C atom to either C, N, or O are especially strong whereas bonds linking N to either N, O, or F are weaker, in line with basic organic chemistry knowledge.

As noted above for APC(HC) and illustrated in Figure 3, aromatic C:C and N:N bonds are somewhat more stable than

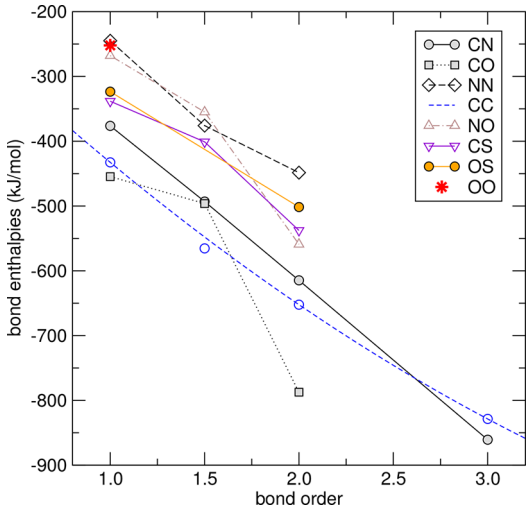


Figure 5. Atom type and bond order dependence of the bond contributions to $\Delta_f H°$ for the APC model.

expected from an extrapolation of the values obtained for integer bond orders. This is to be expected in view of the enhanced delocalization energy in aromatic systems compared with the corresponding (hypothetical) bond-alternant structures. For instance, from the values of the C∼C bond enthalpies, a resonance energy of 137 kJ/mol may be estimated for benzene, in qualitative agreement with the empirical value of about 150 kJ/mol obtained from hydrogenation experiments.[74] Similarly, a value of 175 kJ/mol is obtained for the resonance energy of a hypothetical N_6 aromatic ring, arising from the fact that the bond enthalpy for N:N is more negative than the average of the N—N and N═N bond enthalpies, as is clear from the convexity of the relationship shown in Figure 5.

On the other hand, this figure also shows that in contrast to C∼C and N∼N bonds, covalent bonds between distinct atoms (C∼N, C∼O, C∼S, or N∼O) exhibit a concave relationship between bond enthalpy and bond order. This may be explained by their polarity arising from the large electronegativity differences between the bonded atoms. In fact, the larger the difference between the electronegativities χ_A and χ_B of the bonded atoms A and B, the more concave the bond enthalpy−bond order relationship (i.e., the weaker the aromatic bond compared with the average of the single and double bonds). This is made clear in Figure 6, where Pauling electronegativities[50] are used and the concavity is defined from the bond enthalpies involved in eq 1 as follows:

$$\text{concavity(A∼B)} = \frac{\Delta_f H°(\text{A:B}) - \frac{1}{2}[\Delta_f H°(\text{A—B}) + \Delta_f H°(\text{A═B})]}{\Delta_f H°(\text{A—B}) - \Delta_f H°(\text{A═B})} \quad (2)$$

The correlation shown in this figure is especially clear if only first-row atoms are considered. However, it should be kept in mind that it critically depends on the values obtained for the bond enthalpies (especially for aromatic bonds), which could be somewhat unreliable for aromatic bonds involving N and especially O atoms, as the latter are mostly encountered in five-membered rings, whose strain energy is crudely described through a single R_{5a} correction.

Table 3 reports the contributions associated with geminal pairs. Except for F..H and F..F, they are all positive, in line with

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

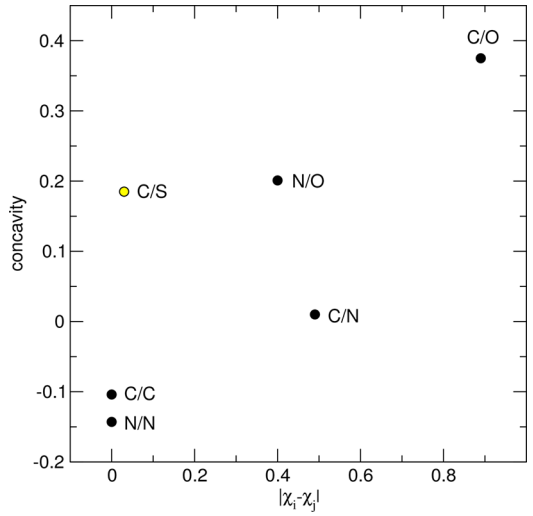


Figure 6. Concavity of the bond enthalpy defined according to eq 2 and considered as a function of bond order. The correlation between the concavity and the absolute electronegativity difference $|\chi_i - \chi_j|$ between the bonded atoms i and j is shown. The yellow symbol denotes a bond involving a second-row atom.

the repulsion that may be expected between atoms maintained within a van der Waals distance from each other. Finally, the structural corrections are listed in Table 4. Not surprisingly, rings with fewer than six atoms (especially aromatic ones) get increasingly strained as they become smaller. Large values are obtained as well for $\Delta_fH°(aa)$ and $\Delta_fH°(3a)$. However, they partially cancel each other in compounds involving the $3a$ correction. Finally, Table 4 also indicates as $N_{test}(k)$ the number of occurrences of each structural correction k in the external test sets. This points to the fact that the present external test sets do not sample the possibly ill-defined structural correction R_{4a}. However, the very high value obtained for this parameter is clearly consistent with the high strain associated with four-membered aromatic rings.

The two a posteriori corrections $\Delta_fH°(2/3/NO_2)$ and especially $\Delta_fH°(CA)$ exhibit large values as well, as expected from the significant errors preliminarily obtained in the absence of such terms. The especially large value of $\Delta_fH°(CA)$ contributes to the good performance of the explosive 5-nitro-1,2,4-triazol-3-one (NTO), presumably without increasing its sensitivity. Therefore, the present $\Delta_fH°(CA)$ parameter at least partially explains the exceptionally good performance−sensitivity trade-off offered by this energetic material as discussed recently.[58]

Interestingly, the cage corrections $\Delta_fH°(444)$ and $\Delta_fH°(666)$ prove to be negative. For the former, this may be explained by the fact that $\Delta_fH°(R_4)$ corrections already account for the strain energy associated with cubane-like structures.

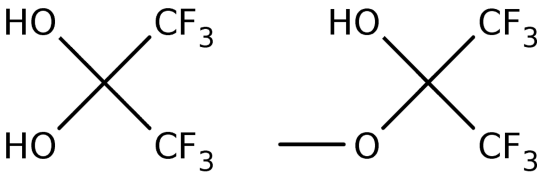
On the other hand, the values obtained for $\Delta_fH°(NO_2)$ and $\Delta_fH°(N_3)$ can be easily rationalized in terms of bond orders. Indeed, because nitro groups are viewed as involving two N═O double bonds, their stability is overestimated, and a positive $\Delta_fH°(NO_2)$ correction is needed. Conversely, the present description of the azide group as involving two N═N double bonds should underestimate its stability in view of the actual triple-bond character of the terminal N∼N bond.[75] Therefore, a negative correction is expected.

Validation of the APC Model. The predictions of the general APC models are compiled in Tables S4 and S5. The following discussion relies on these data and especially on the detailed statistical performance indicators compiled in Table S6. It should be noted that all of the statistical data were obtained after two refrigerants initially included in FRIG-32 were discarded because their measured enthalpies proved to be clearly erroneous, as pointed out recently by Demenay et al.[65] and detailed below. For convenience, we focus on RMSE data, as represented in what follows using bar charts.

Fitting Ability and Overall Predictive Value. Although the fit of the present model yields reasonable values of the parameters, it reveals a small number of outliers (Tables S4 and S6). There is little doubt that some of the largest differences between APC and observed values arise as a result of erroneous experimental data. For instance, the third most negative value of δ (−314) is obtained for perfluorobiphenyl. However, the APC enthalpy of −1579 kJ/mol is in fair agreement with the AE2 value of −1611 kJ/mol, hence strongly suggesting that the experimental value of −1264 kJ/mol is too high.

Similarly, the formation enthalpy of 165 kJ/mol predicted for thiophene is much lower than the most recent value of 218 kJ/mol reported in the NIST Chemistry WebBook. However, this database indicates that three earlier independent measurements agreed on a much lower value of 116 kJ/mol, in perfect agreement with present AE2 predictions and with high-level ab initio calculations.[76] Therefore, the upper experimental value is probably erroneous.

Other very negative values of δ were obtained for the two compounds shown in Figure 7. The experimental data



677-71-4 ($\delta = -152$) 662-16-8 ($\delta = -177$)

Figure 7. Two examples of molecules with dubious experimental data: the APC model and quantum-chemical calculations closely agree on much lower values of their formation enthalpies.

measured in the same study for the diol compound and its derivative are −1551 and −1506 kJ/mol, respectively, whereas corresponding APC values are −1703 and −1683 kJ/mol. The latter are in excellent agreement with the AE2 values of −1701 and −1696 kJ/mol, and the PM3 values are very similar. Again, as observed above for perfluorobiphenyl, the two experimental values might be erroneous for these two fluorinated molecules. This is all the more likely as they were measured in the same study and have not been subsequently confirmed.[77]

Actually, by far the most negative δ is observed for the indigo isomer 17352-37-3 shown in Figure 8. The validity of the experimental value is supported by the fact that it is fully consistent with the value reported for 578-95-0, also shown in Figure 8, assuming that the APC error for the central subunit is additive. Indeed, the error is twice as large for 17352-37-3, in line with the fact that it exhibits two such subunits. This demonstrates the value of additive increments associated with large molecular subunits, as used in advanced GC methods.[70] The reasons for such severe errors are not clear. In addition to the above-mentioned limitations of simple additivity schemes for fused aromatic systems, another factor that might be relevant for these two molecules is the potential delocalization

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

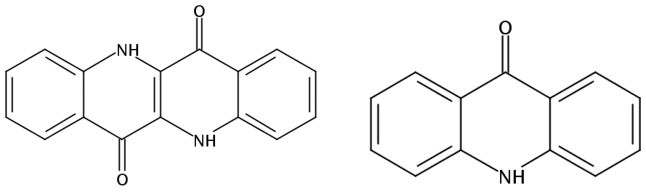


Figure 8. APC errors showing their approximate additivity for a specific six-membered cyclic subunit. These two compounds exhibit some of the most significant deviations from experiment observed using APC.

of the labile protons that can bind to either nitrogen or oxygen atoms. Although the canonical structures shown in Figure 8 are probably predominant, alternative tautomer forms ignored by the present approach might contribute to the special stability of such compounds.

The quality of fit and predictive value of APC and other reference methods are summarized in Figure 9. The lower part

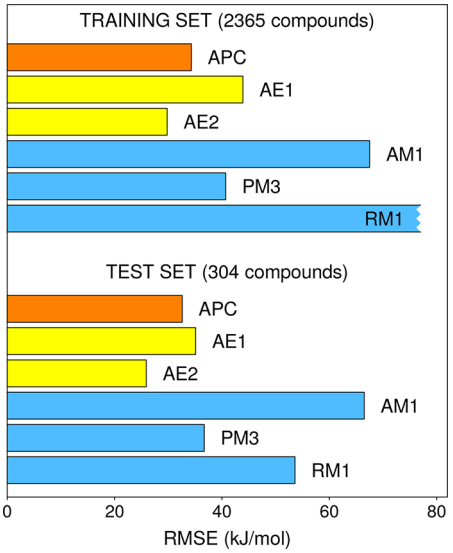


Figure 9. Performance of APC and quantum-chemistry-based procedures for the present training and test sets. The orange, yellow, and blue bars show RMSE data for empirical/additivity methods and procedures based on DFT and semiempirical Hamiltonians, respectively.

of this figure summarizes statistics derived from the results for the 304 compounds included in the overall test set, which was obtained by merging all of the individual test sets. These statistics are fully consistent with the results obtained for the training set and summarized in the upper part of the figure, as expected from the large numbers of compounds considered in both sets and the relatively small number of fitting parameters. In fact, the RMSE values tend to be somewhat lower for the test set. This is understandable since the 304 compounds in this test set were previously used to validate previously existing schemes. In many cases, special care was taken by the authors to avoid dubious experimental values and compounds for which the predictive methods used were deemed unreliable.

In short, Figure 9 demonstrates that when it comes to predicting formation enthalpies of general organic compounds, the AM1 and RM1 Hamiltonians lead to especially poor results compared with PM3, APC, and atom-equivalent schemes based on DFT. In term of reliability, AE2 ≫ APC > AE1 ≈ PM3. The superiority of AE2 is clearly to be expected in view of the relatively high level of description of the electronic structure, involving an extended basis set and an advanced (i.e., hybrid) functional combined with dispersion corrections.

It is interesting to observe that in spite of its extreme simplicity, APC proves to be more reliable than any NDDO Hamiltonian or even simple DFT procedures like AE1, which is based on a medium (i.e., double-ζ) basis set and a pure (i.e., nonhybrid) gradient-corrected density functional. Many such procedures have been put forward in the last two decades.[21,57,78−80]

The present NDDO results are consistent with the extensive calculations reported in the literature. For instance, the AAE values (in kJ/mol) obtained for a training set of 1480 species were reported to be 46.6, 33.4, and 24.1 for AM1, PM3, and RM1. The corresponding data derived from the present calculations are 45.8, 28.8, and 46.1, respectively, for the training set and 44.4, 25.3, and 31.1, respectively, for the test set. Interestingly, in contrast to the present results, this earlier study reported better results using RM1 instead of the earlier AM1 and PM3 methods. This is probably the case because they compared the methods on the basis of the RM1 training set. The present work shows that this good performance of RM1 significantly deteriorates as compounds outside its training set are considered. Therefore, PM3 clearly emerges as a better choice to predict the formation enthalpies of new compounds. In what follows, a more detailed comparison between APC and alternative methods is carried out on the basis of the individual test sets.

Hydrocarbon Compounds. The overall APC predictions for hydrocarbon compounds (i.e., for the three test sets HC-100, PH-103, and CUB-20) are sketched in Figure 10. The results

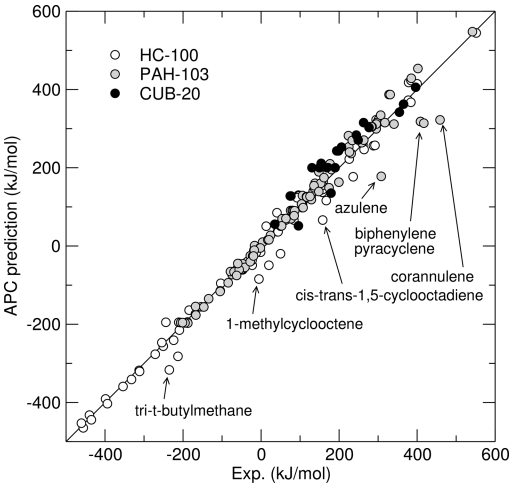


Figure 10. Formation enthalpies calculated for the HC-100, PAH-103, and CUB-20 data sets: APC vs reference values (experimental for HC-100 and PAH-103, G2MP2B3 for CUB-20).

reveal significant errors for structures with multiple bonds within unusual rings (1,5-cyclooctadiene, $\delta = -92$; 1-methylcyclooctene, $\delta = -79$) or crowded structures (tri-*tert*-butylmethane, $\delta = -81$). However, the most significant ones are systematically associated with compounds with partial aromatic or antiaromatic character, especially as conjugated structures involving five-membered aromatic rings are present (corannulene, $\delta = -136$; azulene, $\delta = -130$; pyracyclene, $\delta =$

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

−90). Such large underestimations are expected in view of the preliminary APC(HC) findings.

The predictive value of APC for hydrocarbon compounds is compared to that of previous models in Figure 11. Interestingly,

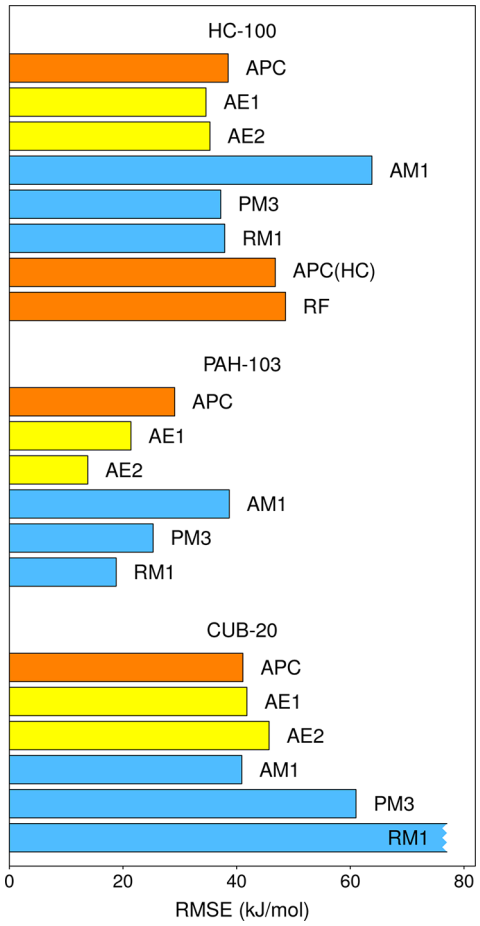


Figure 11. Performance of various procedures for hydrocarbons: HC-100, PAH-103, and CUB-20. The color code is the same as in Figure 9.

despite the fact that it must accommodate a much larger amount of $\Delta_f H°$ data for a wide variety of chemical structures, APC proves to be better than APC(HC). This result dismisses the idea that a more specialized model should necessarily perform better on its applicability domain than a more general one. This reflects the benefit of explicitly introducing enthalpy contributions for every interaction deemed to be relevant, especially in the present case geminal interactions.

Although the RF model performs about as well as APC(HC), its performance is poorer than that of APC, presumably because it was trained using only hydrocarbons. This result is a clear illustration of the potential interest of using large and general data sets, even if one is only interested in a specific subset of compounds. In the present case, the advantage of the larger data set stems from the fact it makes possible the introduction of physically relevant parameters (geminal interactions) that cannot be unambiguously derived using only the hydrocarbon data set.

Figure 11 indicates that APC performs slightly worse than AE1 or AE2 for general (HC-100) and polyaromatic (PAH-103) hydrocarbons. It should be noted, however, that although AE1 and AE2 naturally include strain energies and therefore do not yield severe underestimations of $\Delta_f H°$ similar to those obtained using APC due to a current lack of strain corrections, these two procedures severely overestimate $\Delta_f H°$ for large molecules. For instance, the AE2 estimate for benzerythrene ($C_{24}H_{18}$) is overestimated by 158 kJ/mol, in line with the well-documented failure of B3LYP for such large molecules.[81,82]

On the other hand, despite the fact PAHs are extended structures with the possibility of long-range interactions mediated by delocalized electrons, the accuracy of the APC results for PAH-103 is similar to the average, with a typical RMSE of about 30 kJ/mol, while HC-100 and CUB-20 exhibit larger errors, with RMSEs close to 40 kJ/mol. For HC-100, the most serious errors are associated with large (benzerythrene, 1-methylcyclooctene, etc.) or strained (tri-*tert*-butylmethane, *cis*,*trans*-1,5-cyclooctadiene) molecules that are absent from the other data sets, as is clear from Figure 10. On the other hand, HC-100 might exhibit larger experimental errors compared with PAH-103. Indeed, while the authors of ref 45 were especially careful to compile high-quality data, those of ref 4 were more concerned with the comparison of learning methodologies. Similarly, the larger errors observed for CUB-20 might stem from the uncertainties associated with the reference values for this data set, due to possible errors inherent in the G3MP2B3 procedure employed. These explanations are supported by the fact that the $\Delta_f H°$ values estimated using AE1 or AE2 are also in better agreement with experiment for PAH-103 than for HC-100 and CUB-20.

It is of course possible to get better $\Delta_f H°$ estimates using higher-level procedures. For instance, by the use of a combination of high-level DFT calculations (B3LYP/cc-pVDZ) and a group-based correction scheme, near chemical accuracy (RMSE = 6.4 kJ/mol) was obtained for an extended set of PAHs.[45] Because of the introduction of group corrections, that approach performs much better than all of the present methods. However, it is obviously much more costly than APC and exhibits a more limited scope.

The cubane derivatives in CUB-20 were anticipated to be challenging compounds for APC because of the strain energy associated with the cagelike structures. Since APC considers only the topology of the molecular graph and not the geometry, strain energy must be explicitly included through specific corrections. Therefore, this model might prove unreliable for this data set, as the present implementation accounts only for strains associated with atoms belonging simultaneously to three rings sharing a common size of four or six atoms (through the $\Delta_f H°(444)$ and $\Delta_f H°(666)$ parameters), whereas many compounds in CUB-20 contain atoms involved simultaneously in three rings of various sizes.

Indeed, APC performs especially poorly for this test set, with RMSE > 40 kJ/mol with respect to the G3MP2B3 reference values.[69] However, these compounds seem even more challenging for quantum-chemistry-based procedures. This is especially the case for AE2, although in principle it is the most reliable procedure in this comparison. PM3 and especially RM1 yield even worse estimates. Finally, APC yields the best predictions for this challenging test set among all of the presently considered procedures. The deviations between the APC and G3MP2B3 values may be attributed to the lack of corrections like $\Delta_f H°(445)$ and $\Delta_f H°(466)$, which should improve predictions for such cage structures. However, these results must be considered with caution in view of possible inaccuracies inherent in the G3MP2B3 recipe.

Organic Compounds and Refrigerants. APC predictions for general organic compounds (ORG-20 and ORG-10) and

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

refrigerants (FRIG-32) are shown in Figure 12. For these relatively simple compounds, the APC predictions are mostly in

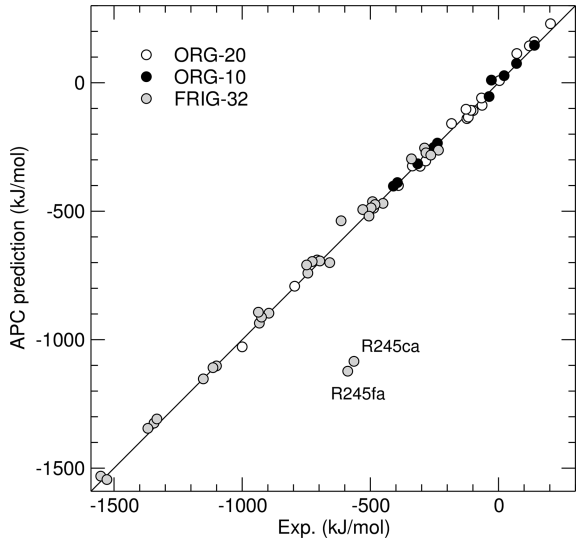


Figure 12. Formation enthalpies for the ORG-20, ORG-10, and FRIG-32 test sets (organic compounds and refrigerants): APC vs experimental values.

very good agreement with experiment, except for two pentafluoropropane refrigerants (R245ca and R245fa) as a result of probable experimental errors previously noted in the light of DFT calculations (B3LYP-cor procedure).[65] The anomalously large experimental values measured for these $C_3H_3F_5$ isomers are reminiscent of those reported for the two molecules shown in Figure 7. In fact, the APC value perfectly agrees with the DFT prediction reported in ref 65 for R245ca. However, although both methods predict a decrease in $\Delta_f H°$ in going from R245ca to the other isomer R245fa, this decrease is predicted to be more significant by DFT (−71 kJ/mol) compared with APC (−38 kJ/mol). Since the corresponding experimental values are probably wrong, R245ca and R245fa were discarded from all statistical analyses reported in this work.

The average performance of APC is compared to those of other methods in Figure 13. All in all, the APC predictions are somewhat better for general organic compounds (RMSE < 20 kJ/mol) than for refrigerants (RMSE = 27 kJ/mol). For organic compounds, APC performs especially well, i.e., approximately as well as all of the DFT-based atom-equivalent schemes considered (i.e., AE1/AE2 and the similar procedures B3LYP and BP86 considered in ref 71). In contrast to NDDO methods, APC yields consistently good predictions.

For ORG-20, a recent and heavily parametrized GC method yields predictions with chemical accuracy.[70] It is clear that this achievement is critically dependent on the specific fragment contributions introduced in the model and cannot be universally obtained for arbitrary organic compounds. Indeed, such a GC method exhibits the same fundamental limitations as APC (especially associated with the assignment of formal bond orders necessary to define the groups). Nevertheless, it is also clear that within its applicability domain, any GC method should outperform APC because the larger number of parameters enables better distinction among the various environments of a chemical bond. The issue with GC methods stems from the fact that they usually exhibit parameters derived

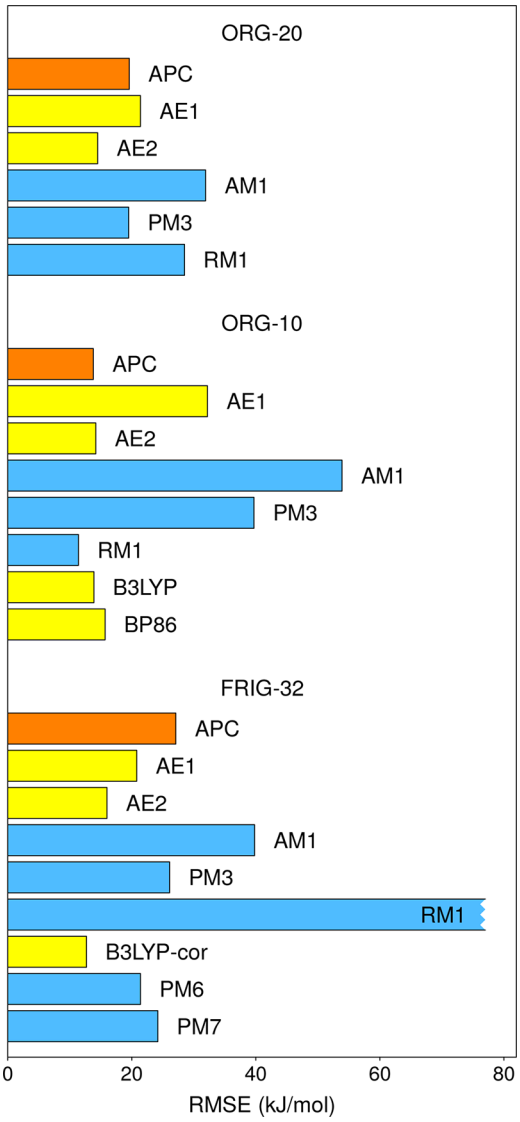


Figure 13. Performance of various procedures for organic compounds, including refrigerants. The color code is the same as in Figure 9.

from a very small number of compounds, whose transferability is therefore not firmly established.

As is clear from Figure 13, APC exhibits a loss of accuracy in going from organic compounds to refrigerants. This is understandable because the latter compounds are F- or Cl-rich molecules. The many halogen atoms in such compounds induce a rather unusual environment for the atoms, especially F atoms because of their very high electronegativity. The specificity of such an environment is naturally taken into account by quantum methods. Therefore, it comes as no surprise that the AE2 results are only slightly worse than for ORG-20/ORG-10. In fact, even some NDDO methods tested (PM3, PM6, and PM7) perform quite well for this test set, despite the poor performance of PM3 for ORG-10. In contrast, RM1 yields exceptionally large errors for this test set due to a systematic underestimation of $\Delta_f H°$ for high-fluorine compounds. Thus, it clearly appears that the accuracy to be expected from NDDO methods is critically dependent on the compounds studied.

The best predictions for refrigerants were obtained using DFT-based methods. The B3LYP-cor procedure used in ref 65

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

provides significantly improved results compared with the present AE1 and AE2 procedures. This illustrates the benefit of including vibrational contributions to $\Delta_f H°$ explicitly[59] and developing atom equivalents specific to given chemical families. For instance, the C equivalent in the refrigerant parametrization of B3LYP-cor is about 2.3 kJ/mol larger than the value previously optimized for energetic materials.[83] This contributes a variation of about 10 kJ/mol for a molecule with six C atoms, which partially explains why the RMSE decreases from 16 to 12.7 kJ/mol in going from AE2 to B3LYP-cor.

High-Energy Compounds. APC results for HEDM-45 are shown in Figure 14 and summarized in Figure 15. As outlined

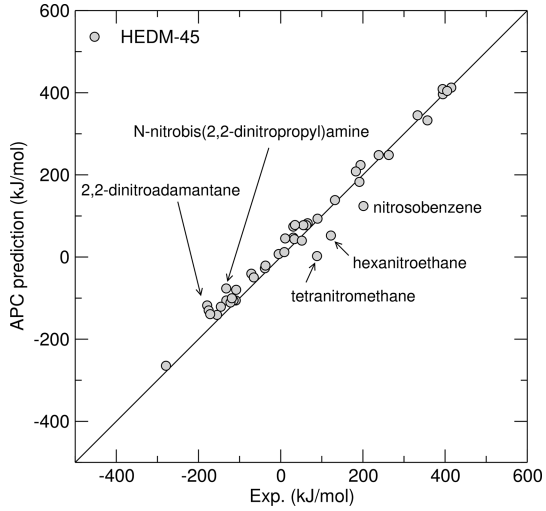


Figure 14. Formation enthalpies of the high-energy compounds in HEDM-45: APC predictions vs experiment.

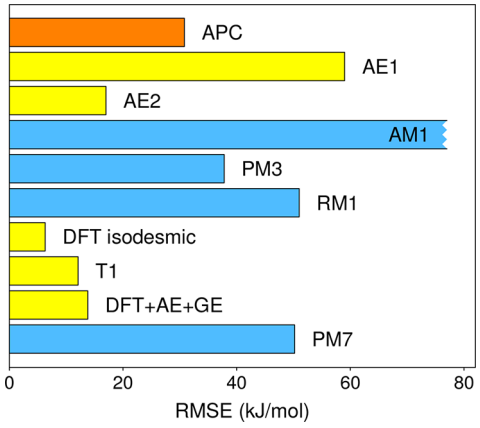


Figure 15. Performance of APC and quantum-chemistry-based procedures for the high-energy compounds in HEDM-45. The color code is the same as in Figure 9.

in Figure 14, the most significant APC errors are associated with underestimated predictions for nitrosobenzene, tetranitromethane, and hexanitroethane. For the two latter molecules, this error is to be expected in view of the extreme crowding due to the close proximity of the nitro groups. Therefore, it may be repaired through an ad hoc crowding correction. The overall APC results are surprisingly good considering the fact that high-energy compounds exhibit a much larger fraction of N/O atoms compared with the typical organic compounds used to train the model. This demonstrates a fairly good transferability of the APC parameters. In fact, with an RMSE of 30.8 kJ/mol, APC performs significantly better than the simple DFT-based scheme AE1 or recent NDDO schemes like RM1 or PM7 (Table S6). Although they were extensively parametrized and validated on large sets of general organic compounds, RM1 and PM7 lead to unusually large errors for such high-energy-density materials (HEDMs). For instance, the formation enthalpy of hexahydro-1,3,5-trinitroso-1,3,5-triazine is underestimated by as much as −246 and −226 kJ/mol at the RM1 and PM7 levels, respectively.

On the other hand, the calculations reported by Elioff et al. in ref 41 based on the recent M06-2X hybrid functional or the T1 thermochemical recipe provide significantly improved results with respect to APC, as expected from the fact they combine more rigorous theoretical bases with fined-tuned parameters. Of course, this comes at an incomparably higher computational cost. Elioff et al. obtained especially good results by combining M06-2X calculations with isodesmic reaction schemes ("DFT isodesmic" in Figure 15). These encouraging results warrant further investigations. Indeed, relatively few studies have considered this specific procedure, as the vast majority of DFT calculations using isodesmic reaction schemes are done using the B3LYP functional.

DISCUSSION

The present study provides a clear answer to the question raised in the Introduction. It is quite obvious that GC methods or procedures based on molecular mechanics should provide more accurate predictions than APC, with this better performance coming at the cost of more restricted applicability. However, this work shows that APC definitely outperforms current QSPR and NDDO approaches when it comes to predicting $\Delta_f H°$ data for general organic compounds. In fact, notwithstanding specific chemical families like PAHs and halogen-rich refrigerants, it appears to be even better than simple atom-equivalent schemes based on a pure (gradient-corrected) DFT functional. In other words, among generally applicable procedures, only methods based on hybrid DFT functionals or higher theoretical levels provide clearly superior average performance.

This conclusion was not necessarily anticipated, as quantum-chemical methods are more physically grounded, accounting for a much broader range of physical interactions contributing to molecular energetics (including their dependence on molecular conformation), and, for semiempirical methods, involve a larger number of adjustable parameters (about 200 for NDDO Hamiltonians vs 68 for the present APC implementation). However, it must be kept in mind that quantum procedures predict a wealth of molecular properties in addition to $\Delta_f H°$. Therefore, the superior performance of APC is understandable. In principle, a semiempirical quantum-chemical method developed with a stronger focus on $\Delta_f H°$ might outperform current popular NDDO procedures. Indeed, given a semi-empirical formalism, a special-purpose parametrization usually yields more accurate predictions than a general-purpose approach, as recently demonstrated in the context of semi-empirical Hamiltonians.[84,85]

Considering the relatively good performance of the APC and APC(HC) methods and the extreme simplicity of the working equations, it is reassuring that advanced QSPR techniques can predict $\Delta_f H°$ with a similar accuracy, as demonstrated by the RF model of Teixeira et al.[4] for hydrocarbons. However, for extended sets of general organic compounds, QSPR models

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

systematically fail to match the reliability of APC. For instance, recent investigations resorting to various regression methods (multilinear, support vector, and neural network) combined with advanced feature selection techniques did not lead to any model with RMSEs significantly less than 100 kJ/mol.[6,9]

These results should be interpreted with care, as these models were trained and validated using different data sets. However, they may illustrate an issue inherent in QSPR techniques. While the latter are invaluable to estimate complex properties for which rigorous theory-based approaches are unthinkable, especially in biology-related fields, their heavy reliance on statistical criteria to select a suitable model is problematic for simple properties like $\Delta_f H°$. Indeed, it is clear from eq 1 that many mathematical expressions are likely to approximate $\Delta_f H°$ fairly well. For instance, simply dropping a seldom-encountered geminal repulsion does not significantly affect the average performance of APC. Therefore, it is easy to realize that any QSPR procedure will have difficulties in identifying the optimal model among many acceptable candidate expressions. Accordingly, it might prove more fruitful to develop a QSPR to predict the APC residuals, which may be viewed as the hard part of $\Delta_f H°$.

In a production context, APC is clearly to be preferred over current QSPR and NDDO methods in view of its simplicity and superior performance. Nevertheless, QSPR and NDDO methods might be useful for comparison purposes or as components of a consensus model. Another asset of APC is the fact that its predictions do not depend on the input conformations. Procedures that require 3D geometries as input are likely to provide sporadically overestimated $\Delta_f H°$ values due to irrelevant conformations. Finally, a further advantage is the fact that probable outliers for which APC predictions are deemed unreliable naturally emerge from the present analysis to be compounds with ambiguous bond orders or strained molecules. By comparison, defining the applicability domain of a QSPR procedure is an important concern that remains an active research topic.[86]

With regard to future development, the present study demonstrates that improving the performance of currently available additivity schemes for $\Delta_f H°$ is not simply a matter of increasing the number of fragment parameters, as is done in modern GC methods. Indeed, the fact that they rely on a definition of the transferable fragments involving formal bond orders is a fundamental limitation of such approaches. It might be overcome through the development of a simple classical procedure to assign fractional bond orders, similar to charge equilibration schemes. This would enable the development of an APC procedure that takes advantage of explicitly bond-order-dependent covalent contributions to $\Delta_f H°$. Alternatively, developing a QSPR model for the APC residuals might appear as another attractive option.

CONCLUSION

This work demonstrates that the formation enthalpies of organic molecules in the gas phase can be predicted on the basis of the newly introduced APC additivity scheme with an accuracy better than that obtained using preexisting fast and general predictive methods, including popular semiempirical Hamiltonians or advanced QSPR procedures. This points to the fact that their reliance on formal bond orders is a major limitation of current additivity models for large conjugated systems with partial aromatic/antiaromatic or diradical character. On the other hand, the compilation of extended databases is a necessary prerequisite for the development of general and reliable additivity methods. In this respect, the continuous progress in high-level quantum-chemical methods should lead to the growth of large databases of accurate $\Delta_f H°$ data, thus opening new perspectives and promising a bright future for more heavily parametrized and ultrafast procedures to predict gas-phase formation enthalpies.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00613.

Worked-out examples of application of the model (PDF)
Tables S1−S6, including new atom equivalents derived in this work, experimental and calculated $\Delta_f H°$ data for all compounds studied, and corresponding statistics (XLSX)
Python script for easy application of the model (TXT)

AUTHOR INFORMATION

Corresponding Author

*E-mail: didier.mathieu@cea.fr. Phone: +33 (0)2 47344185. Fax: +33 (0)2 47345158.

ORCID

Didier Mathieu: 0000-0003-3832-2286

Notes

The author declares no competing financial interest.

REFERENCES

(1) *Computational Thermochemistry*; Irikura, K. K., Frurip, D. J., Eds.; ACS Symposium Series, Vol. *677*; American Chemical Society: Washington DC, 1998.

(2) Vatani, A.; Mehrpooya, M.; Gharagheizi, F. Prediction of Standard Enthalpy of Formation by a QSPR Model. *Int. J. Mol. Sci.* **2007**, *8*, 407−432.

(3) Zhang, Y. An Improved QSPR Study of Standard Formation Enthalpies of Acyclic Alkanes Based on Artificial Neural Networks and Genetic Algorithm. *Chemom. Intell. Lab. Syst.* **2009**, *98*, 162−172.

(4) Teixeira, A. L.; Leal, J. P.; Falcao, A. O. Random Forests for Feature Selection in QSPR Models - An Application for Predicting Standard Enthalpy of Formation of Hydrocarbons. *J. Cheminf.* **2013**, *5*, 9.

(5) Albahri, T. A.; Aljasmi, A. F. SGC Method for Predicting the Standard Enthalpy of Formation of Pure Compounds from their Molecular Structures. *Thermochim. Acta* **2013**, *568*, 46−60.

(6) Borhani, T. N.; Bagheri, M.; Manan, Z. A. Molecular Modeling of the Ideal Gas Enthalpy of Formation of Hydrocarbons. *Fluid Phase Equilib.* **2013**, *360*, 423−434.

(7) Cao, C.-T.; Yuan, H.; Cao, C. New concept of organic homo-rank compounds and its application in estimating enthalpy of formation of mono-substituted alkanes. *J. Phys. Org. Chem.* **2015**, *28*, 266−280.

(8) Toropova, A. P.; Toropov, A. A.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. CORAL: QSPRs of enthalpies of formation of organometallic compounds. *J. Math. Chem.* **2013**, *51*, 1684−1693.

(9) Bagheri, M.; Yerramsetty, K.; Gasem, K. A.; Neely, B. J. Molecular Modeling of the Standard State Heat of Formation. *Energy Convers. Manage.* **2013**, *65*, 587−596.

(10) Xiao, F.; Peng, G.; Nie, C.; Yu, L. Predicting Thermodynamic Properties of PBXTHs with New Quantum Topological Indexes. *PLoS One* **2016**, *11*, e0147126.

(11) Allinger, N. L.; Schmitz, L. R.; Motoc, I.; Bender, C.; Labanowski, J. K. Heats of Formation of Organic Molecules. 2. The Basis for Calculations Using Either Ab Initio or Molecular Mechanics Methods. Alcohols and Ethers. *J. Am. Chem. Soc.* **1992**, *114*, 2880−2883.

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

(12) Langley, C.-H.; Lii, J.-H.; Allinger, N. L. Molecular Mechanics Calculations on Carbonyl Compounds. IV. Heats of Formation. *J. Comput. Chem.* **2001**, *22*, 1476−1483.
(13) Lii, J.-H.; Liao, F.-X.; Hu, C.-H. Accurate Prediction of the Enthalpies of Formation for Xanthophylls. *J. Comput. Chem.* **2011**, *32*, 3175−3187.
(14) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RM1: a Reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27*, 1101−1111.
(15) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.
(16) Wu, Y.-Y.; Zhao, F.-Q.; Ju, X.-H. A Comparison of the Accuracy of Semi-Empirical PM3, PDDG and PM6Methods in Predicting Heats of Formation for Organic Compounds. *J. Mex. Chem. Soc.* **2014**, *58*, 223−229.
(17) Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. Comparison of SCC-DFTB and NDDO-Based Semiempirical Molecular Orbital Methods for Organic Molecules. *J. Phys. Chem. A* **2006**, *110*, 13551−13559.
(18) Dral, P. O.; Wu, X.; Spörkel, L.; Koslowski, A.; Weber, W.; Steiger, R.; Scholten, M.; Thiel, W. Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Theory, Implementation, and Parameters. *J. Chem. Theory Comput.* **2016**, *12*, 1082−1096.
(19) Byrd, E. F. C.; Rice, B. M. Improved Prediction of Heats of Formation of Energetic Materials Using Quantum Mechanical Calculations. *J. Phys. Chem. A* **2006**, *110*, 1005−1013.
(20) Byrd, E. F. C.; Rice, B. M. Improved Prediction of Heats of Formation of Energetic Materials Using Quantum Mechanical Calculations. *J. Phys. Chem. A* **2009**, *113*, 5813−5813.
(21) Rousseau, E.; Mathieu, D. Atom Equivalents for Converting DFT Energies Calculated on Molecular Mechanics Structures to Formation Enthalpies. *J. Comput. Chem.* **2000**, *21*, 367−379.
(22) Wu, J.; Xu, X. The X1Method for Accurate and Efficient Prediction of Heats of Formation. *J. Chem. Phys.* **2007**, *127*, 214105.
(23) Ohlinger, W. S.; Klunzinger, P. E.; Deppmeier, B. J.; Hehre, W. J. Efficient Calculation of Heats of Formation. *J. Phys. Chem. A* **2009**, *113*, 2165−2175.
(24) Cherkasov, A.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.
(25) Karton, A. A Computational Chemist's Guide to Accurate Thermochemistry for Organic Molecules. *WIRES Comput. Mol. Sci.* **2016**, *6*, 292−310.
(26) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212−225.
(27) Dinca, N.; Dragan, S.; Dinca, M.; Sisu, E.; Covaci, A. New Quantitative Structure-Fragmentation Relationship Strategy for Chemical Structure Identification Using the Calculated Enthalpy of Formation as a Descriptor for the Fragments Produced in Electron Ionization Mass Spectrometry: A Case Study with Tetrachlorinated Biphenyls. *Anal. Chem.* **2014**, *86*, 4949−4955.
(28) Barbiric, D.; Tribe, L.; Soriano, R. Computational Chemistry Laboratory: Calculating the Energy Content of Food Applied to a Real-Life Problem. *J. Chem. Educ.* **2015**, *92*, 881−885.
(29) Gupta, S.; Basant, N.; Singh, K. P. Three-Tier Strategy for Screening High-Energy Molecules Using Structure-Property Relationship Modeling Approaches. *Ind. Eng. Chem. Res.* **2016**, *55*, 820−831.
(30) Halls, M. D.; Tasaki, K. High-Throughput Quantum Chemistry and Virtual Screening for Lithium Ion Battery Electrolyte Additives. *J. Power Sources* **2010**, *195*, 1472−1478.
(31) Laidler, K. J. A System of Molecular Thermochemistry for Organic Gases and Liquids. *Can. J. Chem.* **1956**, *34*, 626−648.
(32) Reid, R. C.; Prausnitz, J. M.; Sherwood, T. K. *The Properties of Gases and Liquids*; McGraw-Hill: New York, 1976.
(33) Türker, L.; Bayar, C. C. A Computational View of PATO and its Tautomers. *Z. Anorg. Allg. Chem.* **2012**, *638*, 1316−1322.
(34) Zauer, E. A. Enthalpies of Formation of Polycyclic Aromatic Hydrocarbons. *Russ. J. Gen. Chem.* **2012**, *82*, 1135−1144.
(35) Zauer, E. A. Enthalpy of Formation of Five-Membered Nitrogen-Containing Aromatic Heterocycles. *Russ. J. Gen. Chem.* **2015**, *85*, 2268−2276.
(36) Korth, M. Large-Scale Virtual High-Throughput Screening for the Identification of New Battery Electrolyte Solvents: Evaluation of Electronic Structure Theory Methods. *Phys. Chem. Chem. Phys.* **2014**, *16*, 7919−7926.
(37) Sahin, S.; Bleda, E. A.; Altun, Z.; Trindle, C. Computational Characterization of Isomeric C_4H_2O Systems: Thermochemistry, Vibrational Frequencies, and Optical Spectra for Butatrienone, Ethynyl Ketene, Butadiynol, and Triafulvenone. *Int. J. Quantum Chem.* **2016**, *116*, 444−451.
(38) Rajagopal, K.; Ahón, V. R.; Moreno, E. Estimating Thermochemical Properties of Hydroprocessing Reactions by Molecular Simulation and Group Contribution Methods. *Catal. Today* **2005**, *109*, 195−204.
(39) Ju, X.-H.; Li, Y.-M.; Xiao, H.-M. Theoretical Studies on the Heats of Formation and the Interactions among the Difluoroamino Groups in Polydifluoroaminocubanes. *J. Phys. Chem. A* **2005**, *109*, 934−938.
(40) Whiteside, T. S.; Priest, M. A.; Padgett, C. W. Enthalpies of Formation of Methyl Substituted Naphthalenes. *Thermochim. Acta* **2010**, *510*, 17−23.
(41) Elioff, M. S.; Hoy, J.; Bumpus, J. A. Calculating Heat of Formation Values of Energetic Compounds: A Comparative Study. *Adv. Phys. Chem.* **2016**, *2016*, 5082084.
(42) Mathieu, D. Formation Enthalpies Derived from Pairwise Interactions: A Step toward More Transferable Reactive Potentials for Organic Compounds. *J. Chem. Theory Comput.* **2012**, *8*, 1295−1303.
(43) Grela, M. A.; Colussi, A. J. Quantitative Structure-Stability Relationships for Oxides and Peroxides of Potential Atmospheric Significance. *J. Phys. Chem.* **1996**, *100*, 10150−10158.
(44) Benson, S. W. *Thermochemical Kinetics*, 2nd ed.; Wiley: New York, 1976.
(45) Allison, T. C.; Burgess, D. R. First-Principles Prediction of Enthalpies of Formation for Polycyclic Aromatic Hydrocarbons and Derivatives. *J. Phys. Chem. A* **2015**, *119*, 11329−11365.
(46) Daylight Theory Manual. http://www.daylight.com/dayhtml/doc/theory/ (accessed August 2017).
(47) Landrum, G. RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org.
(48) *Encyclopedia of Spectroscopy and Spectrometry*, 3rd ed.; Lindon, J. C., Tranter, G. E., Koppenaal, D. W., Eds.; Elsevier: Amsterdam, 2017.
(49) Karthikeyan, M.; Vyas, R. *Practical Chemoinformatics*; Springer India: New Delhi, 2014.
(50) *CRC Handbook of Chemistry and Physics*, 95th ed.; Haynes, W. M., Ed.; CRC Press: Boca Raton, FL, 2014.
(51) Granovsky, A. A. Firefly version 8.0. http://classic.chem.msu.su/gran/gamess/index.html.
(52) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14*, 1347−1363.
(53) Neese, F. The ORCA Program System. *WIREs: Comput. Mol. Sci.* **2012**, *2*, 73−78.
(54) Lagorce, D.; Villoutreix, B.; Miteva, M. A. Three-Dimensional Structure Generators of Drug-Like Compounds: DG-AMMOS, an Open-Source Package. *Expert Opin. Drug Discovery* **2011**, *6*, 339−351.
(55) Dewar, M. J. S.; Storch, D. M. Comparative Tests of Theoretical Procedures for Studying Chemical Reactions. *J. Am. Chem. Soc.* **1985**, *107*, 3898−3902.
(56) Ibrahim, M. R.; Schleyer, P. v. R. Atom Equivalents for Relating Ab Initio Energies to Enthalpies of Formation. *J. Comput. Chem.* **1985**, *6*, 157−167.
(57) Mathieu, D.; Pipeau, Y. Formation Enthalpies of Ions: Routine Prediction Using Atom Equivalents. *J. Chem. Theory Comput.* **2010**, *6*, 2126−2139.

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26

(58) Mathieu, D. Sensitivity of Energetic Materials: Theoretical Relationships to Detonation Performance and Molecular Structure. *Ind. Eng. Chem. Res.* **2017**, *56*, 8191−8201.
(59) Paulechka, E.; Kazakov, A. Efficient DLPNO-CCSD(T)-Based Estimation of Formation Enthalpies for C-, H-, O-, and N-Containing Closed-Shell Compounds Validated Against Critically Evaluated Experimental Data. *J. Phys. Chem. A* **2017**, *121*, 4379−4387.
(60) Tetko, I. V.; Sushko, Y.; Novotarskyi, S.; Patiny, L.; Kondratov, I.; Petrenko, A. E.; Charochkina, L.; Asiri, A. M. How Accurately Can We Predict the Melting Points of Drug-like Compounds? *J. Chem. Inf. Model.* **2014**, *54*, 3320−3329.
(61) Mathieu, D.; Bouteloup, R. Reliable and Versatile Model for the Density of Liquids Based on Additive Volume Increments. *Ind. Eng. Chem. Res.* **2016**, *55*, 12970−12980.
(62) Transtrum, M. K.; Machta, B. B.; Brown, K. S.; Daniels, B. C.; Myers, C. R.; Sethna, J. P. Perspective: Sloppiness and Emergent Theories in Physics, Biology, and Beyond. *J. Chem. Phys.* **2015**, *143*, 010901.
(63) *NIST Chemistry WebBook*; Linstrom, P. J., Mallard, W. G., Eds.; NIST Standard Reference Database Number 69; National Institute of Standards and Technology: Gaithersburg, MD, 2011; http://webbook.nist.gov.
(64) Verevkin, S. P.; Emel'yanenko, V. N.; Diky, V.; Muzny, C. D.; Chirico, R. D.; Frenkel, M. New Group-Contribution Approach to Thermochemical Properties of Organic Compounds: Hydrocarbons and Oxygen-Containing Compounds. *J. Phys. Chem. Ref. Data* **2013**, *42*, 033102.
(65) Demenay, A.; Glorian, J.; Paricaud, P.; Catoire, L. Predictions of the Ideal Gas Properties of Refrigerant Molecules. *Int. J. Refrig.* **2017**, *79*, 207−216.
(66) Suntsova, M. A.; Dorofeeva, O. V. Use of G4 Theory for the Assessment of Inaccuracies in Experimental Enthalpies of Formation of Aromatic Nitro Compounds. *J. Chem. Eng. Data* **2016**, *61*, 313−329.
(67) Dorofeeva, O. V.; Ryzhova, O. N.; Suntsova, M. A. Accurate Prediction of Enthalpies of Formation of Organic Azides by Combining G4 Theory Calculations with an Isodesmic Reaction Scheme. *J. Phys. Chem. A* **2013**, *117*, 6835−6845.
(68) Suntsova, M. A.; Dorofeeva, O. V. Use of G4 Theory for the Assessment of Inaccuracies in Experimental Enthalpies of Formation of Aliphatic Nitro Compounds and Nitramines. *J. Chem. Eng. Data* **2014**, *59*, 2813−2826.
(69) Novak, I. Ab initio vs Molecular mechanics Thermochemistry: Homocubanes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 903−906.
(70) Hukkerikar, A. S.; Meier, R. J.; Sin, G.; Gani, R. A Method to Estimate the Enthalpy of Formation of Organic Compounds with Chemical Accuracy. *Fluid Phase Equilib.* **2013**, *348*, 23−32.
(71) Li, J.; Kim, C. K. Comparative Study on the Gas Phase Heats of Formation. *Bull. Korean Chem. Soc.* **2015**, *36*, 1536−1538.
(72) Wodrich, M. D.; Corminboeuf, C.; Schreiner, P. R.; Fokin, A. A.; Schleyer, P. v. R. How Accurate Are DFT Treatments of Organic Energies? *Org. Lett.* **2007**, *9*, 1851−1854.
(73) Diogo, H. P.; Kiyobayashi, T.; Minas da Piedade, M. E.; Burlak, N.; Rogers, D. W.; McMasters, D.; Persy, G.; Wirz, J.; Liebman, J. F. The Aromaticity of Pyracylene: An Experimental and Computational Study of the Energetics of the Hydrogenation of Acenaphthylene and Pyracylene. *J. Am. Chem. Soc.* **2002**, *124*, 2065−2072.
(74) Wiberg, K. B.; Nakaji, D. Y.; Morgan, K. M. Heat of Hydrogenation of a Cis Imine. An Experimental and Theoretical Study. *J. Am. Chem. Soc.* **1993**, *115*, 3527−3532.
(75) Chen, F.-F.; Wang, F. Electronic Structure of the Azide Group in 3′-Azido-3′-deoxythymidine (AZT) Compared to Small Azide Compounds. *Molecules* **2009**, *14*, 2656−2668.
(76) Lo, P.-K.; Lau, K.-C. High-Level ab Initio Predictions for the Ionization Energies and Heats of Formation of Five-Membered-Ring Molecules: Thiophene, Furan, Pyrrole, 1,3-Cyclopentadiene, and Borole, $C_4H_4X/C_4H_4X^+$ (X = S, O, NH, CH_2, and BH). *J. Phys. Chem. A* **2011**, *115*, 932−939.
(77) Rogers, F. E.; Rapiejko, R. J. Thermochemistry of Carbonyl Addition Reactions. I. Addition of Water and Methanol to Hexafluoroacetone. *J. Am. Chem. Soc.* **1971**, *93*, 4596−4597.
(78) Habibollahzadeh, D.; Grice, M. E.; Concha, M. C.; Murray, J. S.; Politzer, P. Nonlocal Density Functional Calculation of Gas Phase Heats of Formation. *J. Comput. Chem.* **1995**, *16*, 654−658.
(79) Mole, S. J.; Zhou, X.; Liu, R. Density Functional Theory (DFT) Study of Enthalpy of Formation. 1. Consistency of DFT Energies and Atom Equivalents for Converting DFT Energies into Enthalpies of Formation. *J. Phys. Chem.* **1996**, *100*, 14665−14671.
(80) Delley, B. Ground-State Enthalpies: Evaluation of Electronic Structure Approaches with Emphasis on the Density Functional Method. *J. Phys. Chem. A* **2006**, *110*, 13632−13639.
(81) Redfern, P. C.; Zapol, P.; Curtiss, L. A.; Raghavachari, K. Assessment of Gaussian-3 and Density Functional Theories for Enthalpies of Formation of C_1-C_{16} Alkanes. *J. Phys. Chem. A* **2000**, *104*, 5850−5854.
(82) Lu, L.; Hu, H.; Hou, H.; Wang, B. An Improved B3LYP Method in the Calculation of Organic Thermochemistry and Reactivity. *Comput. Theor. Chem.* **2013**, *1015*, 64−71.
(83) Osmont, A.; Catoire, L.; Gökalp, I.; Yang, V. Ab Initio Quantum Chemical Predictions of Enthalpies of Formation, Heat Capacities, and Entropies of Gas-Phase Energetic Compounds. *Combust. Flame* **2007**, *151*, 262−273.
(84) Grimme, S.; Bannwarth, C. Ultra-Fast Computation of Electronic Spectra for Large Systems by Tight-Binding Based Simplified Tamm-Dancoff Approximation (sTDA-xTB). *J. Chem. Phys.* **2016**, *145*, 054103.
(85) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1−86). *J. Chem. Theory Comput.* **2017**, *13*, 1989−2009.
(86) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Definition. *SAR QSAR Environ. Res.* **2016**, *27*, 865−881.

DOI: 10.1021/acs.jcim.7b00613
J. Chem. Inf. Model. 2018, 58, 12−26