Estimation of Molecular Properties by High-Dimensional Model Representation

Michael Y. Hayes, Baiqing Li, and Herschel Rabitz*

Department of Chemistry, Princeton University, 207 Frick Laboratory, Princeton, New Jersey 08544-1009

Received: June 14, 2005; In Final Form: September 7, 2005

Additivity models have been widely employed to approximate unknown molecular properties based on previously measured or calculated data for similar molecules. This paper proposes an improved formulation of additivity, which is based on high-dimensional model representation (HDMR). HDMR is a general function-mapping technique that expresses the output of a multivariate system in terms of a hierarchy of cooperative effects among its input variables. HDMR rests on the general observation that, for many physical systems, only relatively low-order input variable cooperativity is significant. A molecule is expressed as a multivariate system by defining binary-valued input variables corresponding to the presence or absence of a chemical bond, with the molecular property as the output. Conventional additivity decomposes a molecular property into contributions from nonoverlapping subcomponents of fixed size. On the other hand, HDMR decomposes a molecular property into the exact contributions from the full hierarchy of its variable-sized subcomponents and contains additivity as a special case. The complete hierarchical structure of HDMR can in many cases lead to a much more accurate estimate than conventional additivity. Also, when full group additivity is not possible, HDMR gives an expression for a lower-order approximation for the missing group additivity value, greatly expanding the scope of HDMR compared to additivity. The component terms in an HDMR approximation have well-defined physical significance. Moreover, HDMR gives an exact expression for the truncation error in any given HDMR approximation, also with a well-defined physical significance. The HDMR model is tested for the enthalpy of formation of a broad range of organic molecules, and its advantages over additivity are illustrated.

1. Introduction

The conceptual reduction of a molecule to a sum of its components is central to much of chemistry. At the simplest level, a chemical formula reduces a molecule to an algebraic sum of its atoms, and a two-dimensional structural representation reduces a molecule to a geometric sum of its atoms and bonds. Extending this idea, one can numerically express a molecular property as a sum of contributions from molecular subcomponents. Many empirical models are built upon this concept and are known as additivity models.[1] These models approximate the property of a given molecule by taking combinations of properties of smaller molecules which represent fixed-size, nonoverlapping substructural fragments of the molecule under consideration. The general assumption of additivity is that the property being modeled is dominated by local interactions among the molecular subcomponents. Many properties have been modeled by additivity, including enthalpy, entropy, heat capacity,[2] boiling point,[3] liquid state thermal conductivity,[4] and critical temperature and pressure,[5] among others.

The utility of additivity models lies in their simplicity compared to performing a new experimental measurement or theoretical calculation. Some applications require knowledge of a vast amount of molecular property data and necessitate a general interpolation/extrapolation method that is extremely fast. One example of such an application is an optimization problem in chemical engineering, where a molecule with a predetermined molecular property is desired. In other applications, the complexity of calculating a molecular property scales with the size of the molecule being studied. In ab initio quantum chemistry, for example, detailed electron correlation methods become prohibitive for large molecules. In this case it is preferable to compute full ab initio calculations on smaller fragments and to approximate the molecular property as some function of the properties of the smaller fragments. An elegant and systematic approach to this type of quantum chemistry was recently proposed by Deev and Collins.[6]

Additivity models have been formulated in various ways, but they all share a common theoretical framework. Cox and Pilcher[1] show that the major models of Benson, Laidler, and Allen are mathematically equivalent. The theory of additivity is well formalized by Benson and co-workers,[2,7,8] who represent the additivity approximation in terms of the disproportionation reaction:

$$\mathrm{RNR} + \mathrm{SNS} \rightarrow 2\mathrm{RNS} \qquad (1)$$

Here R and S are atoms and N is a variable number of atoms. The assumption of additivity is that the property change of the disproportionation reaction is negligible for sufficiently large N, meaning that atoms R and S separated by N atoms have no significant effect on one another. If N vanishes, eq 1 implies the additivity of atomic properties; if N is one atom, the additivity of diatomics is implied, etc.[2] The most common order of approximation is "group additivity" (GA), where N is two atoms. For example, using GA, a property Φ of butane may be written as a sum of contributions from its constituent "groups", which are two end groups $\mathrm{C{-}(C)(H)_3}$ (i.e., a carbon bonded to one other carbon and three hydrogens), and two middle groups $\mathrm{C{-}(C)_2(H)_2}$ (i.e., a carbon bonded to two other carbons and two hydrogens). The GA expression of the property is then

10.1021/jp053197w CCC: $33.50

$$2[\mathrm{C-(C)(H)_3}] + 2[\mathrm{C-(C)_2(H)_2}] = \Phi_{\mathrm{H_3CCH_2CH_2CH_3}} \quad (2)$$

By setting up similar equations for similar molecules of known Φ, one can construct a solvable set of linear equations to calculate the group contributions $\mathrm{C-(C)(H)_3}$ and $\mathrm{C-(C)_2(H)_2}$. These calculated group contributions can then be used to predict the properties of larger molecules made up of these groups. Oftentimes the problem of solving for a set of group contributions is overdetermined, in which case a regression analysis may be invoked to incorporate all available input data.[7]

It is important to note that group additivity is used in the literature to loosely refer to a more elaborate system of approximations that include many extensions to the formal theory described above.[7] In particular, R and S in eq 1 do not refer to generic atoms, but rather to atoms in a particular environment (some examples include primary, secondary, and tertiary carbons, carbons in a benzene ring, doubly or triply bonded carbons, etc.). Furthermore, there are a large number of "correction terms" designed to account for such effects as tertiary, gauche, and ring interactions. For the remainder of this paper, "group additivity" will refer to the formal theory implied by the disproportionation reaction, unless indicated otherwise.

Although the additivity approach has proven its utility in many applications, it still suffers from some conceptual and practical limitations. Considering the diversity of factors that can contribute to a molecular property, which include conjugation, steric crowding, rings, hydrogen bonding, etc., it is apparent that the assumption of group additivity is only applicable for a special subset of molecules. In fact, there exists a well-known hierarchy of intramolecular phenomena that may contribute to a molecular property, in which the local interactions among "groups" are only among the lowest order effects. Consequently, it is natural to use an additive type approximation that accounts for the full hierarchical progression of intramolecular phenomena. The importance of these systematic higher-order effects is evidenced by the proliferation of "correction terms" in the GA literature,[7] which are not accounted for by the additive formulation. Also, the domain of GA is dictated by the availability of the specific input data needed to derive group values, and for many molecules, group values are simply not available. In these cases, it is desirable to perform a lower-order approximation only where necessary, and to retain a GA level or higher approximation for the remainder of the molecule. This procedure is also outside of the scope of the additive formulation. Finally, the component additive terms generally do not have an easily understandable physical significance, and there is no well-defined expression for the error of the additive approximation.

This paper proposes an improved theoretical framework for the estimation of molecular properties based on high-dimensional model representation (HDMR)[9,10] that contains additivity as a special case, and which addresses the aforementioned limitations of simple additivity. HDMR is a general function mapping technique that expresses a multivariate system in terms of the hierarchical cooperative effects among its variables, starting with each variable acting independently, followed by variables acting in pairs, groups of three, etc. Each level of cooperativity yields exact effects to its particular order. The HDMR input → output function maps have been used to accurately model a wide variety of physical phenomena, including chemical kinetic systems,[11] quantum dynamic phenomena,[12] and semiconductor material properties.[13] By defining a molecule as a multivariate system in which its bonds are the input variables and its property is the output, the task of molecular property modeling becomes a high-dimensional problem that may be well-described by HDMR. HDMR accounts for the full hierarchy of subcomponent contributions to the molecular property and has a clear physical interpretation. For example, the effect of an *n*-membered ring is precisely defined as an *n*th-order HDMR effect. The formulation of HDMR naturally allows for a mixed order approximation, so structures that are expected to require a higher order treatment may be easily incorporated into an otherwise low-order model. Furthermore, the error in any HDMR approximation has a well-defined expression, and a clear physical meaning.

We propose two types of HDMR for the estimation of molecular properties. Random-sampling HDMR (RS-HDMR) derives a multivariate function from a random selection of experimental data.[14] Cut-HDMR expresses a function with respect to a particular reference point in the input variable space, as a collection of lines, planes, subvolumes, etc. that cut through the reference point. The choice between RS-HDMR and cut-HDMR will depend in part on the data available for the property being modeled. For the property under consideration in this paper, enthalpy of formation, the experimental data for the small molecules is thorough, so the focus will be on cut-HDMR. An advantage of cut-HDMR is that the physical significance of the model is precisely defined in terms of specific chemical components contributing to the property value.

The technique of expressing a molecular property as a hierarchical sum of subcomponent contributions appears at least as early as 1947 when Platt[15] described a model of alkane molar refractivity, molar volume, heat of formation, and boiling point. In the Platt model, the properties are decomposed into contributions from bonds, and the bond contribution terms are modified by adjacent atoms, atoms once removed, atoms twice removed, etc. The procedure of decomposing a group contribution into lower-order terms was employed by Pedley, Naylor, and Kirby[16] to allow for a mixed-order prediction. HDMR offers a fully unified framework for these methods.

This paper is structured as follows. The general foundations of HDMR for molecular properties are introduced in section 2. In section 3 we present the results of a comprehensive test of the HDMR for enthalpy of formation. HDMR is then compared with additivity in section 4. We first demonstrate that HDMR contains additivity as a special case and proceed to illustrate the advantages of HDMR over additivity.

2. High-Dimensional Model Representation

2.1. Formulation. HDMR has a generic structure that breaks down a multivariate system in terms of the cooperative effects among its variables.[9,10] First-order HDMR maps the effect of each variable acting independently, second-order HDMR maps the cooperative effect of pairs of variables, etc. For a full mapping of a system to all orders, the complexity of sampling the multivariate space would scale exponentially with the number of variables. However, for well-defined physical systems, typically only low-order correlations among the input variables are significant with respect to the output. Therefore, the input → output behavior of a multivariate system often can be accurately expressed as a low-order truncated HDMR expansion, with sampling complexity scaling only polynomially with the number of variables.

For a system of n input variables x_1, x_2, ..., x_n and output g, a complete HDMR expansion maps the input → output behavior of the system described by $g(x_1,x_2,...,x_n)$ as a hierarchical sum of terms from zeroth to nth order in the following form:

$$g(x_1,x_2,...,x_n) \equiv f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{i<j}^{n} f_{ij}(x_i,x_j) + ... + f_{12....n}(x_1,x_2,...,x_n) \quad (3)$$

In this expression f_0 is a constant, the term $f_i(x_i)$ describes the effect of variable x_i acting independently, the term $f_{ij}(x_i, x_j)$ describes the effect of variables x_i and x_j acting cooperatively, etc. Cut-HDMR defines f_0 as the system output at a reference point $\bar{\mathbf{x}}$ in the input space. The choice of reference point is arbitrary, but in many cases a logical choice can be made on physical grounds. When taken to convergence, the cut-HDMR formulation is independent of the choice of $\bar{\mathbf{x}}$. Each term in the HDMR expansion represents the unique contribution from its particular order of variable cooperativity, and the first few functions have the following form:

$$\begin{aligned} f_0 &= g(\bar{\mathbf{x}}) \\ f_i(x_i) &= g(\bar{\mathbf{x}}^i, x_i) - f_0 \\ f_{ij}(x_i,x_j) &= g(\bar{\mathbf{x}}^{ij}, x_i, x_j) - f_i(x_i) - f_j(x_j) - f_0 \\ &\vdots \end{aligned} \quad (4)$$

The notation $\bar{\mathbf{x}}^i$ means that all variables are set to their reference values except for x_i, $\bar{\mathbf{x}}^{ij}$ means that all variables are set to their reference values except for x_i and x_j, etc.

Because only low-order effects (i.e., up to order l where $l \ll n$) are expected to significantly influence the output, knowledge of a small set of low-order contributions is often sufficient to accurately represent a function in high dimensions with many variables. By observing the behavior of the various low-order functions, one can identify regions of sensitivity in the input space and reduce the HDMR expansion to include only terms that are significant with respect to the output. Moreover, the behavior of the HDMR function terms often leads to insight into the nature of the model system.[10] The validity of the HDMR hypothesis that many physical systems are dominated by low-order effects is supported by successful applications to a diverse set of problems.[9−14,17,18]

2.2. HDMR of Molecular Properties. The hierarchical structure of HDMR suggests its suitability for the modeling of molecular properties, considering a molecule as naturally broken into its subcomponent fragments. With HDMR one is able to mathematically express this notion through the construction of a hierarchical map of a molecular property in terms of contributions from its atoms, diatomic subcomponents, triatomics, etc. up to the fully constructed molecule. Consistent with the HDMR hypothesis, it will be shown that the enthalpy of formation is largely determined by the low-order effects of a small number of subcomponents. That is, one can accurately represent the enthalpy of formation of a large set of molecules in terms of a small subset of molecular subcomponents.

The input variables are defined as a chosen set of bonds in a molecule, and the output as the molecular property. The variables can take on one of two discrete values, either "bonded" ($x = 1$) or "nonbonded" ($x = 0$). When a variable is "bonded", the corresponding bond is intact, and when a variable is "nonbonded", the bond is dissociated. Because this dissociation would generally result in two unstable species, we define the nonbonded species to be capped with a chosen saturating agent to fill the vacated valencies, and those added saturating agent atoms are subtracted from the overall sum in such a way as to balance the stoichiometry. To illustrate the meaning of the variables, consider the variable x associated with the C−O bond of methanol, with hydrogen as the saturating agent. This system can take on one of the following two states:

$$\begin{aligned} g(x{=}1) &= \Phi_{CH_3OH} \\ g(x{=}0) &= \Phi_{CH_4} + \Phi_{H_2O} - \Phi_{H_2} \end{aligned} \quad (5)$$

where Φ indicates the known property values of the molecules in the subscripts. In the $g(x{=}0)$ state, hydrogen atoms are added to the dissociated fragments $-CH_3$ and $-OH$ to form CH_4 and H_2O, and the hydrogen molecule is subtracted to remain stoichiometrically consistent. Because the variables take on only the discrete binary values of 0 and 1, the role of HDMR in describing molecular properties is to decompose the molecular property into its subcomponent contributions in eq 3. This treatment of bonds as either present or absent amounts to expressing a molecular structure in terms of its connectivity. By characterizing a molecule in this way, the nonbonded interaction of atoms through space, such as in sterically crowded clusters, is only captured at high orders. However, a modified definition of the variables could directly include such nonbonded effects, although this treatment is not exploited in the present work. Another consequence of a connectivity representation is the lack of cis−trans and optical isomer specificity. It is possible to apply an extra variable index to distinguish these isomers, but for some properties this specificity may be unimportant. For example, in the case of enthalpy of formation, optical isomerization is generally not specified in reporting experimental enthalpy values. The bonds defined as variables, as well as the saturating agent, may be flexibly chosen. However, in general, it is natural to define all bonds between non-hydrogen atoms as variables, and the saturating agent as hydrogen.

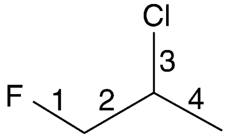

Figure 1. 1-Fluoro-2-chloropropane with labeled variable indices.

The system as a whole is the entire molecule fragmented to the extent prescribed by the variable values, and the output is the property value of the associated collection of fragments. For example, if all variables are equal to 1, the molecule is fully assembled. If all but one variable are equal to 1, the molecule is broken into two fragments at the bond i for which $x_i = 0$. If all variables are equal to 0 the molecule is fully fragmented. The reference term f_0 is naturally defined as the molecule in this fully fragmented state (i.e., $x_i = 0$ for all i). In this way the HDMR expansion progresses from the fragmented to the fully assembled representation of the molecule, which generally leads to a progressively more accurate representation of the molecular property. The first-order terms will include all groups bonded in pairs, the second-order terms will include all sets of three groups bonded, and so on to the fully assembled molecule. Only those bonds that are present in the fully constructed molecule will be formed to generate the fragments. This hierarchy will generally not need to be taken to the level of $x_i = 1$ for all i to accurately represent the property of the original molecule, as high-order cooperative terms are expected to be negligible.

To illustrate the application of HDMR to molecular properties, consider the molecule 1-fluoro-2-chloropropane shown in Figure 1. All bonds between non-hydrogen atoms are defined as variables, and the saturating agent is hydrogen. The first few terms of the HDMR expansion for this molecule are as follows:

$$
\begin{aligned}
f_0 &= \Phi_{\mathrm{HF}} + \Phi_{\mathrm{HCl}} + 3\Phi_{\mathrm{CH_4}} - 4\Phi_{\mathrm{H_2}} \\
f_1 &= \Phi_{\mathrm{FCH_3}} + \Phi_{\mathrm{HCl}} + 2\Phi_{\mathrm{CH_4}} - 3\Phi_{\mathrm{H_2}} - f_0 \\
&= \Phi_{\mathrm{FCH_3}} - \Phi_{\mathrm{HF}} - \Phi_{\mathrm{CH_4}} + \Phi_{\mathrm{H_2}} \\
f_2 &= \Phi_{\mathrm{CH_3CH_3}} + \Phi_{\mathrm{HF}} + \Phi_{\mathrm{HCl}} + \Phi_{\mathrm{CH_4}} - 3\Phi_{\mathrm{H_2}} - f_0 \\
&= \Phi_{\mathrm{CH_3CH_3}} - 2\Phi_{\mathrm{CH_4}} + \Phi_{\mathrm{H_2}} \\
&\vdots \\
f_{12} &= \Phi_{\mathrm{FCH_2CH_3}} + \Phi_{\mathrm{HCl}} + \Phi_{\mathrm{CH_4}} - 2\Phi_{\mathrm{H_2}} - f_1 - f_2 - f_0 \\
&= \Phi_{\mathrm{FCH_2CH_3}} - \Phi_{\mathrm{FCH_3}} - \Phi_{\mathrm{CH_3CH_3}} + \Phi_{\mathrm{CH_4}} \\
&\vdots
\end{aligned}
\tag{6}
$$

The first-order term f_1 represents the unique effect of the fragment FCH_3, formed by bond 1, and the term f_2 represents the unique effect of the fragment CH_3CH_3, formed by bond 2. The second-order term f_{12} represents the unique effect of the fragment FCH_2CH_3, formed by bonds 1 and 2 acting together. In the HDMR of molecular properties, variable cooperativity can be interpreted as bonds acting cooperatively to form a unique molecular fragment. Higher-order terms for subcomponents of any size are treated in analogy with those in eq 6, and represent the unique effect of their associated collection of atoms and bonds acting cooperatively upon the output. Note that some terms might represent noncontiguously bonded fragments, such as f_{13}. However, such terms are already fully expressed as two first-order terms, and from the defining relations in eq 4, the lower-order terms are subtracted from each higher-order term. Therefore the term f_{13} will be zero, as there is no direct cooperativity between the fragments FCH_3 and $ClCH_3$ formed by bonds 1 and 3. This is true in general for noncontiguous terms. In contrast, a term like f_{123} will likely be small on physical grounds but will not be zero because bonds 1, 2, and 3 form a contiguous molecular fragment FCH_2CH_2Cl with a corresponding unique cooperative impact.

The individual HDMR terms can be included in any combination, as each term represents the unique contribution from its particular order. This feature of the HDMR allows for the flexible inclusion of terms that are physically relevant and quantitatively significant. Also, because the HDMR terms correspond to stoichiometrically balanced summations, they are each described by the property change of an associated chemical reaction.

Higher-order effects of rings or sterically crowded subcomponents can be incorporated by taking the relevant higher-order fragment to local convergence, meaning the HDMR terms for all fragments containing the special structure are included. For example, in the case of methylbenzene, one would include all first through sixth-order terms associated with the isolated benzene ring, along with the standard low-order terms for fragments containing the methyl substituent. This treatment only requires one additional data point for each higher-order fragment (the property value of that higher-order fragment), because when taken to convergence the HDMR expansion is by construction the exact property value. That is, the sum of all terms associated with an isolated benzene ring is by definition the exact property value of benzene. This treatment is analogous to the "difference method" sometimes used in the context of additivity.[7] This work treats higher-order effects in this way.

Another way to treat special higher-order structural effects is to define the relevant structure (e.g., the ring or sterically crowded fragment) as an irreducible structural entity. This may be achieved by excluding the internal bonds of the structure in the definition of the variables. Then even at zeroth-order, the special group would not be reduced to smaller subcomponents, and the previously high-order effect becomes a zeroth-order effect. In the case of methylbenzene, the benzene ring may be treated as an irreducible structure, and the benzene−methyl bond becomes the only variable for the molecule. This system could then take on one of the following two states:

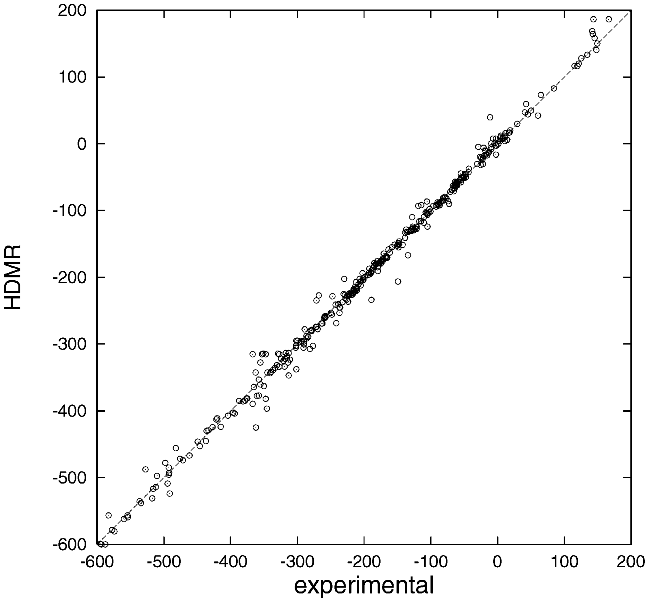

Figure 2. Results of HDMR predictions of heats of formation plotted against experimental values (kJ/mol).

TABLE 1: Statistical Analysis for the HDMR of Enthalpy of Formation

type	no. of molecules	average absolute error (kJ/mol)
acyclic	257	7.7 ± 13.1
monocyclic	106	8.5 ± 11.3
polycyclic	12	16.9 ± 11.7
total	375	8.2 ± 12.6

$$
\begin{aligned}
g(x{=}1) &= \Phi_{\mathrm{PhCH_3}} \\
g(x{=}0) &= \Phi_{\mathrm{PhH}} + \Phi_{\mathrm{CH_4}} - \Phi_{\mathrm{H_2}}
\end{aligned}
\tag{7}
$$

where Ph is a phenyl group. Because both states incorporate the experimental property value of benzene, the previously high-order effect of the benzene ring becomes a zeroth-order effect. In all molecules containing benzene, the benzene ring can thereby naturally be incorporated into all subcomponent fragmentations of the molecule. This treatment of high-order effects requires more experimental data of large molecules and was not tested in this work.

As for additivity, the HDMR of molecular properties should ideally be applied under a consistent set of conditions such as temperature, pressure, phase, solvent, and concentration. Because no interaction is assumed between nonbonded subcomponents, the model is expected to be most accurate under conditions where intermolecular interactions are negligible, such as in a dilute gas.

3. Application of HDMR to Heats of Formation

The HDMR of molecular properties was tested on the enthalpy of formation, which has been a central focus of

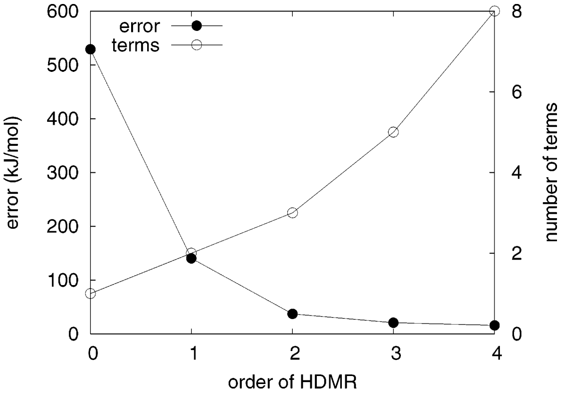

Figure 3. Average absolute error of different order HDMR predictions of heats of formation (kJ/mol), and the number of terms required for alkanes at different orders.

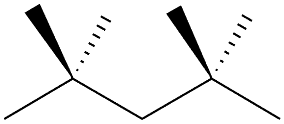

Figure 4. Illustration of steric crowding effect in a 2,2,4,4-substitution, incorporated into the HDMR by taking this structure to convergence.

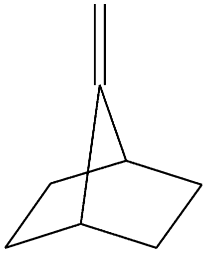

Figure 5. 7-Methylenebicyclo[2.2.1]heptane with strained methylene substituent.

additivity models. The enthalpies of formation of a broad range of organic molecules were calculated by HDMR and compared with the measured enthalpy value. All compounds compiled in *Thermochemical Data of Organic Compounds*[16] were tested, excluding those used as input data and those for which the input data was insufficient. A total of 375 test molecules were examined, ranging in size from six non-hydrogen atoms to over 30, with average molecular formula $C_{8.0}H_{16.0}O_{0.5}N_{0.1}S_{0.1}Cl_{0.05}Br_{0.06}$. Of these molecules, 257 were acyclic, 106 were monocyclic, and 12 were polycyclic. The predictions were made for compounds in the gas phase at 298.15 K and 1 atm. The variables were designated as all bonds between non-hydrogen atoms, and the saturating agent was hydrogen. The HDMR was taken to fourth-order for all molecules, meaning that terms for groupings of up to five non-hydrogen atoms are included. This level of approximation is equivalent to a GA-level approximation, with the addition of terms across three or four bonds (such as f_{123} and f_{124} for 1-fluoro-2-chloropropane), although in older GA literature many quaternary carbon (fourth-order) GA values were not available. These terms across several bonds are especially important in conjugated molecules, where conjugation is a unique effect that spans several contiguous bonds. Also, rings and sterically crowded clusters in the form of 2,2,4,4-substitutions were taken to local convergence. Terms for fragments with *cis*−*trans* isomers were calculated by taking the average property value of the two isomers; in general this caused a disparity of only a few kJ/mol. The results are summarized in Table 1 and Figure 2. Figure 3 shows the average absolute errors for these molecules taken to different orders, as well as the number of HDMR terms required for alkanes at different orders to indicate the scaling of the sampling complexity. We note that even at third order, the HDMR model performs very well. Consistent with the HDMR hypothesis, the errors become small by fourth order. A compilation of HDMR terms derived from ref 16 is given in the Appendix.

Several factors might be expected to contribute to the errors in the HDMR predictions. First, systematic errors in the experimental heats of formation of the input subcomponents will tend to distort the HDMR prediction. However, there is no way to quantify such errors if they exist. Also, the data used in the HDMR calculations are verified to be thermodynamically consistent in terms of Hess's Law,[16] so we expect that these systematic errors do not play a large role in the HDMR errors. Secondly, each experimental heat of formation has an associated statistical error, and for HDMR expansions with a large number of terms, these errors will propagate to produce uncertainty in the HDMR estimate. We note that when higher-order structures such as rings are taken to local convergence, the resulting HDMR sum actually includes a smaller total number of input terms, because the converged ring terms simply sum to the exact experimental value of the ring. For the low order of HDMR that we employ, the statistical uncertainties are generally expected to be much smaller than the disparity between the HDMR estimate and the experimental value, so we do not discuss this point further. Finally, each HDMR estimate has a truncation error that may be traced to higher-order effects not included in the HDMR expansion. Systematic high-order effects can be incorporated into HDMR predictions as discussed below. One such effect is a ring structure, which involves all bonds acting cooperatively to form the ring. Another is steric crowding, which can oftentimes be anticipated from experience with HDMR predictions of this type. For example, in this work we find that a 2,2,4,4-substitution shown in Figure 4 systematically contributes to the heat of formation, so we include this term in all relevant molecules.

Acyclic molecules are straightforward to approximate with HDMR, because they involve few systematic higher-order effects. The only higher-order effect incorporated into the acyclic HDMR expansions was the dense cluster of a 2,2,4,4-substitution, which is an eighth-order steric crowding effect, as shown in Figure 4. Monocyclic molecules were treated by taking the ring subcomponent to convergence, which produces an HDMR of comparable accuracy to that for acyclic molecules. The success of this treatment of rings suggests that in these test molecules a substituted ring is generally not greatly distorted from its unsubstituted conformation, and therefore forms a good basis to describe substituted rings.

Polycyclic molecules were found to be least tractable. Because of the broad variety of polycyclic rings, experimental knowledge of many special polycyclic subcomponent molecules is required to take the simple polycyclic subcomponents to convergence. The increased error of these predictions indicates that even taking the simple polycyclic ring to convergence, the enthalpy values will have relatively large additional high-order effects. This indicates that substituents on the polycyclic rings do not behave as they do in isolation. For example, in the molecule 7-methylenebicyclo[2.2.1]heptane, shown in Figure 5, the contribution from the methylene substituent is approximated on the basis of molecules such as ethylene, propylene, and isobutene. However, the methylene substituted on 7-methylenebicyclo[2.2.1]heptane is actually substituted onto a strained carbon that can be expected to behave differently from the simple unstrained fragments. As expected, this level of approximation yields a high error of 17.7 kJ/mol for this molecule. The

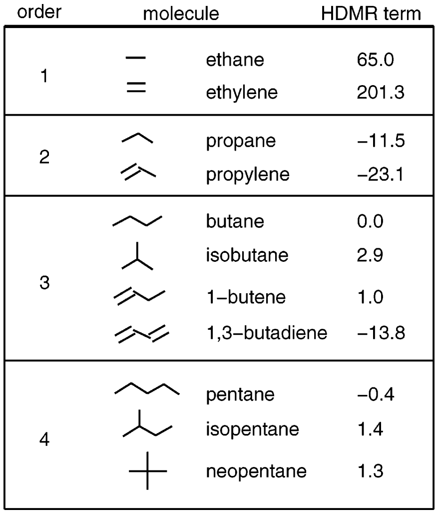

Figure 6. Illustration of some HDMR terms (kJ/mol) for common hydrocarbon structures.

polycyclic molecules would be more accurately treated by defining the polycyclic rings as irreducible structural entities. However, the experimental data for these molecules was not sufficient for a thorough test of this approach.

To illustrate the physical interpretation of HDMR, we examine the HDMR terms for common hydrocarbon structures shown in Figure 6. Each HDMR term represents the unique contribution from the structure shown. As expected by the HDMR hypothesis, the terms generally decrease in magnitude with increasing order. The first-order term for ethylene is larger than that for ethane, because the cooperative effect of a double bond is larger than that for a single bond. This illustrates the familiar heuristic notion of π electrons contributing to the bond strength of a double bond, in addition to the σ bond, which increases the cooperativity of the ethylene bond compared to the ethane bond. The second-order term for propylene is larger than that for propane, because the π-electrons in propylene interact with the end methyl group. There is no such effect in propane, which explains why the propylene term is greater in magnitude. The third-order terms vary considerably in magnitude due to the different physical effects associated with each structure. The butane term is zero, indicating that the three carbon−carbon bonds do not act together to influence the heat of formation. The isobutane term is slightly larger, because there is a small crowding effect from having the three methane substituents in close proximity. 1-Butene has a small effect, but 1,3-butadiene has a much larger one, because its π-electrons are conjugated, which is a unique third-order effect not captured by any of its smaller subcomponents. Finally, the fourth-order terms are generally small. The pentane term is nonzero, however, while the butane term is zero. This may be because as the carbon chain becomes longer, the end carbons begin to interact with other regions of the molecule through the relatively unhindered torsional motion about the carbon−carbon single bonds. The isopentane and neopentane terms are slightly larger due to steric crowding.

4. Comparison between HDMR and Group Additivity

4.1. Group Additivity as a Special Case of HDMR. As mentioned before, HDMR contains the popular GA scheme as a special case. To illustrate this point, consider a generic ethane CabcCxyz (Figure 7), where abc and xyz are arbitrary substituents. The GA expression for this molecule is

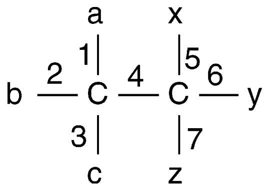

Figure 7. Generic ethane with labeled variable indices.

$$\phi^{\mathrm{GA}} = [\mathrm{C{-}(C)(a)(b)(c)}] + [\mathrm{C{-}(C)(x)(y)(z)}] = \Phi_{\mathrm{CabcCH_3}} + \Phi_{\mathrm{CxyzCH_3}} - \Phi_{\mathrm{CH_3{-}CH_3}} \quad (8)$$

where

$$\mathrm{C{-}(C)(a)(b)(c)} = \Phi_{\mathrm{CabcCH_3}} - \frac{1}{2}\Phi_{\mathrm{CH_3{-}CH_3}} \quad (9)$$

and

$$\mathrm{C{-}(C)(x)(y)(z)} = \Phi_{\mathrm{CxyzCH_3}} - \frac{1}{2}\Phi_{\mathrm{CH_3{-}CH_3}} \quad (10)$$

by analogy with eq 2. More generally, a generic *n*-membered substituted hydrocarbon, CabcCde...Cxyz (Figure 8), is in the GA approximation

$$\phi^{\mathrm{GA}} = [\mathrm{C{-}(C)(a)(b)(c)}] + [\mathrm{C{-}(C)_2(d)(e)}] + ... + [\mathrm{C{-}(C)(x)(y)(z)}] \quad (11a)$$

$$= \Phi_{\mathrm{CabcCH_3}} + \Phi_{\mathrm{Cde(CH_3)_2}} + \cdots + \Phi_{\mathrm{CxyzCH_3}} - (n-1)\Phi_{\mathrm{CH_3{-}CH_3}} \quad (11b)$$

The equivalent HDMR expansion for CabcCxyz involves each substituted carbon taken to local convergence, where

$$f_{1234} + f_{123} + ... + f_0 = \Phi_{\mathrm{CabcCH_3}} \quad (12)$$

and

$$f_{4567} + f_{456} + ... + f_0 = \Phi_{\mathrm{CxyzCH_3}} \quad (13)$$

because the converged HDMR summation for a given fragment is the exact property value of the fragment. When the sums in eqs 12 and 13 are added together to approximate the property of CabcCxyz, the connecting ethane (f_4) will be counted in both fourth-order methane expansions, so $\Phi_{\mathrm{CH_3{-}CH_3}}$ is subtracted off. Similarly, for an *n*-membered substituted hydrocarbon, $n - 1$ ethanes will be subtracted. Hence, the HDMR and GA expansions are equivalent.

Similar reasoning can be employed to show that HDMR also contains bond additivity, atomic additivity, etc. as special cases. We note that the exact equations presented above are slightly different if a carbon is replaced by an oxygen, nitrogen, etc. In these cases, it is not possible to define unique group contributions from the C−(O)(x)(y)(z) and the O−(C)(x) groups, as they always occur together in pairs. Therefore in GA one can only solve for the sum of the two groups, as mentioned in ref 7. This is equivalent to going to a higher-order HDMR approximation, and the logic of the proof is identical to that presented above.

4.2. Advantages of HDMR Over Additivity.

4.2.1. Practical Advantages. The main practical advantage of HDMR over additivity is that HDMR accounts for the full hierarchy of subcomponent contributions, which allows for the most efficient and flexible use of the input data. In some cases, the thorough hierarchy of the HDMR formulation can lead to a

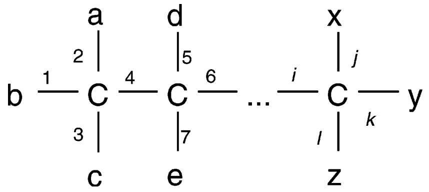

Figure 8. Generic hydrocarbon with labeled variable indices.

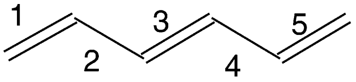

Figure 9. 1,3,5-Hexatriene with labeled variable indices.

more accurate estimate. In others, HDMR allows for an estimate that is not possible with GA due to lack of specific input data.

To illustrate the more efficient use of the input data by HDMR, consider the molecule 1,3,5-hexatriene, shown in Figure 9 with labeled HDMR variable indices. In this example, for the sake of comparison, we will apply the common extension to GA of differentiating between singly and double bonded carbon atoms. Furthermore, we assume that the input data is restricted to methanes, ethanes, propanes, and butanes. The GA approximation for this molecule is

$$\phi^{\mathrm{GA}} = 4[\mathrm{C_d{-}(C_d)_2(H)}] + 2[\mathrm{C_d{-}(C_d)(H)_2}] \tag{14}$$

The $\mathrm{C_d{-}(C_d)(H)_2}$ group values can be derived from ethylene, viz.

$$\Phi_{\mathrm{C_2H_4}} = 2[\mathrm{C_d{-}(C_d)(H)_2}] \tag{15}$$

and the $\mathrm{C_d{-}(C_d)_2(H)}$ group values can be derived from 1,3-butadiene, viz.

$$\Phi_{\mathrm{C_2H_4}} = 2[\mathrm{C_d{-}(C_d)_2(H)}] + 2[\mathrm{C_d{-}(C_d)(H)_2}] \tag{16}$$

giving

$$\mathrm{C_d{-}(C_d)(H)_2} = \frac{1}{2}\Phi_{\mathrm{C_4H_6}} \tag{17}$$

and

$$\mathrm{C_d{-}(C_d)_2(H)} = \frac{1}{2}[\Phi_{\mathrm{C_2H_4}} - \Phi_{\mathrm{C_4H_6}}] \tag{18}$$

As shown in section 4.1, the GA approximation

$$\phi^{\mathrm{GA}} = \phi^{\mathrm{HDMR}} = f_0 + \sum_{i=0}^{5} f_i + \sum_{i<j} f_{ij} + f_{123} + f_{345} \tag{19}$$

is a special case of the HDMR approximation. However, a term that is excluded from eq 19 is f_{234}, which takes 2-butene as input. Inclusion of this term brings the HDMR approximation within the uncertainty of the experimental value. In contrast, GA will never make use of 2-butene as input data for this molecule, because it does not share any groups with 1,3,5-hexatriene. Rather it consists of two $\mathrm{C_d{-}(C_d)(C)(H)}$ groups and two $\mathrm{C{-}(C_d)(H)_3}$ groups. This example demonstrates the fundamental advantage of a comprehensive hierarchy of effects in HDMR, compared to the more limited group-based formalism of additivity.

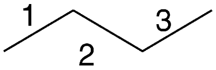

Figure 10. Butane with labeled variable indices.

In general, it holds that for an unconverged GA estimate, i.e., one for which the estimated value lies outside the uncertainty of the true value, it is *always* possible to improve upon the GA estimate using HDMR, because the HDMR is an exact decomposition of the molecular property into its constituent subcomponent contributions. Although in some cases this improvement might not arise until the trivial case where all HDMR terms are needed to obtain convergence, the above example demonstrates that in other cases, additional low-order HDMR terms are sufficient to obtain convergence.

4.2.2. Conceptual Advantages. The HDMR formulation is especially attractive compared to additivity because of its decomposition of the property value into physically interpretable units. This occurs naturally in HDMR by first defining all terms in reference to the $g(\bar{\mathbf{x}})$ state, and then decomposing the property value into the unique contributions from fragments corresponding to each bond, pair of bonds, triplet of bonds, etc. in reference to that state. For each collection of bonds i, j, k, etc. the term $f_{ijk...}(x_i,x_j,x_k,...)$ expresses the unique contribution from the fragment formed by bonds i, j, k, etc., as the lower-order contributions are subtracted off. Furthermore, because the complete HDMR expansion is by definition the exact property value, the error of a truncated HDMR is precisely defined as the sum of the finite number of terms excluded from the expansion. The error of an HDMR expansion can therefore be rationalized through a physical interpretation of those excluded terms. All HDMR terms are also described by the property change of associated chemical reactions, which provide for an alternative chemical interpretation of the term values.

In contrast, additivity derives its contribution values directly from the solution of a given set of linear equations, like eq 2, which gives no related expression for higher or lower order effects. Once a given order of additivity is chosen, it is difficult to derive expressions for additive effects of different orders. Furthermore, there is no natural procedure to write an expression for the error in an additive approximation.

As an illustration, consider the GA and HDMR decompositions of butane, shown in Figure 10. The GA approximation, denoted ϕ^{GA}, is as follows:

$$\begin{aligned}
[\mathrm{C{-}(C)(H)_3}] &= \frac{1}{2}\Phi_{\mathrm{CH_3CH_3}} \\
\mathrm{C{-}(C)_2(H)_2} &= \Phi_{\mathrm{CH_3CH_2CH_3}} - \Phi_{\mathrm{CH_3CH_3}} \\
\phi^{\mathrm{GA}} = 2\mathrm{C{-}(C)(H)_3} + 2[\mathrm{C{-}(C)_2(H)_2}] &= \\
2\Phi_{\mathrm{CH_3CH_2CH_3}} &- \Phi_{\mathrm{CH_3CH_3}}
\end{aligned} \tag{20}$$

The above group contribution values $\mathrm{C{-}(C)(H)_3}$ and $\mathrm{C{-}(C)_2(H)_2}$ are derived from eq 2. In comparison, the equivalent second-order HDMR expansion of butane, denoted ϕ^{HDMR}, becomes

$$\begin{aligned}
f_0 &= 4\Phi_{CH_4} - 3\Phi_{H_2} \\
f_1 = f_2 = f_3 &= \Phi_{CH_3CH_3} - 2\Phi_{CH_4} + \Phi_{H_2} \\
f_{12} = f_{23} &= \Phi_{CH_3CH_2CH_3} - 2\Phi_{CH_3CH_3} + \Phi_{CH_4} \\
\phi^{HDMR} &= f_0 + \sum_{i=1}^{3} f_i + \sum_{i<j}^{3} f_{ij} \\
&= 2\Phi_{CH_3CH_2CH_3} - \Phi_{CH_3CH_3}
\end{aligned} \tag{21}$$

As expected, because HDMR is constrained to the GA level of approximation, the final HDMR and GA summations are identical, but their respective decompositions demonstrate the conceptual difference between HDMR and GA. The GA values $C-(C)(H)_3$ and $C-(C)_2(H)_2$ can be interpreted as the property contributions from the fragments H_3CC and CCH_2C, and are the result of a particular solution of the chosen set of linear equations. On the other hand, the HDMR first-order terms f_i describe the unique contribution of ethane fragments to the property value, and the second-order terms f_{ij} describe the unique contribution of propane fragments to the property value. Moreover, the error of this second-order HDMR expansion is clearly identified as the unique contribution from the entire butane fragment, expressed by f_{123}.

The complete hierarchy of HDMR also provides a physical basis for the numerous correction terms commonly employed in GA. As mentioned before, the ring corrections used in GA may be expressed as a unique *n*th-order HDMR effect, for an *n*-membered ring. Similarly, gauche interactions may be expressed as third-order HDMR terms across three contiguous bonds. In general, any GA correction term may be expressed as a unique HDMR effect, thereby providing a physical and consistent framework for the hierarchy of intramolecular effects.

TABLE 2: First-Order HDMR Terms (kJ/mol)

acetylene	377.0	ethylene	201.3	ethane	65.0
toluene	42.2	methylamine	97.3	formaldehyde	207.6
methanol	114.7	chloromethane	84.8	bromomethane	75.2
iodomethane	62.6	biphenyl	16.2	aniline	50.4
benzenethiol	50.4	fluorobenzene	74.7	chlorobenzene	61.7
bromobenzene	59.1	iodobenzene	55.8		

TABLE 3: Second-Order HDMR Terms (kJ/mol)

propyne	−33.9	allene	11.1	propene	−23.1
propane	−11.5	phenylethylene	−29.4	ethylbenzene	−11.1
diphenylmethane	−36.2	ketene	−65.8	acetaldehyde	−48.1
ethanol	−24.3	benzaldehyde	−52.9	benzenemethanol	−23.7
ethylamine	−15.0	benzylamine	−14.0	ethanethiol	−14.0
chloroethane	−20.8	bromoethane	−17.0	iodoethane	−12.8
dichloromethane	−6.2	*dimethylbenzene	−0.1	dimethylamine	−18.5
diphenylamine	−0.8	*iodomethylbenzene	−3.3	*difluorobenzene	11.4
dichlorobenzene	4.7				

5. Conclusion

We have presented an improved formulation for the additive estimation of molecular property values using HDMR. HDMR provides a complete hierarchical decomposition of a molecular property into exact contributions from molecular subcomponents. In the example application of enthalpy of formation, we have shown that there exists a systematic hierarchy of important subcomponent contributions which should be included in the HDMR estimate. In this way, we have accurately reproduced the experimental enthalpies of formation of a broad range of organic molecules.

HDMR was shown to contain the conventional additive schemes as special cases. However, the additive formulation does not account for the full hierarchy of subcomponent contributions, even to a given order, due to the rigid definition of groups as nonoverlapping subcomponents. Many limitations of additivity are improved by means of the more general theory of HDMR.

Going beyond the ordered sampling of cut-HDMR, future work could employ RS-HDMR to flexibly take input data distributed randomly over a range of molecules. For example, the first-order term for an ethane fragment could be derived from any number of molecules containing a carbon−carbon bond. This is similar to the way in which GA makes use of the experimental data, but RS−HDMR retains the physically intuitive form of eq 3, decomposing the property output into a hierarchy of components with clear physical significance.

This paper only investigated enthalpy of formation, but prior work demonstrating the broad utility of HDMR[13,17,18] suggests that many nonthermodynamic properties might also be reliably treated by cut- or RS-HDMR. Other properties for which additive empirical models have been applied are especially promising for reanalysis with an HDMR treatment. One example is the modeling of carbon-13 NMR peaks.[19] Given the generic nature of the HDMR formulation, extension to other properties is straightforward.

We finally note that, in common usage, additivity uses experimental data as input. However, as high-level ab initio calculations are becoming more accessible, it is natural to extend the scope of HDMR to take ab initio data as input, with the aim of calculating the properties of large molecules. The promising work of Deev and Collins in this field[6] suggests that

TABLE 4: Third-Order HDMR Terms (kJ/mol)

isobutene	7.1	isobutane	2.9	isopropylbenzene	6.1
triphenylmethane	79.8	acetone	17.8	2-propanol	7.6
diphenylmethanone	55.9	isopropylamine	−0.5	2-chloro-1-propene	−5.0
2-chloropropane	8.9	2-bromopropane	0.4	1,1-dichloroethylene	−13.3
1,1-dichloroethane	18.9	trichloromethane	11.9	2-iodopropane	0.9
*trimethylbenzene	1.2	1,1-diphenylethylene	38.5	trimethylamine	9.0
acetophenone	18.6	1-butyne	1.2	2-butyne	4.1
1,2-butadiene	4.3	1,3-butadiene	−13.8	1-butene	1.0
2-butene	3.2	butane	−0.0	propylbenzene	−1.1
ethylmethylbenzene	1.2	*1,2-diphenylethylene	0.9	1,2-diphenylethane	−0.7
propanal	1.4	1-propanol	1.0	glyoxal	36.4
1-propylamine	−1.9	1,2-ethanediamine	−6.6	1-propanethiol	−0.7
1,2-ethanedithiol	−0.9	3-bromo-1-propene	3.3	*1-bromo-1-propene	−4.3
3-iodo-1-propene	−4.8	1-chloropropane	1.1	1-bromopropane	−4.2
1-iodopropane	−1.9	*1,2-dichloroethylene	−17.3	1,2-dichloroethane	13.5
1,2-dibromoethane	2.5	1,2-diiodoethane	−2.0	cyclopropene	226.7
cyclopropane	116.0	oxirane	115.1	aziridine	110.1
thiirane	82.5				

TABLE 5: Fourth-Order HDMR Terms

neopentane	1.3	tetraphenylmethane	−129.0
tert-butylamine	8.4	tetrachloromethane	−1.8
1,1,1-trichloroethane	−25.9	2,2-dichloropropane	−19.7
tert-butylbenzene	1.8	isoprene	1.4
2-methyl-1-butene	1.5	3-methyl-1-butene	0.8
2-methyl-2-butene	1.1	isopentane	1.4
isobutylbenzene	1.2	*sec*-butylbenzene	0.6
isobutylaldehyde	−2.1	2-butanone	−1.9
2-butanol	−0.2	1-phenyl-1-propanone	−1.4
methylglyoxal	−9.2	2-butylamine	1.7
isobutylamine	2.9	1,2-propanediamine	2.3
2-butanethiol	0.9	2-methyl-1-propanethiol	0.8
1,2-propanedithiol	10.6	1-chloro-2-methylpropane	0.9
2-chlorobutane	3.5	2-bromobutane	4.2
1,2-dichloropropane	−4.2	1,2-dibromopropane	7.7
1,2-diiodopropane	3.5	trichloroethylene	39.1
1,1,2-trichloroethane	−22.2	1,2-pentadiene	−1.7
1,3-pentadiene	−2.0	1,4-pentadiene	0.7
2,3-pentadiene	−1.0	1-pentene	−0.5
2-pentene	−0.6	pentane	−0.4
butylbenzene	−0.1	1-butanol	1.0
1,3-propanediol	13.4	1-butylamine	−0.9
1-butanethiol	0.7	1-chlorobutane	−1.8
1-bromobutane	0.8	1,3-dichloropropane	−0.1
cyclobutene	122.0	cyclobutane	112.0
thietane	80.1	methylenecyclopropane	57.4
1-methylcyclopropene	0.2	cyclopropylbenzene	−9.3
methyloxirane	−6.9	cyclopropylamine	6.6
methylthiirane	−6.9		

this approach might prove to be a successful and efficient method to calculate the properties of large molecules.

Appendix

Tables 2−5 give the HDMR terms of first to fourth order derived from ref 16. Benzene is treated as an irreducible structure, so methylbenzene is defined as a first-order term, ethylbenzene as a second-order term, etc. Terms containing cis−trans isomers are derived by taking the average of the two isomers, and are indicated with an asterisk.

Acknowledgment. We acknowledge support from the National Science Foundation.

References and Notes

(1) Cox, J. D.; Pilcher, G. *Thermochemistry of Organic and Organometallic Compounds*; Academic Press: New York, 1970.

(2) Benson, S. W.; Buss, J. H. *J. Chem. Phys.* **1958**, *29*, 546−572.

(3) Stein, S. E.; Brown, R. L. *J. Chem. Inf. Com. Sci.* **1994**, *34*, 581−587.

(4) Nagvekar, M.; Daubert, T. E. *Ind. Eng. Chem. Res.* **1987**, *26*, 1362−1365.

(5) Jalowka, J. W.; Daubert, T. E. *Ind. End. Proc. Design Dev.* **1986**, *25*, 139−142.

(6) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.

(7) Cohen, N.; Benson, S. W. *Chem. Rev.* **1993**, *93*, 2419−2438.

(8) Benson, S. W. *Thermochemcial Kinetics*, 2nd ed.; Wiley: New York, 1976.

(9) Rabitz, H.; Alis, O. *J. Math Chem.* **1999**, *25*, 197−233.

(10) Li, G. Y.; Rosenthal, C.; Rabitz, H. *J. Phys. Chem. A* **2001**, *105*, 7765−7777.

(11) Shenvi, N.; Geremia, J. M.; Rabitz, H. *J. Phys. Chem. A* **2002**, *106*, 12315−12323.

(12) Geremia, J. M.; Rabitz, H. *Phys. Rev. A* **2004**, *70*, 023804.

(13) Rabitz, H.; Shim, K. *J. Chem. Phys.* **1999**, *111*, 10640−10651.

(14) Alis, O. F.; Rabitz, H. *J. Math Chem.* **2001**, *29*, 127−142.

(15) Platt, J. R. *J. Chem. Phys.* **1947**, *15*, 419−420.

(16) Pedley, J. B.; Naylor, R. D.; Kirby, S. P. *Thermochemical Data of Organic Compounds*; Chapman and Hall: New York, 1986.

(17) Shorter, J. A.; Ip, P. C.; Rabitz, H. *J. Phys. Chem. A* **1999**, *103*, 7192−7198.

(18) Geremia, J. M.; Rabitz, H.; Rosenthal, C. *J. Chem. Phys.* **2001**, *114*, 9325−9336.

(19) Fürst, A.; Pretsch, E.; Robien, W. *Anal. Chim. Acta* **1990**, *233*, 213−222.