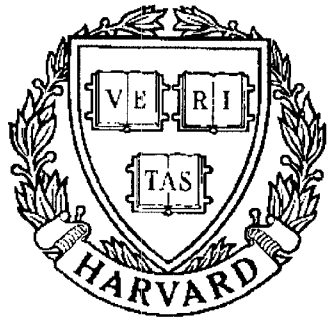
TECHNICAL RESEARCH REPORT

SYSTEMS RESEARCH CENTER



Supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), Industry and the University

Computer-Manipulation of Conjugate Forms in Proper Estimation

by M. L. Mavrovouniotis and L. Constantinou

TR 91-5

Computer-Manipulation of Conjugate Forms in Property Estimation

Submitted for presentation at PSE '91
and publication in the special issue of Computers and Chemical Engineering

Michael L. Mavrovouniotis and Leonidas Constantinou
Department of Chemical Engineering and Systems Research Center
A.V.Williams Bldg. (115), Room 2149
University of Maryland
College Park, MD 20742
Internet: mlmavro@phoenix.src.umd.edu Tel. (301) 405-6620

Abstract

Physical and chemical properties of pure compounds and mixtures are essential for the analysis and design of chemical processing systems. A method for the estimation of properties of organic compounds from their molecular structure is presented, based on the contributions of Atoms and Bonds in the properties of Conjugate forms of a molecular structure (ABC). A real chemical compound can be considered the hybrid of a number of conjugates, which are alternative formal arrangements of the valence electrons of the molecule. The property-estimation method generates all conjugate forms of the molecule and assigns properties to each conjugate, simply by summing contributions from atoms and bonds in the particular electronic arrangement of the conjugate. The properties of the actual compound are then derived from the properties of the conjugates.

The generation and analysis of conjugates is based on symbolic computation and Object-Oriented Programming (OOP). Atoms, bonds, molecules, electron pairs, and other entities can be represented as interconnected objects within OOP. The generation, comparison, and analysis of conjugates can be carried out through computer-based manipulation of the objects and their interconnections. One needs to encode operators which generate the conjugates, as well as rules for pruning the generation so that only the most important conjugates are considered.

The nature and connectivity of atoms within a molecule determine the physical and chemical properties of the molecule. Traditional group-contribution methods eliminate much of the detailed molecular-structure information at an early stage of the property-estimation effort. Through symbolic computation and the concept of conjugation, the ABC approach aims to use molecular structure information more effectively.

1. Introduction

A variety of physical and chemical properties of pure compounds and mixtures, under specified conditions, must be used in the modelling, design, and analysis of chemical process systems. When experimental values for the necessary parameters are not available, the properties must be estimated from information on the molecular structures of the compounds involved. If several systems or alternatives must be evaluated or designed quickly and approximately, the initiation of complex and time-consuming quantum-mechanical computations (or experimental measurements) cannot be justified. Consider for example the selection of a compound, based on a set of specifications. In order to narrow the field of candidates and focus on sets of promising compounds for detailed study, one must have very fast evaluation of computer-generated candidate compounds. One must also be able to determine trends which allow the elimination of whole sets of compounds; only simple estimation methods can make such trends apparent. Thus, simple and efficient methods for the approximate estimation of thermodynamic, physical, and chemical properties from the molecular structure of substances are essential for the preliminary modeling, analysis, and design of chemical processes.

In additive group contribution methods [Joback and Reid, 1987, Benson, 1968, Domalski and Hearing, 1988, Mavrovouniotis *et al.*, 1988, Mavrovouniotis, 1990a, Reid *et al.*, 1987], a property is estimated as a summation of contributions of the groups comprising the structure of the molecule. The advantage of these methods is that they can give quick estimates without requiring substantial computational resources. They also readily provide trends within sets of compounds. For example, the variation of a property with elongation of a side-chain is apparent in the contribution of the elongating group; substitutions that would have a specific effect on a property are apparent in the sign and magnitude of the relevant contributions,

permitting systematic generation of molecules [Joback and Stephanopoulos, 1989]. The simplified view of molecular structures inherent in group-contribution methods leads to significant limitations in their scope and accuracy [Mavrovouniotis, 1990]. Group contributions methods largely ignore concepts such as delocalized bonds, resonance, and other electronic interactions among groups. For example, the properties of ions are strongly dependent on charge dispersion which cannot be readily modelled by group contributions.

The objective of this work is the development of a new computer-based method for the estimation of thermodynamic, physical, and chemical properties of organic compounds from their molecular structure. The method should retain some of the simplicity and additive character that have made group-contribution methods so useful, while using more chemical concepts to achieve wider applicability and better accuracy. The approach presented here is based on the contributions of Atoms and Bonds in the properties of Conjugate forms (ABC) of a molecular structure. Conjugate forms are alternative formal arrangements of valence electrons in a molecule; a real chemical compound can be considered the hybrid of all its conjugates. Conjugates are extensively used for qualitative comparisons but have been largely ignored in property estimation for chemical engineering purposes. In ABC, we start by generating all conjugate forms of the molecule whose properties we wish to estimate. Properties are assigned to each conjugate, simply by summing contributions from atoms and bonds in the particular electronic arrangement of the conjugate. Then, these properties of the conjugates are combined to derive the properties of the compound.

The computer-based generation and analysis of conjugates is facilitated by symbolic computing environments and Object-Oriented Programming (OOP), which allow the flexible representation and manipulation of molecular structures. Atoms, bonds, molecules, electron pairs, and other entities can be represented as interconnected objects, whose attributes and interconnections can be manipulated.

Although our initial focus entails simple thermodynamic properties, such as the enthalpy of formation, the approach will be used for the estimation of other physical properties, fractional charges on individual atoms of the compound, and electron densities and strengths of individual bonds. These properties of molecular fragments will in turn lead to analysis of intermolecular interactions and chemical properties of compounds.

In Section 2, an overview of the ABC approach is provided; some aspects of the technique are covered only briefly, because they are explained more comprehensively in [Mavrovouniotis, 1990]. In Section 3, the empirical function which combines the energies of conjugates to produce the energy of the real compound is analyzed and shown to possess attractive properties. Section 4 provides an illustrative example, and the discussion in Section 5 raises some of the computational issues that influence the practicality of the method.

2. Conjugation and Property Estimation

Whenever a compound can be described by several electronic arrangements which are compatible with the basic topology of the molecule, its stability is enhanced, in the sense that the energy of the compound is lowered. The effect is more pronounced when the alternative structural formulae are energetically similar. The real compound is viewed as a *hybrid* of all its *conjugates*. Each conjugate can be viewed as an arrangement of atoms which are connected, in pairs, by single, double, or triple bonds; each electron pair belongs to a specific bond that connects a specific pair of atoms; each charge present on a conjugate must be integer, because it is due to the deficiency or surplus of an integer number of electrons. The hybrid, however, cannot be represented in this way, because its electronic structure involves delocalized electrons, which bond different pairs of atoms in different conjugates. If we want to view the compound (hybrid) without referring to the original conjugates, we must represent its bonds as being *partially double* and partially single (more precisely, there are delocalized bonds distributed over more than two atoms); we may also need to represent fractional charges. Consider the two isomeric dienes below:

$CH_3—CH=CH—CH=CH—CH_2—CH_3$ (2,4-heptadiene) [I]

$CH_2=CH—CH_2—CH_2—CH=CH—CH_3$ (2,5-heptadiene) [II]

Any of the carbon-carbon double bonds in these compounds can be depicted as having three conjugates:

$>C^{1+}—C^{1-}< \leftrightarrow >C=C< \leftrightarrow >C^{1-}—C^{1+}<$

However, the proximity of the two double bonds in 2,4-heptadiene allows two additional conjugates, involving both double bonds:

$CH_3—CH^{1+}—CH=CH—CH^{1-}—CH_2—CH_3$ [III]

$CH_3—CH^{1-}—CH=CH—CH^{1+}—CH_2—CH_3$ [IV]

These additional conjugates make the enthalpy of 2,4-heptadiene lower than that of 2,5-heptadiene by 13 kJ/mol [Morrison and Boyd, 1973]. The conjugates also affect the properties of individual bonds. In this case, the $^3C—^4C$ bond has a partial double bond character, because it appears as a double bond in the forms [III] and [IV]; naturally, its single bond character is more prevalent, because it occurs in the lowest-energy structure [I]. Similarly the $^2C—^3C$ and $^4C—^5C$ bonds are weaker than a normal double bond [Morrison and Boyd, 1973]. It is important to note that conjugation can include interactions among atoms that are separated by several bonds (Fig. 3), but even one single bond can be viewed as a hybrid of three forms — one purely covalent and two purely ionic conjugates:

$C^{1+}\cdots H^{1-} \leftrightarrow C—H \leftrightarrow C^{1-}\cdots H^{1+}$ [V]

The prevalent conjugate is the covalent one, but the presence of the ionic conjugates implies that the single bond is polar and can be viewed as partly ionic.

In the examples of conjugation shown so far, electrons were moved only in pairs. It is actually possible to separate a pair of electrons and create a di-radical structure, but such structures are actually not considered in the current framework, because they would require consideration of electron-spin. Thus, our attention is hereafter confined to paired-electron structures (neutral molecules as well as ions), and their paired-electron conjugate forms.

It should be emphasized that conjugate forms are imaginary structures. The energies of individual conjugates do *not* represent distinct energy states of the molecule; and a real compound is *not* a mixture of conjugate forms in equilibrium. Each individual molecule of the compound (even at a fixed time point) has an electronic structure which is a hybrid of all the conjugate forms. A consequence of this is that one cannot apply statistical mechanics to relate the hybrid to the conjugate forms. For example, one cannot use partition functions to describe the distribution of conjugates; in fact no such distribution exists. Contrary to statistical-mechanical formulae, the internal energy (or the enthalpy) of the hybrid must be lower than that of any individual conjugate. The fact that conjugation results in lower energy is apparent in quantum-chemical analysis [Mavrovouniotis, 1990].

Many empirical rules used in organic chemistry are based on the idea of conjugation. Consider first the widely-used rule stating that branched alkanes have enthalpies lower than their straight-chain isomers [Morrison and Boyd, 1973]. Such differences cannot be due to steric hindrance among side chains, because such steric interactions should lead to the reverse from the observed order of enthalpies. The basis of the rule lies with an electronic conjugation effect. One would expect lower-energy conjugates that affect as few bonds as possible to exert the strongest influence on the enthalpy of a compound. One-bond conjugates (such as [V] above) cannot be used in the comparison because the number of carbon-carbon bonds and the number of carbon-hydrogen bonds does not vary within a set of isomeric alkanes. The next set of candidates would involve three bonds and would have the form:

$X—C—C—Y \leftrightarrow X^{1-}\ldots C=C\ldots Y^{1+}$

where each of X and Y can be either a carbon or a hydrogen. Since carbon is less electronegative than hydrogen, the most important conjugates are the ones in which X is a hydrogen and Y is a carbon, as can be inferred from the fact that hydrogen generally bears a fractional negative charge in hydrocarbons [Fliszar, 1983]. These will be called HCCC-conjugates. Each distinct subchain of the form H-C-C-C, gives a distinct conjugate. Examination of the C_5H_{12} isomers (first part of Table 1) shows that there are 14 HCCC-conjugates for n-pentane, 20 for 2-methylbutane, and 36 for 2,2-dimethylpropane. In agreement with the stated rule and the experimental enthalpies of formation, the higher number of conjugates leads to lower enthalpy. Consider also the three isomeric methyl-heptanes (second part of Table 1). While group contribution methods would predict the same enthalpy of formation for the three methyl-heptanes, inspection of the number of HCCC-conjugates shows that, in agreement with the data, 2-methylheptane has lower enthalpy and the other two are roughly equivalent.

Table 1. Analysis of the effect of HCCC-conjugates on the enthalpies of formation of alkanes.

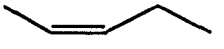
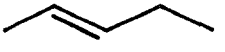
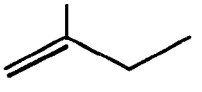
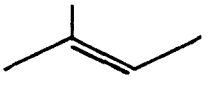
Compound	Molecular Formula	Enthalpy of Formation, kJ/mol [Cox and Pilcher, 1970]	Number of HCCC-conjugates
n-pentane	C_5H_{12}	–146.9	14
2-methylbutane	C_5H_{12}	–154.2	20
2,2-dimethylpropane	C_5H_{12}	–168.5	36
2-methylheptane	C_8H_{18}	–215.3	32
3-methylheptane	C_8H_{18}	–212.5	30
4-methylheptane	C_8H_{18}	–212.0	30
n-octane *(for comparison)*	C_8H_{18}	–208.6	26

A similar empirical rule which states that the enthalpies of alkenes increase with their degree of substitution, i.e., the number of side chains attached on the double-bonded carbons, can also be explained by conjugation. Consider the set of isomeric pentenes with their enthalpies of formation and the number of side chains on double-bonded carbons, shown in Table 2. The empirical rule cannot distinguish between 2-methyl-1-butene and 2-pentene. The HCCC-conjugates discussed for alkanes are relevant for this example, but they take a secondary role because of the important class of HCC=C-conjugates which involve fewer bond disruptions:

$$H—C—C=C \quad \leftrightarrow \quad H^{1+}...C=C—C^{1-}$$

The numbers of HCC=C-conjugates (Table 2) capture the same ordering as the original empirical rule; with the additional evidence from HCCC-conjugates it is also possible to distinguish between 2-methyl-1-butene and 2-pentene. However, the cis and trans versions of 2-pentene remain indistinguishable; this limitation is inherent to the analysis performed here.

Table 2. Analysis of the effect of the number of side-chains and the number of conjugates on the enthalpies of formation of isomeric pentenes.

ΔH_f (kJ/mol)	compound	structure	side chains	HCC=C	CCC=C	HCCC
–20.93	1-pentene		1	2	1	8
–28.09	2-pentene, cis		2	5	1	3
–31.78	2-pentene, trans		2	5	1	3
–36.34	2-methyl-1-butene		2	5	1	7
–42.58	2-methyl-2-butene		3	9	0	6

The property-estimation framework proposed here uses conjugation in a quantitative fashion. In ABC, all compounds are represented as hybrids of conjugates. To estimate the properties of a compound, one must determine and combine formal properties of its conjugate forms. The properties of each conjugate are estimated from the contributions of individual atoms and individual bonds, rather than larger functional groups. The intramolecular interactions among atoms and groups are thus not captured through a large variety of groups (which would require a large number of parameters), but through a variety of conjugates, whose properties can in fact be captured with just a small number of contributions from atoms and bonds.

3. Combination of Energies of Conjugates

We now focus our attention on the quantitative determination of the enthalpy of formation. For brevity, the term "energy" will be hereafter used for ideal-gas enthalpy of formation. Let E_i be the energy of conjugate i, obtained by summing contributions from atoms and bonds. Let F be the function we use to combine the energies of the conjugates:

$$E = F(E_1, E_2, ..., E_{n-1}, E_n) \quad [1]$$

While the function F cannot be fully determined from the qualitative use of conjugates in organic chemistry, certain properties of the function can be postulated. A few basic properties of this function are given in Table 3.

One possible function that satisfies these requirements is:

$$\exp(-E/A)=\Sigma_i \exp(-E_i/A) \Leftrightarrow \Sigma_i \exp[(E-E_i)/A]=1 \Leftrightarrow E=-A \ln [\Sigma_i \exp(-E_i/A)] \quad [1]$$

where A is any *positive* parameter which has molar energy (kJ/mol) units. The apparent similarity of this combination function to statistical-mechanical formulae should not be construed as a statistical-mechanical interpretation of either the properties in Table 3 or conjugation. The similarity is only coincidental, and it should be emphasized that this function is not an exact relationship but only an approximation with certain appealing properties.

It is easy to verify that this function does satisfy the requirements of Table 3. The properties of the derivatives of F are easy to prove, noting that:

$$\partial F/\partial E_i=\exp(-E_i/A)/\Sigma_i\exp(-E_i/A)=\exp[(E-E_i)/A] \quad [2]$$

$$\partial^2 F/\partial E_i^2 = A^{-1} [(\partial F/\partial E_i)^2 - \partial F/\partial E_i] \quad [3]$$

$$\partial^2 F/\partial E_i \partial E_j = A^{-1} (\partial F/\partial E_i) (\partial F/\partial E_j) \quad \text{for } i \neq j \quad [4]$$

Table 3. Desired properties for the energy-combination function

Property	Explanation
$F(E_1, ..., E_n)<E_i$	The hybrid must have energy lower than any of the conjugates
as $E_i \rightarrow +\infty$: $F(E_1, ..., E_{i-1},E_i,E_{i+1}, ...,E_n)\rightarrow F(E_1, ..., E_{i-1},E_{i+1}, ...,E_n)$	If a conjugate has energy much higher than other conjugates, it does not affect the hybrid, i.e., it can be neglected
as $E_i \rightarrow -\infty$: $F\rightarrow E_i$	If a conjugate has energy much lower than all others, then the the hybrid is equivalent to this conjugate
$F(E_1-K, E_2-K, ..., E_{n-1}-K, E_n-K) = F(E_1, E_2, ..., E_{n-1}, E_n)-K$	A change of reference state should alter the energy of the hybrid by exactly the same amount as it alters the energy of each conjugate
$0 < \partial F/\partial E_i \leq 1$	Lowering the energy of any conjugate lowers the energy of the hybrid, but by a lesser amount.
$\partial^2 F/\partial E_i \partial E_j > 0$ $(i \neq j)$	The influence of any conjugate decreases as the energy of another conjugate decreases.
$\partial^2 F/\partial E_i^2 \leq 0$	The influence of any conjugate increases as the energy of the same conjugate decreases.

We take a closer look at one additional property, which is extremely important on physical grounds: Conjugation in distinct non-interacting portions of a molecule should lead to independent effects on the energy of the molecule.

Consider a molecule consisting of a long chain in which conjugation is unimportant. We now substitute at the two ends of the chain two but strongly-conjugating groups, X and Y. Group X occurs has two import ant conjugate forms, X_1 and X_2, and Y has, similarly, the forms Y_1 and Y_2. The non-conjugating middle

portion of the molecule can be divided and lumped with X and Y, so that we have a molecule with only two portions which conjugate but do not affect each other. Based on the procedure outlined in previous sections, we would need to generate all conjugate forms of the molecule XY; these forms are X_1Y_1, X_1Y_2, X_2Y_1, and X_2Y_2. The energy of each form would be computed additively, i.e.

$E(X_1Y_1)=E(X_1)+E(Y_1)$

$E(X_1Y_2)=E(X_1)+E(Y_2)$

$E(X_2Y_1)=E(X_2)+E(Y_1)$

$E(X_2Y_2)=E(X_2)+E(Y_2)$

and the energy of the compound would be obtained through Eq. [1] as:

$$\exp\{-A^{-1}E(XY)\}=\exp\{-A^{-1}E(X_1Y_1)\}+\exp\{-A^{-1}E(X_1Y_2)\}+\exp\{-A^{-1}E(X_2Y_1)\}+\exp\{-A^{-1}E(X_2Y_2)\}$$

The requirement we are introducing states that, since the two portions do not interact we should be able to consider them separately and derive their hybrid energies, denoted as E(X) and E(Y):

$$\exp\{-A^{-1}E(X)\}=\exp\{-A^{-1}E(X_1)\}+\exp\{-A^{-1}E(X_2)\}$$

$$\exp\{-A^{-1}E(Y)\}=\exp\{-A^{-1}E(Y_1)\}+\exp\{-A^{-1}E(Y_2)\}$$

and then simply add them to obtain the energy of the whole compound:

$E(XY)=E(X)+E(Y)$

For this simple example, one can easily verify that the two sets of formulae for E(XY) would yield the same result, suggesting that chosen combination function preserves the intuition of independently conjugating portions of a molecule. Note that both the mathematical statement and the underlying intuition dictate the equivalence only if the conjugate forms of portions can be assembled in any combination, i.e., if the portions are entirely separate and cannot interact. The mathematical statement also presupposes that the energy of each conjugate is obtained in an additive manner from contributions of the building blocks; thus, the empirical function that combines the properties of conjugates is intimately tied to the way these are determined from the structures of the conjugates.

Moving from this example to a general formal statement, consider a molecule with m distinct portions, conjugating independently. Each portion can be in one of several conjugate forms, whose presence and energy are not affected by the conjugate forms of the other portions. Let $E_{pc}(k; i)$ represent the energy of the i conjugate form of the k portion, and $E_p(k)$ the hybrid energy of the k portion of the compound; for a given portion k, $E_p(k)$ is a combination of the energies $E_{pc}(k; i)$ and is independent of the nature and conjugation of other portions. If the k portion has a_k conjugates, the compound as a whole has $a=\Pi_k a_k$ conjugates, because there is one conjugate of the whole compound for each combination of conjugates of the portions. For clarity in the notation for this argument, the index k will always refer to a portion of the compound; the index i to a conjugate form of a portion; and the index j to a whole conjugate, i.e., a conjugate of the whole compound.

Each whole conjugate corresponds to a particular choice (combination) of conjugate forms in the independent portions of the compound. For the whole conjugate j, the symbol w(j, k) will indicate which of the possible forms of portion k occurs in this conjugate. The energy of a whole conjugate is denoted as $E_c(j)$

The requirement that independent portions of the molecule exert their influence on the total energy independently means that the same energy for the compound should be obtainable in two ways:

(a) E is a combination of the whole-compound conjugates:

$$\exp(-A^{-1}E) = \sum_j \exp[-A^{-1}E_c(j)] \qquad [5]$$

with the energy of each whole conjugate being a summation of the energies of the appropriate conjugate forms of the portions,

$$E_c(j) = \sum_k E_{pc}[k; w(j, k)] \qquad [6]$$

because the conjugates of the compound are merely assemblies of the conjugate forms of the portions.

(b) E is a summation of the hybrid energies of the portions of the molecule,

$$E = \sum_{k} E_p(k) \qquad [7]$$

with the hybrid energy of each portion k obtained by combining (using the proposed empirical formula) the energies of the conjugate forms of this portion:

$$\exp [-A^{-1} E_p(k)] = \sum_{i} \exp [-A^{-1} E_{pc}(k; i)] \qquad [8]$$

To prove the equivalence of the two schemes, we will start from recipe (a) above, and transform it to show that it yields the same result as recipe (b).

Combining the two formulae of recipe (a) by substituting $E_c(j)$ from [6] into [5], we have:

$$\exp (-E/A) = \Sigma_j \exp [-A^{-1} \Sigma_k E_{pc} (k; w (j, k))] \qquad [9]$$

$$\Rightarrow \exp (-E/A) = \Sigma_j \Pi_k \exp [-A^{-1} E_{pc} (k; w (j, k))] \qquad [10]$$

Note now that if an summation is to be carried out over all the whole–compound conjugates, it can be denoted as applied for j=1, ..., a, with w (j, k) providing the relationship of the whole conjugate j to each portion–conjugate k=1, ..., m; the summation can also be applied over each combination of indices i_1, i_2, ..., i_m with each i_k indicating which conjugate form of portion k is being used, i.e., i_k = w (j, k). In other words, we can either number whole conjugates with j and derive the appropriate portion–conjugates with w (j, 1), w (j, 2), ..., 2 (j, m) or we can number the whole conjugates directly with the combination of indices i_1, i_2, ..., i_m and omit the index j altogether. Given a function f of the indices i_1, i_2, ..., i_m (the domain of the function essentially being the space of whole conjugates), a summation of f over all whole conjugates can thus be written in two forms:

$$\sum_{j=1}^{a} f (w(j, 1), w (j, 2), ..., w (j, m)) = \sum_{i_1=1}^{a_1} \sum_{i_2=1}^{a_2} \cdots \sum_{i_m=1}^{a_m} f (i_1, i_2, ..., i_m) \qquad [11]$$

It is easy to see that each of the two sums contains precisely $a=\Pi_k a_k$ terms, and each term is an application of the function f on one of the $\Pi_k a_k=a$ possible combinations of indices $i_j=1, ..., a_1$; $i_2=1, ..., a_2$; ...; $i_m=1, ..., a_m$.

Thus, in Eq. [10] the summation over j can be converted to a multiple summation, and i_k can be substituted for w (j, k):

$$\exp (-E/A) = \sum_{i_1=1}^{a_1} \sum_{i_2=1}^{a_2} \cdots \sum_{i_m=1}^{a_m} \prod_{k=1}^{m} \exp [-A^{-1} E_{pc} (k; i_k)] \qquad [12]$$

A lemma provided with its proof in Appendix A allows the conversion of the multiple summation and multiplication into a simpler form. The lemma states that, for any function g (x_1, x_2):

$$\sum_{i_1=1}^{a_1} \sum_{i_2=1}^{a_2} \cdots \sum_{i_m=1}^{a_m} \prod_{k=1}^{m} g (k, i_k) = \prod_{k=1}^{m} \sum_{i=1}^{a_k} g (k, i) \qquad [13]$$

Application of the lemma [13] on the form [12] yields:

$$\exp (-E/A) = \prod_{k=1}^{m} \sum_{i=1}^{a_k} \exp [-A^{-1} E_{pc} (k; i)] \qquad [13]$$

Each factor in the right-hand-side product has the form

$$\sum_{i=1}^{a_k} \exp [-A^{-1} E_{pc} (k; i)] \qquad [14]$$

which represents precisely the combination of the conjugate forms of portion k, isolated from the rest of the molecule (Eq. [8]). Hence:

$$\sum_{i=1}^{a_k} \exp\ [-A^{-1} E_{pc}(k; i)] = \exp [-A^{-1} E_p(k)] \qquad [15]$$

where $E_p(k)$ is the energy of portion k viewed as separate from (and not interacting in any way with) the rest of the molecule. The combination function for the portion is used in precisely the same form as the combination function of the whole molecule. For each k, one can substitute Equation [15] into [14], obtaining:

$$\exp(-E/A) = \prod_{k=1}^{m} \exp(-A^{-1} E_p(k)) \Rightarrow \exp(-A^{-1}E) = \exp [-A^{-1} \sum_{k=1}^{m} E_p(k)] \qquad [16]$$

$$\Rightarrow E = \sum_{k=1}^{m} E_p(k) \qquad [17]$$

The last form completes the proof. It shows that the energy obtained by combining whole-compound conjugates is the same as the energy which is obtained if the conjugates of each portion of the molecule are first combined separately to yield the energy of the hybrid portion $E_p(k)$ and the whole-compound energy is taken as the sum of the portion energies.

The function is written with each conjugate represented by a distinct term. In practice, however, there will be many equivalent conjugates, and it is more convenient to write the combination in terms of *types* of conjugates. If there are N_i conjugates with energy E_i then:

$$\exp(-E/A) = \Sigma_i N_i \exp(-E_i/A) \Leftrightarrow \Sigma_i N_i \exp[(E-E_i)/A] = 1 \qquad [18]$$

If the dominant conjugate are assigned the index 1, the function can be written to show the effect of the recessive conjugates as a correction on the energy of the dominant conjugate:

$$E = E_1 - A \ln \{\Sigma_i N_i \exp[(E_1-E_i)/A]\} \qquad [19]$$

Under the assumption that there is only one dominant conjugate ($N_1=1$), the sum can also be written as:

$$E = E_1 - A \ln \{1+\Sigma_{i\neq 1} N_i \exp[(E_1-E_i)/A]\} \qquad [20]$$

If the effect of the recessive conjugates is small compared to that of the dominant conjugate, the function can be simplified. Specifically, with only one dominant conjugate ($N_1=1$) and under the assumption that:

$$N_i \exp[(E_1-E_i)/A] \ll 1 \quad (\text{for } i\neq 1) \qquad [21]$$

one obtains

$$A \ln [1 + \Sigma_{i\neq 1} N_i \exp[(E_1-E_i)/A]] \approx A \Sigma_{i\neq 1} N_i \exp[(E_1-E_i)/A \qquad [22]$$

$$\Rightarrow E = E_1 - A \Sigma_{i\neq 1} N_i \exp[(E_1-E_i)/A] \Rightarrow E = E_1 - \Sigma_{i\neq 1} N_i \{A \exp[(E_1-E_i)/A]\} \qquad [23]$$

$$\Rightarrow E = E_1 - \Sigma_{i\neq 1} N_i e_i \qquad [24]$$

where:

$$e_i = A \exp[(E_1-E_i)/A] \qquad [25]$$

The parameter e_i is influenced by E_1-E_i (the difference between the recessive conjugate i and the dominant conjugate 1) and A (the scaling parameter for this energy difference). The linearity of Eq. [24] facilitates the regression.

This entirely empirical function, in its linear form [25] and non-linear form [18], yields satisfactory results for the studies that have been carried out to date, but it may be inadequate for more complex cases. It has been shown that other means of combining energies of the conjugates can be developed through quantum mechanical analysis [Mavrovouniotis, 1990]. The quantum mechanical analysis takes into account structural similarity, as evidenced in the overlap of conjugate forms, while only energetic similarity is actually used in the proposed empirical combination function.

4. An Example

The example that is presented here entails the enthalpy of formation of alkanes. In a previous study, [Mavrovouniotis, 1990], data from [Reid *et al.*, 1987] were employed, under the assumption that these are indeed experimental data. A subsequent more extensive literature search revealed that a portion of the data in [Reid *et al.*, 1987] were actually estimated using group-contribution and other additive methods; some of the sources from which Reid *et al.* obtained their data had simply included estimated data along with experimental data, without clearly labelling them as such. This is not an uncommon problem in the literature of thermodynamic data and estimation methods. In the study presented here, only a reliable subset of the experimental data in [Cox and Pilcher, 1970] is used.

As discussed in Section 2, the dominant conjugate and the HCCC-conjugates are the most important for alkanes. Figs. 1 and 2 present evidence that conjugation bears a quantitative relationship to the enthalpy of formation. First, plots of the enthalpy of formation as a function of the number of HCCC conjugates, for alkanes with a fixed number of carbons, show a strong linear correlation (Fig. 1), in accordance with Eq. [24]. Second, if all effects of the degree of branching on the enthalpy are captured by the number of HCCC conjugates, one would expect alkanes with a fixed number of HCCC conjugates but varying number of carbon atoms, n, to have enthalpies of formation linear in n, regardless of their degree of branching. In other words, if the effect of conjugation captured by the number of HCCC conjugates is fixed, differences in enthalpies are only due to differences in the numbers of bonds; hence the enthalpies must be linear in n (Eq. [24]). This is exactly what is observed in Fig. 2.

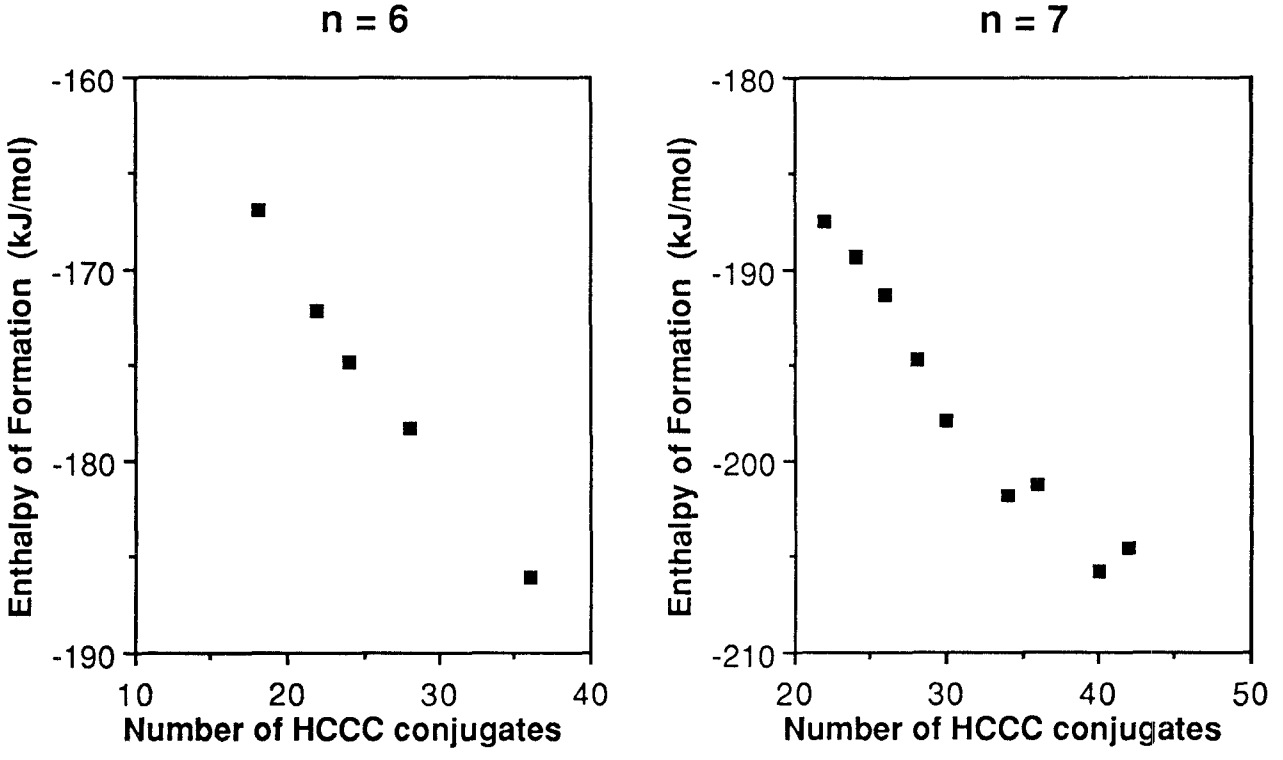


Fig. 1. Enthalpy of formation [Cox and Pilcher, 1970] of classes of isomeric alkanes, as a function of the number of HCCC conjugates. Each graph represents alkanes with a fixed number of carbon atoms (n).

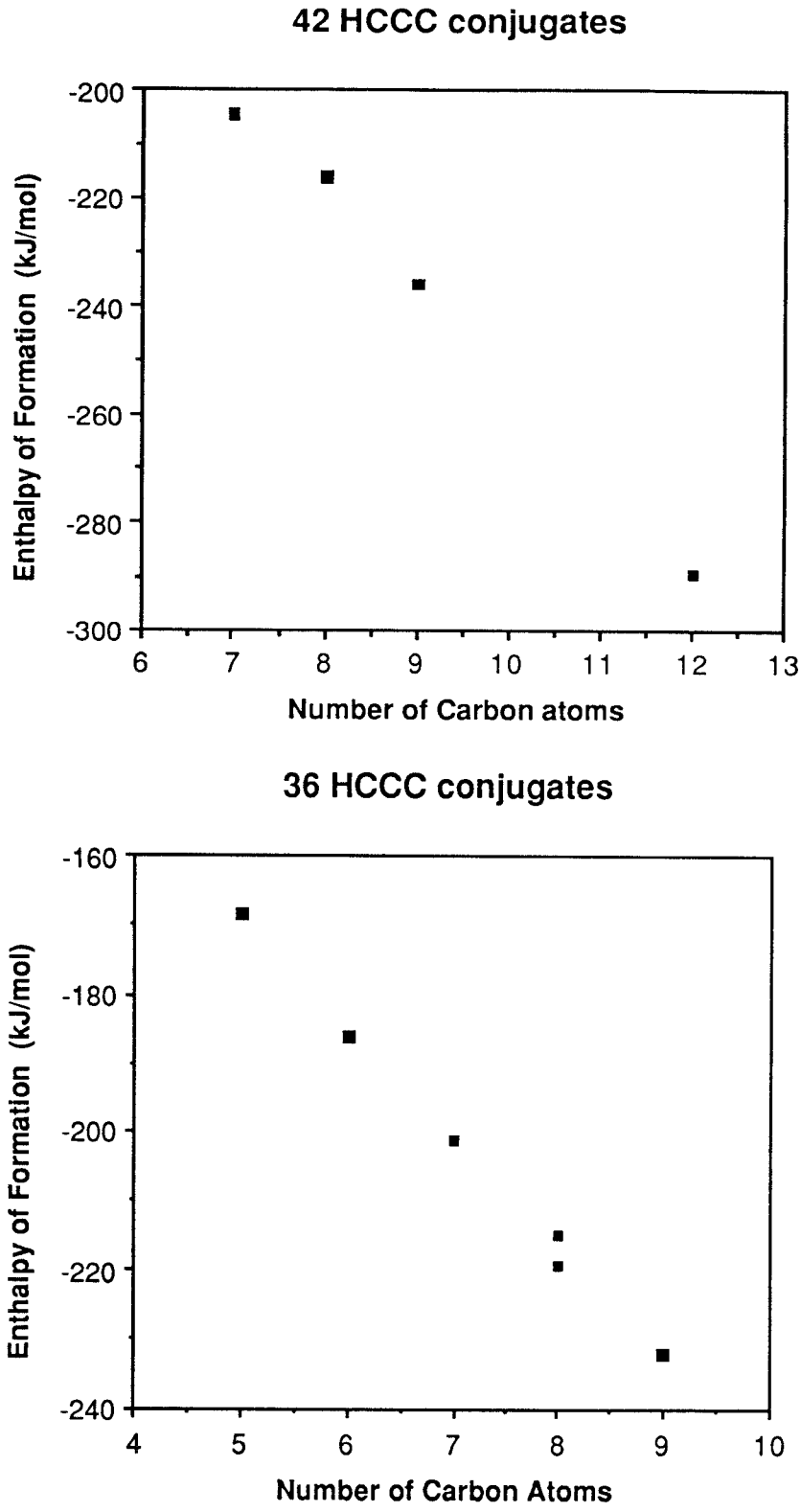


Fig. 2. Enthalpy of formation [Cox and Pilcher, 1970] of alkanes with the same number of HCCC conjugates, as a function of their number of carbon atoms. Each graph represents alkanes with a fixed number of HCCC conjugates.

These observations indicate the potential of a quantitative framework for estimating enthalpies of formation from the number of conjugates. For a class as restricted as alkanes, it is not possible to determine contributions of atoms and bonds, because the contributions are not independent. Thus the enthalpy of formation of the dominant conjugate is simply taken to be a linear function in n, the number of carbons:

$$E_i \text{ (dominant conjugate)} = b\,n + d \qquad [26]$$

The enthalpy of formation of an HCCC-conjugate differs from the dominant conjugate by a constant term independent of the number of carbons:

$$E_i \text{ (HCCC-conjugate)} = b\,n + d + h \qquad [27]$$

If the number of HCCC-conjugates is N, the enthalpy of formation of an alkane is given by:

$$\Delta H_f = d + b\,n - c\,N \qquad [28]$$

where the approximation that the effect of the recessive conjugates is small has been made (Eq. [24]). In this simple expression, the first two terms represent the effect of the dominant conjugate, while the third term represents all the (recessive) HCCC-conjugates.

The determination of the adjustable parameters is based on linear regression with 55 experimental data points from [Cox and Pilcher, 1970]:

$$\Delta H_f = -49.4 - 18.23\,n - 0.587\,N \quad \text{(kJ/mol)}$$

The standard deviation of the regression is 3.7 kJ/mol and the average absolute error 2.9 kJ/mol. For comparison, we also carried out a group-contribution fit on the same set of compounds, using the groups >C<, >CH–, $>CH_2$, and $-CH_3$. This four-parameter group-contribution scheme has a standard deviation of 4.6 kJ/mol and an average absolute error of 3.8 kJ/mol. Even though ABC uses fewer adjustable parameters (only three parameters vs. four for group contributions), it performs better than group contributions in this example. The accuracy of ABC approaches the experimental error, which is generally 1 to 2 kJ/mol [Cox and Pilcher, 1970], even though only the most important three-bond conjugates were used.

5. Discussion

The objective of the work presented here is the development of a new computer-based approach for the estimation of properties of organic compounds from their molecular structure. The proposed technique combines the basic engineering philosophy of group contributions with an empirical organic-chemistry view of compounds as hybrids of conjugate forms and a quantum-mechanical framework [Mavrovouniotis, 1990]. This approach enriches the field of property estimation with new concepts that can lead to techniques with higher accuracy and sound physical interpretation. The approach is based on the contributions of atoms and bonds in the properties of conjugate forms of a molecular structure.

One non-trivial computational aspect of the method is the generation of the conjugates. While the use of contributions from both atoms and bonds in the estimation of properties can be carried out manually in a straightforward fashion, it is not possible to generate manually the vast set of conjugates of any given molecular structure. The generation and analysis of a fairly large number of conjugates must be computer-based. The development and implementation of suitable algorithms for this task is facilitated by symbolic computing environments and Object-Oriented Programming, which allow the flexible representation and manipulation of molecular structures. Atoms, bonds, molecules, electron pairs, and other entities can be represented as interconnected objects within an Object-Oriented Programming (OOP) computation paradigm. The generation, comparison, and analysis of conjugates can be carried out through computer-based manipulation of the objects and their interconnections. One only needs to encode the operators which produce one conjugate from another, the canonical forms for the comparison of conjugates (to avoid duplicates), and the rules for eliminating high-energy conjugates. The Object-Oriented representation and generation of conjugate forms is currently being implemented, using Common Lisp and the Common Lisp Object System (CLOS). OOP will play an important role in the systematic application of the technique in the analysis and design of complex systems. For example, in an investigation of a complex set of reactions (e.g., fluid catalytic cracking), OOP would allow the generation of appropriate reactions and intermediates based on information about the operative reaction mechanisms, as well as the automated application of ABC for the estimation of properties.

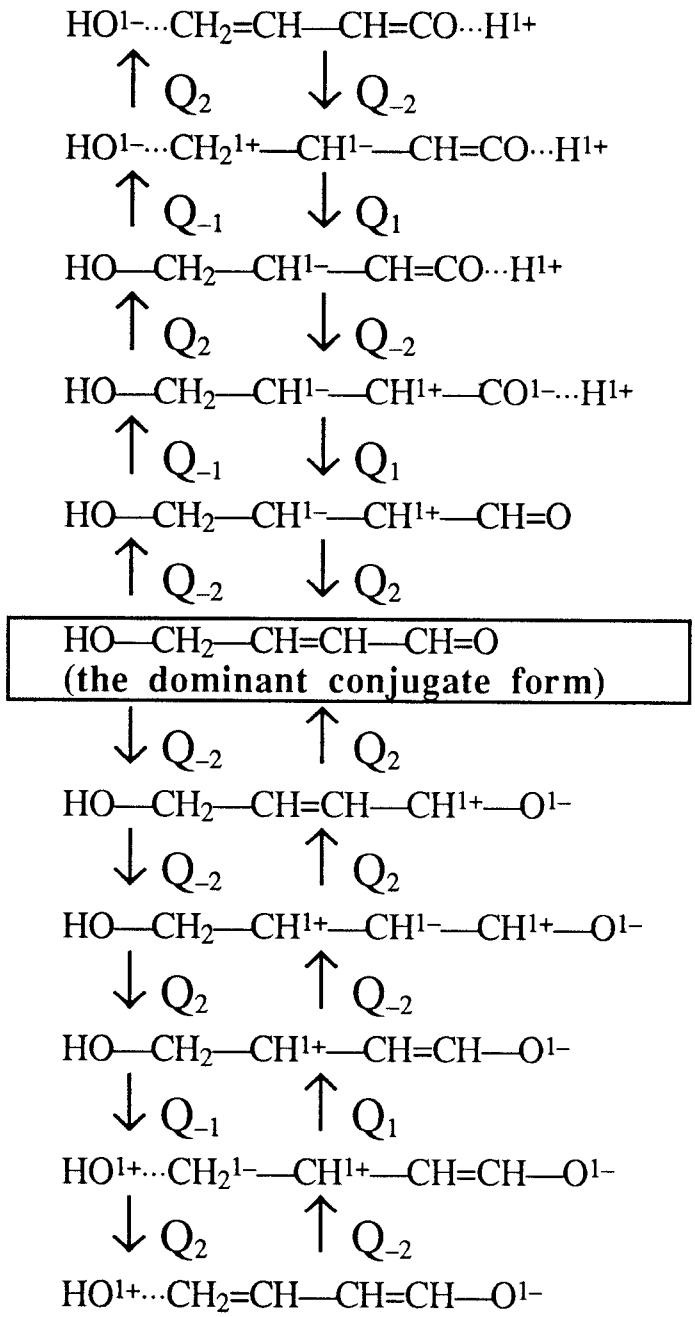


Fig. 3. Conjugates generated by bond-forming operators, Q_i, and bond-dissociating operators, Q_{-i}.

The generation of conjugates can in principle be accomplished through a set of operators each of which constructs at most one bond or dissociates at most one bond. The basic bond-forming and bond-dissociating operators are denoted as Q_i and Q_{-i}. The operator Q_i constructs a bond of order i from a bond of order i–1, between adjacent atoms; the operator Q_{-i} dissociates (heterolytically) a bond of order i into a bond of order i–1. Naturally, each operator is only applicable under certain conditions. As an example, Fig. 3 shows the formation of conjugates through these operators. However, this set of operators is not *monotonic*: It does not proceed from more important conjugates to less important ones. In the algorithms we are developing, we place great emphasis on the order in which conjugates are generated. Ideally, one would like to generate them monotonically, proceeding always from lower-energy conjugates to higher-energy ones. This would allow the generation of precisely that set of conjugates which is necessary for the desired precision of the calculation. This process is akin to the determination of the value of a parameter from a power series, truncating at the appropriate term. The key in accomplishing this

monotonic generation of conjugates lies in restricting the set of operators and the conditions under which they can be applied.

We are also developing a quantum chemical version of the method, based on the preliminary analysis in [Mavrovouniotis, 1990]. In this interpretation, quantities are still determined additively, as summations or products of a small number of parameters describing the energies and overlaps of bonds and atomic charges. It should be noted that the quantum-mechanical treatment entails not only properties of individual conjugates, but also pairwise overlaps of conjugates. This requirement further necessitates an object-oriented implementation, within which algorithms can generate conjugates and determine their properties and overlaps.

Ideal-gas properties are the first focus of the method, because they do not entail intermolecular interactions. The enthalpy of formation is a particularly good starting point, because it depends directly on the distribution of the electron cloud in the molecule. The electron cloud can indirectly account even for the three-dimensional conformation of the molecule and vibrational and rotational contributions to the enthalpy. For example, variations in the effect of bond stretching can be attributed to the electron-density or fractional order of the bond, which can be modelled through conjugation.

To appreciate the significance of the enthalpy of formation, consider its relationship to reaction selectivity. For example, in the cracking reaction of a carbocation (one of the reactions taking place in Fluid Catalytic Cracking), many paths may be possible:

Path 1: $R–CH_2–CH^+–CH_2–R' \rightarrow R^+ + CH_2=CH–CH_2–R'$

Path 2: $R–CH_2–CH^+–CH_2–R' \rightarrow R–CH_2–CH=CH_2 + R'^+$

The reaction selectivity is determined by the structure of the transition state and the energy and entropy barriers for the alternative routes. For such reactions, the transition state has a structure that is, in appearance, an interpolation between the structure of the reactant (which is the same for the two paths) and the products. Thus, the energy of each transition state correlates to that of the respective product, giving rise to the established rule: The path leading to the lowest-energy ion will be preferred; if R^+ is a primary ion and R'^+ a secondary one, path 2 will dominate. Similar observations can be made for many reaction classes, for which the selectivity is determined by the energies of the products. Group contribution methods are especially unsuitable for ions, because the dispersion of the charge (which is formally placed on one atom) to several atoms is a major factor affecting the stability and properties of an ion. The conjugation that leads to the delocalization of charge is not taken into account by group contributions. The ABC approach is quite suitable for estimating properties of ions; these can in turn lead to correlations for chemical properties.

For the estimation of other physical properties, we plan to examine fractional charges on individual atoms of the compound, and electron densities of individual bonds (either empirically or quantum-mechanically). One could then use traditional group-additivity for a variety of properties, but with the contributions considered to be functions of the fractional charges and bond densities. For example, the contribution of a quaternary carbon group >C< to some physical property would not be a constant, but a function of its fractional charge (which depends on the whole structure of the molecule); the contribution of a carbon-carbon bond would be similarly a function of its density or fractional order.

In summary, the modelling of chemical process systems requires the estimation of physical, thermodynamic, and chemical properties of pure compounds and mixtures. With its engineering orientation combined with a strong chemical foundation, the ABC technique may enable simple yet accurate estimation of these properties, and therefore improve the analysis and design of products and processes.

Acknowledgements

This research was supported in part by NSF grant CTS-9010549 and by the Systems Research Center, an engineering research center funded by NSF under grant CDR-8803012.

References

Benson, S. W. *Thermochemical Kinetics;* Wiley: New York, 1968.

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

Domalski, E. S.; Hearing, E. D. "Estimation of the Thermodynamic Properties of Hydrocarbons at 298.15 K". *J. of Phys. and Chem. Ref. Data* **14**, 1637-1678, 1988.

Fliszar, S. *Charge Distributions and Chemical Effects*; Springer-Verlag: New York, 1983.

Joback, K. G.; Reid, R. C. "Estimation of Pure-Component Properties from Group Contributions". *Chem. Eng. Comm.* **57**, 233-243, 1987.

Joback, K.G.; Stephanopoulos, G. "Designing Molecules Possessing Desired Physical Property Values". *Proceedings FOCAPD'89*, CACHE Corporation: Austin, Texas, 1989.

Mavrovouniotis, M. L.; Bayol, P.; Lam, T.-K. M.; Stephanopoulos, G.; Stephanopoulos, G. "A Group-Contribution Method for the Estimation of Equilibrium Constants of Biochemical Reactions". *Biotechnology Techniques* **2**, 23-28, 1988.

Mavrovouniotis, M. L. "Estimation of Properties from Conjugate Forms of Molecular Structures: The ABC Approach". *Industrial and Engineering Chemistry Research*, **29**, 1943-1953, 1990.

Mavrovouniotis, M. L. "Group Contributions to the Gibbs Energy of Formation of Biochemical Compounds in Aqueous Solution". *Biotechnology and Bioengineering*, **36**, 1070-1082, 1990a.

Morrison, R. T.; Boyd, R. N. *Organic Chemistry; 3rd edition;* Allyn and Bacon: Boston, 1973.

Reid, R. C.; Prausnitz, J. M.; Poling, B. E. *The Properties of Gases and Liquids; 4th edition;* McGraw-Hill: New York, 1987.

Appendix A

Lemma: For any function $g(x_1,x_2)$, any positive integer m, and any sequence of positive integers $a_k(k=1,..,m)$

$$\sum_{i_1=1}^{a_1}\sum_{i_2=1}^{a_2}\cdots\sum_{i_m=1}^{a_m}\prod_{k=1}^{m} g(k,i_k) = \prod_{k=1}^{m}\sum_{i=1}^{a_k} g(k,i) \qquad [A1]$$

Proof: Carrying out induction on m, we examine first the case m = 1, for which the lemma trivially holds:

$$\sum_{i_1=1}^{a_1}\prod_{k=1}^{1} g(k,i_k) = \sum_{i=1}^{a_1} g(1,i) = \prod_{k=1}^{1}\sum_{i=1}^{a_k} g(k,i) \qquad [A2]$$

Assuming that the equality holds for m (in the form of [A1]), we will prove the corresponding equality for m+1, *starting from the right-hand side.*

$$\prod_{k=1}^{m+1}\sum_{i=1}^{a_k} g(k,i) = [\prod_{k=1}^{m}\sum_{i=1}^{a_k} g(k,i)]\sum_{i=1}^{a_{m+1}} g(m+1,i) \qquad [A3]$$

We use Eq. [A1] to substitute the m-factor product in the right-hand side of Eq. [A3], and we replace the dummy index i in the m+1 factor with i_{m+1}:

$$\prod_{k=1}^{m+1}\sum_{i=1}^{a_k} g(k,i) = [\sum_{i_1=1}^{a_1}\sum_{i_2=1}^{a_2}\cdots\sum_{i_m=1}^{a_m}[\prod_{k=1}^{m+1} g(k,i_k)]]\sum_{i_{m+1}=1}^{a_{m+1}} g(m+1,i_{m+1}) \qquad [A4]$$

The right-hand side represents a product of two summations and can be expanded to the form of a summation of the pairwise products. In other words, we can use the identity

$$(\sum_i x_i)(\sum_j y_j) = \sum_i\sum_j (x_i y_j) = \sum_j\sum_i (x_i y_j) \qquad [A5]$$

The fact that in [A4] one of the summations is actually a multiple summation does not affect the application of the identity, because a sum of sums is merely a longer sum of the same form, i.e., the same type of terms. Applying the identity [A5] on the right-hand side of [A4], we obtain:

$$\prod_{k=1}^{m+1}\sum_{i=1}^{a_k} g(k,i) = \sum_{i_1=1}^{a_1}\sum_{i_2=1}^{a_2}\cdots\sum_{i_m=1}^{a_m}\sum_{i_{m+1}=1}^{a_{m+1}}[g(m+1,i_{m+1})\prod_{k+1}^{m+1} g(k,i_k)] \qquad [A6]$$

This leads to the desired equality for m+1:

$$\prod_{k=1}^{m+1}\sum_{i_k=1}^{a_k} g(k,i_k) = \sum_{i_1=1}^{a_1}\sum_{i_2=1}^{a_2}\cdots\sum_{i_{m+1}=1}^{a_{m+1}}\prod_{k=1}^{m+1} g(k,i_k) \qquad [A7]$$

completing the induction.