GRAPH-THEORETIC CLUSTER EXPANSIONS. THERMOCHEMICAL PROPERTIES FOR ALKANES

Mary C. McHUGHES and R.D. POSHUSTA
Department of Chemistry, Washington State University, Pullman, Washington 99164-4630, USA

Abstract

Numerical studies of Klein's graph-theoretic cluster expansion ansatz (CEA) are made for several thermochemical properties and graph-theoretic ("topological") indices of alkanes. The ansatz expresses *any* molecular property in terms of unique numerical contributions arising from structural descriptors, subgraphs of the molecular structural formula, called clusters. The collection of cluster contributions comprises a unique fingerprint or signature characterizing each property. Signatures are pattern recognition tools which can be used to (1) identify the most significant structural descriptors for each property, and (2) analyze for similarity and dissimilarity among properties and indices. Visual renderings of CEA signatures display the structural origins of properties in an easily appreciated form.

1. Introduction

What accounts for the physicochemical properties of each pure substance and how may one design a substance with specific desired properties? What distinguishes one property from another and how does "chemical structure" find expression in each property? Two strategies have been followed to answer these fundamental questions. The *ab initio* approach adopts quantum chemistry and statistical mechanics to determine all properties from first principles using laborious computations. Although this approach provides a fundamental mechanistic understanding, it is incapable, with current computer technology, of calculating certain properties – boiling points of hydrocarbons or activities of carcinogens, to mention only two [1]. By contrast, computationally simple and mathematically unsophisticated empirical approaches are sanctioned through long usage in an unrestricted range of applications. It is this second, older approach that we examine here, namely, an empirical but systematic, analytical and objective analysis of extensive data sets seeking patterns which reveal "chemical structure information".

1.1. RECOGNIZING CHEMICAL STRUCTURE

Ultimately, all chemical concepts, including the most fundamental structural ones, are extracted from experimental data by some sort of "pattern recognition" procedure [2]. Experimental data for families of related molecules often exhibit

regularities which reveal the structural origin of physical properties. By this is meant that the properties are related to the arrangement of molecular structural units, e.g. atoms and functional groups. Chemical intuition posits that physical properties arise from contributions made by the constituent parts of a chemical structure. This idea has been extremely fruitful for all branches of chemistry, both for predictive ability and concept development. It provides the foundation of the very concept of a chemical bond.

To discover familial regularities and expose the corresponding structural origins has been a central theme of modern chemistry. The search for a recognizable pattern consists of careful measurements, followed by appropriate data reduction techniques. An early prototype for this paradigm is the expression for heats of formation in terms of "bond energies". Fajans introduced bond energy terms around 1920, assuming such quantities to be *additive* and *transferable* from one molecular structure to another. This met with such success that departures from the rule inspired the definition of new structural concepts: steric strain (destabilization) or resonance (stabilization) energies. In 1934, Zahn refined Fajans' scheme by adding terms associated with nearest-neighbor interactions. Increasingly specialized schemes have been developed, yielding more accurate predictive ability by introducing numerous but less readily interpreted parameters.

1.2. CHEMICAL GRAPH THEORY

As early as 1878, Sylvester recognized the intimate link between chemical structure and mathematical graph theory [3]. Most realizations of this linkage for pattern recognition in physicochemical data employed correlations with various "graph-theoretic invariants". For example, a variety of indices, exemplified by the Wiener number and the Randić index, have been correlated with boiling points, melting points, heats of combustion, carcinogenicity, drug-receptor interactions, etc. [4–7]. Graph theory is by now a respected approach for pattern recognition in chemistry: Trinajstić et al. [8] suggested that graph theory may offer more insight than a computerized numerical study (meaning quantum theory and statistical mechanics) when considering relationships between particular structural features and a single physicochemical property of a molecule; Gordon and Kennedy [9] state that the theory of graphs provides "a single systematic definition [which] contains . . . practically all that is useful in previously proposed additivity schemes for predicting standard thermochemical data". The extent of applicability of graph theory in chemistry is recounted in numerous reviews and monographs [10–13].

1.3. STRUCTURAL DESCRIPTORS

The outcome of pattern recognition analysis is a model in which to correlate and interpret the data. The traditional model for chemical structure contains descriptive elements: bonds, strain, resonance and the like. Each of these descriptors was introduced individually, on more or less subjective grounds, to refine the structural model. No mathematical imperative dictates these particular descriptors; rather, they are inspired

jointly by regularities in the data and subjective reasoning. Nevertheless, it is hardly imaginable that many of these traditional concepts could be eliminated from chemistry.

By adopting graph theory as the basis of structural models, we naturally include the bond descriptors and all the connection relations that follow with it. The possibility is open that new descriptors will appear in the course of graph-theoretic analyses. This prospect is highly dependent upon the specific form of the graph-theoretic model. For example, the various indices (Wiener or Randić) are an attempt to summarize all the structural relationships of a graph into a single number (descriptor) or a small collection of numbers (descriptors) [14]. Other, more refined descriptors, either systematic or arbitrary, have been suggested: the linear combinations of graph invariants (LCGI) of Gordon and Kennedy [15] are important examples.

Recently, efforts have been made to refine the structural paradigm in a rigorous mathematical manner. The goal is to introduce structural descriptors systematically as they are required to account for the data. Acceptable descriptors must satisfy the following criteria.

1.4. SELECTION CRITERIA

Comparison between pattern recognition models can be made rational by agreeing upon a set of selection criteria. For us, the following appear to be reasonable criteria:

- Objective rules are followed to introduce structural descriptors.
- Descriptors are necessary and sufficient, within the model, to account for all experimental data.
- It is also convenient, but hardly necessary, if traditional descriptors are naturally included in the model.

Various graph indices are essentially arbitrary; the collection of such indices does not form a rationally related whole, and it has not been shown how to select independent and complete sets of indices. Similarity indices [16] may follow rigorous mathematical principles, but they too are arbitrary and do not fit the necessary and sufficient criterion.

1.5. GRAPH-THEORETIC CLUSTER EXPANSIONS

The graph-theoretic cluster expansion offers such structural descriptors and provides, as well, the formal mathematics with which to verify and quantify statements comparing chemical structure to physical properties.

Graph-theoretic cluster expansions have been used to investigate a variety of physicochemical properties, and even biogenic activities [17,18]. Our approach is in the spirit of those just mentioned, using a graph-theoretic cluster expansion ansatz (CEA) (see the following sections for an explanation) advanced by Klein [19], which satisfies the criteria listed above. Successful previous applications of the CEA indicate a broad range of applicability for the method. Schmalz et al. [20] applied the method to the Hückel molecular energy of acyclics; similarly, Poshusta et al. [21] have applied it to

eigenvalues of the Pariser–Parr–Pople, Heisenberg, and Hubbard model Hamiltonians; other applications have been made to hydrogen atom chains and to metallic sodium [22,23]. All previous applications were restricted to only small clusters limited to graphs with small diameters and never with more than seven subgraphs. Our present applications are to data for clusters with diameters up to ten and with no restrictions on the induced subgraphs included. Thus, we are able to study the convergence of the CEA for very large expansions.

The CEA has many possible forms, according to the property being studied. Klein [19] has classified properties into constantive, additive, multiplicative and derivative. In each class, a different cluster expansion set is appropriate. The method applies not just to scalar properties, such as heats of formation, but also to Hamiltonian operators and quantum mechanical wave functions. Our applications are to simple scalar properties, thus avoiding the complications of cluster expanded operators or other non-scalar quantities. Our approach, like previous ones, is restricted to the additive type of expansion appropriate for so-called additive properties.

This CEA is recommended by several other useful features described by Klein. The CEA naturally gives rise to a hierarchy of descriptors called clusters, some of which lie in close correspondence to traditional chemical structural elements. Once the cluster expansion method has been selected, any property is rigorously and uniquely resolved into its cluster contributions, which are also necessary and sufficient.

The present work studies the CEA through numerical analysis of several example properties. We choose to cluster expand thermochemical properties of saturated hydrocarbons. Alkanes form a particularly well-documented family of chemical compounds and, as such, provide an opportunity to extend the empirical approach to the deepest level and most detailed structural pattern. We also consider a few mathematical simulated "properties". Our goals are to apply the CEA to large data sets and large clusters, and to possibly identify new, important structural descriptors. Concurrently, we test the CEA convergence rates and compare and contrast properties according to the cluster expansion results.

In the next section we review, very briefly, chemical graph theory to establish notation and to introduce the WAV coding of graphs used to represent them. Following this, we describe the cluster expansion ansatz, identify its structural descriptors and discuss normalization of these descriptors, define additive properties and present convergence criteria. Next, we cite the sources of our data and discuss the data's suitability and limitations for our purposes. Then we display the results of the cluster expansions, identifying significant descriptors and features. Finally, we discuss these results and draw conclusions.

2. Graph-theoretic terminology

Graph-theoretical notation and nomenclature are not uniform; however, the following basic concepts and definitions are well accepted. Let $\Gamma(V, E)$ denote a *graph* whose vertex set is V and whose edge set is E. We then think of a graph as a collection

of vertices (atoms) which are joined together by edges (bonds). We take the term graph to imply a finite, undirected structure without loops or multiple edges.

The *valence* of a vertex v is analogous to the valency of an atom. It is defined as the number of edges incident at that vertex. Our results apply to *trees*, which are graphs without cycles. We say that γ is a *subgraph* of Γ and write $\gamma \subseteq \Gamma$ if $\gamma = (V', E')$ with $V' \subseteq V$ and $E' \subseteq E$. For any set S of vertices in Γ, the induced subgraph $\langle S \rangle$ is defined as that subgraph of Γ with vertex set S such that two vertices are adjacent in $\langle S \rangle$ if and only if they are adjacent in Γ. Whenever the term subgraph appears in this work, induced subgraph is implied. That is, we are concerned only with connected graphs and subgraphs. A graph is said to be *connected* if there is at least one path between all pairs of vertices. A *path* from i to j is a sequence of edges $(i, k), (k, l), \ldots, (m, n), (n, j)$ commencing with i and terminating with j. The *length* of such a path is the number of edges it contains. The length of the shortest path from i to j is the *distance* from i to j. The *diameter* of a graph is the largest distance existing in that graph.

It is convenient for subsequent discussions to define the following special classes of graphs: chain, complete bigraph, star and binary star. *Chains* are trees with no branches. A *complete bigraph* Γ is a graph whose vertex set V can be partitioned into two subsets V_1 and V_2 and contains every edge joining V_1 and V_2 but no edges between pairs of vertices in the same set. If V_1 and V_2 have m and n edges, then $\Gamma = K_{m,n}$. A *star* is a complete bigraph $K_{1,n}$. A *binary star* is two stars with only one common edge.

One mathematical representation of graphs is their adjacency matrix, familiar to chemists from Hückel MO theory. This matrix is formed by taking the vertex labels as the indices for the rows and columns. Entries are either zero, if the row and column vertices are not connected, or one in the case that they are connected, that is, adjacent. Note that the adjacency matrix of a graph is not unique. There are many ways to label the vertices and hence order the rows and columns of the adjacency matrix. This leads to considering graph isomorphism. Two graphs Γ_1 and Γ_2 are isomorphic if there is a one-to-one mapping from the vertices of Γ_1 to the vertices of Γ_2 that preserves the adjacency relationship. Also useful is the distance matrix: D_{ij} is the length of the shortest path between vertices i and j.

We say that γ is embedded in Γ if and only if γ is isomorphic to a subgraph of Γ. It may be that γ is isomorphic to several subgraphs of Γ, in which case we are led to define $n(\Gamma, \gamma)$ to be the number of subgraphs of Γ isomorphic to γ. We also say that γ is embedded in Γ $n(\Gamma, \gamma)$ times. We call $n(\Gamma, \gamma)$ the frequency of γ embedded in Γ.

3. Trees and their WAV codes

We have selected Read's walk around valence code (WAV) [24] to represent the 201 trees with up to 10 vertices used in this work. Table 1 lists the codes used in both theoretical and experimental cluster expansions. The code and its use in our computer program are discussed in more detail in our previous paper [25]. Read describes how to decode the WAV codes. For example, consider the code for 2, 3-dimethylpentane:

Table 1

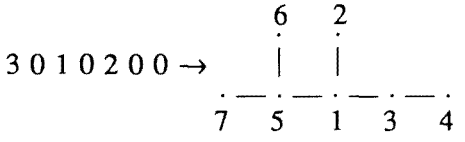
Walk around valency codes (WAV). The first column gives the sequence number for all trees with up to ten vertices, while the second column gives the sequence number for all alkanes up to ten carbons

Sequence numbers		Codes	Sequence numbers		Codes
Trees with up to 10 vertices	Alkanes with up to 10 carbons		Trees with up to 10 vertices	Alkanes with up to 10 carbons	
1	1	0	51	43	2 1 2 0 1 0 1 1 0
2	2	1 0	52	44	2 2 0 1 1 0 1 1 0
3	3	2 0 0	53	45	2 1 1 0 2 1 0 1 0
4	4	2 1 0 0	54	46	3 1 0 1 1 0 1 1 0
5	5	3 0 0 0	55	47	2 1 2 0 0 1 2 0 0
6	6	2 1 0 1 0	56	48	2 2 1 0 2 0 0 1 0
7	7	2 2 0 0 0	57	49	3 0 1 1 0 1 2 0 0
8	8	4 0 0 0 0	58	50	3 0 1 1 0 2 0 1 0
9	9	2 1 1 0 1 0	59	51	3 2 1 0 1 0 0 1 0
10	10	2 1 0 2 0 0	60	52	2 1 2 0 0 2 0 1 0
11	11	3 0 1 0 1 0	61	53	2 2 0 1 0 2 0 1 0
12	12	3 2 0 0 0 0	62	54	2 2 1 0 1 0 2 0 0
13	13	2 3 0 0 0 0	63	55	2 1 1 0 2 0 2 0 0
14		5 0 0 0 0 0	64	56	2 1 1 0 1 3 0 0 0
15	14	2 1 1 0 1 1 0	65	57	2 1 1 0 3 0 0 1 0
16	15	2 1 2 0 0 1 0	66	58	2 3 0 1 0 1 0 1 0
17	16	2 2 0 1 0 1 0	67	59	4 0 0 1 1 0 1 1 0
18	17	3 1 0 1 0 1 0	68	60	4 1 0 1 0 1 0 1 0
19	18	2 2 0 0 2 0 0	69	61	2 2 0 2 0 0 2 0 0
20	19	3 0 1 0 2 0 0	70	62	3 2 0 2 0 0 0 1 0
21	20	2 1 0 3 0 0 0	71	63	3 1 0 2 0 0 2 0 0
22	21	4 0 0 1 0 1 0	72	64	2 1 3 0 0 0 2 0 0
23	22	3 3 0 0 0 0 0	73	65	2 3 0 0 2 0 0 1 0
24		2 4 0 0 0 0 0	74	66	3 1 0 1 0 3 0 0 0
25		6 0 0 0 0 0 0	75	67	3 1 3 0 0 0 0 1 0
26	23	2 1 1 1 0 1 1 0	76	68	3 3 0 0 1 0 0 1 0
27	24	2 1 1 0 1 2 0 0	77	69	4 0 1 0 1 0 2 0 0
28	25	2 1 1 0 2 0 1 0	78	70	2 2 0 3 0 0 0 1 0
29	26	3 0 1 1 0 1 1 0	79	71	2 3 0 0 1 0 2 0 0
30	27	2 2 1 0 1 0 1 0	80		2 1 4 0 0 0 0 1 0
31	28	2 1 2 0 0 2 0 0	81		2 4 0 0 0 1 0 1 0
32	29	3 1 0 1 0 2 0 0	82		5 0 0 1 0 1 0 1 0
33	30	3 2 0 1 0 0 1 0	83	72	4 0 0 2 0 0 2 0 0
34	31	2 2 0 1 0 2 0 0	84	73	3 0 2 0 0 3 0 0 0
35	32	2 2 0 2 0 0 1 0	85	74	2 3 0 0 0 3 0 0 0
36	33	2 1 3 0 0 0 1 0	86	75	4 0 0 1 0 3 0 0 0
37	34	2 3 0 0 1 0 1 0	87		2 2 0 0 4 0 0 0 0
38	35	4 0 1 0 1 0 1 0	88		3 0 1 0 4 0 0 0 0
39	36	3 0 2 0 0 2 0 0	89		5 0 0 0 1 0 2 0 0
40	37	2 2 0 0 3 0 0 0	90		2 1 0 5 0 0 0 0 0
41	38	3 0 1 0 3 0 0 0	91		6 0 0 0 0 1 0 1 0
42	39	4 0 0 1 0 2 0 0	92		4 4 0 0 0 0 0 0 0
43		2 1 0 4 0 0 0 0	93		3 5 0 0 0 0 0 0 0
44		5 0 0 0 1 0 1 0	94		2 6 0 0 0 0 0 0 0
45	40	4 3 0 0 0 0 0 0	95		8 0 0 0 0 0 0 0 0
46		3 4 0 0 0 0 0 0	96	76	2 1 1 1 1 0 1 1 1 0
47		2 5 0 0 0 0 0 0	97	77	2 1 1 1 0 1 1 2 0 0
48		7 0 0 0 0 0 0 0	98	78	2 1 1 1 0 1 2 0 1 0
49	41	2 1 1 1 0 1 1 1 0	99	79	2 1 1 1 0 2 0 1 1 0
50	42	2 1 1 2 0 0 1 1 0	100	80	3 0 1 1 1 0 1 1 1 0

Table 1 (continued)

Sequence numbers		Codes	Sequence numbers		Codes
Trees with up to 10 vertices	Alkanes with up to 10 carbons		Trees with up to 10 vertices	Alkanes with up to 10 carbons	
101	81	2 2 1 0 1 1 0 1 1 0	151	131	2 1 1 0 2 0 3 0 0 0
102	82	2 1 2 1 0 1 0 1 1 0	152	132	2 1 1 0 3 0 0 2 0 0
103	83	3 1 1 0 1 1 0 1 1 0	153	133	4 1 0 1 0 1 0 2 0 0
104	84	2 1 1 2 0 0 1 2 0 0	154		2 1 1 0 1 4 0 0 0 0
105	85	2 1 2 0 1 0 0 1 1 0	155		2 1 1 0 4 0 0 0 1 0
106	86	2 2 0 1 0 2 1 0 1 0	156		2 4 0 0 1 0 1 0 1 0
107	87	2 2 0 1 1 0 1 2 0 0	157		5 0 0 0 1 1 0 1 1 0
108	88	2 2 0 1 1 0 2 0 1 0	158		5 0 1 0 1 0 1 0 1 0
109	89	2 2 0 1 2 0 0 1 1 0	159	134	3 2 0 0 2 0 0 2 0 0
110	90	2 2 0 2 0 1 0 1 1 0	160	135	3 2 0 2 0 0 0 2 0 0
111	91	3 0 1 1 0 2 1 0 1 0	161	136	2 2 0 2 0 0 3 0 0 0
112	92	3 1 0 1 1 0 1 2 0 0	162	137	2 2 0 3 0 0 0 2 0 0
113	93	3 1 0 1 1 0 2 0 1 0	163	138	2 3 0 0 2 0 0 2 0 0
114	94	3 1 1 0 1 1 0 2 0 0	164	139	3 2 0 3 0 0 0 0 1 0
115	95	3 2 0 1 1 0 0 1 1 0	165	140	3 3 0 0 1 0 0 2 0 0
116	96	2 1 2 0 0 2 1 0 1 0	166	141	3 3 0 0 2 0 0 0 1 0
117	97	2 1 2 0 1 0 1 2 0 0	167	142	4 0 1 0 2 0 0 2 0 0
118	98	2 1 2 0 1 0 2 0 1 0	168	143	3 1 0 2 0 0 3 0 0 0
119	99	2 1 1 0 2 1 0 2 0 0	169	144	2 1 3 0 0 0 3 0 0 0
120	100	3 2 1 0 1 0 1 0 1 0	170	145	4 3 0 0 1 0 0 0 1 0
121	101	2 1 1 0 3 0 1 0 1 0	171	146	2 3 0 0 1 0 3 0 0 0
122	102	2 1 1 3 0 0 0 1 1 0	172	147	2 3 0 0 3 0 0 0 1 0
123	103	2 1 3 0 0 1 0 1 1 0	173	148	4 0 1 0 1 0 3 0 0 0
124	104	2 3 0 0 1 1 0 1 1 0	174		2 1 4 0 0 0 0 2 0 0
125	105	4 0 1 0 1 1 0 1 1 0	175		2 4 0 0 0 2 0 0 1 0
126	106	2 3 1 0 1 0 1 0 1 0	176		3 1 0 1 0 4 0 0 0 0
127	107	3 0 1 2 0 0 1 2 0 0	177		3 1 4 0 0 0 0 0 1 0
128	108	2 1 2 0 0 2 0 2 0 0	178		3 4 0 0 0 1 0 0 1 0
129	109	2 2 2 0 0 2 0 0 1 0	179		5 0 0 1 0 1 0 2 0 0
130	110	3 2 1 0 2 0 0 0 1 0	180		2 2 0 4 0 0 0 0 1 0
131	111	3 0 1 2 0 0 2 0 1 0	181		2 4 0 0 0 1 0 2 0 0
132	112	3 0 2 0 1 0 2 0 1 0	182		2 1 5 0 0 0 0 0 1 0
133	113	3 2 1 0 1 0 0 2 0 0	183		2 5 0 0 0 0 1 0 1 0
134	114	3 0 1 1 0 2 0 2 0 0	184		6 0 0 0 1 0 1 0 1 0
135	115	2 2 0 1 0 2 0 2 0 0	185	149	3 0 3 0 0 0 3 0 0 0
136	116	2 2 1 0 2 0 0 2 0 0	186	150	4 0 0 2 0 0 3 0 0 0
137	117	2 1 2 0 0 1 3 0 0 0	187		5 0 0 0 2 0 0 2 0 0
138	118	2 2 1 0 3 0 0 0 1 0	188		3 0 2 0 0 4 0 0 0 0
139	119	2 3 0 1 0 1 0 2 0 0	189		2 3 0 0 0 4 0 0 0 0
140	120	2 3 0 1 0 2 0 0 1 0	190		5 0 0 0 1 0 3 0 0 0
141	121	3 0 1 1 0 1 3 0 0 0	191		4 0 0 1 0 4 0 0 0 0
142	122	3 0 1 1 0 3 0 0 1 0	192		2 2 0 0 5 0 0 0 0 0
143	123	3 3 0 0 1 0 1 0 1 0	193		3 0 1 0 5 0 0 0 0 0
144	124	3 3 0 1 0 1 0 0 1 0	194		6 0 0 0 0 1 0 2 0 0
145	125	4 0 0 1 1 0 1 2 0 0	195		2 1 0 6 0 0 0 0 0 0
146	126	4 0 0 1 1 0 2 0 1 0	196		7 0 0 0 0 0 1 0 1 0
147	127	2 1 2 0 0 3 0 0 1 0	197		5 4 0 0 0 0 0 0 0 0
148	128	2 2 0 1 0 1 3 0 0 0	198		4 5 0 0 0 0 0 0 0 0
149	129	2 2 0 1 0 3 0 0 1 0	199		3 6 0 0 0 0 0 0 0 0
150	130	2 2 1 0 1 0 3 0 0 0	200		2 7 0 0 0 0 0 0 0 0
			201		9 0 0 0 0 0 0 0 0 0

3 0 1 0 2 0 0. The first numeral in the code gives the valence of vertex one, whereas subsequent numerals equal the valence −1 of the respective vertices. Each zero implies a terminal vertex of valence one: a leaf. In this example, three branches were attached to vertex one. The first branch, starting with the second vertex, has valence one and is therefore a terminal vertex and ends this branch. The second branch begins with the third and ends with the fourth vertex. The third branch from vertex 1 starts with vertex 5 in the code and has two (2) further branches, each terminating in a leaf:



In this way, each code in table 1 can be decoded into its graph.

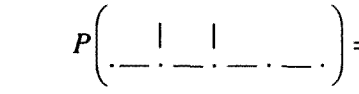
4. The chemico-graph-theoretic-cluster expansion ansatz

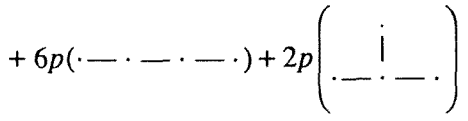
Klein has shown a general cluster expansion ansatz to express properties in terms of graph-theoretic structural elements. The additive case of Klein's ansatz uses a cluster function consisting of the number of times a given cluster appears as a subgraph within a larger graph representing the molecule of interest. Obtaining the cluster function is an enumeration problem and was covered in a previous paper [25]. The graphs considered here belong to alkanes. The hydrogens are suppressed, leaving a graph with vertices representing carbon atoms and edges representing carbon–carbon "bonds". Let the molecular system be denoted by its graph Γ and any property by $P(\Gamma)$. Then the intuitive notion that properties arise from contributions made by parts of the molecule is quantified by the cluster expansion ansatz:

$$P(\Gamma) = \sum_{\gamma \subseteq \Gamma} n(\Gamma, \gamma) p(\gamma). \tag{1}$$

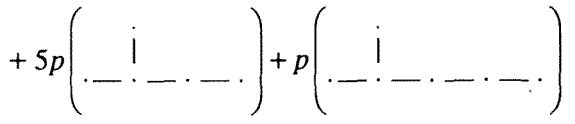
Each cluster, subgraph, structural component, γ, is a descriptor for molecular properties. The cluster expansion resolves $P(\Gamma)$ into a sum of terms arising from all possible clusters. The several $p(\gamma)$ are cluster contributions to P made by cluster γ. Their values are specific to the property being expanded. If $p(\Gamma) = 0$, then the value of $P(\Gamma)$ is entirely accounted for by the proper subgraphs (clusters) of Γ; alternatively, the graphical structure of the molecule itself does not contribute to the property. If $p(\Gamma) < 0$, then the sum of contributions made by $\{ p(\gamma) : \gamma \in \Gamma \}$ exceeds $P(\Gamma)$. Conversely, if $p(\Gamma) > 0$, then the sum falls short of $P(\Gamma)$. In either case, $p(\Gamma)$ is the unique contribution to $P(\Gamma)$ from the structure Γ as distinct from all its substructures. The embedding frequency $n(\Gamma, \gamma)$ counts the number of each cluster in Γ. Frequencies

are independent of the physical properties and are thus invariants for the graph. For example, the graph of 2, 3-dimethylpentane yields the cluster expansion:

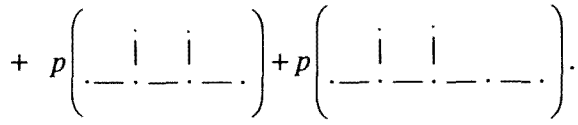
$$P\left(\text{  }\right) = 7p(\cdot) + 6p(\cdot\!-\!\cdot) + 7p(\cdot\!-\!\cdot\!-\!\cdot)$$

$$+ 6p(\cdot\!-\!\cdot\!-\!\cdot\!-\!\cdot) + 2p\left(\text{  }\right)$$

$$+ 2p(\cdot\!-\!\cdot\!-\!\cdot\!-\!\cdot\!-\!\cdot)$$

$$+ 5p\left(\text{}\right) + p\left(\text{  }\right)$$

$$+ 2p\left(\text{}\right)$$

$$+ p\left(\text{}\right) + p\left(\text{  }\right).$$

Embedding frequencies and their properties were reported for all trees with up to ten vertices in our previous paper [25].

One property needed for the CEA is that graphs fall into closed embedding classes. For example, all subgraphs of normal alkanes are themselves normal. Further, the graphs of carbon frameworks which have valences no higher than four are closed under subgraph embedding.

Considering eq. (1) to be a system of linear equations for a collection of embedded graphs, we arrive at the following matrix formulation,

$$\boldsymbol{P} = \boldsymbol{N}\boldsymbol{p}, \tag{2}$$

where $\boldsymbol{N}$ is the lower triangular array labeled by Γ, γ (row, column) and called the embedding frequency matrix. As stated above, graphs fall into closed embedding classes; eq. (2) applies for any such class. For example, restricting the valence of a vertex to four results in a matrix for the class of alkane graphs. Given measured or computed values for the property $P(\Gamma)$ for all graphs of the class, then the cluster contributions are found by inverting the cluster expansion:

$$\boldsymbol{p} = \boldsymbol{N}^{-1}\boldsymbol{P}. \tag{3}$$

We expect the values of $p(\gamma)$ to provide insight into the nature of the property P. The collection of CEA values $\{ p(\gamma)$, all γ in $\Gamma\}$ are a characteristic or signature of the property P in the same way as Fourier coefficients in a spectral analysis.

Errors in the physical measurement are propagated through the inverse cluster function. That is, if the errors in the measured $P(\Gamma)$ are independent with standard deviations ΔP_Γ, then the squares of the standard deviations for the derived cluster coefficients $p(\gamma)$ are given by

$$\Delta p_\gamma^2 = \sum_\Gamma (N^{-1})_{\gamma\Gamma}^2 \Delta P_\Gamma^2. \tag{4}$$

Normalization. It is convenient, when comparing and contrasting properties, to remove from $\boldsymbol{p}$ any dependence upon scale and reference level of measurements. Such dimensionless cluster coefficients we call normalized. Normalization is possible because of the following relationship between two properties which differ in scale and reference level:

$$\boldsymbol{P}' = \alpha\boldsymbol{P} + \beta \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ \vdots \end{pmatrix}. \tag{5}$$

For trees, it follows that

$$\boldsymbol{p}' = \boldsymbol{N}^{-1}\boldsymbol{P}' = \alpha\boldsymbol{p} + \beta \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \\ \vdots \end{pmatrix}. \tag{6}$$

To see this, note that

$$\boldsymbol{N}^{-1} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \Leftrightarrow \boldsymbol{N} \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix}, \tag{7}$$

and

$$\left[N \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \\ \end{pmatrix} \right]_\Gamma = n(\Gamma, \cdot) - n(\Gamma, \cdot — \cdot)$$

$$= |V(\Gamma)| - |E(\Gamma)|$$

$$= 1,$$

since trees have one fewer edges than vertices.

According to eq. (6), the reference level affects only the coefficients of clusters $(\cdot)$ and $(\cdot — \cdot)$, while coefficients of all remaining clusters are multiplied by the same factor with a scale change. Thus, scale may be removed from cluster coefficients using the "normalization" transformation:

$$\boldsymbol{p}' = \frac{1}{|p(\gamma_0)|}\boldsymbol{p} + \frac{p(\cdot — \cdot)}{|p(\gamma_0)|} \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \\ \end{pmatrix}, \tag{8}$$

where γ_0 is any cluster excluding $(\cdot)$ and $(\cdot — \cdot)$ and such that $p(\gamma_0) \neq 0$. With the same choice for γ_0, normalized signatures of properties which differ only in scale and reference level are identical, as may be easily verified.

Klein has classified physicochemical properties into types according to their imagined behavior in relation to bond formation between molecular fragments. Consider a molecule represented by the graph Γ to be constructable from its two fragments represented by disjoint subgraphs A and B. Imagine forming a bond represented by one edge from vertex i_A of A to j_B of B: $\Gamma = A(i_A, j_B)B$. Then, $P(\Gamma)$ is said to be an additive property if

$$\lim_{(i_A, j_B) \to 0} P(\Gamma) = P(A) + P(B), \tag{9}$$

where the limit means that the bond is gradually weakened to zero. In this case, Klein shows that the summation in the cluster expansion (eq. (1)) should extend over all *connected* subgraphs. Non-additive properties which do not follow eq. (9) nevertheless may obey a cluster expansion of the form of eq. (2). In such cases, other classes of subgraphs are appropriate and hence different embedding frequencies N are required.

Convergence. Mathematical convergence of the CEA can be defined in terms of a size function on graphs, e.g. graph diameter. If $S(\gamma)$ is such a function, then we say

the CEA coefficients converge if for every small number ε there exists a size σ_ε beyond which all clusters contribute less than ε:

$$|p(\gamma)| < \varepsilon, \quad \forall\, \gamma \ni S(\gamma) > \sigma_\varepsilon. \tag{10}$$

Only practical convergence can be discerned for experimental properties, since measurements cannot be made for arbitrarily large molecules. Nevertheless, the evidence may strongly suggest convergence or nonconvergence.

5. Experimental data

It is very difficult to find reliable experimental data for graph-theoretic analysis. We have taken experimental data from the TRC Thermodynamic Tables – Hydrocarbons [26]. Occasional checks were made against data tabulated by Cox and Pilcher [27]. We have chosen properties for one hundred and fifty alkanes with up to ten carbons. These will be referred to as the isomeric alkanes data set. Data for the normal hydrocarbons are complete, for many properties, through only twenty carbons. Boiling points and melting points for normal hydrocarbons are available through forty carbons. Although these thermochemical data have been carefully collected and placed in standard reference tables, their reliabilities and uncertainties are not uniform. For example, standard deviations are seldom reported, making it difficult to fully assess the reliability of the CEA. Cox and Pilcher report errors in the form of uncertainty intervals wherever possible. The TRC gives a rough indication of the estimation of uncertainty by the number of significant figures used to display the data, "the uncertainty reflects, in the view of the compiler, the effect of both random and systematic errors". For our purposes, these experimental data suffer not only from unreported or unreliable errors, but also from lacunae. In addition, it is not clear when values have been calculated.

Interpreting the boiling point results is complicated by the lack of consistency from table to table. For example, table 23-2-(1.101)-m [26] tabulates liquid-to-gas phase transition temperatures for normal alkanes with up to twenty carbons; these values often differ by more than the implied error from those data reported in the same reference, table 23-2-(1.101)-a, a tabulation of normal alkane boiling points. For convenience, data from the latter table will subsequently be referred to as nbp. A data set obtained by combining values for alkanes with up to twenty carbons from the former table with values for twenty-one through forty carbons from nbp will be referred to as tvap. The same references are cited in both tables for all but four of the normal alkanes tabulated, yet these two tables concur for only one of these alkanes, *n*-Eicosane. It was not possible to cluster expand the melting points of the isomers due to one or more missing values in the data sets for hexanes, heptanes and octanes. Very few melting points for nonanes and decanes are tabulated.

6. Theoretical data

Certain quantities, variously called graph-theoretic invariants or graph indices, are commonly computed to express the structural characteristics of a graph in a single number. The Wiener number and Randić index are the most common. For a graph Γ, the Wiener number is defined from the distance matrix:

$$W(\Gamma) = \tfrac{1}{2} \sum_{i,j} d_{ij} \tag{11}$$

and the Randić index from the adjacency matrix and valences:

$$R(\Gamma) = \sum_{i,j} \frac{A_{ij}}{\sqrt{v_i v_j}} . \tag{12}$$

These invariants may be regarded as purely mathematical "properties" of the corresponding molecule. Of course, theoretical data are uncorrupted by experimental errors. Cluster expansions of theoretical data have been performed on the 201 trees with up to ten vertices.

Statistical correlations have been found between experimental properties and mathematical indices. In this way, an insight has been gained into the origin of properties in molecular structure.

Finally, new insights might be gained from comparing cluster expansions of experimental and mathematical properties. If two signatures correlate, i.e. show close resemblance, then there may be a genuine underlying relationship between the two properties.

7. Cluster expansion results

7.1. EXPERIMENTAL PROPERTIES

Cluster expansions were performed for heats of formation, boiling points, and enthalpies and entropies of vaporization for all normal and branched alkanes with up to ten carbons. Normal alkane cluster expansions, through twenty carbons, were performed for heats of formation, entropies of vaporization and critical temperatures. Boiling points and melting points were cluster expanded on normal alkanes with up to forty carbons. All results are not displayed due to limited space, but are available elsewhere [28]. Results are shown in the form of signatures in figs. 1–4, with the exception of the heats of formation for normal alkanes, which are given in table 2. The discussion starts with properties of normal alkanes and is followed by the isomers.

Normal alkane boiling points, entropies of vaporization, critical temperatures, melting points and heats of formation exhibit practical convergence, in accordance with our above discussion. As the length of the cluster increases, the magnitude of its

coefficient decreases, within experimental error. This convergence supports intuitive expectation that the primary contributions to these properties arise from clusters which represent carbon atoms C–C bonds and next nearest-neighbor interactions, bond–bond interactions.

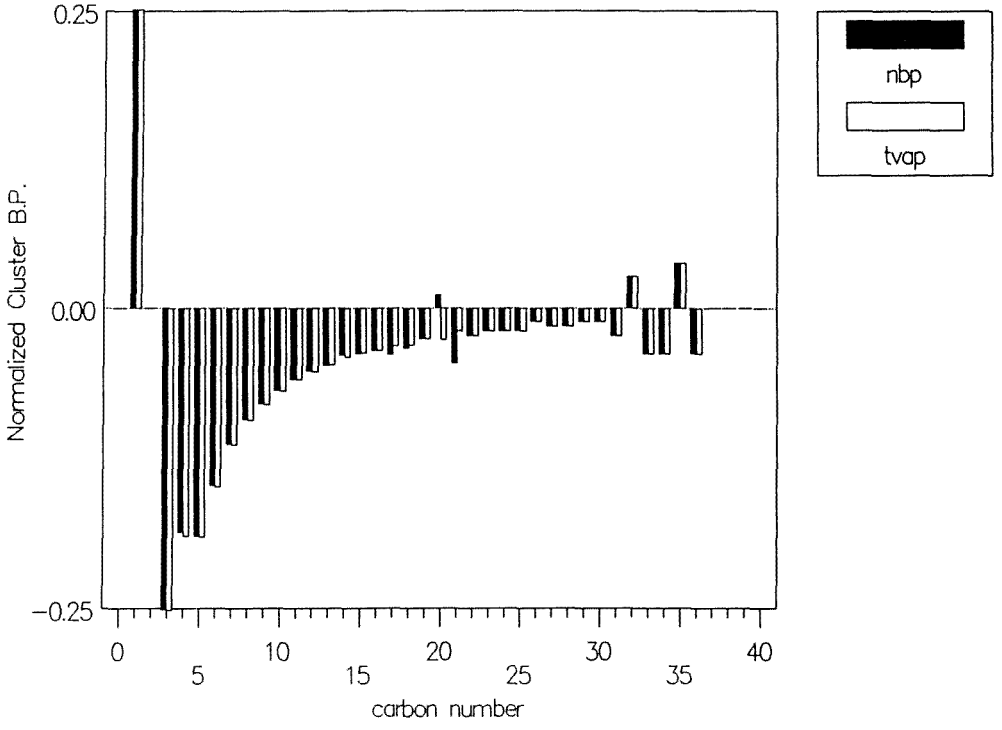


Fig. 1. Boiling point signature for normal alkanes through forty carbons. The cluster b.p. scale is expanded to better display the small coefficients. Values for $p'(\cdot)$ and $p'(\cdot - \cdot - \cdot)$ lie off the scale and are 2.760 and −1, respectively. Two sources for the data are compared: nbp refers to TRC table 23-2-(1.101)-a and tvap contains the twenty boiling points from TRC table 23-2-(1.101)-m with the boiling points for carbons 21–40 from nbp appended. Normalized using $\alpha_{nbp} = 1/26.42\ K^{-1}$, $\beta_{nbp} = -38.7/26.42$, $\alpha_{tvap} = 1/26.3\ K^{-1}$ and $\beta_{tvap} = -38.7/26.5$.

Two normal alkane boiling point signatures are shown in fig. 1, corresponding to experimental data sets nbp and tvap (see above). Note first that CEA coefficients beyond $n = 2$ are negative with few exceptions. An excursion in the cluster coefficients with twenty and twenty-one carbons is seen in the nbp signature, but is absent in the tvap signature. All boiling point cluster coefficients are significantly larger than their propagated errors except for $p(C_{20}) = p(C_{26}) = p(C_{29}) = p(C_{30}) = -0.30 \pm 0.24$ and $p(C_{31}$ through $C_{40})$, which are uncertain by 150% or more. No feature is found near 11 carbons, corresponding to the "change in fractal dimension" detected by Rouvray [29].

CEA signatures for critical temperatures and melting points of normal alkanes were obtained from their respective cluster expansions. The former is very similar to the boiling point signature. Melting point cluster coefficients (fig. 2) are alternately positive and negative, and much more slowly convergent than either boiling points or critical

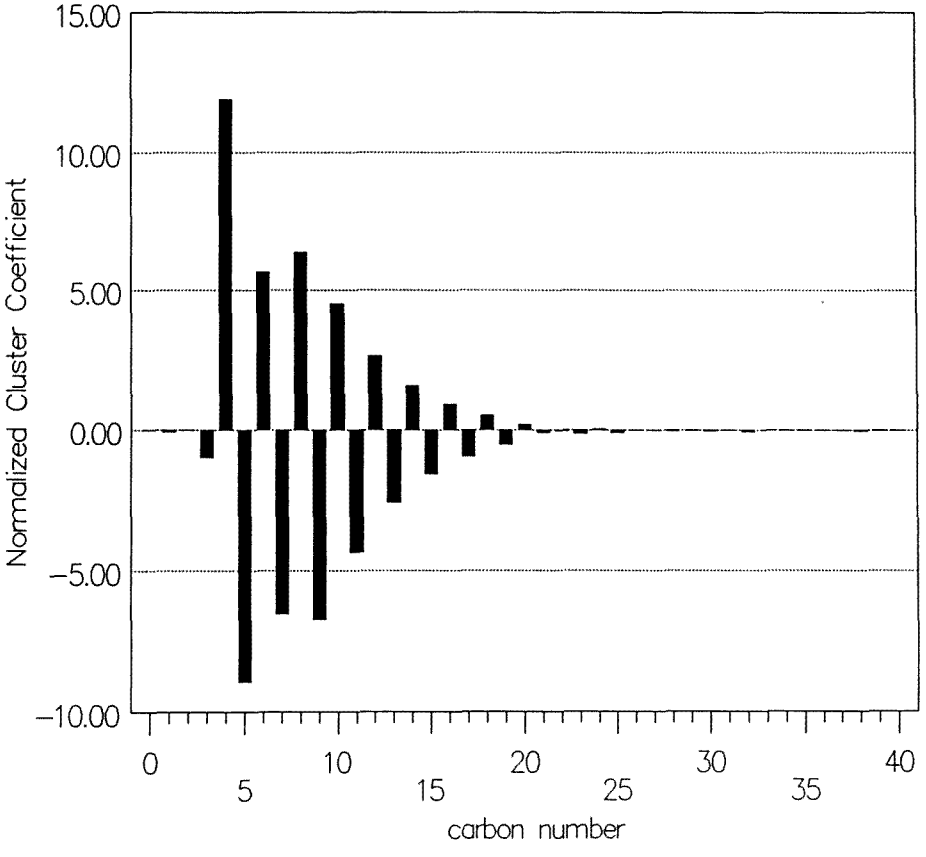


Fig. 2. Melting point signature for normal alkanes through forty carbons. Normalized using $\alpha = 1/4.55\ K^{-1}$ and $\beta = -91.01/4.55$.

temperatures. Even–odd alternation is a well-known phenomenon for hydrocarbon melting points. Broadhurst explains this phenomenon by pointing to the differences in crystal packing of the even and odd chains [30].

The heat of formation cluster expansion coefficients are shown in table 2, together with their errors. These coefficients converge more rapidly than those for the boiling points, a not surprising result in view of the historical success of the additive model for heats of formation. To find the relation connecting bond energies with the CEA coefficient $h(\cdot - \cdot)$, consider the first two cluster expansion equations:

Table 2

Heats of formation for normal alkanes. The number of carbons are given in column one, followed by the heats of formation $(kJ/mol)^{-1}$ with their errors in column two. The cluster coefficients $(kJ/mol)^{-1}$ and their errors are in column three, and the normalized coefficients (dimensionless) with their errors are in column four. Normalized using $\alpha = 1/11.560\ (kJ/mol)^{-1}$, $\beta = 65.220/11.560$

No. carbons	Heat of formation (kJ/mol)	Cluster coefficient (kJ/mol)	Normalized coefficient
1	−74.529 ± 0.010	−74.520 ± 0.010	−0.8045 ± 0.0009
2	83.820 ± 0.010	65.220 ± 0.022	0.0000 ± 0.0019
3	−104.680 ± 0.010	−11.560 ± 0.024	−1.0000 ± 0.0021
4	−125.790 ± 0.010	−0.250 ± 0.024	−0.0216 ± 0.0021
5	−146.760 ± 0.010	0.140 ± 0.024	0.0121 ± 0.0021
6	−166.920 ± 0.010	0.810 ± 0.024	0.0701 ± 0.0021
7	−187.780 ± 0.010	−0.700 ± 0.024	−0.0606 ± 0.0021
8	−208.750 ± 0.010	−0.110 ± 0.024	−0.0095 ± 0.0021
9	−228.740 ± 0.010	0.980 ± 0.024	0.0848 ± 0.0021
10	−249.460 ± 0.010	−0.730 ± 0.024	−0.0632 ± 0.0021
11	−270.430 ± 0.010	−0.250 ± 0.024	−0.0216 ± 0.0021
12	−290.720 ± 0.010	0.680 ± 0.024	0.0588 ± 0.0021
13	−311.770 ± 0.010	−0.760 ± 0.024	−0.0657 ± 0.0021
14	−332.440 ± 0.010	0.380 ± 0.024	0.0329 ± 0.0021
15	−353.110 ± 0.010	0.000 ± 0.024	0.0000 ± 0.0021
16	−374.170 ± 0.010	−0.390 ± 0.024	−0.0337 ± 0.0021
17	−394.450 ± 0.010	0.780 ± 0.024	0.0675 ± 0.0021
18	−415.120 ± 0.010	−0.390 ± 0.024	−0.0337 ± 0.0021
19	−435.790 ± 0.010	0.000 ± 0.024	0.0000 ± 0.0021
20	−456.460 ± 0.010	0.000 ± 0.024	0.0000 ± 0.0021

$$\Delta H^0_{298}(C_s \rightarrow C_g) + 2\Delta H^0_{298}(H{-}H) - 4\Delta H^0_{298}(C{-}H) = \Delta H^0_{f,298}(CH_4) = h(\cdot), \tag{13a}$$

$$2\Delta H^0_{298}(C_s \rightarrow C_g) + 3\Delta H^0_{298}(H{-}H) - 6\Delta H^0_{298}(C{-}H) - \Delta H^0_{298}(C{-}C) = \Delta H^0_{f,298}(C_2H_6(g)) = 2h(\cdot) + h(\cdot\text{—}\cdot). \tag{13b}$$

Solving for $h(\cdot\text{—}\cdot)$ and collecting terms, we find

$$h(\cdot\text{—}\cdot) = -\Delta H^0_{298}(H{-}H) + 2\Delta H^0_{298}(C{-}H) - \Delta H^0_{298}(C{-}C). \tag{14}$$

That is, the (hydrogen suppressed) two-vertex or "bond" cluster coefficient contains not only the traditional C–C bond enthalpy, but also the C–H and H–H bond enthalpies. The right-hand side is equal to 51.04 kJ according to the bond energies tabulated by Pauling [31]. The left-hand side is 65.22 kJ from table 2. The difference arises because bond energies are averaged over a variety of molecules. The third cluster expansion equation yields

$$h(\cdot - \cdot - \cdot) = \Delta H^0_{298}(C_3H_8) - 2\Delta H^0_{298}(C_2H_6) + \Delta H^0_{298}(CH_4), \quad (15)$$

which would be zero (as would all subsequent cluster coefficients) if bond energies were simply additive. The non-vanishing of $h(\cdot - \cdot - \cdot)$ and higher cluster contributions reflects interactions between more or less delocalized aggregates of bonds (usually lumped under the heading "steric effects").

Isomeric alkane boiling points, enthalpies of formation, and enthalpies and entropies of vaporization were cluster expanded, and coefficients do not show convergence analogous to that seen in figs. 1 and 2.

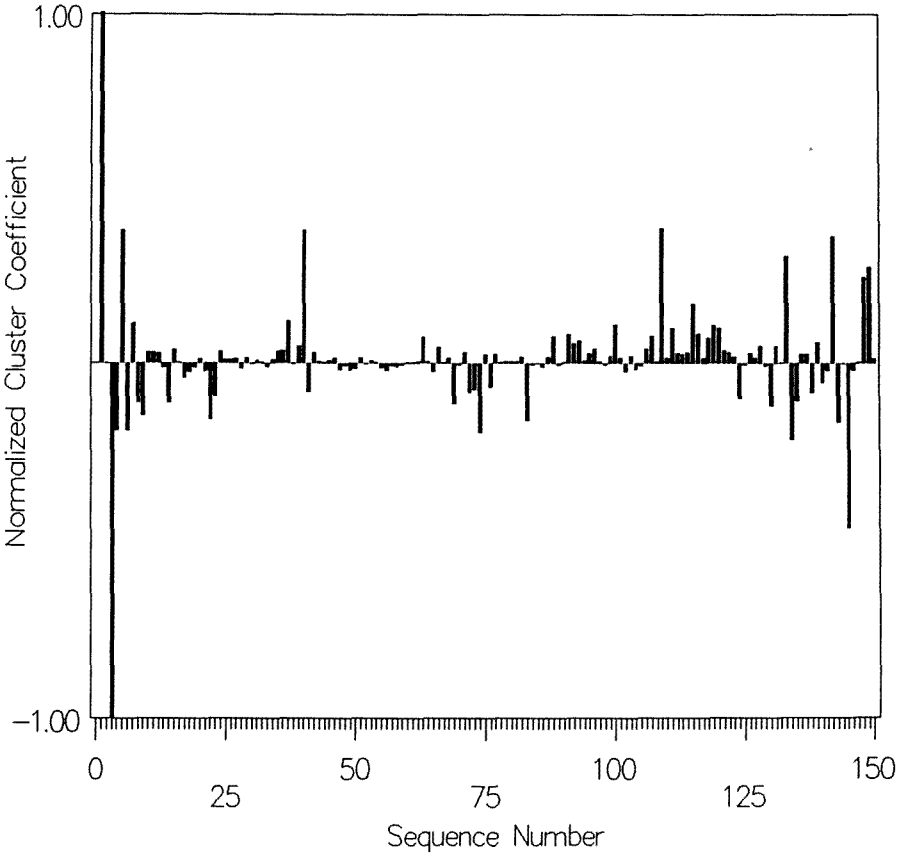


Fig. 3. Boiling point signature for all alkanes through ten carbons. Normalized using $\alpha = 1/26.304\ K^{-1}$ and $\beta = -38.796/26.302$. Sequence numbers correspond to hydrocarbons tabulated by WAV code in table 1.

The boiling point signature for alkane isomers is shown in fig. 3. The first two non-zero terms of the normalized signature are much larger in magnitude than all subsequent terms, but there is clearly no trend for convergence. Conspicuous large contributions are made by several branched clusters, for example:

$bp'(40) = 0.383 \pm 0.004,$

$bp'(109) = 0.389 \pm 0.016,$

$bp'(145) = -0.465 \pm 0.016,$

where sequence numbers are enclosed in parentheses. The data suggest that scattered large cluster coefficients will continue beyond these 150 isomers.

The enthalpy of formation signature is shown in fig. 4. Unnormalized one-, two- and three-carbon cluster coefficients are the large, dominant contributors to the enthalpy

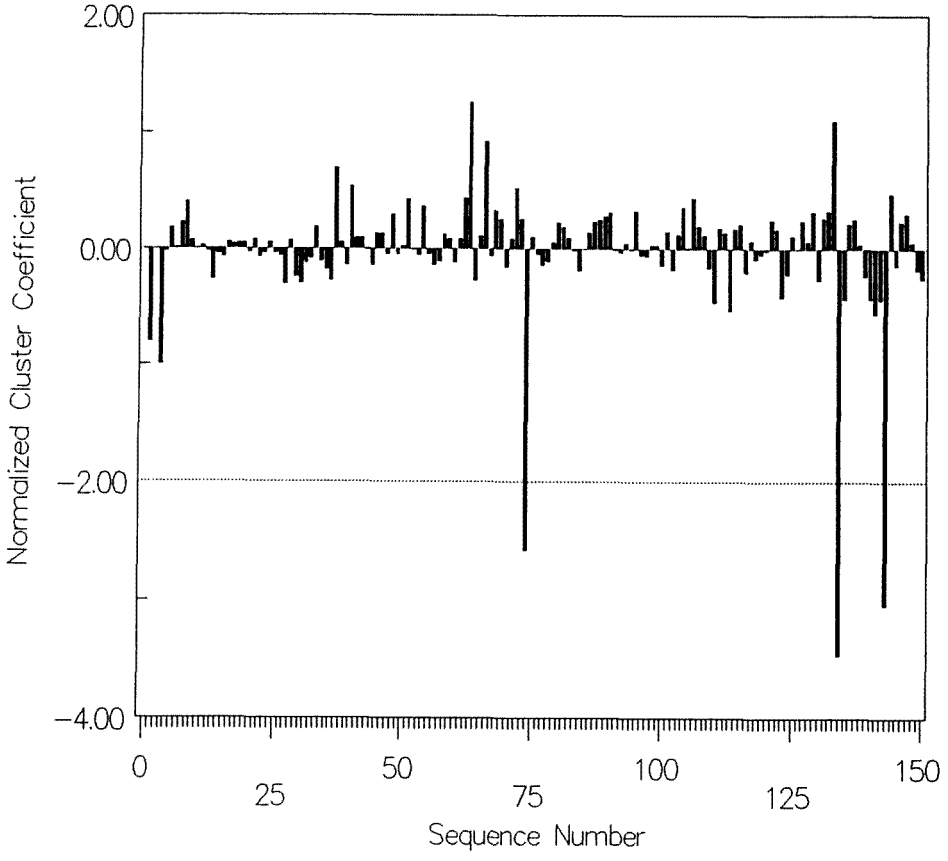


Fig. 4. Enthalpy of formation signature for alkanes through ten carbons. Normalized using $\alpha = 1/11.56$ (kJ/mol)$^{-1}$ and $\beta = 65.22/11.56$.

of formation just as they were for boiling points. Again, conspicuous large contributions are made by several branched clusters not always the same as, and with a range in values that is greater than, those seen for boiling points:

$\Delta h'(74) = -2.582 \pm 0.023,$

$\Delta h'(134) = -3.464 \pm 0.056,$

$\Delta h'(143) = -3.041 \pm 0.038.$

Clusters with negative (positive) contributions confer stability (instability) relative to the sum of enthalpies contributed by smaller clusters.

The signatures of the enthalpies and entropies of vaporization reveal a pattern quite unlike any previously discussed. Since both signatures are similar, only one is shown: the enthalpy of vaporization, in fig. 5. The cluster coefficients for the normal

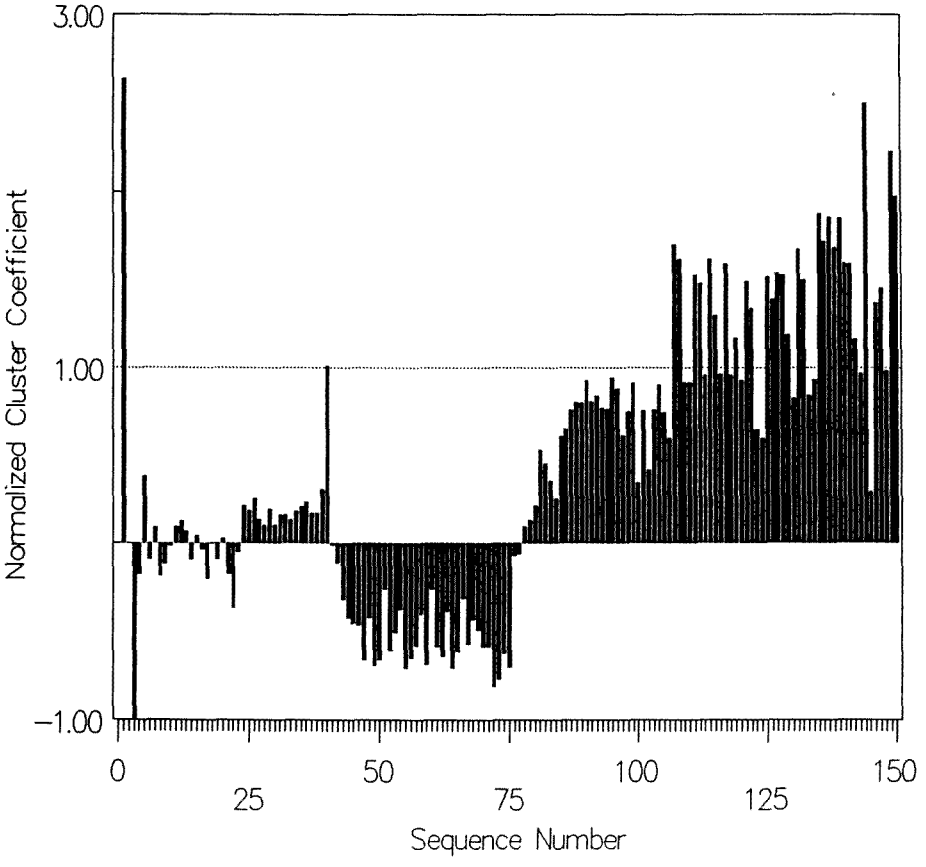


Fig. 5. Enthalpy of vaporization signature for all alkanes through ten carbons. Normalized using $\alpha = 1/0.592$ (kcal/mol)$^{-1}$ and $\beta = 0.393/0.592$.

alkanes are negative; however, the errors in the coefficients for nonane and decane are greater than 400% of the coefficients themselves. There are other 9-carbon and 10-carbon coefficients with large errors, but they are interspersed amid other coefficients

with small errors. What is obvious from the signature is that branched alkane coefficients diverge while alternating in sign for even and odd carbon numbers.

7.2. THEORETICAL PROPERTIES

Since the Wiener number is defined (eq. (11)) on the distance matrix, which in turn contains the path lengths, it follows that the cluster expansion is in terms of chain clusters only (for trees):

$$w(\gamma) = \begin{cases} |E|, & \text{if } \gamma \text{ is normal;} \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

This means that $W(\Gamma)$ depends on branching indirectly as different paths traverse a common vertex (branched vertex). All branched graphs have cluster coefficients equal to zero. From eq. (17), it is clear that the Wiener cluster coefficients diverge.

The Randić index is defined on the adjacency matrix. The cluster expansion of eq. (12) sums only over subgraphs which contain the edge (ij) corresponding to a non-zero value of A_{ij} and also contain those edges which terminate on edge (ij), and thereby determine the valences v_i and v_j. Thus, the only clusters which can contribute have diameter one, two or three. The cluster coefficient of a subgraph with diameter greater than three must vanish. The largest coefficient belongs to the stars.

Any property which correlates well with the Wiener number depends predominantly on chain descriptors. Any property which correlates well with the Randić index depends predominantly on star descriptors and is independent of clusters having diameter greater than three. Correlation between properties and indices can be easily recognized through comparing their CEA signatures.

8. Discussion of results

The boiling point cluster contributions conform to the following physical interpretation. Before normalization, the unit cluster $\gamma = \cdot$ makes a large positive contribution, $bp(\cdot) = 111.658$ K, describing the bulk effect arising from molecular mass (heavier molecules require more thermal energy to reach escape velocity) and molecular "surface area" (attractive Van der Waals forces act through adjacent molecular surfaces). The next cluster descriptor makes a negative contribution, $bp(\cdot - \cdot) = -38.796$, because bond formation lessens the surface otherwise exposed by two monomers: a bond between two units conceals a portion of the surface through which bonding occurs, thus mitigating intermolecular forces. Similarly, each *normal* cluster descriptor, $\cdot - \cdot - \cdot$, $\cdot - \cdot - \cdot - \cdot$, etc., makes a negative contribution because increased length allows more bending of the chain, which again conceals surface. The highly branched clusters, with sequence numbers 40, 109, 145 and others, make unexpectedly large contributions and show no obvious correlation between their sign and either their

graphical structure or graph diameter. Even so, these large contributions are undoubtedly due to physical effects determining the boiling points.

If the boiling point excursion noted previously in the nbp data set is real, it implies that a new mechanism affecting boiling points comes into play at carbon numbers twenty and twenty-one. We prefer to attribute the excursion to experimental errors in the nbp data. Entropy data, which might offer some insight into the nature of this excursion, have not been tabulated, to our knowledge, beyond $n = 20$. The CEA signature of entropies of vaporization on the available data reveal no unusual contributions for the larger graphs. This seems to support the interpretation that the excursion noted for the nbp data is associated with experimental errors.

Similar comments apply to other properties. For example, consider the first three heat of formation cluster coefficients before normalization: $h(\cdot) = -74.520$ kJ/mol, which represents contributions due to formation of 4 C–H bonds, breaking of 2 H–H bonds and atomization of graphite (eq. (13a)); $h(\cdot - \cdot) = 65.220$ kJ/mol represents C–C bond formation, H–H bond formation and 2 C–H bonds breaking (eq. (14)); these first two coefficients comprise the bulk contributions to enthalpies of formation; $h(\cdot - \cdot - \cdot) = 11.560$ kJ/mol represents the destabilizing effect on enthalpy of formation due to the interaction between next nearest-neighboring methyl groups. Similarly, all higher-order cluster coefficients reveal the interactions commonly hidden under the single designation "steric effects" as arising from various cluster descriptors (some stabilizing and others destabilizing).

9. Concluding remarks

The CEA supplies a complete set of independent descriptors for any physico-chemico-pharmacological property or graph-theoretic index. This set of descriptors, clusters, constitutes a signature for each property, and signatures exhibit the difference and similarity among properties and indices. Resolving a property into its CEA signature is a form of pattern recognition. Each term in the CEA or each component of the signature supplies information regarding how structure determines properties, functions and activities. The cluster expansion is a mathematically rigorous and objective extension and refinement of the subjective classification of chemical structure concepts: bonds, steric interactions and resonance. (The latter not having been investigated in this work.)

For example, the cluster coefficients of boiling points for *normal* alkanes converge monotonically (at least within experimental errors), whereas the coefficients of branched hydrocarbon boiling points show unexpected large contributions. Such large cluster coefficients, evident in many of the signatures, point to particular graphs as being noteworthy structural descriptors. In this way, one recognizes the expression of "chemical structure" in physical properties.

In the case of the normal boiling points, a reasonable physical interpretation is possible, but for this property, and for other properties of branched hydrocarbons, interpretation is elusive. Signatures do provide information by which to analyze graph-

theoretic indices and promote the development of new indices with improved correlation to physical properties. Mathematical properties can be contrived to exhibit any desired signature. The Wiener number has the signature: $\{w(\gamma = \text{chain}) = n - 1, w(\gamma \neq \text{chain}) = 0\}$, as shown above. In other words, $W(\Gamma)$ is determined completely from its linear cluster descriptors; its branched clusters do not contribute. On the other hand, the Randić index signature has $r(\gamma) = 0$ if the diameter$(\gamma) > 3$. Then, the only non-zero cluster contributions come from binary stars. In other words, $R(\Gamma)$ is determined completely from branched clusters. Thus, $R(\Gamma)$ and $W(\Gamma)$ express complementary structure information.

Although the CEA supplies a set of descriptors which is both complete and independent, these can only be evaluated if data are available for the complete set of structures involved. Therefore, unlike graph-theoretic indices, the CEA cannot be applied to a single chemical species. Future work modelling properties of alcohols, monocarboxylic acids, halogenated hydrocarbons, etc. could be done using this approach with only minor modifications, viz., using rooted trees. Enumerations of rooted trees have already been accomplished [32]. Other properties, such as double-bond ionization potentials in alkenes [33], are not additive and therefore not susceptible to the kind of cluster expansion we have performed here. In particular, these potentials are constantive and require a different cluster function.

Acknowledgements

Computing resources for this work were supplied by the WSU Academic Computing Facility. Encouragement and valuable suggestions were made by D.J. Klein and T.G. Schmalz, supporting our computations.

References

[1] H. Primas, *Chemistry, Quantum Mechanics and Reductionism* (Springer-Verlag, Berlin, 1983).
[2] G. Nagy, State of the art in pattern recognition, Proc. IEEE 56-5(1968)836–862;
N. Bongard, *Pattern Recognition* (Spartan, 1970) or R. Duda and P. Hart, *Pattern Recognition and Scene Analysis.*
[3] J.J. Sylvester, Nature (London) 17(1878)284.
[4] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research*(Academic Press, New York, 1976).
[5] P.G. Seybold, M. May and U.A. Bagal, J. Chem. Ed. 64, 7(1987)575.
[6] H. Marumi and H. Hosoya, Bull. Chem. Soc. Japan 58(1985)1778.
[7] I. Lukovits, J. Chem. Soc. Perkin Trans. II(1988)1667.
[8] A. Graovac, I. Gutman and N. Trinajstić, *Topological Approach to the Chemistry of Conjugated Molecules* (Springer-Verlag, New York, 1977).
[9] M. Gordon and J.W. Kennedy, J. Chem. Soc. Faraday II 69(1973)484.
[10] A.T. Balaban (ed.), *Chemical Applications of Graph Theory* (Academic Press, 1976).
[11] N. Trinajstić, *Chemical Graph Theory* (CRC, Boca Raton, FL, 1983).
[12] K. Balasubramanian, Chem. Rev. 85, 6(1985)599.

[13] R.J. Hansen and P.C. Jurs, J. Chem. Ed. 65, 7(1988)574.
[14] H. Hosoya, Bull. Chem. Soc. Japan 44(1971)2332.
[15] M. Gordon and J.W. Kennedy, J. Chem. Soc. Faraday II 69(1973)484.
[16] V. Batagelj, *Proc. Int. Symp. MATH/CHEM/COMP*, Dubrovnik (1987).
[17] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976).
[18] M. Gordon and J.W. Kennedy, J. Chem. Soc. Faraday II 69(1973)484.
[19] D.J. Klein, Int. J. Quant. Chem.: Quant. Chem. Symp. 20(1986)153.
[20] T.G. Schmalz, T. Živković and D.J. Klein, Cluster expansion of the Hückel molecular energy of acyclics: Applications to PI resonance theory, *MATH/CHEM/COMP* (Elsevier, Amsterdam, 1987).
[21] R.D. Poshusta, T.G. Schmalz and D.J. Klein, Heisenberg-model cluster expansion for half-filled Hubbard and PPP models, Mol. Phys. 66, 1(1989)317.
[22] R.D. Poshusta and D.J. Klein, Phys. Rev. Lett. 48(1982)1555.
[23] J.P. Malrieu, D. Maynau and J.P. Daudley, Phys. Rev. B30(1984)1817.
[24] R.C. Read (ed.), *Graph Theory and Computing* (Academic Press, 1972), p. 179.
[25] R.D. Poshusta and M.C. McHughes, J. Math. Chem. 3(1989)193.
[26] K. Marsh (Director), *TRC Thermodynamic Tables – Hydrocarbons*, Thermodynamic Research Center, the Texas A&M University System, College Station, TX (loose-leaf data sheets, extant, 1986). (Formerly: Selected values of properties of hydrocarbons and related compounds, Thermodynamics Research Center Hydrocarbon Project.)
[27] J.D. Cox and G. Pilcher, *Thermochemistry of Organic and Organometallic Compounds* (Academic Press, New York, 1977).
[28] M.C. McHughes, Master's Thesis, WSU (1989).
[29] D.H. Rouvray and R.B. Pandey, J. Chem. Phys. 85, 4(1986)2286.
[30] M.G. Broadhurst, J. Res. Natl. Bur. Stds. 66A, 3(1962)241.
[31] L. Pauling, *General Chemistry*, 3rd Ed. (Freeman, 1970), p. 913.
[32] R.D. Poshusta, *Proc. Int. Symp. MATH/CHEM/COMP*, Dubrovnik (1988).
[33] L. Klasinc and S.P. McGlynn, private communication.