Heats of Formation of Polyhex Polycyclic Aromatic Hydrocarbons from Their Adjacency Matrices

Gordon G. Cash†

Environmental Effects Branch, Health & Environmental Review Division (7403), Office of Pollution Prevention & Toxics, U.S. Environmental Protection Agency, 401 M Street, S.W., Washington, DC 20460

Received January 18, 1995⊗

Graph-theoretical indices derived from adjacency matrices were tested as predictors of standard heats of formation for a set of 153 polyhex polycyclic aromatic hydrocarbons (PAHs). In particular, the determinants and principal eigenvalues of the sums of adjacency and distance matrices, which were calculable from the adjacency matrices, were examined. The determinant was equal to zero for approximately 40% of the dataset and was therefore not suitable as a predictor. The principal eigenvalue, however, appeared to encode information not found in the Kekulé structure count and other commonly used predictors of $\Delta H_f°$. A correlation equation using this index, the Kukulé structure count, and one other predictor gave $r = 0.994$ for the entire dataset.

1. INTRODUCTION

Herndon et al.[1] recently published a correlation of certain structural features of a set of 153 polycyclic aromatic hydrocarbons (PAHs) with their standard heats of formation $\Delta H_f°$. The eight independent variables consisted of the logarithm of the number of Kekulé structures (ln K), the number of C−C bonds, the number of C−H bonds, and five structural parameters of the authors' own devising which enumerated bay areas of various shapes and the spatial relationships among them. With multiple $R^2 = 0.997$, this correlation equation was a remarkable discovery. Nevertheless, the values for the five structural parameters appear to be accessible only by manual inspection of each structure, and their use makes one wonder whether addition of new structural features would require additional parameters. Ideally, one would like to use parameters that are independent of structural peculiarities, such as those that are directly calculable from an atom connection table or an adjacency matrix, which is just the connection table in matrix form.

Schulz et al. have successfully correlated such parameters with boiling points[2] and standard heats of formation[3] of alkanes. The present study attempts to determine whether the parameters invented by Schultz (hereafter called Schulz indices) approach the predictive power of Herdon's structural parameters for prediction of PAH heats of formation. In the present author's opinion, this attempt was largely successful.

2. THE SCHULTZ INDICES

For a molecular graph of a molecule comprised of N non-hydrogen atoms, the adjacency matrix **A** is an $N \times N$ matrix whose elements $A_{i,j}$ are directly bonded and zero otherwise. The distance matrix **D** has elements $D_{i,j}$ equal to the number of bonds in the shortest through-bond path from atom i to atom j. A key feature of **D** is that it is mathematically computable from **A** without further reference to the structure. Several algorithms for performing this calculation have been published,[4] but for a molecule of moderate size, say $N \leq 60$, **D** is within the reach of even a modest desktop computer. Schultz found the sum of the **A** and **D** matrices more useful in predicting physical properties than either one alone. This result was counterintuitive because (**A** + **D**) does not distinguish atom pairs separated by one bond from those separated by two bonds. Nevertheless, the natural logarithms of the determinant $|\mathbf{A} + \mathbf{D}|$ and the principal eigenvalue $\lambda_1(\mathbf{A} + \mathbf{D})$ were found to be useful predictors of alkanes' physical properties.

Following the lead of Hall,[5] the author numbered the vertices for each molecular graph in an alternating fashion so that none of the first $N/2$ was directly bonded to any other. This scheme allows **A** to be written in the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix}$$

where $\mathbf{B}^T$ is the transpose of **B**, and **0** is an $N/2 \times N/2$ matrix of all zeroes. Moreover, **B**, unlike **A** and **D**, is not a symmetric matrix, and, for polyhex structures, it is nearly bidiagonal. The main and first upper diagonals are all ones, and the other entries are almost all zeroes. A computer can reliably create the bidiagonal, which can then be modified by hand. By formulating **A** in this way, manual input of data and with it opportunity for input errors are reduced to almost nothing.

The author used the procedure of Bersohn[4a] for computing **D**. This method is slower than the other published algorithms but easier to program. With recent advances in computing machinery, speed is no longer the problem it was when the faster algorithms were published. The procedure is basically this: $D_{i,j} = 0$ if $i = j$ because the distance from an atom to itself is zero. All powers of **A**, i.e., $\mathbf{A}^k$ are computed up to the maximum possible path length. For a general connected graph, k_{max} is $N - 1$, but for polyhexes it is $N/2$. It is convenient to take advantage of such properties of particular structures, since computing the $\mathbf{A}^k$s is the slowest part of the operation. Each $\mathbf{D}_{i,j}$ is equal to the smallest k for which the corresponding element in $\mathbf{A}^k$ is not zero. A corollary is that, for each $A_{i,j} = 1$, the corresponding $D_{i,j}$ is also 1. It is

† Telephone: (202) 260-3900. Fax: (202) 260-1236. Internet: cash.gordon@epamail.epa.gov.

⊗ Abstract published in *Advance ACS Abstracts,* August 15, 1995.

0095-2338/95/1635-0815$09.00/0

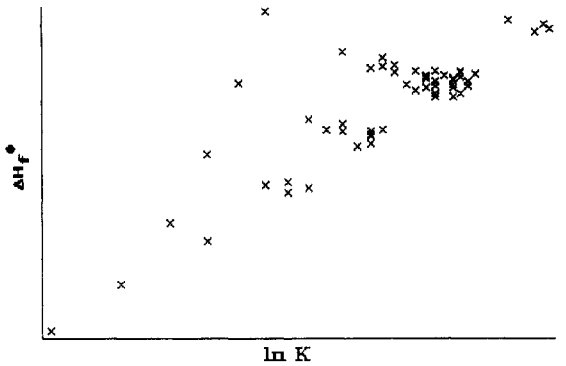

Figure 1. Plot of ln K vs $\Delta H_f°$ for the catacondensed ($n = 62$) subset.

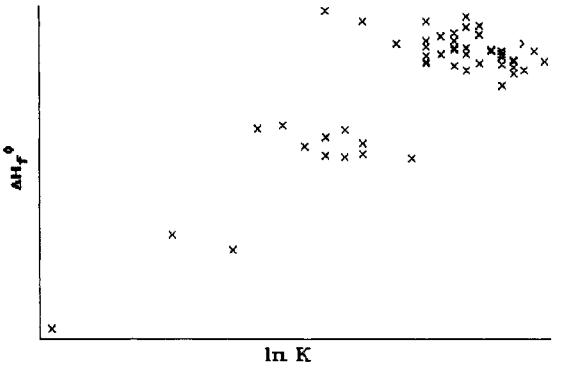

Figure 2. Plot of ln K vs $\Delta H_f°$ for the $n = 53$ subset containing the pyrene nucleus.

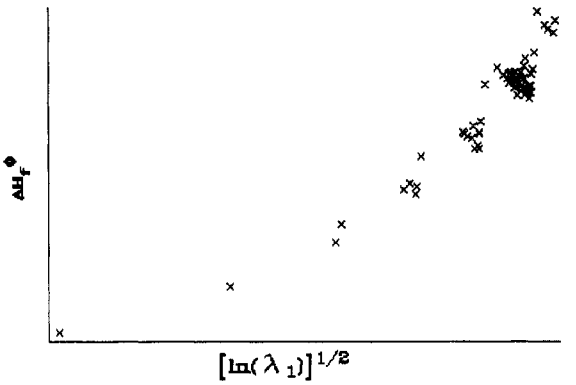

Figure 3. Plot of $\{\ln[\lambda_1(\mathbf{A} + \mathbf{D})]\}^{1/2}$ vs $\Delta H_f°$ for the $n = 62$ subset.

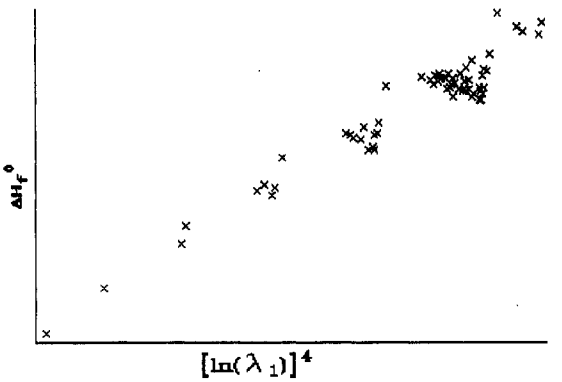

Figure 4. Plot of $\{\ln[\lambda_1(\mathbf{A} + \mathbf{D})]\}^4$ vs $\Delta H_f°$ for the $n = 62$ subset.

possible, by nesting loops, to write very compact computer code to carry out these calculations.

Since only the principal eigenvalue is of interest here, and not the entire spectrum, an iterative method is available for computing $\lambda_1(\mathbf{A} + \mathbf{D})$, namely,

$$\vec{u}_{n+1} = X\vec{u}_n / \|\vec{u}_n\| \qquad (1)$$

where $\vec{u}_0$ is a $1 \times N$ vector of all ones, and $\|\vec{u}_n\|$ is the norm of $\vec{u}_n$, or square root of the sum of the squares of all its elements. $\|\vec{u}_n\|$ quickly converges to λ_1.

The Schultz indices examined in the present study were $\ln \lambda_1(\mathbf{A} + \mathbf{D})$ and $\ln|\mathbf{A} + \mathbf{D}|$.

3. CORRELATIONS

The logarithm of the Kekulé structure count, ln K, has long been known to correlate well with resonance energy for benzenoid hydrocarbons[6] and hence with the overall energy. This parameter, plus mathematical transformations of the Schultz indices, plus another parameter discussed below were the only variables used in the present study to predict $\Delta H_f°$. Of the Schultz indices $\ln \lambda_1(\mathbf{A} + \mathbf{D})$ and $\ln|\mathbf{A} + \mathbf{D}|$, $\ln|\mathbf{A} + \mathbf{D}|$ was found not be very useful for the present study because, for 63 of the 153 compounds examined, $|\mathbf{A} + \mathbf{D}| = 0$. Moreover, statistical exploration of the $n = 90$ subset with $|\mathbf{A} + \mathbf{D}| \neq 0$ showed that $\ln|\mathbf{A} + \mathbf{D}|$ added little or nothing to the predictive power of the models. Therefore, this index was not investigated further.

To simplify the plots for exploratory purposes, the dataset was divided into those compounds with no internal carbon atoms, those with a pyrene nucleus, those with a perylene nucleus, and others. This division later turned out to be particularly fortuitous. Figure 1 shows a plot of ln K vs $\Delta H_f°$ for the $n = 62$ subset with no internal carbons. Clearly, there is some relationship between the two variables, but just as clearly there is some component of $\Delta H_f°$ that is not modeled by ln K alone. The obvious clustering in Figure 1 is according to the number of rings: Benzene is at the bottom left; naphthalene is diagonally above benzene; anthracene and phenanthrene are diagonally above naphthalene. The spreading of each cluster from lower right to upper left appears to be related to the relative extensiveness of the structures. In every cluster the "linear" polyacene member is the farthest to the left. Figure 2 is the same plot for the $n = 53$ pyrene subset, and it shows similar features. Plotting Figures 1 and 2 on the same axes gave a rather messy looking picture in which the pyrene clusters were offset along the lower-left-to-upper-right diagonal from the corresponding no-internal-carbon clusters with equal numbers of rings. This offset turned out to be of considerable significance later in the sudy.

Figure 3 shows a plot of an index actually used by Schultz,[2] namely, $(\ln \lambda_1)^{1/2}$, against $\Delta H_f°$ for the $n = 62$ subset. Again, the correlation and clustering are obvious, but the correlation appears to be exponential rather than linear. Changing the power of the logarithm from $^1/_2$ to other integer and half-integer values quickly revealed that the correct value for this dataset was 4, as illustrated in Figure 4. Similar results were obtained for the $n = 53$ pyrene subset.

Exploratory data analysis revealed that another useful predictor of $\Delta H_f°$ was the number of rings in the structure. (Size of the rings is not meaningful, since all the structures are polyhexes.) Because of the displacement between Figures 1 and 2 cited above, it also quickly became apparent that the entire dataset could be placed on the same coordinates if the number of rings was corrected for the number of internal carbons. The parameter finally used was $R =$ rings $-$ 0.3(internal carbons). The coefficient 0.3 was

Table 1. $\Delta H_f°$ Values from Ref 1 and as Calculated from the Equation 3

compd[a]	ref 1[b]	present study	difference	compd[a]	ref 1[b]	present study	difference
2	34.9	29.5	5.4	78	111.2	110.7	0.5
3	55.6	54.0	1.6	79	112.4	111.6	0.8
4	49.2	48.7	0.5	80	110.1	108.6	1.5
5	78.2	77.6	0.6	81	108.8	107.7	1.1
6	68.1	69.5	−1.4	82	113.4	111.1	2.3
7	65.9	66.4	−0.5	83	111.0	109.9	1.1
8	67.0	65.8	1.2	84	110.5	110.9	−0.4
9	69.3	67.8	1.5	85	110.1	108.2	1.9
10	101.9	100.7	1.2	86	108.4	110.4	−2.0
11	89.8	90.8	−1.0	87	118.2	117.7	0.5
12	81.1	84.8	−3.7	88	108.1	109.2	−1.1
13	81.1	83.6	−2.5	89	110.9	111.4	−0.5
14	85.7	86.0	−0.3	90	116.5	115.9	0.6
15	86.2	88.2	−2.0	91	112.0	112.7	−0.7
16	81.8	81.8	0.0	92	116.4	115.3	1.1
17	85.0	85.3	−0.3	93	106.9	106.0	0.9
18	88.7	88.1	0.6	94	109.6	109.5	0.1
19	84.5	83.9	0.6	95	110.0	108.5	1.5
20	86.3	84.7	1.6	96	106.2	107.4	−1.2
21	86.1	86.0	0.1	97	110.0	109.3	0.7
22	126.2	123.4	2.8	98	107.7	108.0	−0.3
23	113.0	112.2	0.8	99	109.3	109.0	0.3
24	107.7	106.5	1.2	100	110.7	111.1	−0.4
25	105.6	104.8	0.8	101	108.3	110.3	−2.0
26	101.7	103.7	−2.0	102	110.9	110.4	0.5
27	98.4	100.3	−1.9	103	115.3	113.2	2.1
28	107.3	108.2	−0.9	104	108.0	109.2	−1.2
29	99.4	101.4	−2.0	105	110.7	109.4	1.3
30	99.8	103.0	−3.2	106	117.3	114.0	3.3
31	101.7	102.8	−1.1	107	115.5	112.9	2.6
32	100.8	100.1	0.7	108	113.9	114.0	−0.1
33	97.8	98.6	−0.8	109	113.7	113.6	0.1
34	97.8	96.1	1.7	110	101.5	103.9	−2.4
35	102.6	104.1	−1.5	111	106.3	107.3	−1.0
36	98.4	99.9	−1.5	112	100.3	101.8	−1.5
37	106.4	105.5	0.9	113	108.5	108.9	−0.4
38	104.0	102.4	1.6	114	112.6	111.5	1.1
39	100.8	98.4	2.4	115	107.5	106.4	1.1
40	101.2	101.0	0.2	116	111.9	110.4	1.5
41	110.9	108.8	2.1	117	108.6	106.4	2.2
42	101.3	102.6	−1.3	118	103.8	106.3	−2.5
43	100.5	99.2	1.3	119	111.5	111.8	−0.3
44	103.6	103.2	0.4	120	106.7	106.7	0.0
45	108.4	107.9	0.5	121	113.5	111.5	2.0
46	101.2	99.6	1.6	122	112.9	113.7	−0.8
47	106.2	104.6	1.6	123	110.7	111.6	−0.9
48	104.6	103.8	0.8	124	109.1	111.2	−2.1
49	104.5	104.6	−0.1	125	109.6	110.8	−1.2
50	102.4	102.1	0.3	126	115.2	114.4	0.8
51	101.6	101.5	0.1	127	116.6	116.0	0.6
52	104.9	105.5	−0.6	128	126.2	124.8	1.4
53	104.1	103.9	0.2	129	117.0	117.9	−0.9
54	103.8	103.9	−0.1	130	127.3	126.8	0.5
55	105.2	103.1	2.1	131	118.5	119.0	−0.5
56	103.0	100.9	2.1	132	118.7	118.2	0.5
57	105.6	103.2	2.4	133	128.1	125.3	2.8
58	106.3	102.3	4.0	134	130.3	128.2	2.1
59	57.1	57.6	−0.5	135	117.2	118.3	−1.1
60	75.6	77.0	−1.4	136	77.5	82.7	−5.2
61	72.6	73.5	−0.9	137	100.8	101.0	−0.2
62	79.1	79.0	0.1	138	93.6	97.3	−3.7
63	85.4	87.5	−2.1	139	93.5	95.0	−1.5
64	94.0	93.1	0.9	140	96.1	97.5	−1.4
65	95.9	97.1	−1.2	141	100.4	100.7	−0.3
66	90.6	91.7	−1.1	142	100.4	102.6	−2.2
67	90.9	92.0	−1.1	143	134.7	135.8	−1.1
68	96.3	96.0	0.3	144	134.0	134.5	−0.5
69	92.3	93.9	−1.6	145	120.6	123.5	−2.9
70	95.6	94.6	1.0	146	138.2	138.5	−0.3
71	92.8	92.0	0.8	147	120.7	122.8	−2.1
72	90.4	92.2	−1.8	148	130.2	131.6	−1.4
73	90.0	88.6	1.4	149	128.1	130.2	−2.1
74	94.1	95.2	−1.1	150	123.5	123.3	0.2
75	103.8	102.6	1.2	151	119.3	119.2	0.1
76	76.6	79.8	−3.2	152	120.4	119.9	0.5
77	107.4	107.8	−0.4	153	121.7	121.7	0.0

[a] As numbered in ref 1. [b] Values calculated (MM3) in ref 1 plus the enthalpy-energy correction (ref 7).

deduced from graphs and was not optimized statistically. The author believes that mathematical optimization of this coefficient is ill-advised when the dataset contains such a small range of values for the number of internal carbons.

The best equation obtained for all 153 structures is

$$\Delta H_f° = -0.146(0.008)L - 27.511(0.708)\ln K + 45.747(1.078)R - 14.281 \quad (2)$$

$$n = 153, \quad r = 0.994, \quad s = 1.937, \quad F = 4200$$

where $L = \{\ln[\lambda_1(\mathbf{A} + \mathbf{D})]\}^4$ and R is as defined above. This model had one serious outlier, namely, benzene. This point was therefore dropped as a unique anomaly to give

$$\Delta H_f° = -0.157(0.007)L - 27.969(0.590)\ln K + 47.762(0.926)R - 19.027 \quad (3)$$

$$n = 152, \quad r = 0.995, \quad s = 1.606, \quad F = 5241$$

Values of $\Delta H_f°$ predicted by this equation are given in Table 1, along with differences from the MM3 values published by Herndon.[1] To facilitate comparisons, all values of $\Delta H_f°$ used in the present work contain the same energy/enthalpy correction[7] as those in ref 1. For a C_nH_m polyhex, this correction involves subtracting $(1 + n/6 - m/3)$ from the MM3 value. This change is negligible for all structures of the size studied here and zero for the catacondensed structures.

There are a few structures with residuals $> 2\sigma$ in this model, but there is no obvious way to remove them without compromising the generality of the model. Since generality without reference to specific structural features was the primary goal of the present study, the author believes that the $n = 152$ equation is best left alone, at least until further work may reveal a systematic shortcoming of the model.

The correlation coefficient between L and ln K for this dataset was $r = 0.773$. Thus, the two are somewhat intercorrelated, but they are clearly not the same thing. Removing ln K from the $n = 153$ equation reduced r to 0.933. Removing L from that equation only reduced r to 0.982 but increased the standard error of the estimate to 3.33. Thus, it appears that none of the three parameters is dispensable.

4. CONCLUSIONS

Because of the large number of zero values, the determinant of the (adjacency + distance) matrix is not a useful predictor of $\Delta H_f°$ for polyhex hydrocarbons. The principal eigenvalue of this matrix, however, appears to encode information about $\Delta H_f°$ for these compounds not contained in ln K and or other previously used predictors. Excellent correlation ($r = 0.994$) was obtained for the entire $n = 153$ test dataset using the principal eigenvalue, ln K, and one other predictor. Elimination of benzene from the dataset substantially improved the standard error of the estimate (1.937 vs. 1.606).

ACKNOWLEDGMENT

The author wishes to thank Prof. William C. Herndon, University of Texas at El Paso, for providing structure diagrams of these compounds.

REFERENCES AND NOTES

(1) Herndon, W. C.; Nowak, P. C.; Connor, D. A.; Lin, P. Empirical Model Calculations for Thermodynamic and Structural Properties of Condensed Polycyclic Aromatic Hydrocarbons. *J. Am. Chem. Soc.* **1992**, *114*, 41−47.

(2) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological Organic Chemistry. 2. Graph Theory, Matrix Determinants and Eigenvalues, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 27−29.

(3) Schultz, H. P.; Schultz, T. P. Topological Organic Chemistry. 3. Graph Theory, Binary and Decimal Adjacency Matrices, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 144−147.

(4) (a) Bersohn, M. A Fast Algorithm for Calculation of the Distance Matrix of a Molecule. *J. Comput. Chem.* **1982**, *4*, 110−113. (b) Müller, W.R.; Szymanski, K.; Knop, J. V.; Trinajstić, N. An Algorithm for Construction of the Molecular Distance Matrix. *J. Comput. Chem.* **1987**, *8*, 170−173. (c) Mohar, B.; Pisanski, T. How to Compute the Wiener Index of a Graph. *J. Math. Chem.* **1988**, *2*, 267−277.

(5) Hall, G. G. The Bond Orders of Alternant Hydrocarbon Molecules. *Proc. R. Soc. London Ser. A* **1955**, *229*, 251−259; *Int. J. Math. Educ. Sci. Technol.* **1973**, *4*, 233. Enumeration of Kekulé Structures by Matrix Methods. *Chem. Phys. Lett.* **1988**, *145*, 168−172.

(6) (a) Hall, G. G. *Bull. Inst. Math. Appl.* **1981**, *17*, 70; *Int. J. Quantum Chem.* **1991**, *39*, 605. (b) Carter, P. G. Empirical Equation for the Resonance Energy of Polycyclic Aromatic Hydrocarbons. *Trans. Faraday Soc.* **1949**, *45*, 597−602. (c) Gutman, I.; Trinajstić, N.; Wilcox, C. F., Jr. Graph Theory and Molecular Orbitals. X. Number of Kekulé Structures and the Thermodynamic Stability of Conjugated Systems. *Tetrahedron* **1975**, *31*, 143−146. (d) Swinborne-Sheldrake, R.; Herndon, W. C.; Gutman, I. Kekulé Structures and Resonance Energies of Benzenoid Hydrocarbons. *Tetrahedron Lett.* **1975**, 755−758.

(7) (a) Somayajulu, G. R.; Zwolinski, B. J. Generalized Treatment of Aromatic Hydrocarbons. I. Triatomic Additivity Applications to Parent Aromatic Hydrocarbons. *J. Chem. Soc., Faraday Trans. 2* **1974**, 1928−1941. (b) Nelander, B.; Sunner, S. The Interpretation of Ring Strain. *Chem. Phys.* **1966**, *44*, 2476−2480.

CI9500086