REFERENCES AND NOTES

(1) This is paper 11 in the series "Systematic Synthesis Design". For paper 10, see footnote 2.
(2) Hendrickson, J. B.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171.
(3) Zamora, A. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40.
(4) The Morgan algorithm[5] affords similar but not idential numbering since it also numbers out from atom 1, assigning its adjacent atoms as 2, 3, 4, ..., but the algorithm does not necessarily assign atom 1 to a tetravalent one but rather to that which on iteration develops the most extended connectivity. The Morgan numbering for Figure 1A is compared as an example: (maximal) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16; (Morgan) 3 1 7 8 9 2 4 14 5 6 10 16 11 12 13 15.
(5) Morgan, H. L. *J. Chem. Soc.* **1965**, *6*, 107.
(6) An entry in row 2 requires a three-membered ring on atom 1. The earliest row in which a ring of p atoms can appear is row $p - 1$.
(7) All graphs on six points are drawn out in Appendix 1 of Harary, F. "Graph Theory"; Addison-Wesley: Reading, MA, 1969.

(8) Fugmann, R.; Dolling, U.; Nickelsen, H. *Angew. Chem., Intl. Ed. Engl.* **1967**, *6*, 723.
(9) Plotkin, M. *J. Chem. Doc.* **1971**, *11*, 60.
(10) Corey, E. J.; Petersson, G. A. *J. Am. Chem. Soc.* **1972**, *94*, 460.
(11) Bersohn, M. *J. Chem. Soc., Perkin Trans. 1* **1973**, 1239.
(12) Esack, A. *J. Chem. Soc., Perkin Trans. 1* **1975**, 1120.
(13) Wipke, W. T.; Dyott, T. M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140.
(14) Schmidt, B.; Fleischhaver, J. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 204.
(15) Randic, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 36.
(16) Gasteiger, J.; Jochum, C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43.
(17) Roos-Kozel, B. L.; Jorgensen, W. L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 101.
(18) Smith, E. G.; Baker, P. A. "The Wiswesser Line-Formula Chemical Notation (WLN)", 3rd ed.; Chemical Information Management: Cherry Hill, NJ, 1975.
(19) For skeletons containing atoms of connectivity greater than 4, more bits are required since the maximal row sum in either T-list or R-list is presently 4 and so requires only 2 bits. For atom connectivities with $v_i \leq 8$, 3 bits per list entry would be required, or $6n$ bits per skeleton.

Derivation of the Principle of Smallest Set of Smallest Rings from Euler's Polyhedron Equation and a Simplified Technique for Finding This Set

SEYMOUR B. ELK

Department of Computer Science, The William Paterson College of New Jersey, Wayne, New Jersey 07470

Because the arrangement of atoms in a chemical compound often involves many overlapping rings, for purposes of taxonomy and nomenclature, the description that has traditionally been chosen as the geometrical model of this compound is the smallest set of smallest rings (SSSR) that includes all of the chemical bonds present. This paper first shows that SSSR is merely a reformulation of Euler's Polyhedron Equation in topology. Next, the "art" of finding this SSSR is reexamined and a theorem proposed in order to expedite finding the SSSR for a given compound. Finally, limitations, inherent in the topology of Euler's equation, are shown to have their corresponding limitations in the description of three-dimensional chemical molecules.

Chemical compounds have traditionally been classified according to the number and type of closed connected pathways that exist in a given molecule. However, in compounds that contain more than one single continuous closed path between atoms, there often exists many such circuits—which are frequently overlapping.[1] In order to circumvent the descriptive problems associated with overlapping rings, various "quick-fix" methods have been devised that convert such ring systems into topologically simpler systems. The one traditionally chosen,[2] the number of rings in a system is equal to the minimum number of scissions required to convert the cyclic system into an open-chain compound, forms the cornerstone of the method in present usage in most systems of nomenclature, referred to as the smallest set of smallest rings (SSSR).

Such methods give excellent results for compounds that are essentially planar such as polybenzenes,[3] e.g., coronene (Figure 1). In fact, the use of this method removes one element of ambiguity (Zamora's[4] type II and type III ring systems[5]) by mandating that the inner ring of coronene be considered a part of the figure rather than a part of the boundary. This is a highly desirable property for coronene inasmuch as the coplanar "inner" ring contributes to the aromatic stability of the compound. On the other hand, it is of dubious chemical merit for all of the circulenes[6] except coronene.

At this point, however, it should be noted that there exists a much simpler technique for determining the number of rings in a chemical compound, a technique that does not use the idea of bond scission. Instead, only the incidence relationships of the individual vertexes are important: (theorem 1) The number of rings in a chemical compound, R, equals 1 plus half the incidence excess;[9] i.e., $R = 1 + (1/2)n$, when n is the summation over all of the vertexes of their incidence minus 2 [$n = \sum_v (i_v - 2)$].

The mathematical basis for this principle lies in the metonymic[10] usage of "ring" as (1) a closed electronic pathway in a molecule and as (2) the boundary of a face of the geometrical model that represents this molecule.[12] Because we are interested in the number of faces of a simply connected[13] geometrical model in three dimensions, Euler's Polyhedron Formula[14] ($F + V = E + 2$) is applicable. Also, because the projection of a three-dimensional object onto a planar surface, the so-called Schlegel projection,[15] causes exactly one of the planar faces of the polyhedron to become the outer perimeter of the entire figure, $R = F - 1$. Next, we note that each edge of a polyhedron contributes 1 to the sum of the incidence at each edge; i.e., $2E = \sum_v i_v$. Subtracting $2V$ from each side of this equation yields $2E - 2V = \sum_v i_v - 2V = \sum_v (i_v - 2)$, which may be given the symbol n and the name "incidence excess". Finally, by substitution into Euler's equation, this yields $R = 1 + (1/2)n$.

From application of this theorem to coronene, there exists 12 vertexes with incidence = 3 (darkened) and 12 vertexes with incidence = 2 (Figure 2). Each incidence = 3 vertex contributes 1 to the value of n and each incidence = 2 vertex contributes 0. Thus

$$R = 1/2 \times [12 \times (3 - 2) + 12 \times (2 - 2)] + 1 = 7$$

Similarly, for the acyclic compound 2,2,3,3-tetramethylbutane (Figure 3), two carbon atoms have incidence = 4 and six have incidence = 1. Thus

$$R = 1/2 \times [2 \times (4 - 2) + 6 \times (1 - 2)] + 1 = 0$$

Because the traditional technique for finding SSSR is based on finding an appropriate topological simplification, Dyson[16] advises: "Some skill must be exercised in detecting the smallest
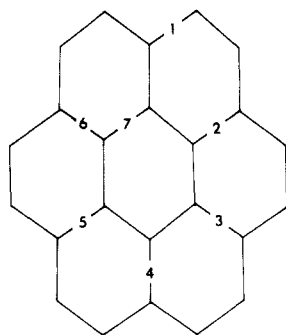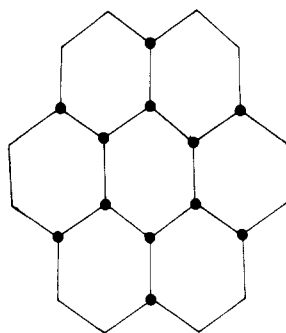
Figure 1.

Figure 2.

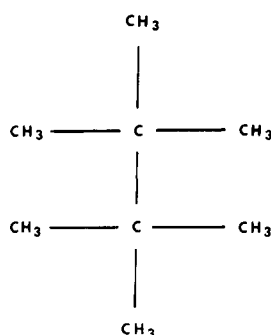**Figure 3.**

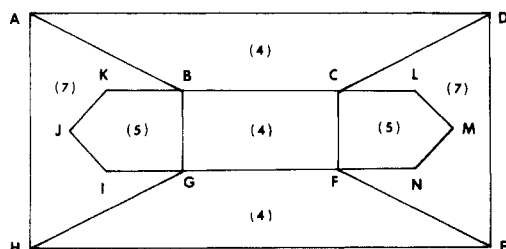**Figure 4.**

Figure 5.

Figure 6.

Figure 7.

rings present". This is illustrated by the seven-ring system shown in Figure 4 and the comment: "A casual glance would access it as made up of two 7-rings, two 5-rings and three 4-rings." Dyson then redraws the system (Figure 5) by using two pentagons and five quadrilaterals. In other words, the technique of finding the SSSR has been relegated to "an art" rather than to "a science".

The missing science is to describe the logic underlying Schlegel projection. A statement of this relationship is as follows: (Theorem 2) If the largest ring illustrated in a connected set of rings in bigger than the boundary ring, the Schlegel projection chosen does not yield the SSSR. There exists a different Schlegel projection which will produce a set of smaller rings.

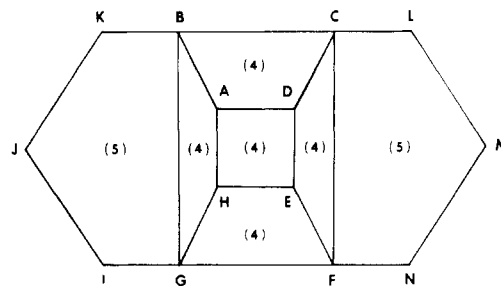As a proof of this theorem, start from a three-dimensional solid and designate any of the faces (polygons) as the "base polygon"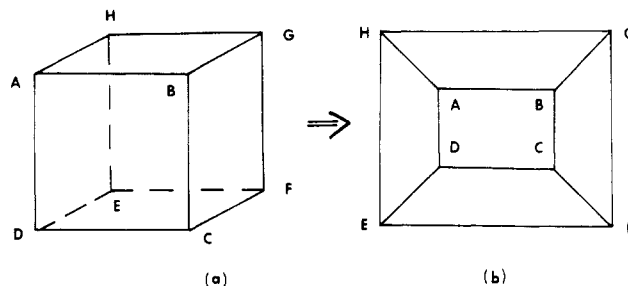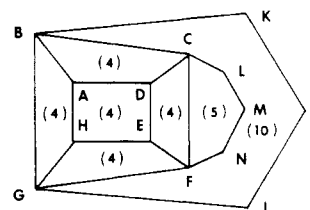. Next, perform a topological transformation[17] in which this base polygon is stretched sufficiently so that, when the remaining vertexes are projected onto this base polygon, each such vertex is interior to the base polygon. In other words, the content-defined[19] "perimeter" polygon exactly covers the union of the projection of the other faces and is thus redundant. This is equivalent to mapping the set of n faces of a polyhedron into a set of $n - 1$ contiguous planar areas—each of which may be described by boundary definition.[19] For example, in three dimensions, a cube (Figure 6a) has six faces, but the Schlegel projection of this cube (Figure 6b) has only five "faces". Next, notice that if there were n polyhedron faces to start with, then there are n fundamentally different Schlegel projections—each one being the union of $n - 1$ contiguous planar segments. Furthermore, each of these projections and its "complementary" base polygon exactly cover the n faces of the three-dimensional solid. Now, since the base polygon is simply the boundary ring and since it is desired to have a set of smallest rings, this is accomplished by choosing as the base polygon the largest polygonal face. Therefore, the remaining set of $n - 1$ rings will have minimum size—which proves the theorem.

By application of this theorem to Figure 4, the outer perimeter is a quadrilateral, while at least one of the inner polygons is larger; therefore, redraw this figure with quadrilateral ADEH in the center of the drawing area. By maintaining the same incidence relationship between the various vertexes, a new figure (Figure 7) is created that has five quadrilaterals, one pentagon, and one decagon.

From a superficial evaluation, the redrawn system appears to be less desirable than the original one, as this system contains an even larger polygon—a decagon vs. the previously largest polygon which was a heptagon. However, the boundary perimeter, which is not included in the description of the SSSR, is now larger. Therefore, this rearrangement is not counter-
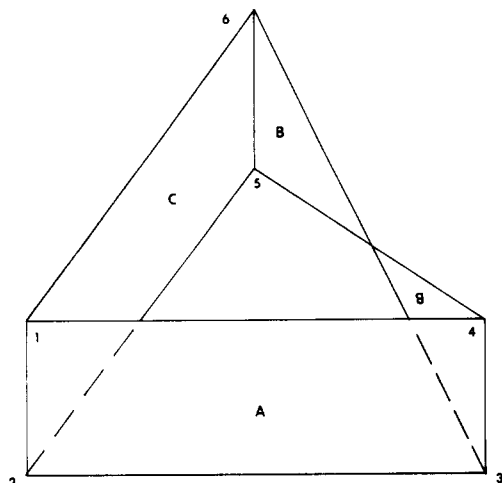
PRINCIPLE OF SMALLEST SET OF SMALLEST RINGS

J. Chem. Inf. Comput. Sci., Vol. 24, No. 4, 1984 **205**

Figure 8.

Figure 9.

Table I. Faces and Rings in Triangular "Prismane"

F_3	A = BCDE = (123); E = ABCD = (456)
F_4	B = ACDE = (1256); C = ABDE = (2345); D = ABCE = (1346)
F_5	AB = CDE = (13256); AC = BDE = (12543); AD = BCE = (12346); BE = ACD = (12546); CE = ABD = (23465); DE = ABC = (13456)
F_6	BC = ADE = (123456); BD = ACE = (125643); CD = ABE = (132546); BCD = (123 – 456)
total	29 faces; 14 rings

Table II. Faces and Rings in Triangular "Moebiane"

F_4	A = (1234); B = (3456); C = (1256)
	ABC = (1236) or (1254) or (1436) or (1456) or (2345) or (2365)
F_6	AB = (123654); AC = (341652); BC = (125436)
	ABC = (123456) or (125634)
	ABCABC = (145236)
total	8 faces (including one two-sided face); 15 rings

productive, as originally feared. The new perimeter, however, is still smaller than the largest interior polygon; therefore, theorem 2 must be applied again. This second reorganization promotes the decagon to the outer perimeter and gives a set of five quadrilaterals and two pentagons as the SSSR, exactly as found by trial and error.

An important limitation to the use of SSSR occurs in compounds that do not have simple[20] geometry. For example, the logical extension and simplification of a Moebius-type compound[21,22] is shown in Figure 8. This C_6H_6 compound, like its untwisted analogue (Figure 9), has six vertexes and nine edges. The number of simple faces in Figure 9 is unambiguously five—in conformity with Euler's equation. Also, the total number of faces is 29 and the number of rings is 14 (Table I). By contrast, the number of faces in Figure 8 is ambiguous inasmuch as the idea of a two-sided face was never intended for use with Euler's equation, or with SSSR. Excluding this two-sided face, there are three single faces, three double faces, and one triple face, as well as 15 rings (Table II).

In summary, although a derivation from first principles of SSSR as well as a simplified technique for generating this set has been presented, there exist many topological problem areas that must be considered when trying to devise a general description of three-dimensional molecules. Although it would be a relatively easy matter to formulate ad hoc modifications of SSSR, such as using a multiplicative weighting factor for multisided faces, etc., every time that the molecule being studied deviates from simple geometry a different set of modifications would be required. I view such a scheme as
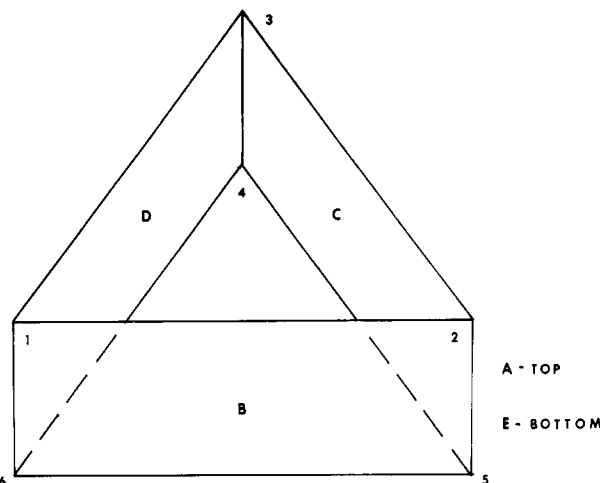
counterproductive. Instead, it would be far better that one not lose sight of the limitations that are inherent in the entire concept of SSSR.

REFERENCES AND NOTES

(1) Corey, E. J.; Peterson, G. A. "An Algorithm for Machine Perception of Synthetically Significant Rings in Complex Cyclic Organic Structures". *J. Am. Chem. Soc.* **1972**, *94*, 460–463.
(2) Patterson, A. M.; Capell, L. T.; Walker, D. F. "The Ring Index", 2nd ed.; American Chemical Society: Washington, DC, 1960; p ix.
(3) The term "polybenzene" refers to those polycyclic aromatic hydrocarbons that may be viewed as the fusion of benzene modules; i.e., every "single" ring is a six-membered closed chain of carbon atoms joined with an alternating sequence of single and double bonds. Examples of polybenzenes include naphthalene, anthracene, coronene, etc. On the other hand, rubicene, although a fused aromatic hydrocarbon, contains pentagonal rings and is thus not a polybenzene.
(4) Zamora, A. "An Algorithm for Finding the Smallest Set of Smallest Rings". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40–43.
(5) Zamora subdivides ring systems into the following three types: (I) ring systems for which no subset of the smallest rings contains all the atoms of the ring system; (II) ring systems for which all the atoms, but not all the edges, of the ring system are contained by a subset of the smallest rings; (III) ring systems for which all the atoms and all the edges of the ring system are contained by a subset of the smallest rings.
(6) The term "circulene" refers to the fusion of benzene rings about a central ring. Greenberg and Liebman[7] advise: "(5)-circulene is a bowl-shaped molecule caught in a tenuous balance between strain and aromaticity". Other possible similarly strained aromatic molecules are illustrated in footnote 8 and are referred to as "(7)-circulene", etc. Note that, in all instances except for the six-member internal ring, the term circulene refers to a strained molecule. Although from a consistency of structure perspective coronene could be referred to as "6-circulene", the usual connotation of strained for a circulene makes this an undesirable name. To the contrary, coronene shows increased, rather than degraded, stability.
(7) Greenberg, A.; Liebman, J. F. "Strained Organic Molecules"; Academic Press: New York, 1978; p 153.
(8) Elk, S. B. "Refinement of Systematic Nomenclature for Polybenzenes and Its Extension to Systems of General Arenes". *MATCH* **1982**, *13*, 265–266.
(9) The term "incidence excess" was chosen since each incidence = 2 vertex merely increases the number of atoms in a given ring without increasing the number of rings. The formation of a multiring system requires the existence of at least one atom with incidence greater than 2.
(10) The word "metonymy" is a figure of speech in English indicating the usage of one word for another word closely related to it; e.g., "the kettle is boiling" is understood to mean the liquid inside the kettle is boiling, rather than there is molten metal bubbling on the stove. This is equivalent to saying that an imprecise usage of a word correctly conveys the desired meaning. In general conversation, this is not a problem, but it can easily become one in scientific use.
(11) "Webster's New College Dictionary"; Merriam: Springfield, MA, 1981; p 718.
(12) Elk, S. B. "Effect of Taxonomy Class and Spanning Set on Identifying and Counting Rings in a Compound". submitted for publication in *J. Chem. Inf. Comput. Sci.*
(13) Courant, R.; Robbins, H. "What Is Mathematics?"; Oxford University Press: New York, 1941; p 243.
(14) Ibid., p 236.
(15) Loeb, A "Space Structure–Their Harmony and Counterpoint"; Addison-Wesley: Reading, MA, 1976.

(16) Dyson, G. M. "Some New Concepts in Organic Chemical Nomenclature". "Subcommittee Report of the Chemical Structure Association", 1983, p D 4/21.
(17) By a "topological transformation" is meant the mapping of one point set (figure) onto another so that the mapping is (1) biunique (each point on the object corresponds to exactly one point of the image) and (2) continuous in both directions (if the distance between two points in the object approach zero, then the corresponding distance in the image will also approach zero).[18] Physically, this is equivalent to saying that an object can be stretched or twisted but not torn or overlapped and joined.
(18) Courant, R.; Robins, H. "What Is Mathematics?"; Oxford University Press: New York, 1941; p 241.
(19) Elk, S. B. "Topologically Different Models To Be Used as the Basis for Ring Compound Taxonomy". submitted for publication in *J. Chem. Inf. Comput. Sci.*

(20) The single adjective "simple" is generally not used in geometry and topology textbooks when referring to figures—as it does not suggest a means of dividing the class of all figures into two disjoint sets: "simple" vs. "nonsimple". However, many geometrical properties, such as closure, connectivity, etc., do allow such a line of demarcation. Thus, any intuitive concept of simple would include simply closed, simply connected, etc. Nevertheless, one important property usually associated with simple figures in geometry—having the boundary set exactly one dimension less than the content set—is not applicable to the chemical model of molecules.
(21) Balaban, A. T. "Valence-Isomerisms of Cyclopolyenes". *Rev. Roum. Chim.* **1966**, *11*, 1097-1116.
(22) Walba, D. M.; Richards, R. M.; Haltiwanger, R. C. "Total Synthesis of the First Molecular Mobius Strip". *J. Am. Chem. Soc.* **1982**, *104*, 3219-3221.

A Priori Estimates of the Elution Profiles of the Pure Components in Overlapped Liquid Chromatography Peaks Using Target Factor Analysis

PAUL J. GEMPERLINE

Department of Chemistry, East Carolina University, Greenville, North Carolina 27834

Received April 5, 1984

Factor analysis is used to detect the presence of overlapping peaks in simulated high-performance liquid chromatography (HPLC) data produced by an ultraviolet/visible (UV/vis) photodiode array detector. The abstract solutions produced by factor analysis are rotated via target tests to produce estimates of the deconvolved elution profiles and the pure spectra of the overlapped peaks. An a priori method of selecting test vectors for target testing is reported here that does not force the elution profiles to fit any predefined functions. Various examples are given to demonstrate the effects of peak separation and random noise on the results. Examples are also given that illustrate the ability of the technique to resolve mixtures of three and four overlapping peaks.

INTRODUCTION

The recent commercial availability of photodiode array detectors for high-performance liquid chromatography (HPLC) has made new and powerful techniques available to detect the presence of overlapping HPLC peaks. One method involves plotting the ratio of two absorbance channels that have been slightly offset in time from each other.[1] This technique works well under some circumstances, but overlapping peaks usually have very similar chemical properties and thus similar spectral properties. In the instance where the spectra are nearly identical, one expects the technique to fail. A more sensitive method involves recording the complete spectrum on the upslope and downslope of the eluting peak.[2] The two spectra are then normalized and superimposed. Any slight mismatch in the two spectra indicates the presence of overlapping or contaminated peaks. Both methods require operator inspection and are not likely to be used routinely. Second, neither technique yields any information concerning the number of contaminating species present.

A form of factor analysis (FA) called principal component analysis (PCA) was first applied to the analysis of chromatographic data by Macnaughtan et al. to deconvolve overlapped mixtures of two components.[3] PCA has since been successfully applied to the detection of overlapping peaks in gas chromatography/mass spectrometry (GC/MS) data.[4,5] Two systems have been described that use PCA to automatically detect overlapping peaks in GC/MS data.[6,7] The technique should be amenable to automation for routine use in conjunction with photodiode array detectors for LC.

Knorr and Futrell were able to separate mass spectra of mixtures by factor analysis when one "pure mass" was available for every component present.[8] Kowalski et al. were able to separate overlapped peaks from GC/MS data for

binary mixtures.[9] Malinowski and McCue were able to identify and quantitate mixtures by target transformation factor analysis of the mass spectra.[10] Malinowski and McCue recently reported the adaptation of target transformation factor analysis to detect the presence of overlapped HPLC peaks.[11] The number of components were calculated from analysis of the ultraviolet/visible (UV/vis) spectra of successive fractions collected under overlapped LC peaks. Qualitative analysis was then performed by target testing the spectra of the components suspected to be in the mixture.

This paper describes the application of FA to detect the presence of overlapped LC peaks by using simulated data. After determining the number of overlapping components present, we apply a new method of selecting test factors that model the elution profiles of the unknown mixtures. Unlike previous work, this technique does not require pure spectra of suspected components or wavelengths that exhibit a unique response for one component. Estimates of chromatographic peak shapes, peak areas, and spectra of the unknown components are produced a priori. The technique is not limited to binary systems, and examples will be given involving three and four overlapping components.

THEORY

Consider the data produced by simultaneously measuring the absorption of light at many wavelengths during the elution of several chromatographic peaks. The total absorbance, $A_{i,j}$, at the jth wavelength during th ith scan is the sum of the absorbance of the n absorbing components present. In eq 1,

$$A_{i,j} = b \sum_{k=1}^{n} (c_{i,k} e_{k,j}) \qquad (1)$$

b is the cell path length, $c_{i,k}$ is the concentration of the kth