LBL-11245 c.2

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

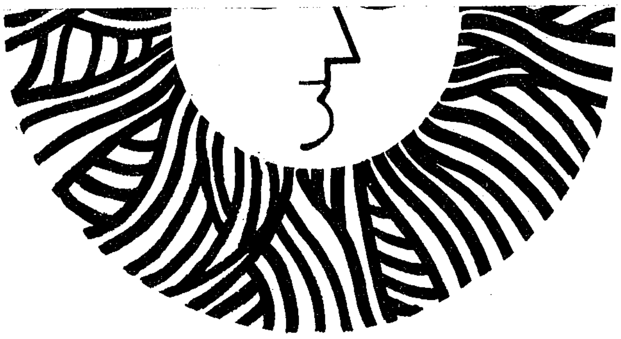
ENERGY & ENVIRONMENT DIVISION

Presented at the U.S. EPA/DOE Symposium on Health Effects Investigation of Oil Shale Development, Gatlinburg, TN, June 24, 1980; and to be published in the Proceedings

TREATMENT OF MULTIVARIATE ENVIRONMENTAL AND HEALTH PROBLEMS ASSOCIATED WITH OIL SHALE TECHNOLOGY

M. J. Kland

July 1980



LBL-11245 c.2

Prepared for the U.S. Department of Energy under Contract W-7405-ENG-48

TREATMENT OF MULTIVARIATE ENVIRONMENTAL AND HEALTH PROBLEMS ASSOCIATED WITH OIL SHALE TECHNOLOGY[*][+]

M. J. Kland

Energy and Environment Division
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

July 1980

[*] Work done under the auspices of the Department of Energy under Contract No. W-7405-ENG-48

[+] Presented at the Symposium on Health Effects Investigation of Oil Shale Development, Gatlinburg, TN, June 24, 1980. Conference sponsored by U.S. EPA/DOE. Proceedings to be published by Ann Arbor Science Pub.

Contents

INTRODUCTION

As the cost of foreign oil derived from traditional sources skyrockets, the economics of developing rich native oil shale deposits of the Green River Basin have become more attractive. Unfortunately, the potential occupational and environmental risks of large scale shale oil production are sufficient to cause concern. In addition, assessment of these risks presents a complex logistic problem. Thus, a preliminary listing of only those identified chemical byproducts found in oil shale and shale oil by EMIC and ETIC, has already yielded some 500 organic compounds, 40 metals and other inorganic species.[1] Included among these substances are potential and known carcinogens. Existing biological data are often inconclusive or inadequate and the logistics and costs of assessing potential environmental and health (E&H) effects by traditional biological methods are prohibitive.

Currently *in vitro* test methods such as the Ames Salmonella tests for mutagenicity are used by the EPA in conjunction with such *in vivo* biological test data as are available to establish a ranking of compounds for future *in vivo* testing of suspect commercial chemicals. A similar approach is being used on shale oil fractions which consist of complex mixtures of neutral, basic and acidic classes of organics. However, *in vitro* testing cannot yet replace long term animal test procedures--themselves often not conclusive because of variations in metabolic pathways with species. The Ames test is not reliable for metals--an important contaminant of shale oils and retort waters.[2a,b]

Under the best of circumstances, there remain unanswered questions of synergisms--the interactions of substances to enhance or inhibit bio-effects; questions of the reliability of mutagenicity as a ranking criterion for testing such a wide variety of substances as are found in raw and spent shales, shale oils and retort waters; and questions also of the relationships, if any, between carcinogens, mutagens, and teratogens.

Since the complex environmental and health problems posed by commercial oil shale development are clearly multitactorial, their solution will require a statistical matrix approach whose components are described in the following sections.

THE MATRIX APPROACH

In the complex chemical environment typical of most oil shale samples, the matrix approach involves parameters which relate biological activity directly or indirectly to chemical composition, and ultimately to molecular structure. In addition to distinctly structural criteria such as rings, double bonds and functional groups, thermodynamic data may also provide useful predictive parameters. Such general properties as spectral and photodynamic behavior, molecular conformation and electro- and nucleophilicities have all been associated with carcinogenicity.[3,4,5]

On the biological side, quantitative mutagenic responses, mammalian cellular transformations, DNA repair and DNA sedimentation analysis are among the useful parameters for correlation.

An overview of matrix methodology applied to the examination of repository samples for potential health effects was of this Symposium given at the opening session by Coffin .

For complex chemical mixtures, quantitative structure-activity and molecular connectivity relationships (QSAR, MC) within classes of compounds, where available, are useful functions, particularly when applied in conjunction with the methods of factor analysis (FA) and pattern recognition (PR). In the remainder of this paper, each of these four methodologies will be considered as it relates to the assessment of environmental and health (E&H) effects in multivariate, multicomponent systems similar to those encountered in the Oil Shale Repository studies reported at this Symposium.

STRUCTURE-ACTIVITY RELATIONSHIPS (SAR)

Historical

The relationship between chemical structure and reactivity of organic compounds was recognized long ago by Crum Brown and Fraser[6a,b] who used the following equation to express biological response (R) in terms of chemical structure (C):

$$R = f(C). \tag{1}$$

However it was Hammett[7] who gave structure-activity relationships of chemical systems more precise definition:

$$\log K_\sigma = \log K_\mu + \sigma\rho. \tag{2}$$

Here K's are rate or equilibrium constants, K_σ refers to a substituted benzoic acid, and K_μ is the unsubstituted parent acid. Rho (ρ) and σ are reaction and substituent constants, respectively.

Still later Hansch[8] and coworkers used the substituent constant π, related to the octanol-water distribution coefficient P, to further quantify these relationships:

$$\pi = \log P_X - \log P_H. \tag{3}$$

Here P_H and P_X are the partition coefficients of the parent compound (H) and its substituted derivative (X). This function and other versions of it have been widely used in drug and pesticide studies.

In 1973 Fahmy[9] and coworkers used the Taft steric parameter[10] E_s to relate LD_{50}'s of DDT derivatives to the size of a substitutent on one of the benzene rings:

$$\log LD_{50} = \alpha + \beta E_s^x + \gamma [E_s^x]^2. \tag{4}$$

Figure 1 shows a typical curve obtained in these studies, indicating that there is an optimum substituent size for maximum LD_{50}, hence the quadratic form of the equation and the resemblance of this plot to a typical potential energy curve.

A number of approaches to SAR have been used. For a recent review of the parameters and methodologies of quantitative structure-activity relationships (QSAR) the reader may consult the chapter by Osman, et al in Ref. 11.

Polynuclear Aromatic Hydrocarbons (PAH)

Polynuclear aromatic hydrocarbons (PAH) are found in the neutral fraction of shale oils. They are among the substances present in shale oils and retort waters which bear close scrutiny, since there are a number of known potent carcinogens among them. It is therefore of interest to consider two simple features, namely ring number and position, as they relate to a health effect: carcinogenicity. In Figure 2 a series of PAHs related to anthracene (I) and phenanthrene (II) are shown. The parent compounds are listed as "inactive." However, it is important to remember that the supposedly inactive lower members of a PAH series may function as initiators, co-carcinogens or synergists to enhance carcinogenic activity of the whole over the sum of its parts. Also at very high doses an occasional papilloma or tumor has been noted at the site of application of the inactive compound itself.

Fusion of a fourth aromatic ring in the 1,2 or [a] position of anthracene enhances carcinogenicity only slightly (III). However, the asymmetric dibenz[a,h]anthracene (IV)is a moderately active carcinogen. The corresponding linear pentacene (not shown) is also classified as inactive.

Early work on the carcinogenic activities of PAHs pointed to the importance of high double bond activity at the K-region as an essential feature (see structure III). Availability of the L-region which is active toward 1,4-addition appeared to diminish carcinogenic activity, at least in the lower fused ring systems. Substitution of methyl groups in the L-region positions has an activating effect, presumably because of the positive inductive (electron releasing) effect of the $-CH_3$ group. Thus, 7,12-dimethylbenzanthracene is a potent carcinogen.

In view of the varied mechanisms of carcinogenisis possible, and the many factors affecting the biological activity of a chemical compound, it is not surprising that no absolute generalizations relating structure to carcinogenicity have emerged. However, many of the more potent PAH carcinogens are relatively good electron donors, have low ionization potentials, form charge-transfer complexes with ease and exhibit "photodynamic activity"--e.g., behave as photochemical sensitiziers.

Other structure-activity relationships which have been used with some success include molecular size and thickness and the relationship of PAH structures to those of the steroids. Perhaps the most universal attribute of the carcinogenic PAHs is their structural relationship to phenanthrene (II). This is consistent with theoretical arguments requiring a region of high double bond activity (K-region).[12]

More recent computer-assisted studies of a PAH data set using pattern recognition techniques by Yuan and Jurs[13] confirm the importance of multidimensional analysis in applying SAR to the prediction of carcinogenicity of PAHs. Among the important determinants mentioned were molecular geometry, structural characteristics, lipophilicity and steric effects.

Aromatic Compounds Containing Nitrogen

Aromatic amines and their alkyl derivatives are prominent contaminents found in the base fractions of shale oils and retort waters. Aniline, all three toluidines and five of the xylidines have been reported, as have some of the mixed ethyl methyl derivatives.[1]

Historically aniline itself was viewed as a suspect carcinogen, because of the clustering of bladder cancers observed by Rehn (1895) in the Swiss dye industry, where commercial aniline was the starting material for magenta and other dyes. Later it was shown that the actual causative agents were aniline derivatives and the 1-and 2-naphthylamines (Figure 3). By the early 1900's bladder cancer was a recognized occupational disease wherever an established chemical industry flourished.[14]

Aromatic amines show much greater species specificity than the PAHs and, unlike the latter, do not induce tumors locally at sites of application. Instead the bladder, liver, or intestines may be affected. Also, 2-naphthylamine, a powerful bladder carcinogen in humans and dogs, is essentially inactive in both rabbits and rats. This highly specific behavior implies that arylamines require metabolic activation to render them carcinogenic. Current evidence is that N-hydroxylation of the substituted arylamine must occur, possibly followed by esterification of the OH group, yielding an unstable intermediate. The latter then breaks down to a positive nitronium ion, an electrophile capable of reacting at a nucleophilic (electron rich) site of the cell (Figure 4).

Although aniline itself does not appear to be a carcinogen in man, o-toluidine (hydrochloride) is, and other derivatives of aniline have also induced cancers in mammals. In extended anilines such as 4-biphenylamine, the position of the amino group is important. In general aromatic amines with a conjugated para substituent are much more active than their isomers with a free para position. Similarly the substitution of a methyl group ortho to an arylamino group often increases carcinogenicity.

With rare exceptions (e.g., o-toluidine) carcinogenicity data of the many isomeric arylalkylamines found in shale oils are missing, incomplete or equivocal. This is unfortunate because recent mutagenicity testing of shale oil fractions[15] indicates that the organic base fraction, because of its magnitude, may prove to be a greater health hazard than the PAHs of the neutral fraction which are present in lesser amounts than encountered in other synfuels and coal.

Among the other known tumor initiators found in shale oils are the benzacridines, benzcarbazoles and substituted phenols. The benzcarbazole series shown in Figure 5 illustrates two possibly significant factors in carcinogenicity--namely hydrogen bonding and steric (electronic) effects. As expected, the addition of benzene rings increases activity, but the degree of enhancement appears to be strongly dependent on the location of the asymmetric benzene(s) with respect to the pyrrole nitrogen. Thus benzcarbazoles I, II and III with benzene rings adjacent to the pyrrole nitrogen are weak carcinogens, whereas IV, in which the pyrrole NH group is sterically unhindered, is a strong carcinogen. Similarly, while replacement of a benzene ring with pyridine enhances the overall activity of a weak carcinogen (II,III) to a moderate one (V,VI), rotation of the pyridine heterocycle in V through 180° produces a strong carcinogen (VII). In the last structure each N-heterocycle can function independently, since H-bonding is sterically impossible.

MOLECULAR CONNECTIVITY (MC)

Chemists have long known that even minor structural variations in molecules can have profound effects on physical properties, chemical reactivity, and on biological toxicity. This knowledge has been applied to the alteration of structures and substituents of molecules intended for a variety of industrial and medical applications. Two of the most successful applications have been in the areas of drug and polymer design.

Molecular connectivity (MC) attempts to relate these structure-property dependencies to the topological characteristics of the molecule. MC is the most recent and probably the least familiar of the predictive methods used in the correlation of molecular structure with chemical and biological activity. It is based on relatively simple topological principles long familiar to the organic chemist, since a structural formula is in reality a topological graph. Figure 6 illustrates in graph form some simple organic structures with their topological descriptions.

Molecular connectivity makes the fundamental assumption that it is possible to differentiate molecular structures by abstract numerical means so that their correlation with physical, chemical, and biological properties become possible. The method is defined by Kier as a "nonempirical derivation of numerical values that encode within them sufficient information to relate (them) to many physicochemical and biological properties."[16]

Some Definitions and Simple Graph Theory

A graph is a set of points (vertices) connected by lines (edges; Figure 6). The Molecular Connectivity method assumes that information essential to a quantitative correlation of organic molecular structure with properties is inherent in a valence-weighted graph (G_v). Secondly, a relationship between the connectivity characteristics of the graph and the specified molecular properties is postulated. This relationship is expressed as a sum of terms, each linearly dependent on the graph characteristics.

The connectivity function $C(\chi)$ for a graph has the form

$$C(\chi) = b_o + \sum_{m,t} b_t(m)^m\chi_t \qquad (5)$$

where b_o is a constant and ${}^m\chi_t$ is the connectivity index. Here $b_t(m)$ depends on the property and may be calculated from a model, from theory, or by multiple regression against experimental data. In the latter case, the experimental values are regressed against $C(\chi)$. The number of edges in G_v determines the highest order of the χ term. Each connectivity index term ${}^m\chi_t$ is defined by its subgraph type, t, of m connected edges and subgraph order m. Subgraphs are of the four types listed in Table I. Connectivity Indices ${}^m\chi_t$ are obtained by summing terms over all distinct subgraphs:

$$^{m}\chi_{t} = \sum_{j=1}^{n_m} {}^{m}S_{j} \qquad (6)$$

where n_m is the number of type t subgraphs of order m. ${}^{m}S_{j}$ terms are calculated for each subgraph as reciprocal square root functions of valency:

$$^{m}S_{j} = \prod_{i=1}^{m+1} (\delta i)_{j}^{-1/2} \qquad (7)$$

where j refers to a particular set of edges. The number of valencies involved depends on subgraph type. Summation terms $^{0}\chi$ through $^{4}\chi$ are shown in Table II. From these it may be seen that the zero order subgraph consists of a single vertex (no edges); $^{1}\chi$ is summed over all edges, appropriately weighted by reciprocal square root valencies. Here we have only one type of graph edge e_s terminating on v_i and v_j, with a total edge number N_e. Second order subgraphs have pairs of adjacent edges of single path type P. Thus, each term will contain the reciprocal square root product of three vertex valencies.

In the third order connectivity index ($^{3}\chi$), path, cluster and chain terms may occur, each to be calculated as a reciprocal square root product of four deltas. Finally, in $^{4}\chi$ all four subgraph types are possible for the first time. Here n_m in the summation term shown refers to the number of type t subgraphs having four edges,

${}^{m}\chi_t$ terms of higher order are calculated in a similar manner, with the term superscript m corresponding to the edges involved in the calculation.

Thus, the connectivity index is a valence-weighted count of connected subgraphs. This weighting process is a key feature of the MC method.

Figure 7 illustrates steps in the calculation of the first order connectivity index ${}^{1}\chi$ for two isomeric branched aliphatic hydrocarbons (n = 7): 2,2,3-trimethylbutane and 2,4-dimethylpentane. Some useful fundamental equations of MC theory are summarized in Table III.

The topological matrices and algorithm for dimethylcyclohexane and subgraphs are shown in Table IV. Connectedness values were determined from edge counts E_s:

$$E_s = 1/2 \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \qquad (8)$$

where A is the adjacency matrix.

Some Uses and Limitations of the MC Method

Using the techniques described, Kier and others have successfully correlated MC with physical and biological properties.[17,18] The method has the advantage of relative simplicity and flexibility. It can be used to represent molecular structure quantitatively at a number of levels of complexity. Each level provides some information uniquely related to the structure (graph, subgraph) and through it to physical, chemical, and biological characteristics.

The C(χ) function is essentially a weighted count of substructures of the molecule, each described numerically with reference to adjacencies within them. Using similar methods MC may incorporate hetero atoms and valence differences between them. ${}^1\chi$ takes account only of adjacent influences on a specific atom. These are modified in the higher order subgraph terms. Also basic to the χ calculation is $\delta_i \delta_j$--the atom product--and use of the reciprocal square roots of this product.

Preliminary attempts have been made to develop an atomic chi value C(${}^1\overline{\chi}_i$) in which each bond term (c_{ij}) is divided equally between the two vertices and the half-bond terms summed. Other aspects of molecular structure which require further refinement are: cis-trans isomerism, nonbonded steric interactions and conformational structure, all of which have either three-dimensional or directional features, or both, that are not included in the original treatment. The fact that reasonable correlations have already been achieved for fairly complex systems largely within the limits of an elementary graphical approach, is encouraging, for MC is a still maturing technique for coping with the multifactorial problems of chemical-biological interactions.

FACTOR ANALYSIS

The complexity of the chemical mixtures and attendant biological problems associated with oil shale technology clearly preclude the use of a simple SAR or MC approach in the calculation and assessment of environmental and health effects. Thus, while the PAHs contained in the neutral fraction of shale oils have high specific mutagenic activities, they represent only a small fraction of the total activity of a given oil sample because of their low concentrations in shale oils. Consequently the base fraction is viewed as posing the greater health risk. Any predictive method must take these factors into consideration.

One method of dealing with such complex multifactorial data utilizes factor analysis (FA), an approach originally applied mainly to the social sciences and only more recently to chemical and biological problems. Using FA as an analytical tool, it is possible to systematize the multifactorial data of complex interactions without knowledge either of the exact number of significant factors involved or of their precise nature. Thus the assumption is made in QSAR analyses that the biological response (R) of a system to a chemical compound is a function of structure, which in turn has electronic, steric, hydrophobic, and polarizibility terms:

$$R = f(\sigma, E_S, \pi, M_R). \qquad (9)$$

The electronic term is further assumed to be factorable into inductive ($\mathfrak{I}$) and resonance ($\mathfrak{R}$) terms. Factor analysis provides a rational approach to handling this multiplicity of interrelationships. One need only represent the observed response data in terms of these parameters, or combinations thereof, and develop a model in conformity with the data.

Some Basic Assumptions of FA

A detailed development of FA is beyond the scope of this chapter, and the reader should consult a suitable reference text such as Rummel's Applied Factor Analysis.[19] What follows is a summary of the basic principles of FA applicable to any multivariate system, including our chemical one.

If we assume a two-dimensional data matrix, two mathematical requirements must be satisfied by the property measured. One requirement is that each data point D be expressed as a linear sum of terms:

$$D = d_1 + d_2 + \cdots + d_n. \qquad (10)$$

The second requirement is that each data point D also be a sum of row and column product terms:

$$D = r_1c_1 + r_2c_2 + \cdots + r_nc_n \qquad (11)$$

where r_k and c_k represent the kth row and column factors, and each is independent of the other. Thus in matrix notation the data matrix may be expressed as the product of a row and a column-related matrix:

$$[D] = [R].[C]. \qquad (12)$$

Matrices may be of the entity-entity or entity-property type. The latter generally has greater relevance in chemical-biological interactive systems. Figure 8 is a diagram of the steps involved in the FA method. The stepwise procedure is described below:

1. Correlation: An experimental data matrix is used to construct a correlation matrix.
2. Decomposition: The correlation matrix is decomposed into a number of linear factors or abstract eigenvectors.
3. Rotation: This is a mathematical operation which relates physically significant parameters to the abstract eigenvectors generated by the preceding operation.
4. Combination: Real factors obtained in step 3 are combined to reproduce a data matrix within the required precision--i.e., to obtain the best solution.
5. Prediction: Good solutions obtained in step 4 are used to predict new data, either by interpolation or by the use of best combinations.

The FA method has been used in a large number of chemical systems and in interactive studies of molecule-biological test pairs.[20-23] It has been used extensively in conjunction with SAR for the study of chemical-biological interactions. The utility of FA and other statistical methods as applied to QSAR has recently been reviewed by Martin.[24]

PATTERN RECOGNITION (PR)

Frequently experimental science also requires the prediction of properties not amenable to direct measurement. Thus we may infer elemental composition from atomic emission or absorption spectra; use quantum theory to provide a rational model of atomic structure, and group theory to describe molecular structure. Chemical interactions with bio-systems constitute yet another domain where cause and effect are often obscured by system noise, and their existence must somehow be indirectly inferred from available data. In addition, direct measurement is becoming increasingly less practical for economic and logistic reasons.[25] Here pattern recognition (PR) is a valuable tool for clarifying and ordering information in an efficient and economic manner. It minimizes the quantity of data required and aids in the selection of suitable parameters for achieving the desired correlations.[26]

PR is an expanding branch of artificial intelligence which has long been familiar to biologists, engineers and psychologists.[26,27] The sole assumption made in PR is that a relationship exists between a set of data and a specified category. For a chemical system this category will be defined in terms of, say, a functional group, or some structural feature (e.g., unsaturation, branching). One then attempts to interpret the experimental data obtained in terms of this classification.

Methodology of Pattern Recognition

Operationally the methods of PR fall into two classes, parametric and non-parametric (Figure 9). Parametric methods assume access to probability density functions not usually available for chemical-biological interaction problems. I will therefore confine my remarks to the non-parametric branch of the PR diagram which is devoid of any *à priori* statistical assumptions concerning data distribution.

First, consider each experimental data point in a collection of measurements as an object in n-dimensional space with coordinates equal to its measurements. The Euclidean distance d_{ij} between any two points, mathematically defined as

$$d_{ij} = [(X_{ik} - X_{jk})^{2}]^{1/2} \qquad (13)$$

is a measure of their similarity. As similarity between data points i and j increases, the distance between them decreases, approaching zero. Therefore we define a new similarity function S_{ij}, such that

$$S_{ij} = 1 - d_{ij}/D_{ij}. \qquad (14)$$

Here D_{ij} is the maximum distance between X_i and X_j and $S_{ij} \rightarrow 1$ when d_{ij} assumes a minimum value.

Next, classification and learning processes operate on the n-space in one of two learning modes--supervised or unsupervised. In supervised learning, some of the points are classified and function as a "training set" which can then be used to classify unknown points. In unsupervised learning there is no training set. Instead, the objective is to locate clusters of points in n-space which serve as clues to possibly significant relationships. In either case the basic aim of the PR method is to classify the patterns obtained into well-defined categories.

Preprocessing involves changing the actual structure of points in n-space, and is minimal in the case of unsupervised learning, generally being confined to the scaling of measurements with different units, so as to obtain equal weighting, regardless of the units employed. In the case of supervised learning, data preprocessing may include algebraic transformations, actual changes in variables via mathematical transforms, and feature selection. These operations serve to enhance pattern discrimination by spreading clusters further apart or by reducing the dimensionality of the n-space.

Display of Data: Mapping

If parameters of a system have been judiciously selected with regard to the property being studied, like objects will have similar measurements, hence their proximity in n-space. For n-space >3 computer techniques can be used to reduce the data to a more manageable 2- or 3-dimensional space. Here the technique of nonlinear mapping (nlm) is often employed to preserve interpoint distances in the ordered space. Figure 10 shows the acid-base separation achieved in a data set abstracted from the periodic table, using the following six properties (n=6) to describe each element: (1) most important valence; (2) melting point; (3) covalent radius; (4) ionic radius; (5) electronegativity and (6) ΔH_{fusion}. None of these properties used alone could achieve such separation.[28]

Another method of displaying n-dimensional data involves their projection onto a selected 2-dimensional plane after appropriate weighting of the data. This method was used by Ting et al in their successful classification of sixty-six drugs as tranquillizers or sedatives.[29]

Pattern recognition techniques have been applied to the screening of prospective anti-cancer drugs[30] and to structure-activity studies of chemical carcinogens.[31] A detailed discussion of the applications of PR to drug design is given in Ref. 32 by Kirschner and Kowalski.[33] Reports of environmental applications of PR are appearing in the current literature with increasing frequency.[34-37] Extensive references to the application of PR methods to chemical, medical and environmental problems are given in at least three recent books dealing totally or in part with computer assisted methods.[32,38,39]

With the wealth of information already available, can application of PR to the multivariate chemical-biological problems associated with the development of new oil shale and other fossil fuel technologies be far behind?

SUMMARY AND CONCLUSIONS

The potentially toxic byproducts of surface and *in situ* oil shale retorting are too large in number to examine individually. Nor is this desirable, since the insults to humans and ecosystems will be in the form of complex mixtures of toxic gases, particulates, unretorted and spent shales, shale oils and retort waters. Shale oils contain the more volatile toxic trace metals (Hg,As, Sb,Se,V, etc.) and a large number of organic contaminants which are separable into three complex fractions: neutral, acid and base. Although the neutral fractions contain the most potent carcinogens - the PAH's - their total amount appears to be low compared with that found in coal and coal-based synfuels, and therefore not of major concern. The base fractions, however, show greater total mutagenic activity and contain a large number of aryl amines and nitrogen heterocyclics which are either known or suspect carcinogens. It is also this fraction which may mobilize the transition metals found in methylene chloride extracts of retort waters by complexing them (Kland, et al).

No single *in vitro* test applied to such complex mixtures can be expected to be of sufficient universal reliability to provide a criterion for the assessment of potential environmental and health effects. Use of a number of *in vitro* tests dependent on different biological mechanisms will enhance the reliability of prediction (Legator, THIS VOLUME). Best of all is a combination of these with judiciously selected *in vivo* testing.

The complex nature of both the contaminants and biological test systems involved in the assessment of environmental and health effects of oil shale technologies requires a statistical matrix treatment. Four methods of dealing with such multivariate biological-chemical systems have been described: quantitative structure-activity and molecular connectivity relationships (QSAR,MC), factor analysis (FA) and pattern recognition (PR). QSAR and MC are useful in the prediction of toxic behavior for individual members of a class of compounds for which much SAR data are already available. The QSAR approach uses mathematical functions based on octanol-water distribution coefficients, electronic, steric, hydrophobic and resonance effects, and molar refractions. QSAR is a statistical method. Only objective data are used. It is therefore an excellent predictive tool. However it is not particularly useful in dealing with chemical mixtures, where complex synergistic effects may be operative, and its greatest successes have been achieved in the fields of drug design and the prediction of new drug behavior.

MC utilizes the topological characteristics of molecules to develop nonempirical numerical values based on the connectivities of the atoms within the molecular structure. These are in turn related to actual physicochemical and biological properties of the molecule using a theoretical model or experimental data. The MC approach holds out the very attractive prospect of reducing the essential empirical components of a matrix, and in this respect has some advantage over SAR. It suffers from the same limitations, however, with respect to complex systems. This leaves the methods of factor analysis and pattern recognition, to cope with complex chemical mixtures perturbing biological-environmental systems.

The methods of FA and PR have both been applied to data derived from SARs and PR techniques have also been used with connectivity functions (Jurs). Thus, both of these methods may be viewed as primarily mathematical techniques for operating on any data to obtain a) more generally applicable solutions in the case of FA and b) reveal clustering or patterns in the data via PR. The biological processes and chemical structures of SARs while implied, are totally irrelevant. PR is a particularly powerful tool for discrimination--e.g., the detection of relationships in categories of data, *regardless of their significance*. It should most certainly be applied to the data matrix for parameters already measured on repository samples, with a view to selecting the most useful information, reducing the number of parameters or measurements required, and defining the relationship(s) between processes and data.

PR should also be applied to occupational and health data where currently available from employee and patient profiles (e.g., Stallard, THIS SYMPOSIUM). In the course of a developing synfuels industry much more occupational health data will become available to health professionals. The early application of PR methods to the growing data based could help anticipate problem areas and enable appropriate preventive action before the industry is burdened with costly employee compensation claims and excessive lost time from job-related causes.

ACKNOWLEDGMENTS

Thanks are due Dr. James C. Bartholomew of the Chemical Biodynamics Division for his review of this paper and helpful comments on the SAR section; to Professor Malcolm M. Renfrew, whose interest in the area of complex environmental systems has resulted in a most stimulating dialog; and to Dr. Christian Daughton of the Shale Group, who reviewed and commented on sections of this chapter.

References

1. R. O. Beauchamp, Jr. and M. D. Shelby, "Chemicals Identified in Oil Shale and Shale Oil. 1. Preliminary list." Environmental Mutagen Information Center, Information Div., Oak Ridge National Lab., Oak Ridge, TN 37830, 19 pp.

2a. M. J. Kland, H. L. Eaton and A. S. Newton, "Characterzation of Trace Contaminants in Oil Shale Retort Waters", Am. Chem. Soc. 35th NW. Regional Mtg., Salt Lake City, Utah, June 14, 1980, Paper No. 28 LBL-10850.

b. ____, ____ and ____, "Trace Contaminants in Oil Shale Retort Waters", in Oil Shale Research: Characterization Studies, 1. E & E Annual Rept. 79. LBL-10486.

3. M. J. Kland, "The VC-PVC Crisis: A systematic Approach to Toxicological Problems", Am. Chem. Soc., Symp. on Environ. Chem., 30th NW Regional Mtg., Honolulu, Hawaii, June 12-13 (1975), Paper No. 84 LBL-3275.

4. R. Foster, "Biochemical Systems", Chap. 12 in Charge Transfer Complexes, Academic Press, London (1970). 335-373. See also Chap. 3, "Electronic Spectra", 33-93.

5. Elizabeth C. and James A. Miller, "Metabolism of Chemical Carcinogens to Reactive Electrophiles and their Possible Mechanisms of Action in Carcinogenesis", Chap. 16 in Chemical Carcinogens, ACS Monograph 173, C. E. Searle, ed. Washington, DC (1976). 737-62.

6a. A. Crum Brown and T. R. Fraser, "On the Connection between Chemical Constitution and Physiological Action. Part I. (On the) Physiological Action of (the) Salts of (the) Ammonium Bases Derived from Strychnia Brucia, Thebaia, Codeia, Morphia, Nicotia", Trans. Royal Soc. Edinburgh 25, 151-203 (1868).

b. "On the Physiological Action of the Ammonia Bases Derived from Atropia and Conia", XX, Part II. ibid, 693-739 (1869).

7. L. P. Hammett, "Effect of Structure upon the Reactions of Organic Compounds: Benzene Derivatives", J. Am. Chem. Soc. 59, 96-103 (1937).

8. C. Hansch and T. Fujita, "ρ-σ-π Analysis: (A) Method for (the) Correlation of Biological Activity and Chemical Structure", J. Am. Chem. Soc. 86, 1616-26 (1964).

9. M. A. H. Fahmy, T. R. Fukuto, R. L. Metcalf and R. L. Holmstead, "Structure-Activity Correlations in DDT Analogs", J. Agric. Fd. Chem. 21 (4) 585-91 (1973).

10. R. W. Taft, Jr., "Separation of Polar, Resonance and Steric Effects in Reactivity", in Steric Effects in Organic Chemistry, M. S. Newman, ed., Wiley, New York (1956) 556-675.

11. R. Osman, H. Weinstein and J. P. Green, "Parameters and Methods in Quantitative Structure-Activity Relationships", Chap. 2 in Computer Assisted Drug Design,

E. C. Olson and R. E. Christofferson, editors, ACS Symp. Ser. 112, Washington, DC (1979). 21-77.

12. A. Dipple, "Polynuclear Aromatic Hydrocarbons", Chap. 5, in Chemical Carcinogens, Chas. E. Searle, ed. ACS Monograph 173 (1976) 245-314.

13. M. Yuan and P. C. Jurs, "Computer-Assisted Structure-Activity Studies of Chemical Carcinogens: A Polycyclic Aromatic Hydrocarbon Data Set", Tox. and Appl. Pharm. 52, 294-312 (1980).

14. D. B. Clayson and R. C. Garner, "Carcinogenic Aromatic Amines", Chapter 8 in Chemical Carcinogens (Ref. 12) 366-461.

15. M. R. Guerin, B. R. Clark, C.-h. Ho, J. L. Epler and T. K. Rao, "Short-Term Bioassay of Complex Organic Mixtures: Part I, Chemistry", Proc. EPA Symp. on Application of Short-Term Bioassay in the Fractionation and Analysis of Complex Environmental Mixtures, Williamsburg, VA. M. D. Waters, et al, editors, Plenum Press, New York (1978) 247-268.

16. L. B. Kier and L. H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York (1976). Preface

17. L. B. Kier and W. J. Murray, "Molecular Connectivity, II: Relationships to Biological Activities", J. Med. Chem. 18 (12), 1272-4 (1975).

18. W. J. Murray, L. H. Hall and L. B. Kier, "Molecular Connectivity, III: Relationship to Partition Coefficients", J. Pharm. Sci. 64 (12), 1978-81 (1975).

19. R. J. Rummel, Applied Factor Analysis, Northwestern University Press, Evanston, IL (1970).

20. M. L. Weiner and P. H. Weiner, "A Study of Structure-Activity Relationships of a Series of Diphenylaminopropanols by Factor Analysis", J. Med. Chem. 16 (6), 655-61 (1973).

21. R. Wooton, R. Cranfield, G. C. Sheppey and P. J. Goodford, "Physicochemical Activity Relationships in Practice. 2. Rational Selection of Benzenoid Substituents", J. Med. Chem. 18 (6), 607-13 (1975).

22. D. G. Howery, "Factor Analyzing the Multifactor Data of Chemistry", Amer. Lab. 8 (2) 14-25 (1976).

23. P. H. Weiner and D. G. Howery, "Factor Analysis of Some Chemical and Physical Influences in Gas - Liquid Chromatography", Anal. Chem. 44(7) 1189-94 (1972).

24. Y. C. Martin, "Other Mathematical Methods of Use in Quantitative Structure-Activity Studies", Chap. 10 in Quantitative Drug Design, Dekker, New York (1978). 233-252

25. M. Kland, "A Priori Predictive Methods of Assessing Health Effects of Chemicals in the Environment", in Water Chlorination: Environmental Impact and Health

Effects, Vol. 2, Ann Arbor Science Pub., Ann Arbor, MI (1978) 451-69.

26. T. L. Isenhour and P. C. Jurs, "Some Chemical Applications of Machine Intelligence", in Report for Analytical Chemists, Anal. Chem., 43 (10) 20A-35A, (1971).

27. J. C. MacDonald, "Time-Shared Pattern Recognition", Amer. Lab. 9 (2), 31-34 (1977).

28. B. R. Kowalski and C. F. Bender, "Pattern Recognition. A Powerful Approach to Interpreting Chemical Data", J. Am. Chem. Soc. 94 (16), 5632-9 (1972).

29. K-L. H. Ting, R. C. T. Lee, G. W. A. Milne, and A. M. Guarino, "Applications of Artifical Intelligence: Relationships Between Mass Spectra and Pharmacological Activity of Drugs", Science 180 417-420 (1973).

30. B. R. Kowalski and C. F. Bender, "Application of Pattern Recognition to Screening Prospective Anticancer Drugs", J. Am. Chem. Soc. 96, 916-918 (1974).

31. D. S. Dierdorf and B. R. Kowalski, "Three-Dimensional Molecular Structure-Biological Activity Correlations by Pattern Recognition", NTIS Rept. No. AD-785863/2SL (1974). 41pp.

32. Medicinal Chemistry, VIII, E. J. Ariens, Ed. Academic Press, New York (1979).

33. G. L. Kirschner and B. R. Kowalski, "The Application of Pattern Recognition to Drug design", Chapter 2 in Ref. 32, 73-131.

34. J. R. McGill and B. R. Kowalski, "Recognizing Patterns in Trace Elements", Applied Spectroscopy 31 87-95 (1977).

35. P. K. Hopke, "Application of Multivariate Analysis for Interpretation of the Chemical and Physical Analysis of Lake Sediments", J. Environ. Sci. Health, All (6) 367-83 (1976).

36. P. D. Gaarenstroom, S. P. Perone and J. L. Moyers, "Application of Pattern Recognition and Factor Analysis for Characterization of Atmosphere Particulate Composition in Southwest Desert Atmosphere, Environ. Sci. Technol. 11 (8), 795-800 (1977).

37. P. L. Briggs and F. Press, "Pattern Recognition Applied to Uranium Prospecting", Nature 268, 125-127 (1977).

38. A. J. Stuper, W. E. Brugger and P. C. Jurs, Computer Assisted Studies of Chemical Structure and Biological Function, John Wiley and Sons, New York (1979). 220 pp.

39. Computer-Assisted Drug Design, E. C. Olson and R. E. Christofferson, Editors, ACS Symp. Series 112, Washington, DC (1979). 619 pp.

Table I. Subgraph Types

Order (m)	Valency (δ)	Path type (t)	Path descriptor (P)
0	1	————	————
1	2	path	P
2	3	path	P
3	4	path, cluster (star),	P,C
		chain (triangle)	CH
4	5	path, cluster	P,C
		path/cluster	P/C
		chain (cycle)	CH

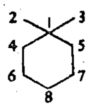
Table II. Connectivity Indices ${}^{m}\chi_t$

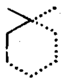
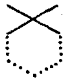
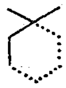
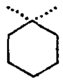
Term order, m	Vertex No., n	Path type(s)	Equation
0	1	P	${}^{0}\chi = \sum_{i=1}^{n} \delta_i^{-1/2}$
1	2	P	${}^{1}\chi = \sum_{s=1}^{N_e} (\delta_i \, \delta_j)_s^{-1/2}$
2	3	P	${}^{2}\chi = \sum_{s=1}^{n_m} (\delta_i \delta_j \delta_k)_s^{-1/2}$
3	4	P, C, CH	${}^{3}\chi_t = \sum_{s=1}^{n_m} (\delta_i \delta_j \delta_k \delta_l)_s^{-1/2}$
4	5	P, C, P/C, CH	${}^{4}\chi_t = \sum_{s=1}^{n_m} (\delta_i \delta_j \delta_k \delta_l \delta_p)_s^{-1/2}$

Table III. Equations of MC Theory

(1)	$\sum_{i=1}^{n} \delta_i = 2m$	$G(v_i)$, $(i = 1 \ . \ . \ . \ n)$ v = vertex δ = vertex valence m = no. of edges
(2)	$m = n-1$	$G^m_{v_n}$ Tree graph
(3)	$\delta_i = \sum_{j=1}^{n} T_{ij}$	Vertex valence from summation of row i, matrix T_{ij}
(4)	$C(\chi) = b_o + \sum_{m,t} b_t(m)^m\chi_t$	Connectivity function
(5)	${}^m\chi_T = \sum_{j=1}^{n_m} {}^mS_j$	Connectivity index
(6)	${}^mS_j = \prod_{i=1}^{m+1} (\delta_i)_j^{-1/2}$	Subgraph term (edge set j, subgraph order m)
(7)	$E_s = 1/2 \sum_{i=1}^{m} \sum_{j=1}^{m} A_{ij}$	Subgraph edges A = subgraph adjacency

Table IV. Subgraph evaluation and enumeration algorithm for 1,1-dimethylcyclohexane[a]

Example Graph	Topological Matrix
	$\begin{pmatrix} 0&1&1&1&1&0&0&0\\ 1&0&0&0&0&0&0&0\\ 1&0&0&0&0&0&0&0\\ 1&0&0&0&0&1&0&0\\ 1&0&0&0&0&0&1&0\\ 0&0&0&1&0&0&0&1\\ 0&0&0&0&1&0&0&1\\ 0&0&0&0&0&1&1&0 \end{pmatrix}$

Vertex Number Sets	Subgraph Matrix	v_i^s	Subgraph	Type
1 2 4 6	$\begin{pmatrix} 0&1&1&0\\ 1&0&0&0\\ 1&0&0&1\\ 0&0&1&0 \end{pmatrix}$	2 1 2 1 $E_s = 3$		Path
1 2 3 4 5	$\begin{pmatrix} 0&1&1&1&1\\ 1&0&0&0&0\\ 1&0&0&0&0\\ 1&0&0&0&0\\ 1&0&0&0&0 \end{pmatrix}$	4 1 1 1 1 $E_s = 4$		Cluster (star)
1 2 3 4 6	$\begin{pmatrix} 0&1&1&1&0\\ 1&0&0&0&0\\ 1&0&0&0&0\\ 1&0&0&0&1\\ 0&0&0&1&0 \end{pmatrix}$	3 1 1 2 1 $E_s = 4$		Path/Cluster
1 4 5 6 7 8	$\begin{pmatrix} 0&1&1&0&0&0\\ 1&0&0&1&0&0\\ 1&0&0&0&1&0\\ 0&1&0&0&0&1\\ 0&0&1&0&0&1\\ 0&0&0&1&1&0 \end{pmatrix}$	2 2 2 2 2 2 $E_s = 6$		Circuit

(E_s = subgraph edge count)

XBL 7711-10514

[a] Adapted from Ref. 16

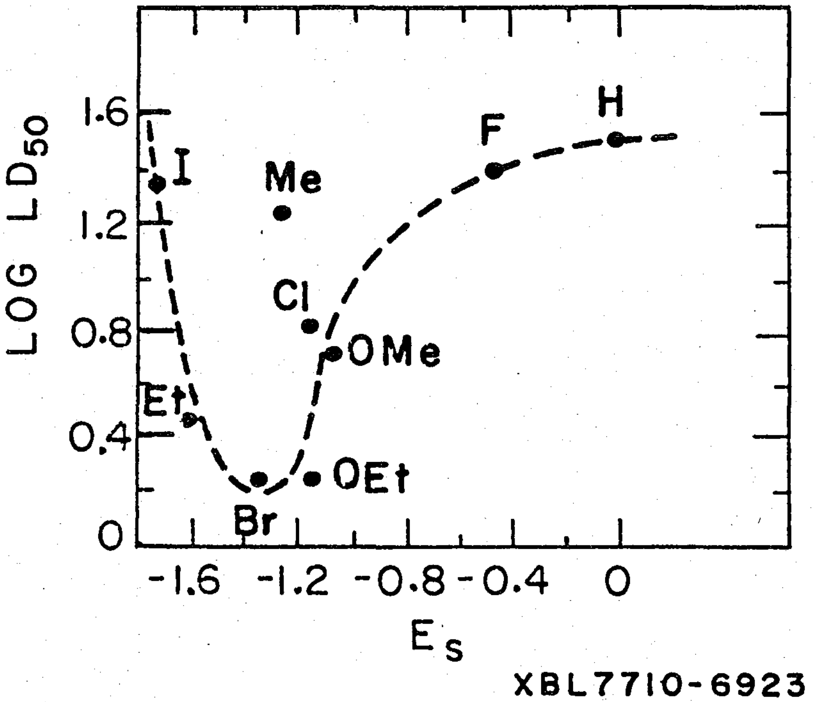


Figure 1. Relationship between toxicity and E_s for 1,1,1-trichloro-p-methyl-p'-x-diphenylethanes.

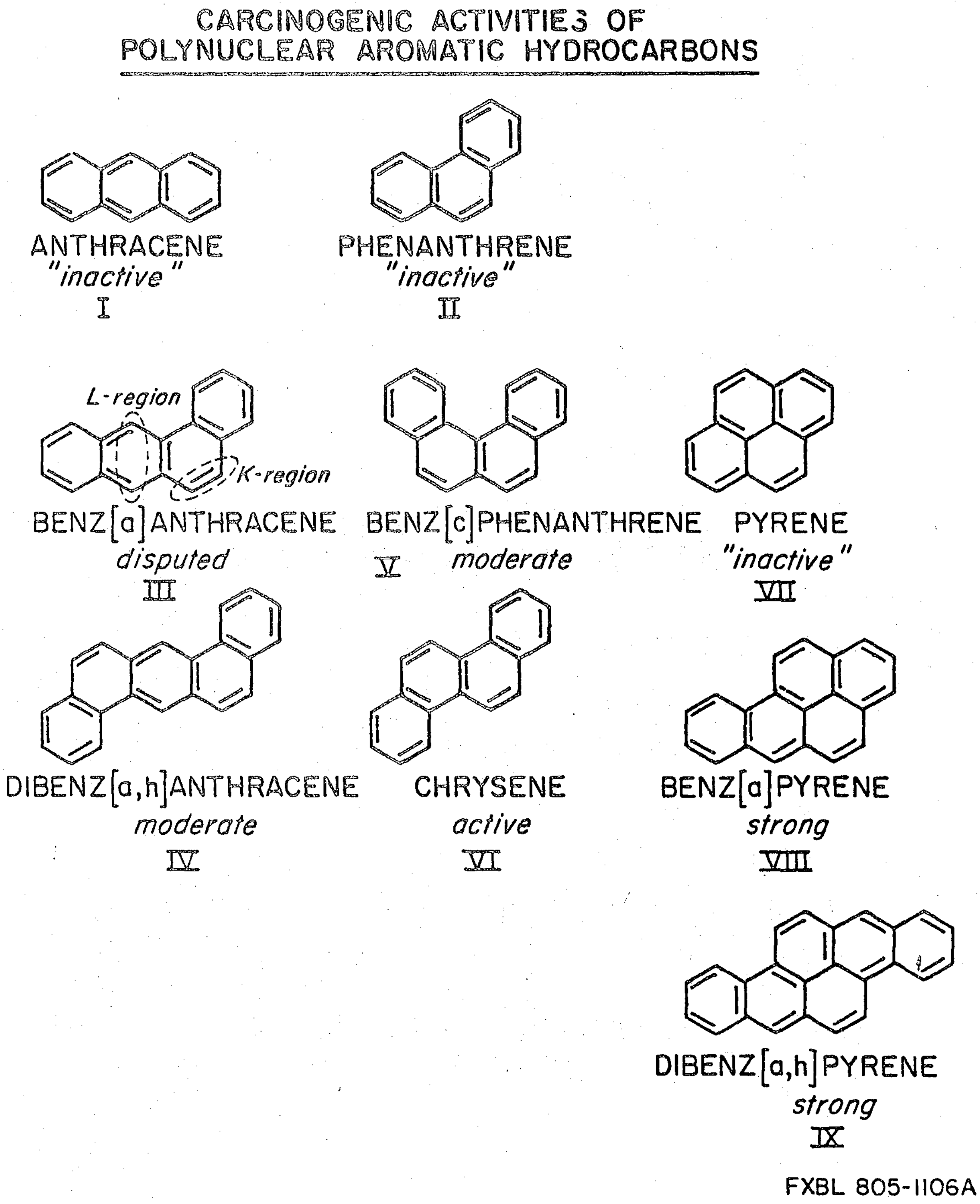
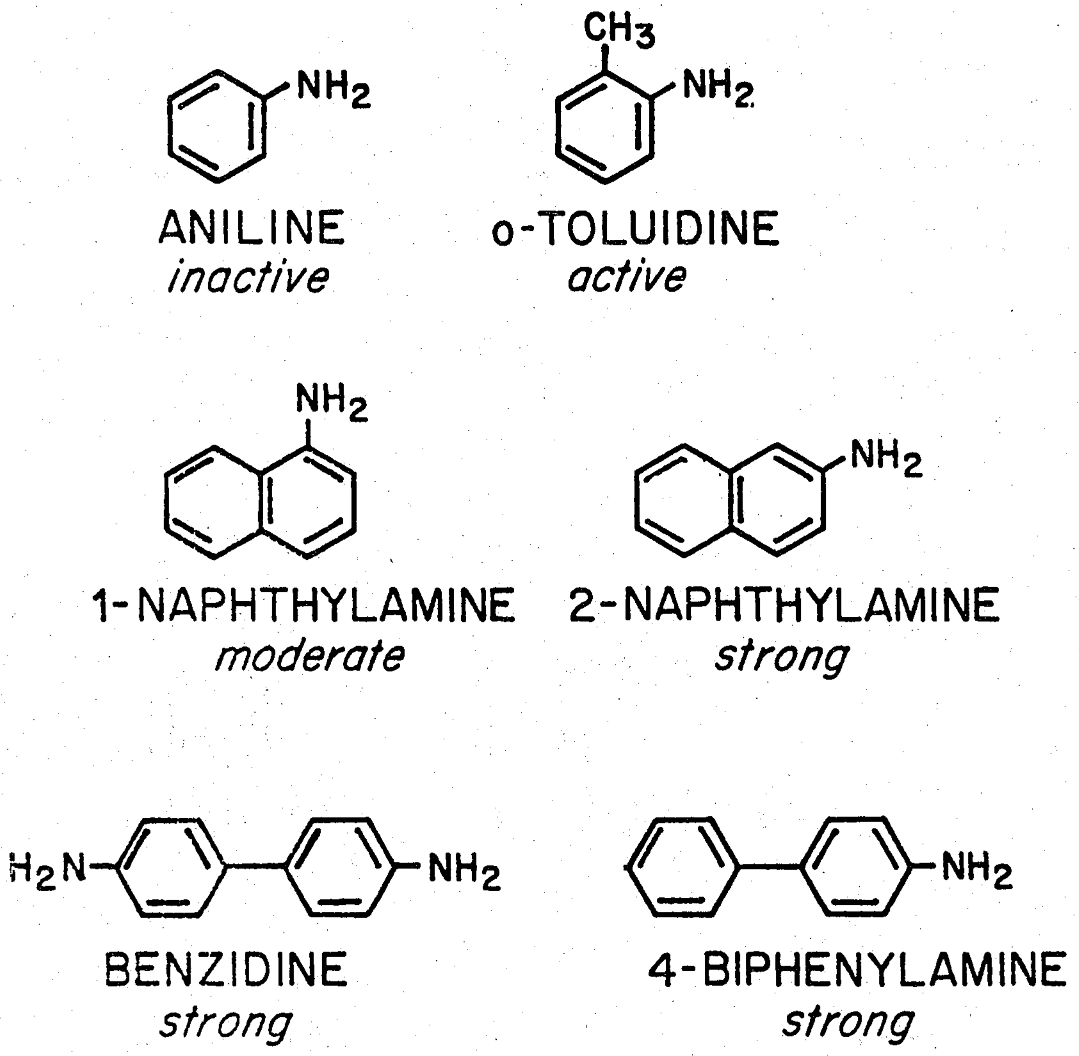


Figure 2. Carcinogenic activities of polynuclear aromatic hydrocarbons.



XBL 805-1107

Figure 3. Carcinogenic activities of aromatic amines

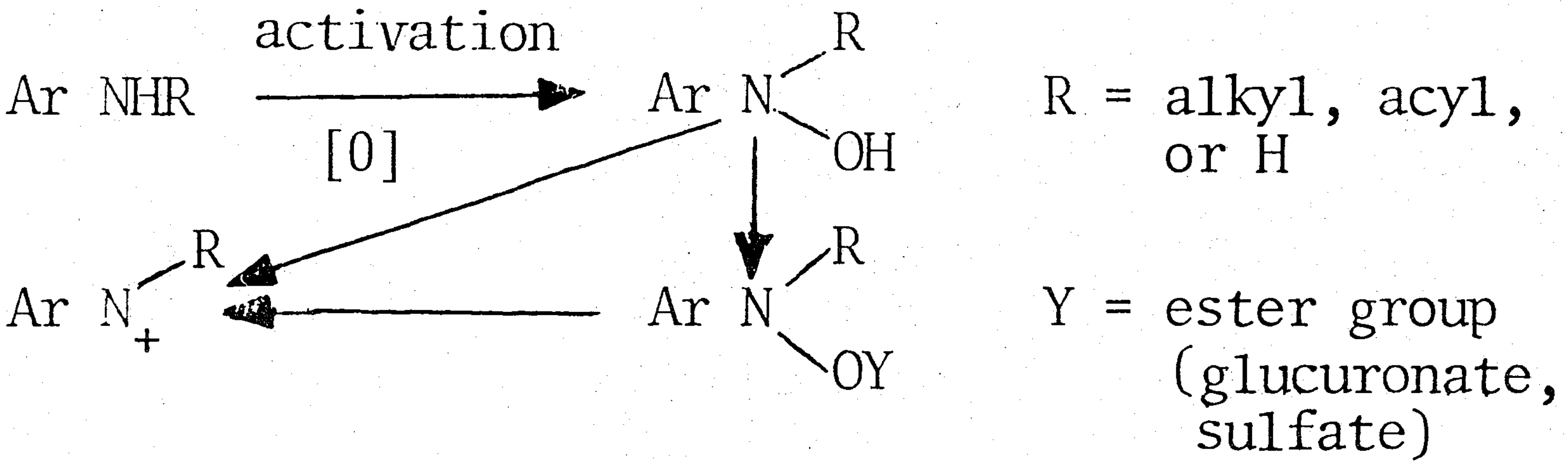


Figure 4. Possible metabolic activation pathways for aromatic amines.

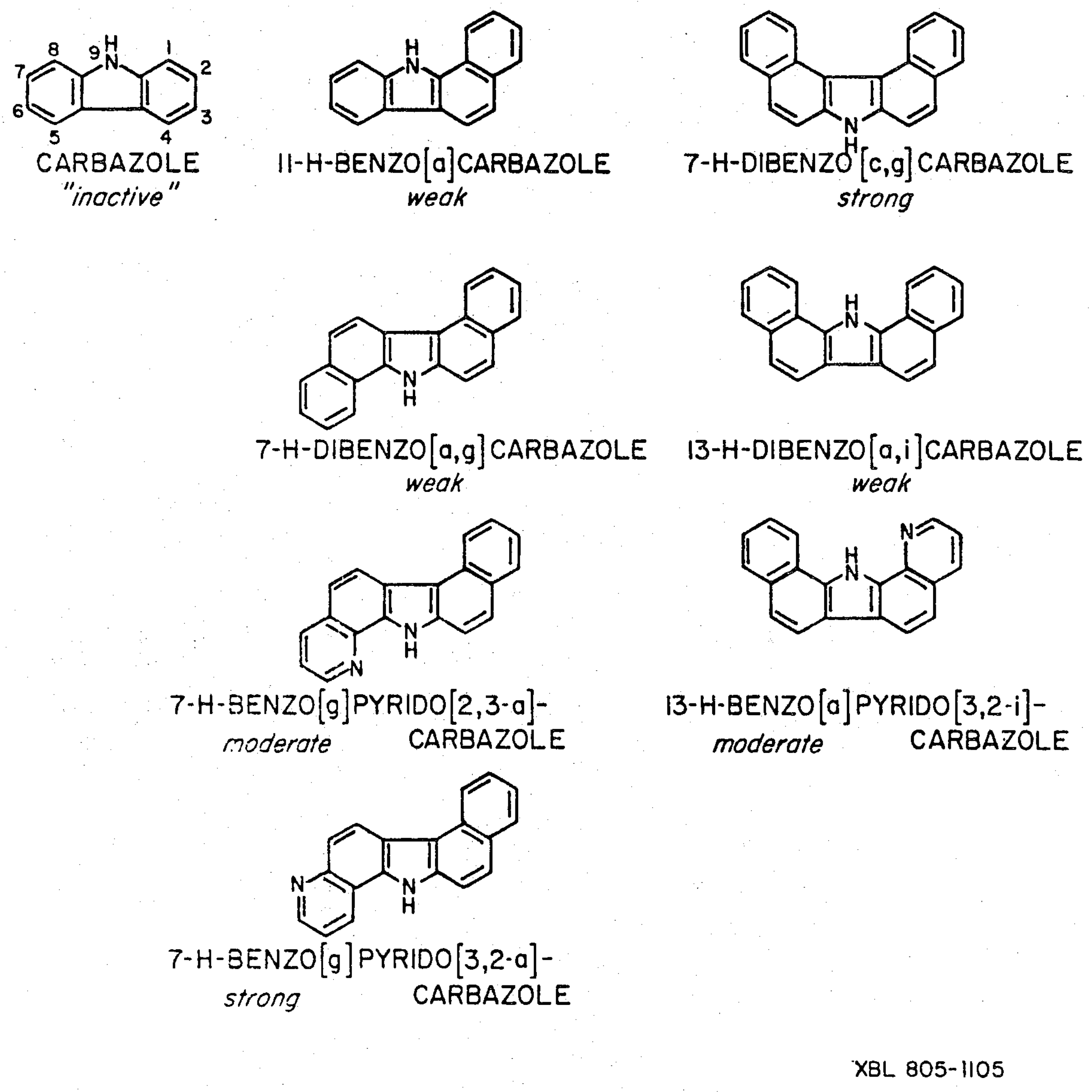
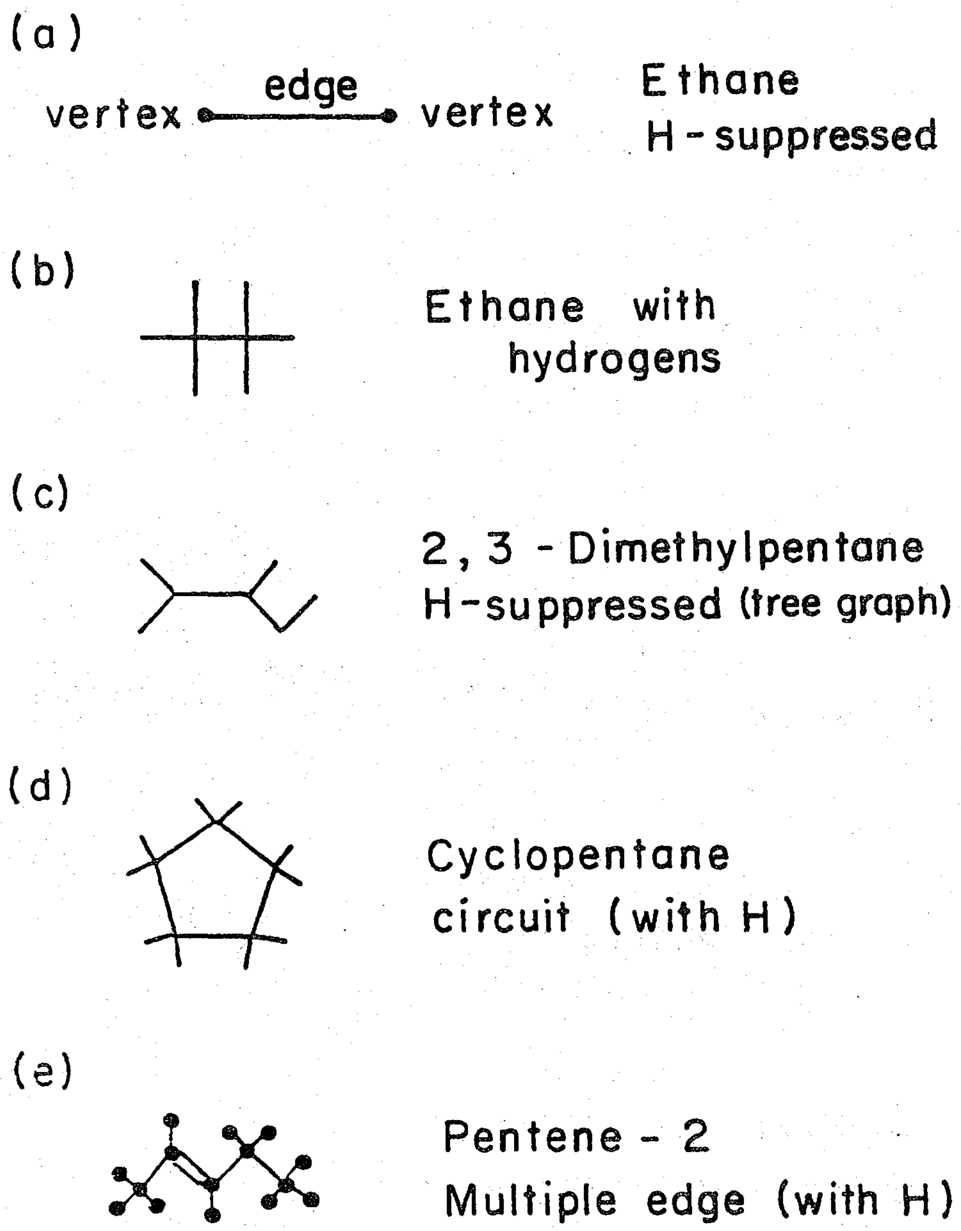
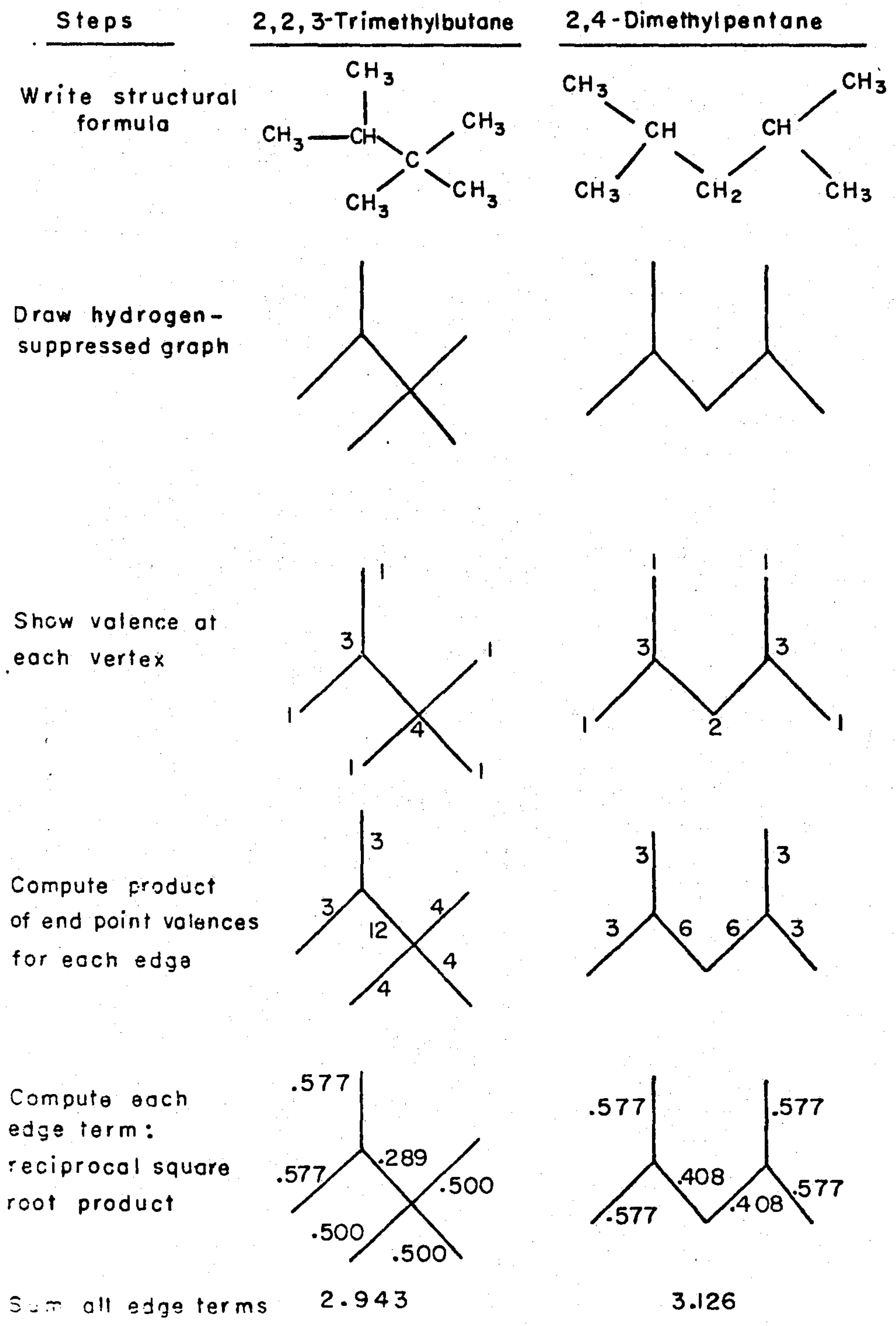

8
9
H
N
1
2
7
6
3
5
4
CARBAZOLE
"inactive"
H
N
11-H-BENZO[a]CARBAZOLE
weak
N
H
7-H-DIBENZO[c,g]CARBAZOLE
strong
N
H
7-H-DIBENZO[a,g]CARBAZOLE
weak
H
N
13-H-DIBENZO[a,i]CARBAZOLE
weak
N
N
H
7-H-BENZO[g]PYRIDO[2,3-a]-CARBAZOLE
moderate
N
H
N
13-H-BENZO[a]PYRIDO[3,2-i]-CARBAZOLE
moderate
N
N
H
7-H-BENZO[g]PYRIDO[3,2-a]-CARBAZOLE
strong
XBL 805-1105

Figure 5. Carcinogenic activities of carbazoles.



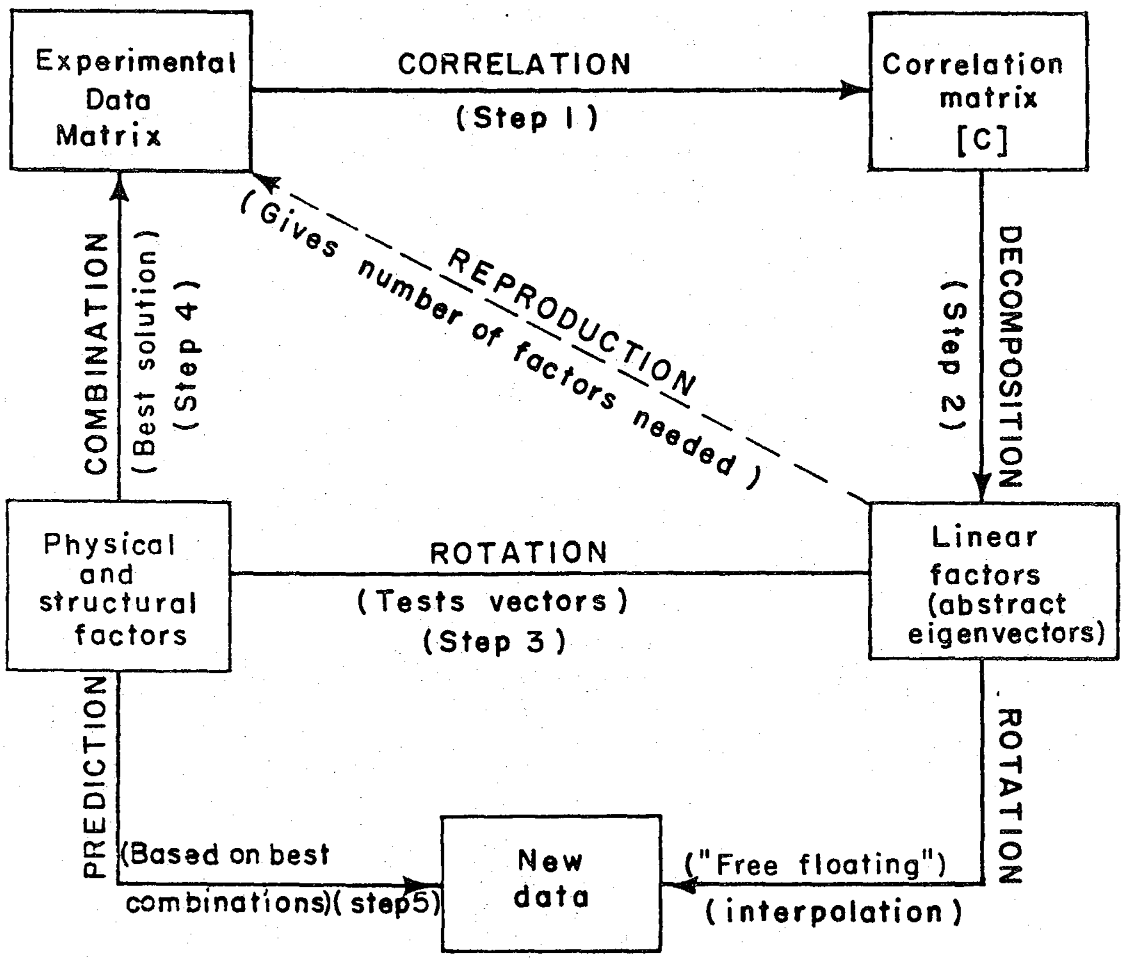
XBL7710-6918

Figure 6. Graph representations of chemical structures.



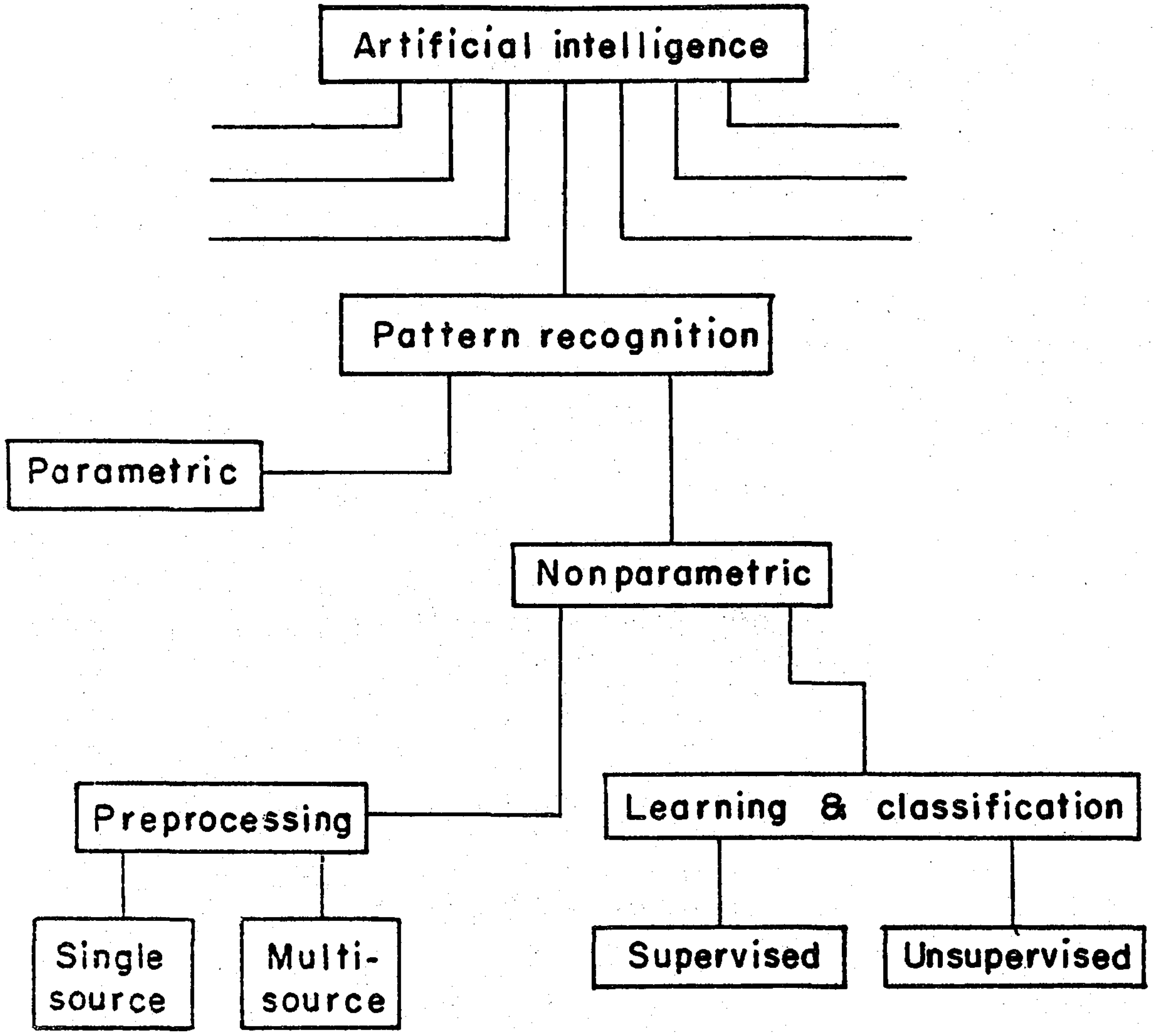
XBL7710-6921

Figure 7. Procedure for calculating connectivity index $^1\chi$.



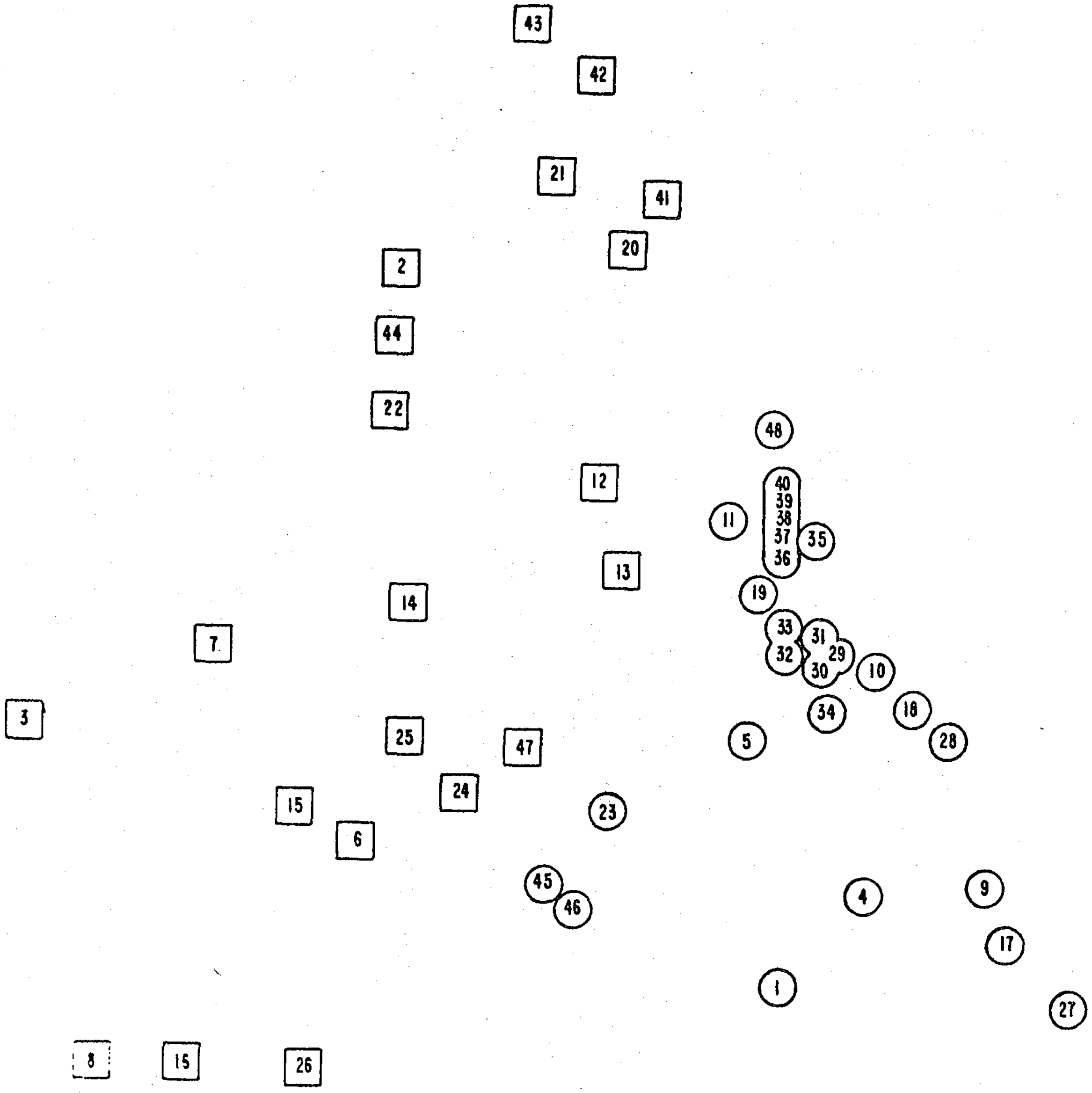
XBL779-6919

Figure 8. Factor analysis: diagram of the stepwise procedure techniques.



XBL7710-6920

Figure 9. Functional analysis of pattern recognition techniques.



XBL7710-6917

Figure 10. Separation of acids □ and bases ○ for 68 elements of the Periodic Table: nℓm from 6-space to 2-space.

This report was done with support from the United States Energy Research and Development Administration. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the United States Energy Research and Development Administration.