Substructural Analysis Techniques for Structure-Property Correlation within Computerised Chemical Information Systems

David Bawden

Ph.D. Thesis, Sheffield University

Postgraduate School of Librarianship and Information Science

February, 1978

Acknowledgements

I have to thank particularly George Adamson, for constant help and encouragement, and the staffs of PSGLIS and (in the latter phases) Pfizer Central Research for support and forbearance. Also, for varied reasons, my thanks go to Val Addis, Barbara Bateson, Pam Bratherton, Phil Briggs, Judy Bush, the Department of Education and Science, Dick Jones, Mike Lynch, Alice McLure, Ian McLure, Brian Pedley, David Saggers, Sheffield University Computing Services, Fiona Strong, Vera Swallow, Iwona Szymanska, Judith Ufton, George Vleduts, Linda Wickham, and Peter Willett.

CONTENTS

Summary

David Bawden

Ph.D. Thesis, 1978

Substructural Analysis Techniques for Structure-Property Correlation within Computerised Chemical Information Systems

Summary

The work described in this thesis involves a novel method of substructural analysis, with potential application for structure-property correlation and information retrieval within computerised chemical information systems.

A review is given of the development of the concept of chemical structure and its representation, its application in computerised chemical information systems, and methods for correlating structure with molecular properties.

A method is presented for derivation of structural features, representing the whole structure, from Wiswesser Line Notation (WLN) by computer program. These features are then used as variables in statistical analysis procedures: in this work multiple regression analysis and cluster analysis are used. This procedure allows for a rapid, convenient and thorough analysis of large data-sets. The type of structural features used may be easily varied, allowing for investigation of factors such as ring substitution patterns, group interactions, and three-dimensional structure. The method is applicable to sets of diverse or structurally related compounds. Statistical tests of the results enable quantitative testing of hypotheses.

Multiple regression analysis allows a direct, quantitative correlation between structure and molecular property, and subsequent property prediction. It is applied to sets of aliphatic, alicyclic,

aromatic, and heterocyclic compounds, including sets of highly diverse structures. Properties examined include biological effects, toxicty, pK, thermochemical properties, boiling point, solubility, and partition coefficient. Some of these properties are highly dependent upon electronic and steric effects, and hence upon relative position of substituents, and on three-dimensional structure. Highly significant correlations are obtained in all cases, and the potential for property prediction is demonstrated.

Cluster analysis is applied to several sets of structures. Intuitively sensible classifications are obtained, and the potential for both property prediction and information retrieval discussed.

Since these techniques involve the widely used WLN, relatively simple COBOL programs, and standard statistical packages, they should be applicable within operational environments.

Chapter One

Introduction

'Book One, Part One, Chapter One, Page One.

What à great start'

(Charles Schulz)

This thesis is divided into four main sections.

Chapter 2 "sets the scene" by reviewing the background to this work. Firstly, the development of the concept of chemical structure, and its representation is discussed. Secondly, the relatively recent incorporation of these ideas within computerised chemical information systems is outlined. Thirdly, the various methods of quantitative structure-property correlation are described.

In Chapter 3, aspects of substructural analysis techniques are outlined. Fragmentation techniques are discussed, and the statistical techniques to be used are described.

Chapter 4 contains an account of the analyses based on WLN structure representation which comprise the practical work of this study. This chapter is divided into two parts, the first dealing with multiple regression analysis, and the second with cluster analysis.

Chapter 5 concludes and summarises the work. Brief details of the computer programs are given in an Appendix.

Chapter Two

Background

'It is hard enough to keep up with life, let alone technical information'

(Anonymous survey respondent)

The purpose of this chapter is to give an outline of the historical development, present state of knowledge, and current practical applications of areas of study relevant to the subject matter of this thesis, in order to put the work described below into context. This involves concentration on two major themes: firstly the development and applications of chemical structure representations, and secondly the development and application of methods for the correlation of chemical structure with molecular properties.

The unifying factor here is the concept of chemical structure. This provides an unambiguous description of a compound, and a basis for rationalising its observed properties, from which may be deduced general principles regarding the relationships between structure and property. It is significant that the first concept of a straightforward quantitative relationship between chemical structure and biological activity was put forward partly by Alexander Crum Brown (CRUM BROWN et al, 1868) who made notable contributions to the development and use of structural representations. This structure concept is fundamental to the greater part of modern chemically related sciences.

The ability to adequately and conveniently represent chemical structure is of obvious importance. It will be seen that, although sophisticated mathematical models of molecular structure have been devised, the chemical structure diagram remains the most widely used and valuable representation, and that computer-readable structure representations are, for the most part, partial or total representations of a structure diagram. These structural formulae are not only the most convenient means for communication

of chemical ideas, but are, as Hammond has pointed out, "the basic vehicle in the search for patterns (in chemical data)" (HAMMOND, 1974). Much of the discussion below will centre on their adequacy in this search.

2.2.

Chemical Structure Representation

In the section below, the development of the concept of chemical structure and its representation will be considered first. The handling of structural representations in chemical information systems will be briefly discussed, and some relevant aspects of the operation of such systems will be considered. The discussion will centre on chemical structure information systems, i.e. those having files of structural representations with which other forms of data are associated. It is with such systems, especially if computer-based, that the techniques of structure-property correlation to be described below are likely to prove of most value.

2.2.1.

Development of the Concept of Chemical Structure and its Representations

2.2.1. Development of the Concept of Chemical Structure and its Representation

The advances made within the past two centuries in those aspects of the natural sciences which depend on an understanding of the properties of material substances, which encompass a large part of modern science and technology, have depended heavily upon the development of two chemical concepts. The first of these is the concept of chemical composition, i.e. the make-up of a substance in terms of numbers of constituent atoms. The second is the concept of chemical structure, i.e. the arrangement of the constituent atoms.

The recognition of the significance of these factors and their elucidation for a great variety of compounds were in themselves important for the development of all branches of chemistry. Their full value could only be realised, however, with the devising of methods for representing composition and structure, either as a nomenclature, a notation, or a diagram. Only when this was achieved could the unifying concepts of composition and structure be used to rationalise, systematise and record the mass of experimental observations produced in the upsurge of chemical research which commenced in the early nineteenth century (PARTINGTON, 1964a).

It has been pointed out that "to write a full description of the origin, growth and misadventures of the language of chemistry is to write a history of the science" (MUIR, 1907). Similarly, in order to describe the development of chemical structural representation it is necessary to touch on the more important steges in the increasing understanding of chemical structure. This subject has been fully treated elsewhere (RUSSELL, 1971a, PORTER, 1965, MASON, 1976, MACHIE, 1954) and will be dealt with very briefly here.

An excellent detailed account of the history of chemical symbolism and nomenclature is available (CROSLAND, 1962a) on which much of the discussion below is based, and reviews of particular aspects of this topic also exist (DYSON, 1953, WISWESSER, 1968, WISWESSER, 1975, WINDERLICH, 1953).

The discussion below will concentrate on the development and use of the chemical structure diagram, which, in addition to its role as a major tool for the communication of chemical ideas, is the basis for structure representation in the majority of chemical information systems (HYDE, 1975).

Early Chemical Nomenclature

Chemistry, in the modern sense of the term, is often considered to have originated in the late eighteenth century, with the introduction of new theories of chemical composition and reaction, most notably by Priestley and Lavoisier (PARTINGTON, 1962). For thousands of years prior to this, however, investigations of chemical substances had been carried out for technical, medicinal, and alchemical purposes (TAYLOR, 1976a, PARTINGTON, 1970) and appropriate terminologies and symbolisms had been developed for the substances used (CROSLAND, 1962a, RUSSELL, 1971g, RUSSELL, 1971H). These could not be systematic, in a modern sense, since the idea of chemical composition was unknown, the distinctions between elements, compounds, and mixtures or alloys were not understood, and adequate methods for the identification of substances did not exist.

Probably the first example of an attempt to produce a form of systematic nomenclature or classification for chemical substances was that of the Sumerians of about the seventh century B.C. (THOMSON, 1936).

An initial term, in their cuneiform language, represented an outstanding property of a class of substances, e.g. ZA (denoting rock or stone), and was followed by suffixes indicating property, e.g. GIN (blue), TW (heavy), AS (hard), AS-AS (very hard), AZTW (effervescent with acid). Thus sapphire was denoted by ZA.GIN.AS.AS, i.e. very hard blue stone.

The majority of other early terminologies were based upon the appearance or an obvious property of the substance, or upon a person or place associated with it.

Some examples of this type of terminology are given here:

i) based on colour (widely used, particularly in early civilisations, and having magical connotations) e.g. plumbum candidum (white lead), denoting tin.

ii) based on consistency
e.g. butter of zinc, denoting zinc chloride.

iii) based on crystalline form
e.g. cubic nitre, denoting sodium nitrate

iv) based on taste or smell
e.g. sugar of lead, denoting lead acetate
bitter salt, denoting magnesium sulphate.

v) based on personal name
e.g. Glauber's salt, denoting sodium sulphate.

vi) based on place name
e.g. Rochelle salt, denoting sodium potassium tartrate.

vii) based on medicinal properties
e.g. diuretic salt, denoting potassium acetate.

viii) based on method of preparation (attained wide use only in the eighteenth century)
e.g. spiritus salis ammoniaci cum sale akali parata, denoting ammonium carbonate.

An additional complicating factor was the use of terminology based on astrological and alchemical-mystical premises (TAYLOR, 1976b). Thus the metals were held to be associated with specific planets, resulting in names such as "lunar nitre" for silver nitrate and "martial chalk" for a carbonate of iron. Mystical terminology was widely used in alchemical texts, often deliberately obscure, and was adapted into early chemical writings, with terms such as "philosophical

water" for aqua regia and "arcanum duplicatum" (twofold secret preparation) for potassium sulphate.

Early Chemical Symbolism

Symbols have been used from the earliest times to represent chemical substances and processes. (BOLTON, 1882, CROSLAND, 1962c, WINDERLICH, 1953). The majority of these were alchemical in origin, and were in no sense systematic, generally being derived from pictograms or abbreviations, or simply convenient, arbitrary signs. Examples of these are given:

⊙ denoting gold (from the hieroglyphic for the sun)

≈ denoting water (from a pictogram)

$\bar{a}\,\bar{a}\,\bar{a}$ denoting amalgama (an abbreviation)

 denoting vinegar (an arbitrary sign)

There was nevertheless a persistent belief that symbols directly represented the properties of a substance (BOERHAAVE, 1735). Thus the symbol for copper ♀ was interpreted to indicate that the metal was partly gold (represented by a circle) and also contained a sharp, corrosive component (represented by a cross).

Perhaps surprisingly the use of these alchemically-derived symbols underwent a revival in the eighteenth century, with the introduction of a considerable number of new symbols, due largely to their usefulness in briefly summarising properties and reactions (NICHOLSON, 1795). Two influential tables were constructed by Geoffroy (GEOFFROY, 1718) and Bergman (BERGMAN, 1775), which made use of such symbols to display relationships between substances. Despite the considerable possibility for misunderstanding involved in the use of these symbols they continued to be used into the nineteenth century.

Reform of Nomenclature

The terminologies and symbolisms described above continued in widespread use to the latter part of the eighteenth century. Up to this time the knowledge of the composition of substances was too limited to make the develcpment of any more systematic nomenclature worthwhile, though criticisms of existing practice had been expressed throughout the seventeenth and eighteenth centuries (CROSLAND, 1962d). With the great increase in the number of substances being identified and investigated, the necessity for a systematic nomenclature capable of describing substances according to their constituents became evident. There is evidence that its development was influenced by the success of the newly-introduced biological classification-nomenclature of Linnaeus (CROSLAND, 1962e). Various nomenclature systems were devised by Macquer (MACQUER, 1766), Bergman, (BERGMAN, 1784) and Guyton de Morveau (GUYTON, 1782). The process culminated with the production of a definitive nomenclature system, derived collaboratively and adopting a good deal from the earlier schemes, which is generally associated with Lavoisier, the 'Methode de Nomenclature Chimique' (LAVOISIER et al, 1787). These schemes all aimed at giving a unique name, of the greatest possible simplicity, to a substance; and to reflect composition in the case of compound substances by compound terms with generic and specific parts (CROSLAND, 1962f, SMEATON, 1954). Lavoisier's 'Methode' pioneered the use of various terminations to the root of a common word to express different compositions.

Lavoisier's scheme rapidly gained a widespread acccptance. It should be noted that this scheme was primarily aimed at inorganic substances. Only a few dozen organic substances were known, mostly

derived from natural sources, and existing techniques of quantitative analysis were not adequate to accurately determine composition. Lavoisier's scheme was entirely suitable at that time for inorganic substances, which could be named adequately by composition alone. It was to be found that this is not so for organic substances.

Reform of Symbolism

Concurrent with the reforms in chemical nomenclature, and largely resulting from them, there came a series of advances in the use of symbols to represent chemical substances. These advances have been described in detail (CROSLAND, 1962c), and will be dealt with briefly here.

As a part of the 'Methode de nomenclature chimique' there was presented a scheme for chemical symbolism to complement the new nomenclature (HASSENFRATZ et al., 1787). This sought to replace the inconsistent and confusing symbols then in use by a system in which simple substances were represented by simple symbols and compound substances by combinations of the appropriate simple symbols.

This was not an entirely new concept: symbol combination had been used in Greek manuscripts, with, for instance, the symbols for stone, ↑, and silver, ☾, being combined to represent litharge. Bergman's scheme (BERGMAN, 1775) also allowed for a good deal of symbol combination. The Hassenfratz-Adet scheme, however, systematised the practice for the first time, with a unique representation for each substance. The system was geometrical in conception, with major classes of substances being represented by geometrical figures, e.g. circles for metals and triangles for alkalis and earths. Finer distinction was made by the use of letters within these figures: thus manganese was represented by Ⓜ, and potash by △P. Compound substances were represented by joining together the appropriate symbols. This system found favour because of its clarity, particularly in representing the course of reactions, but was never used on a large scale in print because of typographical problems. Later

modifications to the system allowed for the use of numbers to indicate proportions of constituents.

John Dalton's atomic theory was introduced in the early nineteenth century. Dalton devised a symbolism to express his theory with atoms denoted by circular symbols, and compounds denoted by such circles in contact (DALTON, 1810).

e.g.  oxygen  carbon  gold

 carbon dioxide

This system attracted much interest, though it also faced typographical problems. It will be noted that the positions of atoms within a compound substance may be represented in this way.

A major step in the development of chemical symbolism came with Berzelius' realisation of the many advantages in using the initial letters of the Latin name of an element as its chemical symbol (BERZELIUS, 1813). Though this system originally had several flaws, particularly Berzelius' introduction of barred symbols and representation of oxygen by dots, it was to find wide acceptance. With modifications introduced by Liebig, notably subscripted numerals, it was found to be ideally suited to the representation of the composition of organic substances.

Concepts of Valency and Chemical Structure

These concepts, and the consequent methods for their representation, were devised by investigators into organic rather than inorganic chemistry, since, as has been stated, the simplicity of the inorganic substances dealt with throughout most of the nineteenth century meant that composition was an adequate description for such substances.

The techniques of quantitative analysis available in the early part of the nineteenth century were adequate for a reasonably accurate determination of the percentage of the elements in organic substances. It was then necessary to determine the molecular weight, and to have accurate atomic weights available, in order to decide on the molecular formula and hence the structure. However,the nature of atoms and molecules was not well understood at this time, and the lack of any agreed set of atomic weights bedevilled chemistry until after 1860 when Avagadro's Hypothesis finally gained acceptance (DEMILT, 1951).

Two theories, or more accurately broad theoretical systems, describing the nature of organic substances flourished during the first half of the nineteenth century. Both were firmly based on Dalton's atomic theory and both, although originally seeming irreconcilable, merged in the later theory of structure.

The first of these, chronologically, was the so-called "Radical Theory", of which the most notable adherents were Berzelius, Kolbe and Frankland. This followed Davy's early suggestion of the association of chemical affinity with electrical attraction, and described molecule formation as a conjunction of groups of atoms of opposing electrical character. It therefore introduced the idea of a number of atoms forming a definite unit within a molecule

(RUSSELL, 1971b, PARTINGTON, 1964b).

The second was the "Unitary" or "Type Theory", supported by Dumas, Laurent and Gerhardt among others. This style of theory in its original form regarded molecules as indivisible entities, so that investigation into their internal structure was valueless (RUSSELL, 1971c, PARTINGTON, 1964c). The properties of organic substances were rationalised by regarding each substance as an example of a particular "chemical type"; ammonia type, water type etc. This resulted in a considerable interest in classification of organic substances (KAPOOR, 1969, FISHER, 1973a, FISHER, 1973b).

From a fusion of modified forms of these two theories, there emerged in the middle decades of the nineteenth century a coherent theory of valency (RUSSELL, 1971a, PORTER, 1965, BROWN, 1959). An early example of the new thought was due to Williamson, an adherent of the Type Theory, who stated "Formulae ... may be used as an actual image of what we rationally suppose to be the arrangement of constituent atoms in a compound" (WILLIAMSON, 1852).

Priority in the introduction of the theory of valency as applied to organic structures has been claimed on behalf of a number of leading workers. A widely-accepted, though perhaps over-simplified, view is that Kekule and Couper independently developed the idea of the tetravalent carbon atom, while Butlerov first used the term "chemical structure" to denote the arrangement of atoms within a molecule, and formally proposed that this arrangement uniquely determined the molecule and its properties. (BROWN, 1959, BYKOV, 1962: RUSSELL, 1971d).

This new understanding of the valency concept led to the

rapid development of two areas of chemical knowledge. The first of these was the impetus it gave to attempts to achieve a satisfactory classification of the elements. The formulation of the Periodic Law by Lothar Meyer and Mendeleef initiated a widespread and profitable use of such classifications (RUSSELL, 1971e, QUAM et al., 1934, MAZURS, 1974, LOACH, 1974).

The second was the widespread adoption of the theory of structure in organic chemistry, which played a considerable part in the subsequent growth of the subject (RUSSELL, 1971f, FINDLAY, 1965b).

This topic will be discussed briefly below.

Development of Organic Structure Representations

Full accounts of this topic are available (RUSSELL, 1971g, CROSLAND, 1962g) and a brief summary will therefore be presented here.

Although the greater part of organic nomenclature depended on the use of trivial names, some attempts at systematisation based on structure were made in the early nineteenth century. Abbreviated word forms to give a more convenient terminology, e.g. "aldehyde" for "alcohol dehydrogenate" were used, and some systematic word endings, e.g. "-one" for ketones, were adopted. The rise of the radical theory resulted in names being devised for common groupings of atoms which remained unchanged through many reactions, e.g. methyl, ethyl, acetyl. The type theory had a considerable effect on nomenclature, and various systems were devised reflecting the new concepts of "homologous series" and "parent nucleus" (FISHER, 1973a, FISHER, 1973b).

With the spreading of structural ideas there came an increasing use of structural formulae and diagrams. Initially used as a short-hand for the representation of composition, such structural representations later came to be regarded as directly portraying the arrangement of atoms with a molecule.

An early kind of notation was the superscript dash, introduced by Odling and used by Kekule and Wurtz among others, to indicate the valencies of atoms and groups. Examples are shown in Figure 1. As valencies became better known, and structural formulae widely used, this notation lost its usefulness.

Brackets were introduced in connection with type theory to link together groups of atoms without implying a corresponding physical arrangement. Later they were used, notably by Kekule and Frankland, to indicate the linking of groups, and this type of notation lasted over

a long period. Examples are shown in Figure 2.

Parentheses were originally introduced by Williamson to enclose groups of atoms remaining unchanged during reaction

e.g. $Ag(NO_3) + KCl = AgCl + K(NO_3)$

but were only used in this way for simple inorganic substances. Widespread use of this notation was later made to represent organic structural diagrams in a linear form (WISWESSER, 1975)

e.g. $CH_3CH(OH)CH_2CH(OH)CH_3$

Notations involving the use of touching and intersecting circles to represent bonded atoms were devised by Loschmidt and Kekule, and examples of these are given in Figure 3. Neither gained wide acceptance: Loschmidt's clear and relatively simple scheme went unnoticed, while Kekule's system was too complex for general use.

Once the representation of atoms by alphabet symbols had been established, the use of a simple linear representation of inter-atomic connections appears, in retrospect, obvious. As early as 1789 such a notation had been used by Higgins, in such representations as S - d for the compound of sulphur with dephylogisticated air, i.e. oxygen (HIGGINS, 1789). An obstacle to its use in structural formulae was the assumption that by such a notation its user was implying the physical reality of the inter-atomic linkage. Although the term "bond" had been coined by Frankland in 1866, the acceptance of this concept, as with "valency" as more than a convenient abstract notation came only slowly (RUSSELL, 1971b, PORTER, 1965).

The first use of lines to represent valencies is due to Couper, who initially used dotted lines and later full lines. Examples of such formulae are shown in Figure 4. This type of notation was

much superior to other structure representations available at that time, but was never developed fully, e.g. lack of provision for denoting multiple bonding. Thus it was not widely adopted, though it was briefly influential.

Chemical structure diagrams of the kind used to the present day were originated by Crum Brown (CRUM BROWN, 1864, LARDER, 1967). These formulae showed each atom separately and indicated all single and multiple bonds unambiguously. They rapidly proved their value by demonstrating that only two isomers of propanol could exist, while other structural notations could not allow rationalisation of this experimental observation. Because of their clarity they rapidly gained acceptance in Britain, and were widely used in the teaching of chemistry, though their acceptance abroad was considerably delayed. Examples of Crum Brown's original formulae are shown in Figure 5. With the omission of the circles around the atomic symbols, which rapidly followed, these graphic formulae have been used essentially unchanged to the present day.

Other forms of graphic formulae were used for a time on the Continent. Lothar Meyer had,independently of Crum Brown, devised a notation with linear inter-atomic connections, while Wilbrand attempted a systematic graphic representation of the variable valency of carbon. Examples of these representations are given in Figure 6. Within a decade, however, graphic formulae of the Crum Brown type had largely superceded all other representations.

Later Development of Structural Representation

Two modifications were made to the original form of Crum-Brown's structural diagrams, following their widespread adoption. Firstly, the bonds to hydrogen atoms were not always specified, with abbreviated notations such as $-CH_3$, $-NH_2$, and -OH being used. Secondly, carbon atoms were not always represented by the C symbol, but were denoted by the intersection of lines denoting bonds. These modifications made the structural diagrams simpler and more convenient, particularly for printing.

Two areas of development of structural theory in the late nineteenth century necessitated amendments and additions to the original form of structure diagram. These were the introduction of the theory of stereochemistry, and developments in the understanding of the problem of variable valency.

The tetrahedral arrangement of the valencies of the tetravalent carbon atom, the foundation of all current stereochemical theory, was discovered independently by van't Hoff and Le Bel in 1874 (RICHARDSON, 1901, DAVIDSON, 1973). This led to an understanding of the nature of geometrical and optical isomerism (ELIEL, 1962b), and of the distinction between isomerism and tautomerism (IHDE, 1959). Methods were then devised, over a long period, for representing isomers both diagramatically and by nomenclature, though stereochemical nomenclature is still far from totally satisfactory (ELIEL, 1962b): this topic has been comprehensively reviewed (MASON, 1976).

The greatest problem in this area is the representation in a structure diagram of compounds containing one or more assymetric carbon atoms. Examples of the most widely used conventions for both cyclic

and acyclic structures are shown in Figure 7.

These are the Fischer projection (FISCHER, 1891) and its abbreviated form (ROSANOFF, 1906) designed to represent assymetric acyclic structures in a systematic diagramatic form: the perspective representation, sometimes termed the "sawhorse", (CURTIN, 1954) and the Newman projection (NEWMAN, 1955), both designed to show molecular conformation more clearly than the Fischer projection: and an example of a class of widely used "wedge/dot" representations (CRAM, 1952).

The development of stereochemical ideas led to the use of molecular models, to directly represent the concept of three-dimensional structure. Models had been used at an earlier stage to exemplify structural ideas by, among others, Dalton, who represented "compound atoms" by spheres joined by pins (DALTON, 1840), Hofmann, who used coloured croquet balls joined by tubes and pins to demonstrate the valency concept (HOFMANN, 1865), Dewar (DEWAR, 1866) and Kekule (KEKULE, 1867), who both used simple models to demonstrate the tetravalency of carbon. Three-dimensional models became widely accepted and many types were devised (PLATT, 1960, PETERSEN, 1970). Such models are still used to a large extent for teaching (WALTON, 1969, ORMEROD, 1970, SANDERSON, 1962, BASSOW, 1968, SAVORY, 1974).

The use of computer display systems for three-dimensional structures (FELDMANN et al., 1972b, LITTLE, 1973, MARSHALL et al., 1974) may be regarded as a natural development of the use of stereochemical models.

The problem of the observed variable valency of carbon in organic compounds was to a large extent solved by Kekule, who formulated the concept of multiple bonding (KEKULE, 1867b), which had been represented diagramatically by Crum Brown. There remained considerable

difficulty in accounting for conjugated molecules, and in particular the aromatic substances typified by benzene (BADGER, 1969, RUSSELL, 1971i). Some of the types of structural formulae used to represent such compounds are shown in Figure 8.

Kekule represented the structure of benzene by alternating single and double bonds, and assumed a rapid oscillation between the two possible forms to account for the known symmetrical nature of the benzene nucleus. Attempts to represent benzene by a single structural diagram were made by a number of workers. The formulae proposed by Claus, and the "centric formulae" of Armstrong, Baeyer and Luthar Meyer adequately represented the nature of benzene, as it was then understood, but conveyed little insight into the nature of the bonding in the molecule. Thiele introduced the concept of a "partial valency" of an unsaturated carbon atom,which provided partial bonding (represented by dotted lines) in conjugated and aromatic systems (THIELE, 1899). Thiele's representation of benzene led to the symmetrical formula used by Thomson (THOMSON, 1916), and the modern representation, more convenient for typography, used as an alternative to Kekule's formula.

More recently,variable valency concepts have been widely used to account for the many relatively unstable species identified in organic chemical research. Structural formulae have been used to represent a variety of charged and free radical structures (KERMACK et al., 1922), and "non-classical" species with delocalisation and partial bonding (BARTLETT, 1965, HUDSON, 1967). The success with which structure diagrams can represent such factors will be discussed more fully below.

The study of variable valency in inorganic chemistry led to the discoveries of Werner in the field of the co-ordination compounds

of metals (WERNER, 1893, SOLOVEICHIK et al., 1967, PARTINGTON, 1964d). This work, an important advance in valency theory, established the necessity for the use of structural and stereochemical concepts in this area of inorganic chemistry. The usefulness of a structural representation with directed bonds between metals and ligands, and the inadequacy of composition as a description of this type of inorganic substance, was demonstrated by Werner, as for example in his discovery of the octahedral nature of the platinum ammine chlorides shown in Figure 9 (WERNER, 1893). The structural diagram has,however, only been found useful to a limited extent in inorganic compared with organic chemistry. This is due partly to the less readily categorisable types of bonding found in inorganic structures. Also inorganic compounds are in general smaller and less complex structurally than organic species, so that molecular formula is often an adequate descriptor.

During the latter part of the nineteenth century organic nomenclature underwent a development paralleling that of diagramatic structural representation. (VERKADE, 1953, CROSLAND, 1962h). Following general agreement at the Karlsruhe Congress of 1860 as to the necessity for a standardised nomenclature and notation, a considerable amount of progress was made internationally. At the Geneva Congress of 1892 the first generally agreed system of systematic organic nomenclature was devised. Subsequent international co-operation in this field ensued; until in 1930 the newly-formed IUPAC organisation took responsibility for the officially-recognised nomenclature system. Nonetheless,even at the present day a number of alternative nomenclature systems are in use (CAHN, 1974a). The use of nomenclature within chemical information systems will be discussed below. It is interesting

to note that the increasing complexity of modern organic nomenclature, and its consequent lack of familiarity to many chemists, has resulted in indexes based on molecular formula, the earliest pre-structural systematic compound identifiers,being widely used for chemical substance searches (HYDE et al., 1975).

Although a link between electricity and chemical structures had been proposed early in the nineteenth century (RUSSELL, 1963a, RUSSELL, 1963b), it was over a century later that the concept of the electron-pair bond was developed (LEWIS, 1916, KOSSEL, 1916, LANGMUIR, 1921).

The introduction of electronic ideas into chemical, particularly organic chemical, thought has been reviewed (BYKOV, 1965).

The principles of this theory were given a more sophisticated form with the application of quantum mechanics to chemical problems SCHRODINGER, 1928). Two forms of quantum theory have been used, generally known as Valence Bond theory and Molecular Orbital theory. A brief account of these methods will be given below. One early and notable result of the introduction of electronic ideas into chemistry was the work on the so-called "English school" of physical organic chemists in the study of the properties of organic compounds (ROBINSON, 1932, INGOLD, 1933). This approach relied heavily on a visualisation on chemical structure and bonding, by means of structural diagrams, and a qualitative, pictorial representation of electronic effects by the well-known "curly arrows" notation (MALKIN et al., 1925). Similarly much of the success of the valence bond formulation of quantum mechanics has been ascribed to the ability to represent the "resonance forms" associated with this approach by means of conventional

structural formulae (PAULING, 1939, WIELAND, 1944).

The most sophisticated description of chemical structure and bonding now available, provided by quantum mechanical calculations, is couched in purely mathematical terms, and is not amenable to any accurate pictorial or physical visualisation. To this extent the most sophisticated model has grown away from its most useful representation, the chemical structure diagram. These diagrams however are still adequate for very many purposes: they are, as Gold has pointed out, "the unambiguous language in which chemists think, formulate and communicate their ideas" (GOLD, 1976).

2.2.2.

Chemical Structure Representation

2.2.2. Chemical Structure Representation

The handling of chemical structure information is fundamental to the operation of most chemical information systems, whether computer-based or otherwise. It is the potentiality for unique and unambiguous description of the items stored in such a system, i.e. chemical structures, in a manner for which conventional information processing techniques are inadequate, which gives chemical structure information systems their distinctive character.

The techniques for representing chemical structure, have been comprehensively reviewed (N.A.S. 1964, N.A.S. 1965, N.A.S. 1969, TATE, 1967, HOLM et al., 1973, WISWESSER et al., 1973, LYNCH, 1968a) and have been the subject of recent monographs (LYNCH et al., 1971a, DAVIS et al., 1974a, ASH et al., 1975) and conference proceedings (WIPKE et al., 1974a) and therefore no full discussion of these topics will be given here. The major features of the various structural representations, will be briefly summarised, with emphasis on the aspects relevant to structure-property correlation, to be discussed later.

A number of techniques are currently used to represent chemical structure within information systems. All of these describe chemical structure by representing, partly or wholly, the structure diagram. Although in some cases the distinction is not clear-cut, four types of representation may be noted, i.e. systematic nomenclature, fragmentation codes, linear notations, and connectivity or topological representations. Each of these will be considered in turn, as will developments in the automatic interconversion of these representations. Finally, methods for input and output of structural representations will be discussed.

Systematic Nomenclature

Systematic nomenclature is the oldest form of structural representation and the most widely used as an alternative to the structure diagram in inter-chemist communication. A number of alternative nomenclature systems exist (CAHN, 1974a), most notably those of the International Union of Pure and Applied Chemistry for organic (IUPAC, 1974a) and inorganic (IUPAC, 1971b) compounds, and of Chemical Abstracts Service (DONALDSON et al., 1974, BLACKWOOD et al., 1975). Attention has also been paid to the development of specialist nomenclature systems, e.g. for biological compounds and for polymers (CAHN, 1974b).

Nomenclature however suffers from several disadvantages for use in chemical information systems (BOWMAN, 1975). The complexities of the truly systematic nomenclature necessary for such application make it too difficult and unwieldy to be an acceptable substitute for the structure diagram for the system's users. Additionally, and perhaps more importantly, systematic nomenclature is generally felt to be unsuitable for computer processing (BOWMAN, 1975, HYDE, 1975a). Few present-day computer-based systems therefore use nomenclature as the sole, or major, representation, although nomenclature has been found to be a valuable component of overall structure handling systems, notably at Chemical Abstracts Service (VANDER STOUW et al., 1974). A form of substructure search procedure using systematic nomenclature has been described (FISANICK et al., 1975).

Fragmentation Codes

A distinction may be made between two types of fragmentation codes; those which are manually assigned, and those which are algorithmically generated from a total structure representation.

Manually assigned fragment codes consist of a series of symbols, often numeric, assigned to structural features considered of importance. In entry of a compound into the system, the appropriate codes are assigned for those structural features present. A file of structures may then be searched for a particular code, i.e. structural feature, or combination of codes. This procedure was well suited for implementation on unsophisticated information processing equipment, e.g. punched card sorters, and hence fragment codes were one of the earliest forms of structural representation to be adopted (N.A.S. 1964). Such codes have been widely used, with several hundred different codes devised and used in conjunction with a variety of equipment from card files to computerised systems (BOWMAN, 1975, N.A.S. 1969).

These codes have the advantages of simplicity in conception and use, and of familiarity to users and system operators. No standardisation of such codes has been attempted, since a major point in their favour is the usefulness of such a code when it is designed for a specific purpose.

Their disadvantages stem from two points (BOWMAN, 1975). Firstly such codes are not a total structural representation. Even if all atoms and bonds of a structure are included in the assigned codes, which is not necessarily the case, little or no information is included, in most cases, about the relative positions of the

structural features. Thus a fragment code description of a structure is both ambiguous and non-unique, often leading to much irrelevant material being retrieved from a search of a large file. Secondly the list of assignable codes is fixed when the system is introduced. If it is found in practice to be inadequate, or the interests of the system users change, an alteration of the code would require re-indexing all the existing file.

For these reaons manually assigned fragmentation codes have not in general been found suitable for structure representation in large systems, although they may well be entirely adequate for specialised files or for particular purposes, e.g. recording structures in patent documentation (BALENT et al., 1975). A small number of such codes have been highly-developed for use in large systems, notably the Smith, Kline and French code (CRAIG et al., 1969), the GREMAS code, used by the IDC group (ROSSLER et al., 1970, FUGMANN, 1975) and the Ring Code, originally developed by the Pharma Documentation Ring and since used in Derwent Publications services (PHARMA-DOK, 1972, NUBLING, 1970). Algorithmic generation of fragments from a total structure representaticn has been described, using connection tables (ASH, 1975)or a linear notation (BOWMAN, 1970), mainly as a screening system for substructure search (LYNCH, 1975a).

Fragmentation codes of both types have been used more widely than other forms of structural representation for structure-property correlation, as will be discussed below.

Linear Notations

Linear notations are complete structural representations, and therefore denote a compound unambiguously. A chemical structure is coded as a string of alphanumeric characters, governed by relatively complex rules of syntax. Most such notations achieve economy by denoting chemically significant groups, ring systems etc. by a very few symbols. Precedence and ordering rules, similar to those for systematic nomenclature, are used to give a unique notation for any given structure. (DAVIS et al., 1974b, LYNCH et al., 1971b). The first proposal for such a notation was made by Dyson (DYSON, 1944), and a number of linear notations have been used operationally (BOWMAN, 1975, N.A.S. 1969).

These include Dyson's IUPAC notation (DYSON, 1975, DAMMERS et al., 1968) and notations due to Hayward (HAYWARD et al., 1965), Silk (SILK, 1963), and Skolnik (SKOLNIK et al., 1964). The only linear notation to have gained wide use outside a single organisation is that orginally proposed by Wiswesser (WISWESSER, 1952), and now generally known as Wiswesser Line Notation (WLN) (BAKER et al., 1975, SMITH et al., 1975).

This notation is currently in use in more than forty information systems of various types, and is applicable both as the sole structural representation in a small file with unsophisticated equipment (GELBERG et al., 1962), and as a main representation in a computerised system (EAKIN, 1975). WLN is also becoming commonly used in handbooks, tables of property values etc. (GRASSELLI, 1973, MARTIN, 1971, CRISTENSEN et al., 1974) partly because of its relative readability, an advantage of most linear notations.

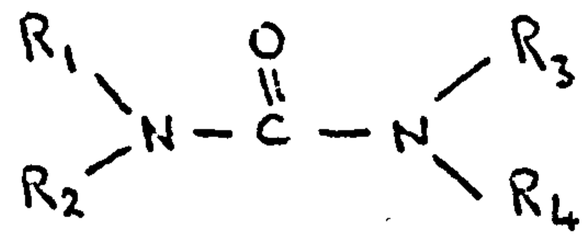
One reason for the success of WLN, in addition to its

intrinsic merits,particularly its compatability with many forms of information processing equipment, is the activity of a user group, the Chemical Notation Association, in publicising and giving advise on the notation, modifying the rules in accordance with experience, and producing a generally-available encoding manual (SMITH et al., 1975). Rules have been devised to deal with particular problems in the practical use of WLN, e.g. representation of stereochemistry, and encoding of polymers. Provision is also incorporated in the WLN coding rules for contracting, i.e. shortening, certain notations in order to save storage space. These contraction rules however increase the complexity of handling WLN to such an extent that they are not universally employed.

A modified form of WLN, known as ALWIN, i.e. algorithmic Wiswesser Notation, with a more rigorous mathematically defined basis, has been described, but has not been adopted as yet by any operational system (KRISHNAMURTHY et al., 1974, SANKAR et al., 1974).

Linear notations are widely use for computerised structure and substructure retrieval, and techniques are well-established for string search, fragmentation, permutation etc. (LYNCH et al., 1971f, CROWE et al., 1973, GRANITO et al., 1965, BOWMAN et al., 1970, EAKIN, 1975). One inevitable problem with the use of most linear notations for substructure search is the variety of possible notation strings which arise for even a simple sub-structure. Thus for a search in a Wiswesser Line Notation file, bearing in mind that the Wiswesser symbols for $-\overset{|}{N}-$, $-\overset{|}{N}H$, $-NH_2$ and $-\overset{O}{\overset{\|}{C}}-$ are N, M, Z, and V respectively, the notations for the

substructure shown



R_{1-4} may or may not be hydrogen

may be ZVZ, ZVM, ZVN, MVM, MVZ, MVN, NVZ, NVM, NVN, and a large number of further alternatives if one or both of the nitrogen atoms are included within a ring system.

The use of linear notations, most notably WLN, for structure-property correlation will be considered later.

Connectivity Tables

Connectivity or topological records are a form of complete, and hence unambiguous, structural representation (ASH, 1975, DAVIS et al., 1974c). These representations are highly specific in that each atom, except for hydrogen in some cases, and/or each bond is explicitly and separately recorded in appropriate detail. Stereochemical information may be included in such representations (ESACK et al., 1975, WIPKE, 1974, BLAIR et al., 1974). Connectivity records are usually not unique, since there is a lack of complex coding rules, making the encoding of structures in this way a purely clerical process. Procedures are available for canonicalisation, i.e. production of a unique form of the connection table (LYNCH et al., 1971c). Although by their nature these representations are not economical in storage space, compacted forms may be produced, contrasting with the redundant forms more generally used.

Some chemical information systems have been described which use connectivity records as the sole structural representation (ASH, 1975, GLUCK, 1965).

In other systems these representations are used for specific purposes, e.g. as intermediates for input-output and interconversion of representations, for atom-by-atom search, which has a precision much greater than that possible with linear notations or fragment codes, and for other applications requiring similar precision, e.g. computer design of synthesis routes (WIPKE et al., 1974b, BERSOHN et al., 1976, COREY et al., 1976) and some types of structure-property correlation.

A considerable number of sophisticated topological

representations have been devised (DAVIS et al., 1974c, N.A.S. 1969) notably the French DARC system (DUBOIS, 1974) and the matrix representations used by Ugi for synthesis planning (BLAIR et al., 1974).

Interconversion of Representations

The capability for the automatic interconversion of structural representations within computer-based chemical information systems has been a topic of active investigation, since this capability has two major advantages for the operators and users of such systems (LYNCH et al., 1971d). Firstly appropriate representations may be produced for specific purposes, e.g. a connection table for atom-by-atom search. Secondly a system is able to accept structural information from external sources, without the need for recoxding.

The production of partial structural representations, i.e. fragment codes, from full structural representations, has been mentioned above, and is relatively straight-forward, though the reverse process is not in general possible. In particular the production of Ring Codes from WLN, of interest because of the wide use of both representations, has been described (GRANITO, 1973, GRANITO et al., 1972).

Conversions from the more complex full structural representations to a connectivity record have been reported, starting from both WLN (GRANITO, 1973, HYDE et al., 1967), and systematic nomenclature (VANDERSTOUW et al., 1974).

The reverse process is more difficult and only limited success has been reported in generating WLN (LYNCH, 1968b, BOWMAN et al., 1968, FARRELL et al., 1971) and nomenclature (CONROW, 1966) from connection tables.

Interconversion of different forms of the same type of representation may be important, and such techniques have been described for connection tables and linear notations (CAMPEY et al., 1970).

Input and Output of Structural Information

Input and output procedures are of considerable importance for chemical information systems (LYNCH et al., 1971e). Choice of an appropriate input technique can have a considerable effect on the costs of information processing, while the form of output chosen can greatly influence the acceptance of the system by its users.

Input of chemical structures was originally achieved by entry, by keyboarding or similar means, of the structural representation. A number of such input media have been compared from a cost-effectiveness viewpoint (MENDENHALL, 1974). More recently the alternative of directly entering a structural diagram by some appropriate technique has been possible. Most commonly, a special form of typewriter or teletype has been used to generate a digital record of the structural symbols and their co-ordinates, from which a connectivity record is derived algorithmically (FELDMAN, 1973, GOTTARDI, 1970, MULLEN, 1967).

A formula reader has been devised with which a structural formula, drawn according to certain conventions, is scanned by photocells to produce a similar digital record (MEYER, 1974). In interactive systems, structures may be entered via a light pen and RAND tablet to give a connectivity record. This has been described for interactive systems designed both for retrieval (FELDMANN et al., 1972a, FELDMANN, 1974) and for synthesis planning (COREY et al., 1972).

There is an evident advantage to any chemical information system, by way of increased acceptability to users, if output from searches can be presented in the form of structural diagrams, rather than registry numbers, names or notations. However, production of

structure diagram output has been found to be both difficult and expensive (ASH, 1975). A structure diagram may be obtained from a structural representation either by a look-up process with a file of stored structure diagrams, or automatically from the structural representation, which gives much greater flexibility. Such automatic structure display methods usually require a connectivity table representation (HYDE et al., 1968), in which case the majority of structures can be dealt with satisfactorily, although it may prove useful to store the structures of the relatively small number of complex structures which cause problems. The generation of structure diagrams directly from WLN has been described (FELDMANN et al., 1971) though this cannot be as efficient as generation via connection tables, because of the problems of WLN syntax.

Structure diagrams are commonly printed on paper or cards (THOMSON et al., 1967, JACOBUS et al., 1970), or may be output on a graphics terminal (FELDMANN et al., 1972, COREY et al., 1972).

If, in addition to a structural representation, some form of atomic co-ordinates are input for a compound, a three-dimensional structural image may be displayed, as have been described for several interactive systems (MARSHALL et al., 1974; MEYER, 1970, FELDMANN et al., 1972b).

Because of these technical advances it is now possible to devise a system in which input and output are performed solely with structure diagrams, while conversion to and from structural representations takes place internally and automatically. Systems of this sort have been described, for operation in both batch (STOCKTON et al., 1974) and interactive (FELDMANN, 1974, FELDMANN et al., 1972a) modes.

2.2.3.

Chemical Structure Information Systems

Chemical Structure Information Systems

A full account of the various aspects of such systems has been recently published (ASH et al., 1975). The discussion below will therefore be brief, and will concentrate on those aspects relevant to the applicability of structure-property correlation and similar techniques in these systems.

The effective and efficient operation of chemical structure information systems depends on the use of appropriate chemical structure handling techniques together with more conventional methods for information retrieval (BOWMAN, 1975). This allows for two means of approach to such systems; the "structure-directed" approach, and an approach via queries directed to properties, uses, or subjects.

Any search may therefore be regarded as either structure-directed, for example "how many compounds contaning this substructure are known", or subject-directed, for example "how many applications of wave mechanics calculations using the PRDDO technique have been published", or as a mixture of the two components, for example "are any examples known of aminoindans as enzyme inhibitors".

The distinction between these two components of a search may not be of any importance in "conventional" information systems, and may involve no difference in search technique. However in the computer-based chemical structure information systems considered here, two difference approaches to the search may be used, and it is convenient to consider these separately.

The first of these, the "structure-directed" approach, aims to deal with queries concerning a particular structure or partial structure. To deal adequately with such qeuries requires a file of structures, coded in one of the representations discussed in the

preceding section, a routine for updating this file as necessary, and some form of substructure search procedure. A variety of such procedures have been developed, varying for highly complex multi-step systems using several types of structure representation (EAKIN et al., 1974, EAKIN, 1975) to the simple, though within limits highly effective, production and use of permuted indexes from linear notations (GRANITO et al., 1965). The advantages of interactive use of the more complex systems have been noted (EAKIN et al., 1974, FELDMANN, 1974).

A related function necessary for some chemical information systems is that of providing a capability for chemical reaction documentation. This is essentially a structure-directed problem, although reaction conditions etc. may be an additional factor, but has been an area of considerable difficulty because of the complexities involved in adequately representing chemical reactions for storage and retrieval (VALLS, 1974, VALLS et al., 1975). No method for automatically indexing and searching reaction information in files of structure representations, in a manner analagous to substructure searching, has been adopted in an operational system, though research to this end has been and is being undertaken (CLINGING et al., 1974, LYNCH, 1975b, OSINGA et al., 1976, WILLETT, 1976, VLEDUTZ, 1963).

In order to deal with the second type of query, that involving a non-structural approach to information or data held in the system, chemical structure information systems will generally have available files of non-structural material, which with appropriate techniques may be searched either alone or in conjunction with a structural search.

Perhaps the best known type is the bibliographic file, which may contain literature references and also, in the case of in-house systems, company reports etc. Examples of the operation of such files have been described (KENNARD et al., 1972, SCHULTZ, 1974). The well-developed methods of computerised information retrieval are applicable to such files (MATTHEWS, 1975, WILLIAMS, 1974, MEADOW et al., 1970).

Machine-readable property files, becoming increasingly widely used, may contain either textual or, more commonly, numerical data. Specialised searching techniques are, in some cases, necessary for the most effective use of such files. Techniques for data analysis and structure-property correlation, which make use of such files, will be discussed in detail in a later section.

Files of biological data are widely used in industrial information systems, particularly in connection with compound screening programmes in the pharmaceutical and agrochemical industries. Such files are particularly useful when used in conjunction with structural information, which may be used to link data in several property files. Subfiles may be readily created, for detailed study of specific sets of data. Descriptions of the design and application of such systems have been given (EAKIN et al., 1974, EAKIN, 1975, BOND et al., 1971, HANSCH et al., 1974a, BROWN et al., 1976, SAGGERS, 1974, WISWESSER et al., 1974).

Machine-readable structure-property files have also found application for spectroscopic data (HELLER, 1974, HELLER et al., 1973, WOODRUFF et al., 1975, HELLER et al., 1972, JAGER et al., 1968) crystallographic data (ALLEN et al., 1973, WILLARREAL et al., 1976) physicochemical properties (LEO et al., 1971), thermochemical data (PEDLEY, 1976), composition-property data (ADLER et al., 1969),

toxicological data (OXMAN et al., 1976) and environmental hazards data (GEISS et al., 1975).

Few detailed descriptions of "total systems" incorporating such files have been published. The ICI CROSSBOW system has a sophisticated structure-handling capability, including interconversion of structure representations, and allows access to property files containing biological screening results. Additionally, files of medical, toxicological and physical property data on the company's compounds, and commercial files of physicochemical properties and bibliographic data are accessible (EAKIN et al., 1974, EAKIN, 1975).

An interactive system, using a graphics terminal, has been described (FELDMANN et al., 1972a). This allows on-line access to files of structural data, bibliographic data, and files of physical and spectroscopic data.

The very considerable problems of attaining effective and efficient storage and retrieval within chemical structure information systems have been overcome, in the ways described above, to such an extent that the operation of the powerful and flexible systems described above is now a matter of routine. A considerable amount of effort, needless to say, is still expended on further improving the retrieval capabilities of such systems, and on maximising their cost-effectiveness.

There is an increasing interest in making use of such computer-based chemical structure systems for information and data analysis and interpretation, in addition to conventional storage and retrieval. These techniques have an aim which Dammers has described as "manipulating data/information as an aid to assimilation and understanding ... to highlight relationships and order in the retrieved data" (DAMMERS, 1975).

Data analysis techniques are finding application in many fields (PERKINS et al., 1975, TUKEY et al., 1966, KOSKINEN et al., 1975), and will be applicable to both of the components of the approach to the contents of a chemical structure information system mentioned above. Relative simple techniques for data analysis and presentation may well be found valuable, as will be discussed in a later section. Conversely, some of the highly sophisticated work on organic synthesis design (WIPKE et al., 1974b, BERSOHN et al., 1976, COREY et al., 1976, BLAIR et al., 1974) and on graph theory as applied to chemical structures (SMITH, 1975, RANDIC, 1974, MASINTER et al., 1974, GUND et al., 1975) could have useful application in the "structure-directed" aspects of systems' operations. Unconventional techniques of information retrieval, devised largely for textual matter (VAN RIJSBERGEN, 1975) could be

applicable to non-structural information within such systems. The possible use of such methods, particularly cluster analysis and related techniques, with structural data will be discussed in detail below. Techniques of statistical analysis and pattern recognition have already found application with both structural information and numerical data. The use of these methods with files of structural and non-structural data in conjunction may be very valuable for structure-property correlation. These aspects will be discussed in detail in a later section.

It should be noted at this point that such techniques are applicable to some extent to structural and non-structural data files, regardless of the type of structural representation used and of the degree of sophistication of the implementation of the system. However, the use of a total structural representation in a computerised system enables a much fuller advantage to be taken of these methods, which is further enhanced by the additional ability to interconvert structural representations or to access property files in conjunction with structural data.

2.3.

Structure-Property Correlation

The discussion of methods of structure-property correlation must inevitably be limited in scope, since much of chemical science may be described in this way, being concerned with the rationalisation of properties by consideration of structure. The term will be taken to mean generalised methods for the investigation of structure-property relationships for a large number of structures and applicable to a number of properties, at least in principle, rather than rationalisations for particular cases. The discussion will centre largely on quantitative techniques, including the calculation of molecular conformations, and on methods involving the classification or categorisation of chemical species. It will be largely, though not exclusively, concerned with methods for dealing with pure, non-polymeric, organic substances.

The historical development of such techniques is outlined below, with consideration of the differing purposes for which various methods have been devised and used. A discussion of the more widely-used techniques follows, inevitably brief and with the aim of illustrating the practical successes, potential limitations, similarities and differences of these methods. No single comprehensive account of the applications of these methods exists, though there are many descriptions of particular procedures and types of procedure, referred to when appropriate below, and useful reviews of the applications of the techniques in particular fields of study, e.g. drug design (REDL et al., 1974, N.C.I., 1974), and thermodynamic property estimation (JANZ, 1967a).

For the discussion below the techniques are categorised, somewhat crudely, according to the nature of the method, rather than by their application. Firstly, theoretical methods are considered.

By this is meant those methods which enable calculation of some property value for one particular chemical species, by mathematical procedures based on a theoretical model, without reliance on large amounts of experimental data. Secondly, semi-empirical techniques, in which the property under investigation is correlated with other measured or calculated molecular properties, or with parameters dervied from such properties, are then considered. Thirdly, fully empirical methods will be discussed. This group comprises a wide range of parametric and non-parametric statistical techniques for analysing quantitative and qualitative data. Most involve structural parameters directly, and it is in this area that substructural analysis techniques have been developed.

Finally the use, or lack of use, of published accounts of property predictions using such methods, and the possible effects of the use of such methods upon patentability will be noted.

2.3.1.

Development of Structure-Property Correlation

Background to the Development of Structure-Property Correlation Methods

The properties of naturally-occurring chemical substances were noted in a qualitative sense from the earliest times, largely for their possible practical value.

Attempts to develop systematic and quantitative relationships became possible with an understanding of chemical composition in the early nineteenth century. These involved the deduction of empirical relationships between properties for a substance, and, as structural ideas gained acceptance, between structure and property (CRUM BROWN, 1869). Such relationships, at that time, were largely derived for physico-chemical properties or organic compounds; boiling point, molar volumes, refractivities etc. (PARTINGTON, 1951).

The first explicit statement of a generalised quantitative structure-activity relation was due to Crum Brown, based on the investigation of the biological effects of alkaloids (CRUM BROWN et al., 1868).

Two lines of development in the study of the properties of chemical substances may be identified, from their beginnings at the end of the nineteenth century. The first of these is the essentially qualitative detailed rationalisation of property, exemplified in more recent years by the elucidation of organic reaction mechanisms and by the investigation of the biochemical basis of medicinal chemistry. The second, of more immediate concern here, is the quantitative empirical or semi-empirical approach to structure-property correlation.

The early development of the correlation of biological activities has been reviewed (HANSCH, 1971b, PURCELL et al., 1971).

In most cases this involved the correlation of a biological activity with a physicochemical property. Typical, and most widely noted, of these studies were the correlations between narcosis and partition coefficient achieved independently by Meyer and Overton (MEYER, 1899, OVERTON, 1897). In a somewhat different field, the study of physico-chemical properties for organic structures was actively pursued, when it was realised that many such properties were additive functions of the atoms and bonds present. This proved to be a valuable tool for structure determination before the advent of sophisticated instrumentation for this purpose: molar refractivity and parachor were widely used in this way (SUGDEN, 1930, VOGEL, 1948, VOGEL et al., 1950).

In the 1920's the newly developed quantum theory was given the wave mechanical formalism which was to enable its application to chemical problems (SCHRODINGER, 1928),and occasioned Dirac's comment that "the underlying physical laws necessary for the mathematical theory of ... the whole of chemistry are thus completely known" (DIRAC, 1929). It was some decades however, before quantum mechanical calculations could achieve widespread use, and Dirac's proviso that "the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble" remains unchallenged. Despite these advances in theory, purely empirical additive property relationships gained wide use, notably in the estimation of thermodynamic properties (PARKS et al., 1932, SIDGWICK, 1933). The increased emphasis on the quantitative aspects of organic chemistry led to the emergence of the specialism termed "physical organic chemistry" (HAMMETT, 1940). This discipline is very largely concerned with the quantitative

rationalisation of chemical reactivity and similar properties in structural terms (HINE, 1975a, HINE, 1962) largely by the development of empirical and semi-empirical relationships. Considerable effort was also expended on development of very simple procedures based on quantum mechanical principles for qualitative and semi-quantitative treatment of organic chemical problems. These included the valence bond theory with its concept of resonance (PAULING, 1939, WHELAND, 1944) and the Huckel molecular orbital theory (STREITWIESER, 1961).

Up to 1950 methods of quantitative structure-property correlation had little practical importance, apart from the use of atom and bond additivities for structure determination and for estimation of physical and thermodynamic properties. Two developments were to change this situation markedly. The first was the increasing investment in those industries producing chemical substances with specific biological effects, coupled with the rising costs of the traditional means of discovering compounds of appropriate activity. The second was the increased availability of data-processing equipment, and especially of digital computers.

The expansion of the pharmaceutical and agrochemicals industries since 1945 is notable (TEELING-SMITH, 1967). This increased investment has however been accompanied by an increase in the costs of the research process, and by a tightening of legal restrictions on the commercial introduction of new products (ROBINSON, 1974, BLOOM, 1971). The conventional means of research has involved identification of an active compound by either extraction from a natural product, random screening of available substances, or biochemical rationalisation, followed by modification of this basic

structure in an attempt to suitably alter the compound's properties (ARIENS, 1971c, BLOOM et al., 1971b). This procedure has become steadily more costly and unproductive: recent estimates suggest that between 3,000 and 10,000 compounds are synthesised and tested for every one which becomes commercially available (SPINKS, 1973, HAHN, 1975, ROBINSON, 1974). There has,within the last twenty years,come a great interest in methods for predicting new active structures, 'lead generation', and for optimising the activity within a particular set of structures, 'lead optimisation'; without the necessity for synthesising all possible compounds. This naturally involves consideration of a wide variety of structure-property correlation methods, and has provided the impetus for the development of such methods (REDL et al., 1974, HAHN, 1975).

The introduction of data-processing equipment, in the form of punched-card handling machinery, coincided with some of the first large-scale screening programmes for biologically active compounds, and greatly ameliorated the problems associated with large volumes of data (SAGGERS, 1974, CRAIG, 1975).

The subsequent wide availability of powerful digital computers in both academic and industrial environments has had three main consequences for the applications of structure-property correlation methods.

Firstly,quantum mechanical calculations may now be regarded as a routine tool in certain areas, though as Cook has suggested regarding the application of rigorous wave mechanical techniques to "the vast bulk of chemistry and biochemistry ... at the present time the only contribution to this field (quantum chemists)

can sensibly make is to wish experimental workers luck in developing empirical theories" (COOK, 1974e). In addition to increased computing power, the development of many methodologies of varying degrees of approximation has contributed to the greater use of this approach. Other theoretical techniques involving extensive calculations have similarly become feasible.

Secondly a variety of multivariate statistical techniques have been applied to structure-activity problems since about 1960 (PURCELL et al., 1973b). These include multiparameter semiempirical correlations, additive modelling, and variety of non-parametric techniques, all of which will be discussed below. Routine application of such techniques is dependent upon adequate computing facilities, and also to a large extent upon the availability of standard statistical programs, in addition to the necessity for appropriate experimental data.

Thirdly the capability of current computer systems to deal with large data-bases allows for manipulation of files of structure and property data in a way not possible without this technology. Such computer-based files were originally developed largely for use within the pharmaceutical and agrochemical industries to control biological test data, and have had a profound effect on the type of structure-property work carried out (HANSCH, 1976, HANSCH et al., 1974).

Applications of such computer-based files have not been widespread in other areas, although some have been reported, as is noted elsewhere.

2.3.2.

Theoretical Methods of Structure-Property Correlation

Quantum Mechanics Calculations

The application of quantum mechanical methods to chemical problems has increased greatly in the past twenty-five years, due largely to the greater availability of large digital computers which has stimulated the development of appropriate techniques of calculation (HALL, 1973).

In the discussion below, some of the more widely used techniques will be outlined and their applications noted. Many detailed accounts of quantum theory and its application to chemical problems are available, and the outline of quantum chemistry below is based on the treatments in some of these (COOK, 1974a, DEWAR, 1969a, KIER, 1971a). The application of quantum techniques to "large" molecules, i.e. those of chemical and biological interest, will be emphasised: this area has been recently reviewed (DUKE, 1975a). The application of these methods in the field of drug research has been described in detail (KIER, 1971a, KIER, 1970).

The basis for all quantum chemical calculations is the well-known Schrödinger equation, with application of the Born-Oppenheimer fixed-nucleus approximation and with the Pauli electron-exclusion principle as a constraint. The most usual representation of this is

$$H\Psi = E\Psi$$

where H is the Hamiltonian operator, Ψ is the molecular wave function, and E represents the eigenvalues of the equation. It is not possible to solve this equation exactly for any system large enough to be of general chemical interest, and it has been pointed out that exact solutions would be unnecessarily complex for most purposes (COOK, 1974b).

The most commonly-used simplification of the quantum mechanical model is to partition the n-electron Hamiltonian operator, in an approximate way, into n one-electron Hamiltonians. The linear combination of products of orbitals given by one-electron Hamiltonians is an approximate solution to the Schrödinger equation. The inherent inaccuracy of this method, generally known as the Hartree-Fock self-consistent field method, due to its neglect of electron correlation, has been noted, and some methods for counteracting this will be mentioned below.

The form of the wave functicn, Ψ, resulting from this approach is given by

$$\Psi = \sum D_i \phi_i$$

where Di represents linear coefficients, and ϕ_i represents the determinants of orbitals. The best wave-function may then in principle be determined by optimising the coefficients and orbitals, using the variation method. However the computational difficulties of full optimisation of both are so great that this technique, known as multi-configurational self-consistent field, is not at present feasible for chemically interesting systems.

Two more approximate approaches have been used to generate wave-functions, starting from some form of approximate atomic orbitals. The first involves choice of a single set of orbitals, represented by a single determinant, and optimisation of these orbitals: this the 'Molecular Orbital' (MO) method. The second involves selection of fixed orbitals, generally atomic orbitals, and optimisation of the coefficients for a multiconfigurational wave function: this is the 'Valence Bond' (VB) method.

The valence bond method gained considerable popularity immediately following its introduction, largely because of its compatability with chemical concepts, bonds, lone pairs, resonance etc., and its findings have been widely used in a qualitative sense (WHELAND, 1944, PAULING, 1939). However it has not subsequently been used extensively for quantitative work, because its unwieldy formalism is not well suited to computation for large molecules (HALL, 1973). Recently there has been a revival of interest in this method of calculation (GERRATT, 1974).

One recent study has suggested that a greatly simplified varient of valence bond theory, considering only the 'chemically sensible' resonance forms, may give useful results (HERNDON, 1973).

The alternative molecular orbital method has predominated quantitative work, particularly for organic and biological molecules. The most commonly used form of this theory enables the calculation of molecular wave-functions by means of the linear combination of atomic orbitals (LCAO) approximation. When carried out directly such calculations are termed "ab initio", by distinction with the "semi-empirical" methods discussed below. Although these calculations contain severe approximations (DEWAR, 1969b, COOK, 1974c) they are the most rigorous which can at present be attempted on chemically useful systems. Although ab initio MO calculations have been refined to the point where they may be regarded as routine, with a considerable number of applications having been described (DUKE, 1975b, CHRISTOFFERSON, 1972, RICHARDS et al., 1974), they remain expensive in computer time, so that

any large-scale application is not at present feasible. Examples have been described where these methods have been used to investigate particular problems, which could not be solved by more economical approximate MO methods (PORT et al., 1974, PULLMAN et al., 1973).

It will be noted that there are two major drawbacks to the ab initio LCAO MO methodology: its uneconomical requirements for computing facilities and time, due in large measure to the necessity for calculating and storing complex integrals representing electron interaction, and its inherent inaccuracy due to the approximations involved in its formulation. A range of methods, generally termed "semi-empirical", have been devised to counteract these problems. These methods avoid the calculation of some or all the integrals resulting from the LCAO MO formulism, either by ignoring them or by deriving empirical values. Such methods may have one of two purposes: they may simply aim to mimic ab initio results more economically, or by use of suitable empirically derived parameters may aim to overcome the inherent inaccuracies of the MO approximations. A variety of such methods have been developed and have found application to chemical and biological systems (DUKE, 1975c, MURRELL et al., 1972, HOYLAND, 1969, HERNDON, 1972). Only a small number of the more widely-used methods can be mentioned here. The techniques described as CNDO (Complete Neglect of Differential Overlap) are simple and economical (POPLE et al., 1970) and have been widely applied to both chemical and biological problems (PULLMAN, 1972, KLOPMAN et al., 1970).

The MINDO (Modified INDO) technique makes extensive use of empirical data, in order to parametrise the method to give results

of the same accuracy as experimental data, and has been used to calculate physical properties of organic compounds (BINGHAM et al., 1975).

The PRDDO (Partial Retention of Diatomic Differential Overlap) method aims to mimic ab initio results as accurately as possible, with a considerable gain in computational economy. It is one of the few such "accurate semi-empirical" techniques to have been applied to large molecules (KIER et al., 1975, DIXON et al., 1976).

At a greater level of approximation than the methods discussed previously is the collection of techniques known generally as Huckel MO theory. The original form of this theory was developed before the general availability of computers, to provide an elementary but useful way of dealing with π-electron systems in terms of MO theory. It is essentially a very simplified form of LCAO MO theory, with the integrals treated as parameters to be determined empirically rather thar calculated. Despite the evident crudity of the method, it gained popularity and has been widely used in organic and biological chemical applications (STREITWIESER, 1961). It is largely a qualitative and semiquantitative technique, and has been found useful for calculating molecular indices within related series of compounds (SCHNAARE, 1971, GREENWOOD et al., 1966).

Various modifications have been made to the basic theory, in attempts to overcome its known defects. In particular the Extended Huckel Theory, which allows for the treatment for σ electrons, has attained wide use, despite criticism of its unsatisfactory theoretical basis (HERNDON, 1972, WOHL, 1971).

A number of alternatives to the LCAO approximation for calculating MO wave functions have been proposed. None has been widely applied, though the so-called Xαscattered-wave method has produced useful results for inorganic and biological molecules (JOHNSON et al., 1973, JOHNSON, 1973c).

Various methods have been proposed to overcome the inherent inaccuracy in the MO formalism, by taking explicit account of electron correlation. The semi-empirical PCILO (Perturbative Configuration Interaction over Localised Orbitals) method is the only such technique to have achieved wide-spread use (DINER et al., 1969). This computationally very economical technique has been applied mainly to the conformations of biological molecules (PULLMAN, 1972).

One feature of molecular orbital theory which should be noted is the invariance of the overall molecular wave function to certain types of transformation of the constituent MOs. Thus it becomes possible to choose between alternative, and entirely equivalent, sets of MOs and select the most useful set for some particular purpose (COOK, 1974d, RVEDENBERG, 1973). One particularly useful procedure is to transform the "canonical" set of MOs, which are generally delocalised over the whole molecule and thus most useful for describing "whole molecule properties" into a set of localised orbitals. Such localised representations may have three advantages (WEINSTEIN et al., 1971, ENGLAND et al., 1971):

i) they may be conceptually preferable, in that they are compatable with conventional chemical structural ideas, e.g. bond energy (VON NIESSEN, 1975).

ii) they may be transferrable between molecules, and could thereby simplify calculation - this has been investigated by a number of workers (O'LEARY et al., 1975) particularly in von Niessen's 'Molecules in Molecules' method (VON NIESSEN, 1973, VON NIESSEN, 1974).

iii) they may form the starting point for a more elaborate treatment including configuration interaction, e.g. PCILO or some more complex treatment.

A number of mathematical criteria have been advanced for localisation (WEINSTEIN et al., 1971).

Localised orbitals may be produced either directly or by transforming canonical orbitals. They may be produced by any ab initio or semi-empirical MO technique: in particular the PRDDO method has been largely used to produce orbital sets of this type (DIXON et al., 1976, KLEIR et al., 1975) and the CNDO/INDO methodologies have been used to this end (FIGEYS et al., 1975).

This localisation concept is very relevant to the subject-matter of this thesis, and is discussed in detail below.

Other Theoretical Methods

In addition to these quantum mechanical techniques, other methods which may be categorised as "theoretical" according to the criteria given above have found application.

Calculations based on the theory of statistical thermodynamics can give very accurate values for the physical properties of some simple compounds. They are however of little value generally, because of the complexity of the calculations for any structure other than the simplest hydrocarbons and because of the lack of the structural parameters needed for such calculations (JANZ, 1967b, PITZER, 1940). A quantum statistical methodology has been suggested for the investigation of certain biological properties (LIN, 1974).

A variety of methods for calculating preferred conformations have been developed, in addition to the use of quantum mechanical methods mentioned above. The use of rigorous techniques based on classical mechanics has been limited (GOLEBIEWSKI et al., 1974), but techniques making use of empirical potential energy functions have found application to both biological molecules (SCHERAGA, 1968) and organic species (ENGLER et al., 1973). It has been suggested that for some applications these methods may give results considerably superior to those obtainable by quantum mechanical methods, in addition to their advantage of computational economy (LILJEFORS et al., 1976). Such techniques have also been applied to the problems of solvent interactions, and to the prediction of partition coefficients (HOPFINGER et al., 1976).

Summary of applications of theoretical methods

The applications of the theoretical techniques discussed above to the investigation of structure-property relationships, and to the prediction of unknown property values may be divided into three categories. Such applications, using the techniques as a means to an end, may be distinguished from the considerable effort put into developments of methodology and testing of the methods on relatively small molecules.

First, and most obvious, is the direct calculation of numerical values for unknown properties of chemical species, using a wave mechanical technique. This is only possible for relatively straight-forward properties directly related to molecular energies, most commonly thermodynamic properties. The calculation by this means of, for example, a pharmacological property is at present out of the question, because of the complexity of biological processes (ARIENS, 1971d).

Because of the inaccuracies implicit in all quantum mechanical formulations at present feasible, as discussed above, the only methodology so far suggested to produce results of chemical accuracy is the heavily parametrised MINDO/3 technique (BINGHAM et al., 1975).

The second manner of application of theoretical techniques involves the calculation of quantum mechanical indices: orbital energies, electron densities, bond orders, atom charges etc. (KIER, 1971b). This has the advantage that, since such indices are usually compared across a series of structures, the inaccuracies inherent in the calculations may be nullified, since the values calculated are treated as purely relative. The simpler, and more economical techniques, are generally used in this way, especially the Huckel and Extended Huckel methods. The main applications of this procedure have been in studies of chemical reactivity and biological activity (SCHNAARE, 1971, GREENWOOD et al., 1966). Indices calculated in this way have been used in multiparameter semi-empirical correlations, generally as a parameter representing electronic effects (PURCELL et al., 1970, ANDREWS, 1972). This will be discussed further below.

The third important aspect of the application of theoretical techniques is the investigation of molecular conformation, which involves both wave mechanical and other methods as described above.

A wide variety of quantum mechanical methods have been applied to this purpose, from the more sophisticated ab initio methods to Huckel and Extended Huckel techniques (KIER, 1971c, KIER, 1972, PULLMAN, 1976). It is not uncommon for alternative methods to give entirely different results, with little rationalisation possible: the authors of a recent study regarded their chosen method, the Extended Huckel procedure, as an essentially empirical technique which fortuitously gave useful results in a certain area (GANELLIN et al., 1973). This may be regarded as a more realistic attitude than is expressed in many theoretical studies.

One major problem with the investigation of conformations is that an accurately calculated minimum energy conformation may be drastically altered due to intermolecular interactions, particularly in a biological environment. Some of the simpler methodologies, e.g. PCILO and some empirical conformation techniques, appear to have so far shown most promise in dealing with this problem (PULLMAN, 1972, HOPFINGER et al., 1976).

2.3.3.

Semi-Empirical Structure-Property Correlation

Semi-empirical Structure-Property Correlation

This part of the chapter will consider methods of structure-property correlation which relate the property under investigation with other measured or calculated molecular properties, or parameters derived from such properties. The term "semi-empirical" is most commonly used for techniques, and so is adopted here, though it tends to obscure the empirical nature of much of this work. The terms "linear free-energy relationship" and "extrathermodynamic relationship" have often been used, referred to the thermodynamic basis of such correlations (EXNER, 1972b), while the self-explanatory term "physicochemical activity relationship" has been suggested for the multiparameter biological applications (NORRINGTON et al., 1975), and the less specific "correlation analysis" for chemical aspects (EXNER, 1972b).

Hansch and co-workers have used "correlation analysis" and also "quantitative structure-activity relationship (QSAR)".

A distinction is often drawn between the two major applications of such correlation methods: application to chemical reactivities and similar aspects of physical organic chemistry, often termed "Hammett correlation", and application to biological activities, generally referred to as "Hansch analysis". Although rather arbitrary, this distinction is convenient for the purposes of discussion and will in general be adhered to below.

Interest in this type of relationship, as has been noted above, began at the end of the last century with observations of quantitative relationships between biological activities and physicochemical properties, notably solubilities, partition coefficients, and boiling points, leading to a rationalisation of these effects in thermodynamic terms.

The Hammett Approach

The development of physical organic chemistry led to the amassing of

much quantitative data, in particular rate and equilibrium constants, and a consequent search for quantitative relationships to summarise and rationalise this data. The relationship which has since found very wide use was originally postulated by Hammett (HAMMETT, 1937), and has provided the basis for virtually all correlation analysis in organic chemistry. Comprehensive reviews of the by now vast literature of the application of Hammett-type analysis are available (CHAPMAN et al., 1972, JAFFE, 1953, JOHNSON, 1973a, HINE, 1976b) and the discussion below will do no more than outline the main points.

Hammett's equation is usually expressed in the form

$$\log k = \log k^{o} + e\sigma$$

where k denotes a rate constant or equilibrium constant for some compound; k^{o} denotes a statistical quantity corresponding to k for an 'unsubstituted' or 'parent' compound; e, the so-called 'reaction constant' depends on the nature and conditions of the reaction or equilibrium process; σ, the so-called 'substituent constant', is characteristic of a particular substituent in a particular position and independent of the reaction. The equation was formulated originally to deal only with meta- and para- substituted benzene derivatives, and most applications have been restricted to this system, although work has been carried out with acyclics, heterocyclics and fused systems (EXNER, 1972c, JOHNSON, 1973b).

The Hammett equation remains essentially empirical, although a good deal of effect has gone into providing a theoretical rationale and establishing its range of validity (RITCHIE et al., 1964, EXNER, 1972d, WOLD, 1974). It is now regarded as well-proven, and in general as "a convenient tool to summarise experimental results and to detect exceptions" (EXNER, 1972e).

The substituent constants of the Hammett equation reflect electronic effects of substituents. It is evident that these will to some extent vary in different reactions, depending on the nature of the interaction between substituent and reaction site. It was found necessary to introduce dual σ values for certain substituents; a "normal" constant, and one or more alternative constants to be used when a strong mesomeric interaction could occur (JAFFE, 1953, STOCK et al., 1963). This was a highly artificial solution, in view of the continuous variation in effects observed, and various alternative and more complex formalisms were proposed (YUKAWA et al., 1959, YUKAWA et al., 1966, HUMFFRAY et al., 1969).

These were equations of the form

$$\log k = \log k^{o} + \rho \left[\sigma + r\,(\sigma^{*} - \sigma)\right]$$

where σ is the normal substituent constant, σ^{*} is an alternative substituent constant and r is a variable depending on the nature of the reaction.

Such equations lead to a "sliding scale" of substituent constant values.

This tends towards the situation where a different σ constant would be required for each reaction, negating the general principles of this type of correlation procedure. As an alternative to this proliferation of substituent constants, procedures were devised for partitioning the electronic effect into inductive and resonance factors (TAFT et al., 1959, SWAIN et al., 1968, EHRENSON et al., 1973). Recent work has involved the "positional weighting" for substituent constants representing these effects, to allow for positional effects in benzene derivatives and thereby increase the generality of the model (WILLIAMS et al., 1976).

The notable work of Taft twenty years ago enabled steric effects to be included in Hammett-type analysis, with the introduction of a steric parameter Es (TAFT, 1952). Totally sterically controlled reactions then conform to the equation

$$\log k = \log k^{o} + SEs$$

where S is a steric susceptibility constant. More general equations involving both electronic and steric parameters have been found useful in some cases. The introduction of the steric parameter is generally regarded as a great advance in Hammett-type work, not least since it enables ortho substituents to be included, though the "ortho-effect" is still a considerable problem (CHARTON, 1971, SHORTER, 1972b).

The partitioning of substituent effects into polar (or field, or inductive), resonance (or mesomeric), and steric factors is one of the major features of this type of correlation analysis (SHORTER, 1972a). A number of authors have however warned against the possibility of drawing unjustified mechanistic conclusions from a successful empirical correlation, particularly in view of uncertainty as to the physical meaning and inter-relationships of the constants (WILLIAMS et al., 1976, DEWAR et al., 1962, RITCHIE et al., 1964, WOLD, 1974).

In addition to an analysis of reactivity in terms of reactant structure, Hammett correlation analysis has enabled studies of the effect of the reagent (PEARSON, 1972) and the solvent (KUPPEL et al., 1972) on reactivity.

Such correlations have also been employed, with mixed success, in the interpretation of spectra (BURSEY, 1972, KATRITZKY et al., 1972, TRIBBLE et al., 1972).

A problem for all correlation analysis of this sort is the choice of the "best" substituent constant or reaction constant. A

notable attempt has been made to put into practice the ideas of Jaffe (JAFFE,1953) by carrying out statistical analysis of large data sets, in order to derive optimal values for the substituent constants, and thereby give a firmer statistical base to such correlation analysis (WOLD et al., 1972, SJOSTROM et al., 1974).

It may finally be noted that the major uses of the Hammett equation and related correlation analysis techniques have been the summarising of data, the detection of anomalous results, and the rationalisation of experimental observations. Little use has been made of such techniques for the prediction of reactivity or similar chemical properties (EXNER, 1972f). Their major application, as discussed above, has been in organic chemistry,but they have also been applied in inorganic chemistry (CHIPPERFIELD, 1972) and in enzymology (KIRSCH, 1972). Their use in the correlation of biological activities will be discussed in the next section.

The Hansch Approach

The methodology generally termed "Hansch analysis" involves the correlation, by means of regression analysis, of biological activities with measured or calculated physicochemical parameters. It has been reviewed in detail by a number of authors (HANSCH, 1971a, VERLOOP, 1972, CAMMARATA et al., 1972, CRAIG, 1972a, MCFARLAND, 1971, VAN VALKENBERG, 1972, GOODFORD, 1973, TUTE, 1971, PURCELL et al., 1973c, HANSCH, 1969) and will only be outlined here.

As mentioned above, a number of early studies in structure-property relationships involved the correlation of biological activities with physicochemical properties. Following the adoption of the Hammett analysis procedure in organic chemistry, this technique was applied to biological problems, but the use of parameters reflecting almost solely electronic effects met with limited success (HANSCH, 1971c).

A major advance was made by Hansch and co-workers, with the use of partition coefficient values for correlation, reflecting the lipophilic character of chemical compounds (HANSCH et al., 1962, HANSCH et al., 1964a). Such parameters have been found to be the most generally useful, of a number investigated, in the correlation of a range of biological properties (LEO et al., 1969). Two main types of relationship have been found to link lipophilicity with biological activity. The first is a linear relationship (HANSCH, 1971d, HANSCH et al., 1972) represented as

$$\log 1/c = a \log P + b$$

and the second is a non-linear relationship (HANSCH, 1971e, HANSCH et al., 1973a)

$$\log 1/c = k_1(\log P)^2 + k_2(\log P) + k_3$$

1/c represents a biological response, P is a partition coefficient, while a, b, k_1, k_2 and k_3 are coefficients, taking particular values

for a given biological effect. Many examples of correlations using both types of relationship have been published.

Because of the additive -constitutive nature of partition coefficient values, it is possible to deal with contributions from parts of a molecule. A quantity generally termed π was defined (FUJITA et al., 1964) as being the logarithm of the partition coefficient for a part of a compound

$$\text{i.e. } \pi = \log P_{px} - \log P_{p}$$

where P_{px} is the partition coefficient of a structure substituted with X, and P_{p} is the partition coefficient of its parent (i.e. unsubstituted) structure. Such π values have been widely used, either to estimate an unknown partition coefficient, or directly in correlation.

Alternative methods for deriving parameters reflecting the lipophilicity of structural fragments have been reported (LEO et al., 1975, NYS et al., 1973, NYS et al., 1974a). These will be discussed in detail below, since they are of considerable relevance to the main subject-matter of this thesis. Other empirically derived lipophilic parameters, essentially similar to π, have been used (ZAHRADNIK et al., 1960, KOPECKY et al., 1967).

It is recognised that π values are at best an approximation, since the lipophilicity of any grouping will differ according to its environment. In particular, deviations can occur due to conformation flexibility and steric or electronic interaction: the only reliable safeguard is to use measured partition coefficient values whenever feasible (LEO et al., 1975, HANSCH et al., 1973b, CANAS-RODRIGUEZ et al., 1972). Problems in the definition and derivation of π values have been noted (DAVIS, 1973, NYS et al., 1973).

The so-called multiparameter procedure involves the use of multiple regression analysis to correlate a biological activity with a number of parameters, intended to represent different relevant aspects of the compound. A great variety of such parameters have been used, which may be divided roughly into three categories (HANSCH, 1971f, VERLOOP, 1972).

Firstly, there are parameters to represent lipophilicity: generally π values or their equivalent. Chromatographic parameters have also been used (BIAGI et al., 1972), because they are in general simpler to measure than are partition coefficients. Limited use of other parameters related to lipophilicity has been reported (MORIGUCHI, 1975, MCGOWAN, 1952).

Secondly, parameters representing electronic effects: a large number of forms of Hammett σ-constants and similar have been extensively used, and recently the Swain and Lupton parameters have been applied. Various quantum mechanical indices have also been utilised.

Thirdly, parameters representing steric factors: the Taft steric parameter was the first to be employed in this way (KUTTER et al., 1969), but simpler measures of bulk, notably molar refractivity, have been more widely used.

It should be noted that these three groups are not entirely independent, since for example molar refractivity, generally used as a steric parameter, is also a measure of electronic effects (HANSCH et al., 1973b). A study of the interrelations between such parameters has indicated that there is a fairly clear division between polar and non-polar parameters, but a considerable amount of inter-correlation otherwise (CRAIG, 1971a).

Collections of self-consistent values for such parameters have been published (HANSCH et al., 1973b, NORRINGTON et al., 1975).

It will be seen that no distinction is made between parameters derived from experimental measurements and from theoretical calculations. The choice of parameters used will largely depend on convenience, on the significance of the correlations attained, and on any mechanistic rationale for the observed activities. The possible dangers of an over-enthusiastic "mode-of-action" interpretation of such correlations have been discussed (VERLOOP, 1972).

It has been found useful in some cases to introduce "dummy variables", representing structural features of the compound, in addition to physicochemical property measures in multiparameter studies (HANSCH et al., 1974b, FUKUNAGA et al., 1976). This indicates a possible overlap of physicochemical parameter correlations with additive modelling techniques. This aspect will be discussed in the section dealing with additive modelling methodologies.

Although physicochemical-activity relationships are in principle applicable to sets of diverse structures, and have indeed been applied to such cases (HANSCH, 1971a), their most extensive use has been within homologous series; i.e. for lead optimisation rather than lead generation (REDL et al., 1974).

The thermodynamic basis of such relationships, as with the Hammett equation, has long been recognised, and was first explicitly employed in a rationalisation of the equivalence of the varied parameters used in early work (FERGUSON, 1939). More recently the analysis of such

correlations in terms of thermodynamics and theoretical models has been investigated (HIGUCHI et al., 1970, DAVIS et al., 1974d, HYDE, 1975b, MARTIN et al., 1976, MCFARLAND, 1971).

These methods are dependent upon adequate computing facilities for statistical analysis, and more recently data-processing techniques have been utilised to assisting in handling structural parameters. These aspects have been discussed (CRAIG, 1971b, HANSCH et al., 1973b, HANSCH, 1972).

The statistical evaluation of such correlations has also been discussed, particularly with regard to the necessity for a sufficiently large ratio of observed property values to parameters in order to minimise the possibility of chance correlations (TOPLISS et al., 1972, UNGER et al., 1973, CRAIG, 1972a).

With the widespread use of physicochemical parameter correlations for optimisation of activity within a set of structures, it is of evident importance to carry out investigations in a rational and systematic fashion, so as to minimise costly synthesis and testing procedures. Methods have been suggested for choosing appropriate derivatives for synthesis in order to give a wide range of physico-chemical parameter values for formulation of useful correlation equations. These have included procedures for the cluster analysis of substituents based on physicochemical properties, with subsequent synthesis of derivatives from each cluster (HANSCH et al., 1973c), and for maintaining minimum distances in multidimensional parameter space between substituents (WOOTOON et al., 1975).

A simpler means to achieve the same aim is the 'scatter diagram' in which substituents are plotted on axes representing physicochemical parameters (CRAIG, 1971a). Other methods for

rationalising synthesis programmes, based on a knowledge of physicochemical parameters, have been reported (BUSTARD, 1974,, SANTORA et al., 1975, DEEMING, 1976, DARVAS, 1974, TOPLISS, 1972, MARTIN et al., 1973g, TOPLISS et al., 1975).

Various means for presenting the results of physicochemical-activity analyses graphically have been devised (VERLOOP, 1972). One such method involves the plotting of 'structure-activity surfaces', i.e. three-dimensional plots of activity against two physicochemical parameters (NEELY et al., 1968, NEELY et al., 1970).

2.3.4.

Empirical Structure-Property Correlation

Empirical Structure-Property Correlation Techniques

Under this heading will be grouped a wide variety of techniques. Following a common convention, a distinction will be made between"parametric"and"non-parametric"methods. By parametric is meant those methods designed to operate on interval or ratio data, which in this field implies the use of regression analysis, or similar techniques. Those parametric techniques falling within the area of pattern recognition, which presume a knowledge of the distributions of the patterns to be studied, and make use of Bayesian strategies in a classification process (HOEL, 1971, NILSSON, 1965b) have not be applied in the structure-property area, and will not be further discussed.

Non-parametric methods include all those techniques which do not require interval or ratio nor any knowledge of the distributions of the variables considered. This section includes a number of methods which are not strictly statistical analysis techniques, since they involve the displaying or listing of data. A distinction is made between classificatory and non-classificatory methods.

Classification methods are divided into supervised learning, where some predetermined categorisation is imposed on the data to be analysed, and unsupervised techniques, where no a priori categorisation is applied. In this latter section are included the various data ordering and display methods.

In the section on non-classificatory methods are discussed several non-parametric statistical techniques which have found application in this area.

"Pattern recognition" is a term which has been used to describe a variety of non-parametric techniques used in this area. It is generally applied to supervised learning, and to some of the unsupervised analysis and display methods (KOWALSKI et al., 1972). An integral part of all the pattern recognition techniques applied in this area is "preprocessing" of the data: this will be further discussed below.

It should be noted that only software techniques, i.e. involving computer programs, are of concern here. Various pattern recognition techniques have been devised for spectral analysis using digital electronics hardware (STONHAM et al., 1975 these will not be further discussed.

A number of the methods described here are included in the "substructural analysis" categories, with automatically derived structural features. This aspect will be noted in the discussion below, but a detailed discussion of structural feature derivation will be left for the following chapter.

Additive models for physical property estimation

As noted in an earlier section, empirical relationships between structure and physical properties have been studied for more than a century.

At the present time such methods are most extensively used for estimation of thermodynamic properties, an area of considerable practical importance. Many such empirical schemes have been devised, of varying degrees of sophistication, all based on the assumption that structural units make constant contributions to property, regardless of the environment of each unit (REID et al., 1966, JANZ, 1967c, HINE, 1975c, COX et al., 1970b, STULL et al., 1969, BENSON et al., 1969, KITAIGORODSKY, 1973). The complexity of the structural units used vary greatly, and various methods have involved atomic contributions, bond contributions, and group contributions, and the more sophisticated techniques allow for substituent interactions and for strain energy contributions.

The performance of some of these techniques will be examined in detail, for comparison with work carried out as part of the study reported in this thesis, in a later chapter.

With the increasing use of computers for handling thermochemical information (ZWOLINSKI et al., 1972b, PEDLEY, 1976) has come a consequent interest in the development of automated methods for estimating such physical properties (MEADOWS, 1965).

The use of substructure search techniques to identify the structural units for estimation procedures has been demonstrated, using connection tables (JOCHELSON et al., 1968) and Wiswesser Line Notation (BRASIE et al., 1965). A fragment code, especially suitable for this purpose, has been devised (ROMANEC, 1974).

Other physical properties which are approximately additive functions of simple structural units, which as noted above were widely used in the past for structure determination, are still a topic of investigation (EXNER, 1966). Properties which reflect molecular bulk are of particular interest (EXNER, 1967a, EXNER, 1967b). A recent study has shown that such relations can be of value in investigating conformational problems, if appropriate structural units are used (KELLIE et al., 1975): this work will be considered in detail in a later chapter.

An additive modelling approach of this sort has been applied to partition coefficients (NYS et al., 1973, NYS et al., 1974): this work will be discussed further below.

Additive Modelling for Biological Properties

This area has been reviewed (PURCELL et al., 1973d, CRAIG, 1972a, CRAIG, 1975).

The first example of this kind of empirical relationship was given in a study of thyroxine analogues (BRUICE et al., 1956), which made use of the equation

$$\log \text{activity} = k \sum_i f_i + c$$

where fi is a coefficient representing the effect of substituents at a particular position in the parent ring system : k and c are constants.

This approach was given a more general form in the so-called Free-Wilson methodology (FREE et al., 1964). This is based on the single assumption that a particular substituent group in a particular position in a parent structure has a constant and additive effect on biological activity. Thus

$$\log \text{activity} = \text{overall average} + \text{(group contributions)}$$

It is important to note that this model can be applied only to sets of structurally similar compounds. The constant term is the overall average of the biological activities because this original form of the model is based on symmetry equations. Various modified forms of the Free-Wilson method have been used (KUBINYI et al., 1976). In one such modification (FUJITA et al., 1971), which dispenses with symmetry equations, the contributions to activity for all substituents are relative to hydrogen as a substituent, assumed to have zero contribution, a modification originally devised by Cammarata (CAMMARATA et al., 1970). The equation then becomes

$$\log \text{activity} = \sum_i a_i \bar{} + \mu$$

where a_i are the activity coefficients for the substituents: the constant, μ, is the theoretically derived log activity value for the unsubstituted parent compound.

The presumption of additivity of substituent effects in this model generally means that substituent interactions must be discounted. An example has been published where a term representing the interaction of two substituents was included in such an analysis, and shown to significantly improve the correlation (FUJITA et al., 1971).

More generally, an additive model has been proposed which includes terms representing possible intersubstituent interaction, generally termed the Docek-Kopecky model (BOCEK et al., 1964, KOPECKY et al., 1965, BOCEK et al., 1967). In this model the activity of a compound with substituents X and Y would be given by

$$\log \text{activity} = b_x + b_y + e_x e_y + k$$

which may be regarded as a Free-Wilson relationship with the additional term $e_x e_y$ representing substituent interaction.

Additive empirical models of this kind will give different activity coefficients for each set of data analysed. This, as has been pointed out, may be both a strength and a weakness (CRAIG, 1975). The coefficients are not generally applicable, as are eg π values, but are a more accurate representation of a particular data set than would in general be possible using generalised parameters. Since the coefficients embody all factors, known and unknown, which affect activity, there is no necessity to prejudge which type of parameter

to use. Examination of empirical coefficients may give sequences correlating with physical property and so give an indication of the physicochemical factors involved, perhaps leading to the formulation of a useful Hansch-type correlation equation (CRAIG, 1972a, KUBINYI et al., 1976a). The physicochemical and empirical correlation methods are closely related: it has been suggested that the Free-Wilson model is equivalent to the linear Hansch approach, while the Bocek-Kopecky model is related to the parabolic Hansch method (SINGER et al., 1967, CAMMARATA, 1972, KUBINYI et al., 1976).

A mixed approach has been described, essentially a generalisation of the use of dummy variables in Hansch analysis as discussed above, which combines the empirical and physicochemical parameters approaches (KUBINYI, 1976). This yields equations of the form

$$\log i/c = k_1\pi^2 + \sum_i a_i + \sum_j k_j\phi_j + k'$$

$\sum a_i$ are the Free-Wilson activity coefficients, $\sum_j k_j\phi_j$ is the "Hansch part" for physicochemical parameters ϕ, the term $k_1\pi^2$ allows for a parabolic dependence on partition coefficient, k' is a constant.

Empirical methods of the Free-Wilson type are particularly valuable for complex structures with several possible positions of substitution, where the possible permutations of a relatively small number of substituents increase drastically. Thus in one such study a series of structures with six substituent positions, substituted by 16, 4, 7, 4, 5 and 4 groups respectively, with a resultant possible 35,840 analogues were subjected to Free-Wilson analysis which gave a good structure-activity relation based on property values for sixty-nine

structures (CRAIG, 1972b). Such complex systems are the only area in which additive modelling techniques have been used to anything like the same extent as physicochemical activity relations. This may perhaps be ascribed to their restricted applicability to certain kinds of closely related sets of structures, and to the relatively large number of compounds for which property values must be available before the Free-Wilson approach is practicable (CRAIG, 1975).

The statistical bases for this type of analysis have been discussed (HUDSON et al., 1970, CRAIG, 1972a). It is suggested that a compromise is necessary between the requirements of statistical rigour and those of convenience and economy, particularly as regards the ratio of measured property values to variables in the analysis. The number of variables produced may be a particular problem with the Bocek-Kopecky type models including interaction terms.

Recent work has demonstrated the applicability of sub-structural analysis to additive modelling of biological properties (BUSH, 1976, ADAMSON et al., 1974, ADAMSON et al., 1976a). Property values are assumed to be an additive function of structural features automatically derived from a connection table representation of structure, and correlated by multiple regression analysis. This method may be applied to groups of structurally diverse compounds. This work will be discussed in detail in the next chapter.

Classification Techniques - Supervised

Supervised classification techniques, often termed supervised learning, form a branch of pattern recognition. The published literature of this field, generally dealt with from a mathematical viewpoint, is considerable. It will not be reviewed here, since it has been very adequately dealt with in discussions of the applications of these techniques in chemistry-related fields (KOWALSKI et al., 1972, JURS et al., 1975a, BUSH, 1976, KOWALSKI, 1974, ISENHOUR et al., 1974).

The basis of supervised learning involves a set of entities which are to be categorised in some pre-determined fashion. The categorisation may be binary, e.g. active or inactive for a pharmacologically-tested compound, or may be multicategory, e.g. type of functional group present. The majority of applications have used binary classifications. A "learning set", i.e. a set of entities whose categorisation is known, is used to create a decison-making process, so that subsequently input entities will be classified on the basis of their attributes.

Two forms of decision-making method have been largely used for supervised learning in chemistry-related areas. The first of these is the so-called "learning machine", or "trainable classifier" (NILSSON, 1965, KOWALSKI, 1974). These are discriminant functions which operate on an input pattern, i.e. set of attributes of an entity to be classified, to produce a numeric value, which indicates the classification. A simple and widely-used example of this is the binary pattern classifier using a linear discriminant function (JURS et al., 1975b). This is essentially a weighting function which when multiplied by a pattern vector gives a scalar result, the sign of which indicates the classification. Such discriminant functions are

"trained" by inputting the pattern vectors of the entities comprising the "learning set" one by one. As each is entered the function is altered according to various criteria (KOWALSKI et al., 1969, NILSSON, 1965c) so as to give the correct classification. The classifier, thus trained, is used to classify other entities, whose categorisation is unknown.

The use of such learning machines has been criticised on various grounds (KOWALSKI et al., 1972b). In particular these authors suggest that these methods are poorly adapted for anything other than two-class, linearly-separable data, that they generate non-unique solutions, and that their ill-defined statistical foundation makes their applicability to many cases difficult to assess. It has been suggested that in order to use these methods with any confidence, a ratio of entities to attributes of at least 3 to 1 is required (FOLEY, 1972).

The second decision-making method which has been applied is the "K-nearest neighbour" methodology (KOWALSKI et al., 1972b, KOWALSKI, 1974). This has been suggested to be superior to the learning machine because of a firmer statistical basis, and its ability to generate a unique solution. Entities are projected as vectors in n-dimensional space, according to the value of their n attributes, and classified according to a "majority vote" of the categorisation of their K nearest neighbours in n-dimensional space, where K is usually a small number, generally 1, 2 or 3. The learning set entities are used to categorise the unknowns in this way.

Other related pattern recognition techniques have been suggested for application in drug design (HILLER et al., 1973) and chemical reactivity (IOFFE et al., 1969), but do not appear to have attained practical use.

An integral part of supervised learning procedures is pre-processing of data before it is input as features by which the entities are to be classified (KOWALSKI et al., 1972a, JURS et al., 1975c, KOWALSKI, 1974).

This may be necessary to allow the simultaneous use of varying kinds of input, e.g. different forms of spectral data. It may make classification more distinct, e.g. by weighting appropriate features, or may make the process more economical computationally by reducing the dimensionality of the problem, either by omitting certain attributes, or by combining attributes by some transformation procedure. Most pattern recognition feature selection techniques are purely statistically based, and may therefore be inappropriate for chemical problems, since the frequency of occurrence of structural features, spectral peaks etc. is no indication of their importance. One widely-used method to overcome this problem is the so-called "weight-sign" feature selection procedure (JURS, 1970), whereby features are discarded from the classification procedure so long as their omission does not affect the overall decision process.

The great majority of chemical applications of supervised learning techniques have been in the area of analytical chemistry, where data from analytical measurements, often spectroscopy, are used to categorise the nature of the substance (JURS et al., 1975a, KOWALSKI, 1974, ISENHOUR et al., 1974). Applications involving the input of features representing molecular structure to deduce properties have been more limited.

Learning machine techniques have been used to classify groups of compounds having pharmacological activity as hypnotics

into subgroups of tranquillisers and sedatives. One such study used connection table fragments to represent molecular structure (CHU, 1974), while another, on a larger group of structures used a variety of structural descriptors as input (STUPER et al., 1975).

Learning machines and K-nearest neighbour methods have been applied to the categorisation of potential anti-tumour drugs as active or inactive. One such study used a variety of structural descriptors to classify a set of purine and pyrimidine nucleoside derivatives (KOWALSKI et al., 1974), while another used a complex set of structural features, derived automatically, to deal with a set of structurally diverse compounds (CHU et al., 1975).

The use of learning machines to predict mass spectra, i.e. the presence or absence of a strong peak in each relevant position, based on input structural descriptors has been described (JURS et al., 1974d).

Features representing molecular structure other than structure diagram fragments have been used as input for supervised learning procedures. Descriptors including electronegativities, dipole moments, and intramolecular dimensions were used with learning machine and K-nearest neighbour techniques to classify substituted benzoic acids as more or less reactive towards hydrolysis than the parent structure (KOSKINEN et al., 1974). In a further learning machine distinction between sedatives and tranquillisers, three-dimensional co-ordinates were obtained from crystallographic data or from standard bond dimensions, and rendered suitable for input by a molecular transform (SOLTZBERG et al., 1976). A similar study used three-dimensional data from compilations of bond dimensions plus atom electronegativities to

classify aliphatic and polycyclic aromatic structures as active or inactive carcinogens using learning machine and K-nearest neighbour techniques, following appropriate molecular transforms (DIERDORF et al., 1974).

Criticisms have been levelled at these pattern recognition techniques, and others to be described below, largely on the grounds that the data sets used have been unbalanced so as to make the results either trivial or misleading (CLERC et al., 1973, PERRIN, 1974, MATHEWS, 1975).

Classification Techniques - Unsupervised: Clustering, Display etc.

A wide variety of methodologies will be considered in this section. Firstly,the techniques known generally as "unsupervised learning" or "cluster analysis" will be discussed. Secondly,a variety of display and mapping methods will be considered, together with other simple data analysis techniques.

The terms "unsupervised learning" and "cluster analysis" are for practical purposes synonymous, except perhaps for "unsupervised learning" to refer to mathematically simpler techniques, and they will be used interchangably here. Other essentially synonymous terms are "numerical taxonomy", generally applied to biological studies, and "automatic classification", used in the context of information retrieval.

The underlying rationale of the many techniques of this kind is straightforward in principle. Starting from the vector in n-dimensional feature space representing each of the objects to be classified, a matrix is computed giving a coefficient of the similarity between each pair of objects, according to some measure of similarity of dissimilarity. From this matrix some form of clustering procedure is used to identify clusters within the data set and to output this result in some appropriate format. A great variety of such techniques have been devised (SNEATH et al., 1973, BUSH, 1976).

Since cluster analysis is one of the techniques used for the work described in this thesis, a detailed discussion of aspects of this procedure is necessary. This will be given in a later chapter. It is sufficient to note at this point that there is relatively little theoretical statistical basis for preferring one clustering method

over another, and that therefore many investigations are empirical in nature (EVERITT, 1974).

Relatively few examples of the applications of these methods to structure-property correlation have been published.

They have been used to a limited extent as complements to supervised learning (KOWALSKI, 1974), as in certain of the studies cited in the previous section (CHU, 1974, KOSKINEN et al., 1974). An early study of this kind used unsupervised methods to distinguish between sedatives and tranquillisers on the basis of the position of peaks in the mass spectrum (TING et al., 1973). This publication was one of those included in the general criticisms of pattern recognition applications mentioned above.

Hierarchical clustering procedures have been applied to chemical structure problems. An early study of this sort involved a classification of the twenty naturally occurring amino-acids, using a large set of structural descriptors and physicochemical parameters (SNEATH, 1966), giving intuitively sensible results.

Substructural analysis methods have been used in conjunction with hierarchical clustering techniques, using structural features automatically derived from connection table representations (BUSH, 1976). These methods may be used for property prediction, taking an unknown property value as equal to that of its nearest neighbour, to the average of the values of the other structure in its cluster, or some similar measure - this procedure is obviously similar to the K-nearest neighbour supervised learning procedure. Examples have been given of the application of this methodology to pK values for amino-acids (ADAMSON et al., 1973) and anaesthetic activity of diverse structures (ADAMSON et al., 1975a). This work will be discussed in detail below.

The application of hierarchical cluster analysis based on the biological activity values of the compounds to be classified, thereby possibly giving an insight into similarity of mode of action, has been described (SAGGERS, 1974).

The use of cluster analysis procedures for selection of substituents for Hansch-type analysis has been discussed in an earlier section.

The procedures described as "display" or "mapping" have as their general aim the representation of points in n-space by the same number of points in m-space, where m is less than n, and will usually be 2, for a graphical representation. Some information loss is inevitable in this process, and a variety of methods have been devised to minimise this loss (KOWALSKI et al., 1973). These methods may be divided into two categories: linear methods, for which the co-ordinate axes of the points in m-space are linear combinations of the co-ordinates of the n-space points; and non-linear methods, where this is not the case. It is obviously advantageous if such display methods can be used with an interactive graphics computer system, for "informal exploration of a given data set" (BALL et al.,1970): the application of such a system to chemical problems has been described (KOSKINEN et al., 1975).

It has been suggested that methods of this kind may be the most valuable of the "pattern recognition" techniques, in allowing the investigator to examine a complex data set from several respects, and thereby either solve a problem directly or formulate a more complex analysis approach (KOWALSKI et al., 1973). Although these

techniques have been found useful for analytical problems (KOWALSKI, 1974), few applications to structure-property relationships have been reported (KOSKINEN et al., 1974, TING et al., 1973).

The use of a similar technique to deal with data of the olfactory properties of substances has been described (SCHIFFMAN, 1975). Each substance was represented as a point in n-space based on judged olfactory attributes, and multidimensional scaling used to give a two-dimensional representation, for a subsequent rationalisation by consideration of structural and physicochemical factors.

A simple but highly effective way of displaying structure-activity data is the permuted list. Generally used with linear notations, such lists allow all significant notation symbols to appear as "headings", with appropriate property data listed alongside. Alternatively, if several types of property data are available, these may be permuted, thus giving an overview of their interrelatedness. The use of this admittedly crude technique has been described by several authors (SAGGERS, 1974, GRANITO et al., 1965, EAKIN, 1975, OYNACKER et al., 1970).

Under the general heading of display techniques should be mentioned the methods of graphically representing physicochemical factors, the "scatter diagram" and the "structure-activity surface", mentioned in a previous section.

Non-Classificatory Methods

One relatively simple method of quantitative structure-property correlation, suitable for application to large files of structures and properties, involves calculation of "substructure-activity frequencies", i.e. approximate measures of the contribution to a particular activity by a given structural feature over a wide range of structures (EAKIN, 1975, REDL et al., 1974, CRAMER et al., 1974). These procedures may be used with qualitative biological data, e.g. active/inactive, and the results may be assessed by standard non-parametric statistical tests, e.g. chi-squared. Rough measures of the likely activities of substances for which quantitative measurements are not available may be made in this way, and smaller groups of structures may be identified for more detailed study. This technique is particularly suited for automatic substructural analysis, and the published examples cited above have described the use of structural features derived from fragment codes.

Discriminant analysis is a non-parametric statistical technique, which, when dealing with entities arbitrarily divided into groups, can determine which of the properties or combination of properties of the entities gives the best classification of the entities into their assigned groups. This technique has been applied to a group of enzyme inhibitors, divided into groups on the basis of semiquantitative activity measurements: physicochemical properties and structural descriptors were assessed for discriminatory power, i.e. effect on activity (MARTIN et al., 1974). It is suggested that this method may be valuable as an initial tool to determine the major factors in

causing molecular activity, since it can deal with qualitative data, and inactive compounds can be included in the analysis, which is not the case with parametric methods.

Factor analysis and principal component analysis are closely related non-parametric techniques,devised to simplify the description of the variance observed in a complex data set. They have been used in this way in certain of the strcture-activity investigations mentioned above (SNEATH, 1966, SCHIFFMAN, 1975), and for preprocessing of data for pattern recognition (DIERDURF et al., 1974). A study of eleven different biological activity values for structure in a set by factor analysis enabled the investigation of the interrelatedness of the biological tests, and, by identification of the abstract factors with structural features and physicochemical properties, allowed an insight into the causes of activity (WEINER et al., 1973).

A recent study has used factor analysis techniques to classify structures into groups corresponding to pharmacological activities on the basis of structural descriptors and physicochemical properties (CAMMARATA et al., 1976).

Assessments of the interrelations of molecular properties and of various parameters have featured largely in many investigators of structure-property relationships. A recent study has demonstrated an unusually systematic approach (DOVING, 1973). Similarity measures between the physiological effects of a number of substances causing olfactory stimuli were correlated with Euclidean distances calculated from a large number of physicochemical parameters for each pair of substances, to enable some insight into the factors associated with physiological effect.

2.3.5.

Property Prediction

Use of Quantitative Structure-Property Correlation Procedures for 'Prediction' of Biological Activity

Despite the many published accounts of the application of structure-property relationships to biological problems (detailed in the review articles cited above), very few property predictions have been reported (REDL et al., 1974). This excludes simulated predictions made solely to test the performance of a particular technique.

Such predictions, followed by synthesis and testing, have been described resulting from Free-Wilson analysis (BEASLEY et al., 1969, COUSSE et al., 1973) from Hansch-type analysis (FULLER et al., 1968, GOODFORD et al., 1973) and from regression analysis with parameters derived from Huckel M.O. theory (MARTIN et al., 1973b).

The potential application of these methods for prediction of biological activity has been reviewed in detail (GOODFORD, 1973).

It may be noted at this point that attention has been drawn to the possible effects of the use of quantitative structure-activity correlation technique on the patentability of compounds with pharmacological activity (HUMBER et al., 1975). It is suggested that prediction of a biological activity of an as yet unsynthesised compound may preclude its subsequent patenting, and may thus inhibit the synthesis and testing of such compounds. These authors feel that the problem will be more likely to arise with predictions for novel types of structures, i.e. lead generation, than with optimisation of a parent structure.

It is conceivable that such considerations may affect the nature of what appears in the published literature concerning the application of structure-property correlation techniques.

Chapter Three

Substructural Analysis Techniques

(The scientists) can analyse the face -

but will they not lose the smile?

(Antoine de Saint-Exupery)

The term "substructural analysis" was first used only recently (CRAMER et al., 1976) to denote empirical methods of structure-property correlation, using structural features derived automatically from computer-readable representations of the chemical structure diagram. Since all the work described in this thesis falls within this category, the examples of these, and related, methods so far reported in the literature will be described in detail, with emphasis upon the types of structural features used.

Substructural analysis techniques may be used in conjunction with a variety of statistical procedures. Multiple regression analysis and cluster analysis were used in the work described below, and in related work (ADAMSON et al., 1973a, ADAMSON et al., 1974, ADAMSON et al., 1975a, ADAMSON et al., 1976a, BUSH, 1976), and these statistical techniques will be described in detail in later chapters. Applications of pattern recognition techniques (CHU et al., 1975) and of other non-parametric methods (CRAMER et al., 1974, EAKIN, 1975) have also been reported.

Representation of Chemical Structure

The choice of the type of chemical structure representation is of evident importance for any substructural analysis procedure, and the various possibilities will now be considered.

The simplest representation is molecular formula, a representation of composition rather than of structure. It is too crude to be of use as the sole structural representation for correlation, although descriptors derived from molecular formula, e.g. atom counts and molecular weight, have been used in conjunction with more specific structural features, particularly in applications of pattern recognition techniques. This will be discussed further below.

The chemical structure diagram is the most widely used structural representation for correlation purposes, and the term "substructural analysis" generally implies the derivation of structural features from this form of representation. Three main reasons for this may be noted. Firstly the structural diagram is known, by definition, for all compounds whose structure has been determined. Additionally it may be used to represent hypothetical compounds, for which no experimental data is available. Secondly, as discussed in an earlier chapter, the structure diagram has been, since its inception, the major conceptual tool for the rationalisation of the properties of chemical species. Thirdly the techniques devised for the computerised representation and handling of the chemical structure diagram for storage and retrieval may be readily adapted for substructural analysis.

Since the work described in this thesis involves the use of computer-readable representations of the chemical structure diagram, the adequacy of such diagrams as a structural representation is of considerable importance, and will be discussed in detail below.

In order to represent explicitly three-dimensional structure, three approaches are possible. Structure diagrams, and their computer-readable forms, will represent certain stereochemical factors, e.g. cis-trans isomerism, configuration about asymmetric carbon, and equatorial-axial conformations; and appropriate structural features may be derived. Experimental data on molecular dimensions may be utilised, though this is available only for a few thousand compounds as yet. Use may alternatively be made of known standard bond dimensions, or of some form of calculation of conformation etc., as described in an earlier chapter.

Calculated molecular wave functions are potentially the most sophisticated structural representation thus far envisaged. However, the twin problems of lack of accuracy, and considerable cost have so far prevented their widespread use. Indices calculated from the simpler quantum mechanical procedures have been used as alternatives to physicochemical property values in semi-empirical correlation techniques, as discussed in an earlier chapter.

Experimentally measured molecular properties, used in semi-empirical correlation, may also be regarded as representations of particular aspects of molecular structure. It has been suggested that physicochemical properties are intermediate in the correlation between biological properties and chemical structure, represented in an appropriate fashion (MURRAY et al., 1976). The use of substructural features and property values are in any sense mutually exclusive: it is likely that they will achieve maximum usefulness if they are regarded as complementary. The possible manner of use of substructural analysis will be discussed at a later stage.

Adequacy of the Chemical Structure Diagram Representation

In this section the adequacy of the structural diagram as a representation of chemical structure, suitable for the derivation of structural features for correlation, will be discussed. These factors will apply equally to the computer-readable forms of the structure diagram.

The chemical structure diagram serves two functions. On the one hand it acts as a means of communication of structural ideas: such diagrams, and their computerised forms, have been analysed in terms of linguistic concepts (RANKIN et al., 1971, TAUBER et al., 1972, MUNZ, 1968). On the other hand the structure diagram is a representation of a model of an underlying physical reality, and it is this aspect which will now be examined.

Three points must be considered: firstly the adequacy with which the model, the chemical structure concept, represents physical reality, secondly the adequacy with which the chemical structure diagram represents the model, and thirdly the adequacy with which the computer-readable forms represent the structure diagram. The term "model" is here used to mean the conceptual basis of the theory of chemical structure (HESSE, 1963, RUSSELL, 1971j).

There is little point in any detailed consideration of the deficiencies of the theory of chemical structure, i.e. the adequacy with which the model represents reality. Although in some cases the nature of bonding etc. is still a matter for debate, the successful rationalisation of the great amount of data associated with chemical compounds indicates the overall adequacy of the structural concept. Additionally, as discussed below, deficiencies in the model need not

necessarily negate the usefulness of its representation.

As noted in an earlier chapter, the chemical structure diagram, in essentially its present form, was devised soon after the middle of the last century, at a time when the existence of the electron had not been demonstrated, the nature of the forces within molecules was entirely unknown, and even the physical existence of atoms was still a matter for debate (RUSSELL, 1971j). That the structure diagram should still be of use at the present time, when the underlying model of chemical structure has changed so drastically, is a measure of the remarkable power of this simple representation, and the associated concept of the chemical bond.

In general terms the adequacy of the structural diagram is shown by its wide use, in both qualitative and quantitative applications. Its major potential failing appears to be its unsophisticated representation of atoms and bonds: thus a $-CH_2-$ linkage or a $-C=O$ grouping are represented as the same, regardless of their environment, an evident, though not necessarily gross, approximation. A more sophisticated form of structural representation may be envisaged, with quantities from wave mechanical calculations replacing or supplementing the conventional atom and bond representations. It is likely however that for most structure-property correlation applications such sophistication, regardless of any practical problems in deriving the values to be used, would not be justified by the accuracy of the data available.

Two particular potential problems with the use of features from the structure diagram for correlation should be noted: these are electronic molecular properties and effects due to three-dimensional structure.

As described in an earlier chapter, the molecular wave

function may be built up from either delocalised or localised molecular orbitals. Those molecular properties termed "additive, constitutive", i.e. those related to bond energies, are most readily accounted for in terms of localised orbitals, which extend over groupings of atoms and bonds closely resembling chemically significant groupings from the structure diagram (ENGLAND et al., 1975, VON NIESSEN, 1975). This gives theoretical support to the many bond- and group- contribution energy schemes in thermochemisty, and indeed to the general principles of substructural analysis for such properties. It is however generally necessary to use delocalised orbitals in order to deal with molecular electronic properties, and this indicates that substructural analysis techniques could be inappropriate for such properties. In practice however,this limitation may be circumvented, either by choosing appropriate structural features so as to avoid the necessity of fragmenting a possibly delocalised system, or by dealing only with derivatives of a single system. In this latter case only a single orbital will be affected, to a reasonable degree of approximation, by a change in substituent, and the problem could then be dealt with, at an approximate empirical level, by substructural analysis. This type of approach has led to some success in the application of the Hammett relationship to the electronic spectra of organic substances (KATRITZKY et al., 1972).

The three-dimensional structure of a molecule is a second potential problem for substructural analysis techniques, using features derived from the structure diagram. Certain stereochemical factors, e.g. cis-trans isomerism and equatorial-axial conformation, may be dealt with explicitly. Additionally,information on molecular size and shape is included implicitly in such descriptors as number of atoms and bonds, type of branching etc., and deviations from non-additivity with simple structural features may indicate steric or conformational

factors. Although some workers in this field have used explicit three-dimensional descriptors, using information other than the structure diagram, others have considered that molecular size and shape may be adequately accounted for using the structure diagram alone (MURRAY et al., 1976). These points will be further considered below.

One particular aspect of molecular structure affecting the applicability of these techniques, is the possible difficulty of correlating structure with vector properties, e.g. dipole moments. A recent study has shown a successful correlation of dipole moment with structure for benzene derivatives, using structural features representing relative positions of substitution (KELLIE et al., 1975). This indicates that such properties are amenable to substructural analysis techniques, though the derivation of appropriate structural features would pose considerable difficulty for more complex compounds.

The general conclusion appears to be that the structural diagram is an adequate representation of chemical structure, within the limits required to allow correlations to the accuracy of most of the data available, and with possible reservations for particular structural types and particular properties.

For substructural analysis techniques of the type studied in this thesis, it is necessary to consider also the adequacy of the possible computer-readable forms of the structure diagram.

Some information is inevitably lost with the use of a fragmentation code, since this is not a total structural representation. However, fragmentation codes have generally been designed for information storage and retrieval in a specific environment, and therefore usually reflect the interests of a particular user group, or the characteristics of a particular data base. It may well be that such a code will include

most of the substructures necessary for adequate structure-property correlation in its specific environment. Nonetheless, use of such a representation inevitably involves loss of flexibility in the choice of structural features: this applies whether the code is fixed or open-ended, since in either case the type of features are essentially pre-ordained.

The use of a total structural representation, either a connectivity table or a linear notation, gives a greater flexibility to structural feature derivation, since these are in general total representations of the structure diagram. The only notable general failing of these representations is an inability to represent partial, i.e. non-classical, bonding: however, this is not a great drawback for the majority of substructural analysis applications which may be envisaged. The techniques of structural feature derivation from such computer-readable representations will be discussed in detail below.

The use of systematic nomenclature for substructural analysis has not yet been reported. In view of the development of techniques for substructure searching with nomenclature, there seems no obstacle to its use for substructural analysis, at least for limited types of structure.

Structural Feature Derivation

In this section there will be given a brief outline of the various structural features used in published reports of substructural analysis and related correlation methods. These will be divided into three types: topological descriptors, geometrical (three-dimensional) descriptors, and single indices representing structure.

Topological Descriptors

The simplest such descriptors are features such as molecular weight, and numbers of particular types of atoms and bonds. These are not generally used alone, but rather in conjunction with structural fragments; functional groups, types of ring systems, substitution patterns etc., usually derived non-algorithmically. Several examples of the use of descriptors of these sorts have been reported (KOWALSKI et al., 1974, JURS et al., 1975d), sometimes in conjunction with property values for structural features or for the compound (KOSKINEN et al., 1974, SCHIFFMAN, 1975, SNEATH, 1966). The "indicator variables" used by Hansch, either alone (HANSCH et al., 1975) or in conjunction with physicochemical properties (HANSCH et al., 1974b), are structural fragments of this sort.

Such structural fragments may be applied in a more systematic fashion by use of a fragment code. Algorithmic derivation of such features has been described from a fixed fragment code (CRAMER et al., 1974) and from an open-ended code derived from a linear notation (EAKIN, 1975). A fragmentation code devised specifically for biological structure-property correlation has been described (AVIDON et al., 1974).

A variety of schemes for systematic derivation of features from the structure diagram have been devised for the estimation of thermochemical properties (BENSON et al., 1969, JANZ, 1967c, REID et al., 1966). The structural units are of varying levels of complexity,

from simple atom and bond fragments to more complex functional group fragments, and including in some cases intergroup interactions.

Several examples have been reported of automatic structural feature derivation from total structural representations. Two alternative procedures are possible: firstly the generation of pre-selected structural features by substructure search techniques, and secondly the systematic algorithmic fragmentation of the structure.

Both procedures have been applied to connection table representations of structure. One pattern recognition study used an interactive substructural search system to generate structural features including augmented atoms, heteropath fragments and ring fragments (CHU et al., 1975), following an earlier study in which augmented atom descriptors were assigned manually (CHU, 1974). A similar system has been used to produce simple atom and bond counts, substructural descriptors, and features representing the environment of particular groups (BRUGGER et al., 1976, STUPER et al., 1975). As noted in a previous chapter, substructure search procedures on connection table records have been used to generate structural features for the estimation of thermochemical properties (JOCHELSON et al., 1968).

Systematic algorithmic feature derivation from connection table representations, with the generation of a variety of atom- and bond-centred fragment types, has been reported (ADAMSON et al., 1973g, ADAMSON et al., 1976, ADAMSON et al., 1975a, ADAMSON et al., 1976a, BUSH, 1976).

Less use has been made of linear notations for this purpose. The derivation of a fragment code from WLN has been noted earlier in this section, and substructural analysis techniques for this notation have been used for thermochemical estimation procedures (BRASIE et al., 1965)

Geometrical Descriptors

Relatively few structure-property correlation studies have made use of descriptors explicitly representing three-dimensional structure. Simple manually assigned descriptors representing configuration and conformation have been used in studies described in the previous chapter (KELLIE et al., 1975, SCHIFFMAN, 1975), as have standard values for structural dimensions (DOVING, 1973, KOSKINEN et al., 1974, DIERDORF et al., 1974).

Pattern recognition studies have used structural descriptors derived from three-dimensional molecular structure, deduced from experimentally established atomic co-ordinates (SOLTJBERG et al., 1976) or from a molecular mechanics calculation procedure (BRUGGER et al., 1976). In the latter case the geometrical descriptors were used together with automatically derived topological descriptors.

A procedure for searching a file of connection table structural representations for three-dimensional "pharmacophoric patterns" has been described (GUND et al., 1973), though no application of such a procedure for quantitative correlation has been reported.

Single Index Structure Representation

A highly desirable development in the area of structure-property correlation would be the introduction of a single numerical index, derived solely from the structure diagram, capable of representing chemical structure sufficiently for correlation with a large number of properties. The only index of this sort to have achieved any extensive application is the "branching index" or "connectivity index" devised by Randic, from graph theoretical considerations (RANDIC, 1975).

This index, calculated by a simple arithmetic formula allowing for the degree of branching, appears to be a measure of molecular size and shape. It has since been applied to the correlation of a variety of physicochemical and biological properties, with considerable success (KIER et al., 1975a, HALL et al., 1975, MURRAY et al., 1975, KIER et al., 1975b, MURRAY et al., 1976). It has been suggested by these authors that direct correlation between structure, represented by the connectivity index, and biological properties, without resort to semi-empirical calculations (MURRAY et al., 1976).

An index representing molecular structure based on the concepts of information theory has been devised (RASHEVSKY, 1955, KARREMAN, 1955). This sort of index does not appear to be suitable for structure-property correlation, and no such application has been reported.

Feature Selection and Preprocessing

These topics, which have been briefly discussed in the previous chapter, are relevant to all substructural analysis and related procedures, although this terminology is not always used.

The first problem is the choice of structural features to be used, out of many possible for any given structure. With the majority of the techniques previously described this is an unsystematic and intellectual procedure, with descriptors being chosen on an intuitive judgement of their suitability. The use of structural features from fragmentation codes (EAKIN, 1975, CRAMER et al., 1974) reduces the extent of possible choice, since in general the appropriate structural features will be assigned algorithmically. A major amount of intellectual effort has of course been put into the construction of the code, i.e. the choice of fragments, at an earlier date.

Those procedures involving algorithmic generation of structural features from connection tables (ADAMSON et al., 1973a, ADAMSON et al., 1974, ADAMSON et al., 1975a, ADAMSON et al., 1976a, BUSH, 1976) involve the choice of the appropriate level of description: augmented atom, simple pair etc. Derivation of appropriate structural features representing the total structure is then automatic.

A rationale for one form of substructural analysis has been given, by a set theoretical study of molecular structures (KIHO, 1971).

A related problem is the reduction in number of structural features to give a smaller set to be used in the statistical analysis procedures. This is often regarded as desirable in order to reduce the sample: feature ratio, or to allow concentration on a relatively small number of features considered of particular importance.

Some pattern recognition studies have used feature reduction techniques based solely on statistical distribitions of features (CHU, 1974). This has been suggested to be unsuitable for the non-parametric data sets generally found in chemical applications. Alternative techniques based on the discriminating power of each feature, generally known as weight-sign feature selection, have been more widely adopted (JURS, 1970, JURS et al., 1975e).

If the data set is being dealt with by multiple regression analysis, the same effect may be obtained by carrying out the analysis so as to omit structural features insignificant at some predetermined level of confidence. This will be discussed further in the next chapter.

A variety of other preprocessing methods may be applied to the structural features before analysis. The features may be coded as binary (i.e. present/absent) or numerically, or may have some arbitrary numeric coding, to indicate one of several possibilities. Preprocessing methods such as weighting, normalising, autoscaling, and creation of new variables play an important role in virtually all pattern recognition applications (JURS et al., 1975a, KOSKINEN et al., 1975, STUPER et al., 1976). Much the same ends, with parametric techniques, may be achieved by the use of non-linear functions, i.e. including higher order terms.

It should be noted that no firm distinction can be drawn between the various topics discussed above. Thus any feature reduction procedure necessary will depend very much on the initial feature derivation. Similarly the use of non-linear functions in multiple regression analysis may be regarded as an alternative to generation of larger, and hence more complex, structural features (ADAMSON et al., 1976a).

Substructural Analysis techniques used in this study

The substructural analysis procedures used for the work described in succeeding chapters involved algorithmic generation of structural features from Wiswesser Line Notation (WLN) representation of structure. They resemble most closely, of all previous work in this field, the structure-property correlations carried out with features derived from connection tables by algorithmic means (ADAMSON et al., 1973a, ADAMSON et al., 1974, ADAMSON et al., 1975a, ADAMSON et al., 1976a). Like these studies, the work described in this thesis involves algorithmic generation of features representing the total structure, from computer-readable representations of the chemical structure diagram, by methods generally compatible with the operations of chemical structure handling information systems. The differences between these studies and this work are very largely due to the differences in the structural representations used, and this aspect will now be considered.

WLN, as noted in an earlier chapter, is widely, and increasingly, used in chemical information systems and in data compilations, handbooks etc. A demonstration of its usefulness for structure-property correlation, and related data analysis procedures, could therefore be of considerable practical importance. The only such application of this notation to date, apart from relatively simple techniques such as permuted lists and the generation of a fragment code mentioned at an earlier stage, has been the derivation of structural features for thermochemical property estimations by substructure search techniques (BRASIE et al., 1965). Although these authors suggested the use of algorithmically generated WLN fragments for direct structure-property correlation, no such application has been reported.

WLN, as a total structural representation, shares with connection tables the potential for allowing generation of any desired aspects of the structure diagram as structural features. It should be emphasised here that any such structural feature may be generated from any total structural representation. However, it is certainly true that some forms of structural feature are very much more simply generated from WLN than from connection tables, and vice versa. It may well be that it will be regarded as preferable to choose an appropriate structural representation, rather than to use highly complex structural feature derivation procedures with a less suitable representation. This is particularly so in view of the possibilities for interconversion of representations, as discussed in the previous chapter.

The coding rules of WLN include a good deal of allowance for ideas of "chemical significance", e.g. the coding of functional groups and ring systems (BAKER et al., 1975), and advantage may be taken of this in devising structural feature derivation procedures. It should be recognised that this is a possible source of bias, in that if only structural features corresponding to conventional chemical groupings are generated, possible insight may be lost. The use of connection table fragments, which do not correspond to "chemically sensible" groupings, may help to rectify this. This is one of many examples of cases which indicate that the use of different structural representations, and hence of different types of structural feature, are complementary.

The procedure adopted was to select a particular specificity of structure feature, e.g. simple functional groups, whole ring systems

and whole side-chains, interaction terms between ring substituents etc., and to allow these to be produced algorithmically from the WLN. This produced structural features of a wide range of size: for example Cl and SO_2NH_2 could both be functional groups. This is by contrast to connection table fragmentation of the Adamson and Bush type, where features of a more nearly constant size are produced.

The type of WLN fragmentation used involved no overlap of fragments (except for bond overlap). This again is in contrast to Adamson-Bush connection table fragmentation, where considerable overlap occurs, since each atom and bond in turn is taken as the centre of a structural feature (BUSH, 1976). The results of an analysis using structural features derived from WLN are therefore usually more easily interpreted, since WLN features correspond more immediately to chemically significant groupings.

There is thus very considerable control over the type of structural features produced from a WLN analysis, unlike a corresponding connection table analysis, where only the centre and extent of fragments may be easily specified. The advantages of this are obvious, although the possibility of bias mentioned above, i.e. of seeing only that which is looked for, should again be noted.

The derivation of particular substructures as structural features was not in general carried out in this work, since it was thought that, if required in a practical situation, as a complement to the algorithmic form of feature derivation, this could be most easily achieved by a standard substructural search procedure.

It was considered that there were three major areas in which WLN could be used to particular advantage in substructural analysis, in addition to the practical importance of assessing the usefulness of the widely-used representation. These were functional groupings, cyclic systems, and stereochemistry.

Although the generation of functional groups from connectivity representations has been described (ESACK et al., 1975), WLN appeared to offer a particularly ready approach to this sort of structural feature.

A WLN representation appeared likely to allow a convenient derivation of structural features representing whole ring systems, by contrast to Adamson-Bush connection table fragmentation in which ring systems were fragmented into subunits. Further, the WLN ring locant rules seemed to make possible the ready derivation of features accounting for substituent patterns, and for interactions between substituents, heteroatoms etc. Although a good deal of work has been carried out on the analysis of ring systems using computerised representations (ADAMSON et al., 1973b, ADAMSON et al., 1973c), no such systematic substructural analysis procedure has been reported.

The rules of WLN include sufficient stereochemical information for it to seem likely that structural features explicitly accounting for stereochemical factors could be relatively simply derived.

It should be noted here that one major objective was to develop procedures which could, if proved to be useful, have wide applicability. To this end a widely-used structural representation, WLN, and a widely-used programming language, COBOL, were adopted. Additionally the substructural analysis procedures were designed to be as straightforward as possible, in some areas at the expense of neatness and economy of programming.

The statistical analysis procedures adopted were multiple regression analysis, for quantitative structure-property correlation, and cluster analysis, potentially useful for both correlation and more general data analysis and information retrieval. These were used for three reasons. Firstly the performance of WLN structural features with these methods was thought likely to give a good indication of their overall performance. Secondly a direct comparison was possible with the work of Adamson and Bush, where these methods were applied with connection table fragments. Thirdly these are widely-used techniques, which in the case of this work involved standard commercially-available program packages, which should tend to aid the general applicability of the whole procedure.

As a general principle, preprocessing of the feature sets before analysis was minimised, and the various statistical procedures were carried out as straightforwardly as possible. This was firstly to maintain the overall simplicity of the process, and secondly because it was felt that alteration of the feature generation algorithm was the preferable way of achieving the best results. For example it has been shown in one case (ADAMSON et al., 1976a) that use of larger structural features is slightly more effective than application of a non-linear regression function. Since WLN is well suited to the derivation of large structural features, this was in general the preferred method, and higher-order regression analyses were only rarely used.

Statistical Techniques

Multiple regression analysis procedure

Multiple regression analysis is a parametric technique for statistical analysis of multivariate data. In recent years it has been widely applied in a variety of fields of study, largely due to the increased availability of standard computer programs.

The technique will not be discussed in detail here. Very full descriptions of such methods (SNEDECOR et al., 1967a) and of their application in substructural analysis (BUSH, 1976) are available. Accounts of these statistical procedures as used in semi-empirical structure-property correlation have been given (SHORTER, 1973, VERLOOP, 1972).

The assumptions underlying multiple regression analysis are those applying, in general, to all parametric procedures (SNEDECOR, 1967b, SIEGAL, 1956). It is presumed that the observations are independent, and are drawn randomly from normally distributed populations having the same variance, or a known ratio of variances. The multiple regression model is only fully valid if the observations are measured on an interval or ratio scale. It is generally assumed that these conditions are adequately fulfilled for quantitative property measurements, although the underlying distributions will not in fact usually be known.

The simplest form of regression analysis, linear regression, seeks to calculate the best linear relationship between two variables. This is achieved using the method of least squares, in which the sum of the squares of the deviations between observed and estimated values of the dependent variable are minimised. The linear relationship fulfilling this condition is taken as the best such relationship.

This relatively simple procedure is the basis of multiple

regression analysis, in which a number of explanatory variables are tested for their effect in accounting for the variance of the dependent variable. Deriving the "best" linear relationship for a relatively large number of explanatory variables is a highly complex process, and correspondingly costly in computing time. The majority of computer programs for multiple regression analysis therefore use a "stepwise" procedure. Variables are introduced one at a time in decreasing order of their effect on the variance of the dependent variables. At each stage variables whose effect on the overall regression falls below some pre-determined level of significance are excluded. This process continues until no significant improvement in correlation can be brought about.

It should be noted that the results produced by such stepwise techniques will not necessarily show the best possible correlation. An analysis involving simultaneous consideration of the effect of all explanatory variables, impracticable in most cases because of the computing power required, would be necessary to guarantee an optimal solution.

The standard program package used for the work described below makes use of such a stepwise multiple regression procedure (I.C.L. 1971).

In the work described below, the property value for each structure was the dependent variable, while the structural features present in any of a set of structures were the explanatory variables. The numerical values for the occurrence of structural features were used directly, without normalisation, weighting or any other preprocessing. In most cases the logarithmic (base 10) value of the property was used, as is conventional in both the semi-empirical and additive modelling techniques discussed above.

For the most part, a simple linear function was adopted. The property value was assumed to be an additive function of the structural

units present, so that its value, y, for the ith compound is given by:

$$yi = \sum_{j=1}^{n} bjxij + constant$$

where there are a total of n types of structural feature in the set of structures, and xij is the number of times that the jth feature occurs in the ith structure; bj is the regression coefficient for the jth feature and represents the effect of that feature in increasing (positive coefficient) or decreasing (negative coefficient) the measured property values for those compounds in which it occurs.

Higher order functions were in some cases also considered. A function including squared terms takes the form:

$$yi = \sum_{j=1}^{n} bjxij + \sum_{j=1}^{n} cj(xij)^2 + constant$$

where cj is the regression coefficient for the squared term for the jth structural feature, and other terms are as defined above.

A full quadratic function takes the form:

$$y = \sum_{j=1}^{n} bj\ xij + \sum_{j=1}^{n} \sum_{k=k+1}^{n} djk(xij.xik) + \sum_{j=1}^{n} cj(xij)^2 + constant$$

where xik is the number of times that the kth feature occurs in the ith structure, where $j<k<n$, djk is the coefficient for the cross-product term for the jth and kth structural features, and other terms are as defined above.

The use of higher order functions of this sort in substructural analysis has been shown (ADAMSON, 1976a, BUSH, 1976). However, as discussed in the preceding chapter, larger structural features were generally used in preference to more complex functions.

The use of multiple regression analysis enables a number of parametric tests of significance to be carried out on the results (SNEDECOR, 1967a, BUSH, 1976).

The basic statistical quantity denoting the goodness of fit, i.e. the success of the correlation, is the multiple correlation coefficient, R. R^2 gives the fraction of the variance in the dependent variable which is accounted for by the regression. This coefficient, and similar quantities, have been used in many cases as the sole criteria of the success of structure-property correlations (SHORTER, 1973). This is not however a satisfactory state of affairs. These quantities take no account of the degrees of freedom of the regression, and in most cases an increase in the number of parameters in the analysis will of itself increase the correlation coefficient. Such coefficients are in general insensitive measures of goodness of fit, but may be greatly affected by the inclusion or exclusion of a regression constant (BUSH, 1976).

The residual error, r, of a regression, defined as the residual error sum of squares divided by the number of degrees of freedom, gives a more realistic assessment of the success of an analysis. The significance of the difference between any two regressions may be assessed by the F-test. An F-value is calculated as the ratio of the squares of the residual errors of the regressions, and its significance found from tables of F-values with the same numbers of degrees of freedom as the regressions.

The F-value for a single regression is given by:

$$F = \frac{R^2.n}{(1-R^2)m}$$

where R is the multiple correlation coefficient, n is the number of degrees of freedom, and m is the number of explanatory variables in the analysis. The statistical significance of the correlation may then be assessed by comparison with tables of F values with n and m degrees of freedom.

The multiple regression programprovides t statistics for each regression coefficient, so that the significance of individual coefficient values may be assessed by comparison with tables of t statistics with the same number of degrees of freedom as the analysis.

The statistical significance of the difference between pairs of coefficient values may be assessed by calculation of a t value:

$$t(b_1, b_2) = \frac{\sqrt{S_1^2 + S_2^2 - 2r^2C_{12}}}{r}$$

where S_1 and S_2 are the standard errors of the coefficients, r is the residual error of the regression, and C_{12} is the corresponding term from the inverse cross-product matrix. The value of $t(b_1 b_2)$ is then compared with values in tables of t statistics with the same number of degrees of freedom as the regression.

The ability to make such statistical tests on the results of the analysis is a valuable feature of this procedure. It acts as a caution against drawing firm conclusions from particular values, or differences between pairs of values, of low statistical significance. It may be of course that useful information may be gained from statistically insignificant results, particularly if trends are observed among individually insignificant values.

It may be noted at this point that statistical significance is not an additive quantity. Thus, to take a simple example, it may be that the inclusion of a structural feature a, will bring about no significant improvement in correlation, and the same will be found for another feature a_2. However, inclusion of both a and a_2 may give a significant improvement in correlation. Effects of this sort, although sometimes appearing anomalous, are to be expected.

In the majority of cases the regression analyses were carried out so as to include as many structural features as possible, within the accuracy of the calculation. This is termed analysis at the 99.99% level. The cut-off point of the calculation is governed by the setting of the pivot size (I.C.L. 1971). This was not varied during the work described below, but was setat the value found to be the best compromise to allow reasonable accuracy with stability of the analysis (BUSH, 1976).

Features were omitted by the regression program in analysis at the 99.99% level for any of three reasons. Firstly, because they had no effect, within the accuracy of the calculation, on the property value under consideration. Secondly, because they occurred to the same extent in all structures of the set. Thirdly, because they were perfectly correlated, i.e. a number of types of structural feature occurred only in a fixed ratio in the same structures. Only one of any group of perfectly correlated features could be included in the regression, and its coefficient value reflected the effect of the group of features as a whole.

On occasions the analyses were carried out so as to omit from the regression any structural features whose effect on the property value was insignificant at some pre-determined level of confidence. This is termed analysis at the n% level, where features whose coefficient values are insignificant at the n% level are excluded. This may be useful in concentrating attention on the more important factors affecting the variance of the property values.

A number of authors have commented on the need for a sufficiently large ratio of observations to variables in multiple regression analysis, the suggestions ranging from an acceptable minimum of 5 to a statistically desirable 20 (TOPLISS et al., 1972, CRAIG, 1975, UNGER et al., 1973).

Attention has been drawn to the possibility of chance correlations with large numbers of explanatory variables (TOPLISS et al., 1972): these authors' suggestions for guarding against this eventuality are essentially equivalent to the use of the F values for comparing the significance of regression results.

It has been suggested that "the statistical methodology must be used as a guide, but need not be rigidly adhered to. Thus, there may frequently be fewer compounds than is desirable, but one may still try to obtain correlations. Common sense must always be applied to the consideration of correlations" (CRAIG, 1975). This attitude may well be appropriate in many practical situations, and these principles were borne in mind in carrying out the analyses required below.

In addition to examining the overall regression results and the individual coefficient values, the estimated values for each observation, i.e. the residuals of the regression, were studied. These indicated if any type of structure were particularly badly estimated, and also allowed examination of the effects of a different set of structural features or a different type of analysis on the estimations for particular structures.

The extent of agreement of observed and estimated values is however not a good indication of the likely success of property prediction, where the compound under consideration is not included in the analysis since its property value is unknown. In order to simulate this situation the "hold-one-out" method was used. This involves exclusion of one structure from the set to be analysed, so that its property value may be "predicted" by summation of the regression coefficients from the analysis of the remainder of the set of structures for those structural features in the excluded structure (REDL et al., 1974,

ADAMSON et al., 1976a, BUSH, 1976). Each structure in a set may be excluded in turn, so as to gain an overall impression of the predictive power of the method.

Cluster Analysis Techniques

Cluster analysis, as described here, is a general term for techniques for the analysis of multivariate data, which aim at grouping objects, on the basis of variables describing the objects, into classes, where the number and characteristics of the classes is not known *a priori*. These techniques fall within the area termed "unsupervised pattern recognition". A variety of other terminology is applied to these and related techniques, e.g. "numerical taxonomy", "clumping", and "automatic classification", perhaps due to their use in many different areas of study (EVERITT, 1974). These techniques have been widely used, most notably in the biological and social sciences.

A number of full accounts and critical reviews of this subject are available (CORMACK, 1971, EVERITT, 1974, SNEATH *et al.*, 1973, JARDINE *et al.*, 1971a) and those aspects of particular relevance to the classification of chemical structures have been discussed in detail (BUSH, 1976). No detailed account will therefore be attempted here. Rather, a brief survey of the major points will be attempted, based largely upon the treatment by Everitt referred to above.

The clustering process, as noted above, consists of the grouping of entities on the basis of a number of attributes. There are a number of distinguishable types of clustering technique: the methods used here fall into the category termed *hierarchical, agglomerative*. The set of entities (structures) are fused into groupings at varying levels of inter-entity similarity. The results of these procedures are generally expressed as a dendrogram or "classification tree".

The first step in these methods is the calculation of some measure of similarity (or dissimilarity) between each pair of entities on the basis of their attributes (structural features). From this

"similarity matrix" groups of the most similar entitites are formed, by some clustering method. This process is repeated at decreasing levels of similarity to form the dendrogram.

It is evident that there is scope for investigation of three factors in the classification of any group of structures: firstly, choice of attributes, secondly, choice of similarity measure, thirdly, choice of clustering method. These will now be considered individually.

i) Choice of attributes

N.B. The terms "attribute" and "variable" are here used synonymously.

Since in this work the attributes are algorithmically generated structural features, the choice of attributes amounts to the choice of appropriate WLN structural feature set. The effect on the classification of the use of different feature sets was investigated empirically, as will be described below.

It is known that standardisation of input data may considerably affect the classification (SNEATH et al., 1973) and standardisation was carried out in the examples below to investigate this effect. Standardisation involved the transformation

$$Zik = Xik/\sigma k$$

where Xik is the value for the kth variable of the ith structure, Zik is its standardised value, and σk is the standard deviation of the kth variable.

Weighting of variables, in order to bias the classification towards the effect of particular attributes, is a common pre-processing step, though it has been criticised on theoretical grounds (SNEATH, 1957). Weighting can only be potentially useful if there is some purpose in mind for the classification, and its effect was therefore not investigated in the work described below.

ii) Choice of similarity measure

A wide variety of similarity/dissimilarity measures have been used in cluster analysis (SNEATH et al., 1973). A detailed study has been made of the performance of a number of such coefficients in the classification of chemical structure (ADAMSON et al., 1975a, BUSH, 1976). This study indicated that there is little advantage in the use of the more complex coefficients. It was not thought worthwhile to further investigate this point, and it was decided to make use of one simple measure of similarity in the work described below. The measure chosen was the Euclidean distance squared coefficient. This coefficient, dij, for two entities. i and j is given by

$$dij = \sum_{k=1}^{p} (Xik - Xjk)^2$$

where Xik and Xjk are the values of the kth variable for the entities i and j respectively, and there are a total of p variables.

This distance measure has the advantages of relative conceptual simplicity and computational economy, and may be used with any of the hierarchical clustering procedures to be considered. It is a widely used coefficient, of generally high reliability.

iii) Choice of clustering procedure

A considerable number of procedures for hierarchical agglomerative clustering based on a distance coefficient are available, and a number of these were investigated. In general terms, all of the methods calculate the distance, measured in terms of the Euclidean distance matrix, between clusters, consisting of one or more entities, and fuses those clusters with the smallest inter-cluster distances. This process continues until all the entities are fused into a single cluster.

The difference between the methods lies in their different definitions of inter-cluster distances.

In single link or nearest neighbour clustering the distance between groups is defined as the distance between their closest members (SNEATH, 1957). In addition to its wide use in biological taxonomy (SNEATH et al., 1973), this type of classification has been applied to chemical structures (ADAMSON et al., 1973a, ADAMSON et al., 1975a, BUSH, 1976, SNEATH, 1966), and to problems of document retrieval (VAN REIJSBERGEN, 1976). It has been suggested that this is the only clustering technique, of the kind considered here, to satisfy certain theoretical criteria (JARDINE et al., 1971a). Other workers have argued this is too restrictive a viewpoint, and have suggested that single link clustering is rarely able to reveal a clear data structure (WILLIAMS et al., 1971).

Furthest neighbour or complete link clustering defines the distance between clusters as the distance between the most remote pair of members, while in group average clustering the inter-cluster distance is defined as the average of the distance between all pairs of entities in the two groups (SOKAL et al., 1958). The method of McQuitty is essentially the same as group average (McQUITTY, 1966).

In centroid clustering, and the very similar median clustering, the entities are depicted in Euclidean space and the co-ordinates of the centroid of each group calculated. Inter-cluster distances are then taken to be the distances between the centroids of the clusters (GOWER, 1967, SOKAL et al., 1958).

Ward's method of clustering is derived from the assumption that the loss of information which inevitably occurs when entities are grouped into clusters may be measured by the total sum of squared deviations

of every point from the mean of the cluster to which it belongs. This information loss is minimised at each step of the analysis, by fusing those two clusters whose fusion results in the minimum increase in the sum of squares deviations (WARD, 1963).

A number of empirical comparisons of the performance of various clustering procedures have been reported (PRITCHARD et al., 1971, HODSON et al., 1966, EVERITT, 1974, JARDINE et al., 1971b).

These have been subject to the problems implicit in the judgement of classifications mentioned above, and have involved generally intuitive assessments. No clear indication of the superiority of any particular method emerges from these reports. Rather, the usefulness of the methods appears to depend upon the data set being analysed, and the purpose of the classification.

Chapter Four

Analyses

'Computerised means of frustration, and electronically encoded ineptitude are no better than the manual variety'

(Hans Wellisch)

Chapter Four, Part 1

Multiple Regression Analyses

Serum Binding Activity of Penicillins

The work described in this section was aimed at investgating the possible use of structural features derived from Wiswesser Line Notation for structure-property correlation. The data set consisted of values for serum binding activities for a series of 79 penicillin structures with diverse side-chains. The logarithm of the ratio of bound: free compound was used as the property value. The structures and properties are listed in Table 1.

A structure-activity study of this data set using Hansch analysis has been reported (BIRD et al., 1967). This study has been reassessed, using similar semiempirical methods, by other workers (TUTE, 1971, NYS et al., 1974). Substructural analysis techniques, using connection table features, have also been applied to this data set (ADAMSON et al., 1974, BUSH, 1976). The possibility of comparison with the results of these studies made this series particularly suitable for the initial trial of WLN features.

The structural features used were derived manually, so that their usefulness could be assessed before any effort was put into programming. Structural features were devised with a view to a relatively simple algorithmic generation from WLN representation.

Two sets of structural features were originally used in the analysis.

Set A, a relatively simple set, included as structural features whole ring systems, hydrocarbon fragments, halogen atoms and other simple functional groups.

Set B, a more complex set, distinguished between halogen, $-NH_2$, -OH, and -O- features according to whether they occurred in an aliphatic chain, or substituted on an aromatic ring. Substitution patterns of ring systems were also distinguished: this distinction

has been reported using connection tables (ADAMSON et al., 1973c). but would be particularly convenient using WLN locants.

Examples of structural feature derivation using these sets are shown in Figures 10 and 11. The structural features from these sets are listed in Tables 2 and 3 (together with the results of the regression analyses to be discussed below). The corresponding WLN symbols are included to illustrate the ready derivation of such features from WLN. It should be noted that these WLNs are illustrative and do not enumerate possible permutations, alternative ring locant sets etc.

Two other structural feature sets were tested, to assess the effect of a more complex treatment of particular types of structural features.

Set C distinguished naphthalene rings according to substitution pattern and environment. The structural features in this set are listed in Table 4.

Set D distinguished the substitution patterns of halogen substituted benzene rings, and also made a more complex distinction than set B between the environments of halogen atoms. The structural features of this set are listed in Table 5.

The overall results of the regression analyses with these sets of structural features are given in Table 6. All the analyses show reasonably good correlations, statistically significant at the 1% level. The F-test shows that the correlation with set B structural features is statistically superior to that with set A features at the 5% level. This indicates that the more complex structural features are accounting better for the variation in property value, and that the improvement in correlation is probably not simply a result of the introduction of a larger number of parameters.

Sets C and D show correlations very little different, as assessed by the residual error, from set A. This indicates that the use of more detailed structural features to represent particular structural types has, in this case, little effect on the overall correlation.

The individual coefficient values for the analyses with structural feature sets A and B, together with their t statistics, are listed in Tables 2 and 3. In general these results indicate that lipophilic groups, i.e. hydrocarbons, aromatic rings, and halogens, tend to increase serum binding activity, while hydrophilic groups, e.g. -OH, $-NH_2$, $-SO_2NH_2$, tend to decrease activity. This is in agreement with the results of a study of this data set using semi-empirical methods (BIRD et al., 1967), and also with those of a substructural analysis using connection table fragments (ADAMSON et al., 1976).

There is no statistically significant difference between the overall regression results of the analyses using augmented atom connection table fragments and structural feature set B described here.

It should be noted that, although a large proportion of the coefficient values are individually statistically significant, there is, in the great majority of cases, no significant difference between pairs of coefficient values. Only in cases of pairs of widely differing coefficient values, e.g. $-SO_2NH_2$ and 2-naphthyl in feature set B, was the difference significant at the 10% level. This suggests that it may be unwise to draw firm conclusions from particular pairs of coefficient values from these analyses.

Several interesting points may be noted in these results.

For all those groups for which comparisons are possible, a more positive contribution to activity is observed for substitution on an aromatic ring compared with the same group in an alkyl chain.

Although the differences between pairs of values are not statistically significant, a trend such as this may be regarded with more confidence. This effect has been noted in studies of group contributions to partition coefficients (NYS et al., 1974), and is in agreement with the generally accepted close connection between serum binding activity and lipophilicity.

The analyses with both set A and set B features show that the effect in increasing serum binding activity of hydrocarbon fragments increases in the order $-CH_3 < -CH_2- < -\overset{|}{C}H- < -\overset{|}{\underset{|}{C}}-$, i.e. increasing with increasing degree of branching. This is the opposite of what might be expected from the generally accepted effects of chain branching on lipophilicity (HANSCH 1971a, NYS et al., 1973, NYS et al., 1974), though the lack of statistical significance in the differences should be borne in mind. It is possible that this may be due to the lack of distinction between these fragments in hydrocarbon chains and attached to hydrophilic groups: this factor has been found to obscure assessment of the lipophilicity of hydrocarbon groups (NYS et al., 1974). Such a distinction was made in choosing π values in the original Hansch analysis of these compounds (BIRD et al., 1967). This idea is supported by examination of the coefficients of the augmented atom fragments from the analysis of this data set using connection tables (ADAMSON et al., 1974). These structural features include several atoms, and for purely hydrocarbon fragments branching is found to consistently decrease serum binding activity, in accordance with expected trends in lipophilicity. This suggests that larger structural features derived from WLN could perhaps be used with advantage in this case.

Considering the coefficient values for the various benzene substitution patterns, the 1,2,3 substituted ring has a value

considerably, though not significantly, lower than the remainder. It has been suggested that steric factors will force some rings of this kind from planatary with the amide linkage, preventing any mes meric interaction (NYS et al., 1974), and this may be a contributing factor to the discrepancy.

The estimated log (b/f) values from the residuals of the regression analysis with set B structural features are listed in Table 1. No particular type of structure appears to be particularly badly predicted. Only structure 14, with an unsubstituted benzene ring directly attached to the amide linkage, has a discrepancy between observed and estimated log (b/f) values greater than 0,5, perhaps indicating some aromatic interaction (NYS et al., 1974).

As noted above the distinction between the different modes of attachment of the naphthalene ring to the penicillin nucleus, feature set C, made a negligible difference to the overall correlation achieved. To assess what improvement is brought about for those few compounds affected, the estimated values for the seven structures containing naphthalene rings from the analyses with structural feature sets A and C were compared. The discrepancy for all except one structure are much reduced by use of structural features of set C. Structures 65, 69 and 70 contain a unique set C feature, and their discrepancy is therefore essentially zero.

It seems that use of complex structural features applicable only to a relatively small number of compounds is likely to produce this form of results for many data-sets; a negligible change in overall correlation, with greatly improved estimations, perhaps tending towards triviality, for those compounds directly affected.

It may therefore not be a generally useful approach.

All the regression analyses described above were carried out at the 99.99% level, i.e. so as to include as many structural features as possible. In order to assess the usefulness of analyses including fewer variables, the analysis with set B features was carried out excluding variables not significant at the 10%. The overall regression result is included in Table 6. 9 structural features included at the 99.99% level analysis are excluded at the 10% level, and to this extent consideration of these results enables a concentration on the more important aspects. Those structural features included in the 10% level analysis are listed in Table 7, together with regression coefficients and t statistics.

It appears that these results do not greatly aid interpretation of the data set. The coefficient values are in no case greatly altered from those from the 99.99% level analysis. Those structural features excluded as insignificant in the 10% level analysis were nonetheless useful in confirming tends among coefficients in the results of the 99.99% level analysis.

It may be that analysis excluding individually insignificant variables will be of more use with larger sets of structural features, where interpretation is confused by a reletively large number of insignificant features. This type of analysis may therefore be less valuable for feature sets derived from WLN than those derived from connection tables.

The work described in this section demonstrated, for the first time, the usefulness of structural features derived from Wiswesser Line Notation for structure-property correlation. It showed that this form of substructural analysis can give good correlations, and results readily interpretable in conventional chemical terms. In this case the analysis brought out points of considerable interest in the data, and the results were largely in accordance with the generally accepted nature of serum binding. The results were for the most part in line with those obtained using connection table fragments, as had been anticipated, since both types of analysis use structural diagram features. If the results had not been similar, it would have cast doubt on the reliability of these forms of substructural analysis.

These results suggested that two general types of WLN structural features might be of use. The first type, similar to the set A used here, would be a simple, basic set, applicable to a wide range of structural types. Structural features of this sort would correspond to chemically significant units, ring systems, functional groups etc. The size and complexity of such structural features would be of obvious importance. These results suggested that the hydrocarbon fragments used may have been too small, while the more complex naphthalene features may have been too large.

The second, and more complex, type of structural feature would be of the sort in set B. They might include ring substitution patterns, and a generally more specific description of fragment environment. Such feature sets would probably be to some extent dependent on the type of structures under consideration.

These findings were encouraging, in that they indicated the likely value of the sort of structural features which could be readily derived algorithmically from WLN; in particular functional groups and ring systems with their substitution patterns. It was decided not to investigate this data set further, since it had been analysed to a reasonably satisfactory extent, but rather to apply these findings to other data sets with a view to devising algorithmic feature generation procedures.

The work described in this section has been described in a paper published in the Journal of Chemical Information and Computer Sciences.

Partition Coefficients of Diverse Compounds

Following on the work described in the previous section, it was thought worthwhile to further investigate structural feature derivation from Wiswesser Line Notation representations. Initially relatively simple and straightforward structural features were to be considered, with a view to algorithmic generation.

Two papers had been published, describing a form of substructural analysis closely related to the methods considered here, allowing the correlation of partition coefficient with structure for diverse structural types (NYS et al., 1973) and for aromatic and heterocyclic compounds (NYS et al., 1974). The studies used manually derived structural fragments, whose contribution to partition coefficient was calculated by multiple regression analysis, giving "hydrophobic fragmental constants" or "f values". The structural fragments which might be derived from WLN, e.g. CH_3-, -CH_2-, -NH-, -O-, -COOH. Both alicyclic and aromatic rings were broken down into smaller fragments, much as is done in substructural analysis of connection table representations. Distinctions were made between the effect of a given substructure substituted on an aromatic ring and in a non-aromatic environment, and proximity effects were included for hydrophilic groups.

This approach has been suggested to have considerable advantages over use of the π constants, since problems of definition of parent structures and corrections for branching and chain-folding are avoided (NYS et al., 1974, LEO et al., 1975).

It was decided to re-examine the data for the set of diverse structures (NYS et al., 1973), with the aim of arriving at a useful algorithmic technique based on Wiswesser Line Notation. This investigation had the advantages of allowing comparison with the original study, and of testing the applicability of the methods to a property of considerable practical importance.

In the analyses in the original paper several property values were included for each of a considerable number of the structures considered. For the work described here these values were averaged, so as to give a single property value for each of the 84 structures. The data set used in these analyses are shown in Table 8.

Structural features were initially derived manually from the structure diagrams. Three structural feature sets were used, differing in the way in which alicyclic ring systems were fragmented. In feature set A these rings were fragmented into smaller units, which were not distinguished from similar fragments in aliphatic structures or side-chains on aromatic rings. Feature set B fragmented the rings similarly, but distinguished ring fragments from others. Set C treated whole ring systems as single structural features.

All three feature sets involved fragmentation of aliphatic structures and side-chains into smaller units, and treated the benzene ring as a single structural feature.

Examples of structural feature derivation are given in Figure 12.

The overall results of the multiple regression analyses using the three structural feature sets, with log P as dependent variable, are shown in Table 9. All give correlations statistically significant at the 1% level. There is essentially no difference in correlation, assessed by the residual errors, in the analyses using the three different sets. This insensitivity to the nature of the structural features used to represent the alicyclic rings suggests that any differences in the effect on partition coefficient of particular fragments, according to their occurrence in rings or chains, are not

sufficient to be shown by data of this accuracy.

The structural features in each set are listed, together with their regression coefficients and t statistics in Tables 10, 11 and 12.

The coefficient values are in general consistent for similar structural features in each of the analyses. The values for set A structural features, equivalent to the structural fragments of Nys and Rekker, are generally in line with the f values (NYS et al., 1973). The discrepancies are presumably due to the different method of handling the raw data, i.e. inclusion of several values for one compound as against averaging before inclusion in the analysis. The use of different computer programs for the analysis may also have affected the results. The only notable difference concerns the nitrogen atom fragments. Nys and Rekker's original analysis gave a lower f value for a tertiary N atom than for the secondary -NH- fragment, an unexpected result, which was reversed on inclusion of a proximity effect variable. However, in the analyses performed here, the -N- fragment coefficients were found to show the expected trend, indicating that the data handling method may affect the results to some extent.

The coefficient values from the three analyses are all much as might be expected from the generally-accepted principles of lipophilic/hydrophilic forces (FUJITA et al., 1964, HANSCH, 1971a). Virtually all the coefficient values in the three analyses were of high statistical significance. However, when the statistical significance of the difference between pairs of coefficient values was assessed, by calculation of a t statistic, in no case was the difference significant at the 10% level or higher. This suggests that it may well be unwise to draw firm conclusions from comparison of particular coefficient values in this data set.

In the analysis with set B structural features, the values for hydrocarbon fragments within acyclic rings appear to be anomalous. Thus the value for secondary $-CH_2-$ in a ring is considerably lower than for the corresponding chain fragment, while for the tertiary $-\overset{|}{C}H-$ fragment with one non-ring bond the coefficient value shows an extremely high lipophilicity, approaching that of the benzene ring. Perfect correlation unfortunately prevents comparison with the $-\overset{|}{C}H-$ fragment with three ring bonds. Although all the coefficients concerned are of high statistical significance, the cautionary remarks above regarding the lack of significance of the difference between coefficients apply here. Also, the lack of overall improvement in correlation on distinguishing ring and chain fragments may indicate that any such effect, even if real, is not of great importance in the set as a whole. It is nonetheless possible that some difference in lipophilicity is to be found between chain and alicyclic ring fragments, perhaps associated with ring strain.

In general, the effect of branching on the lipophilicity of hydrocarbon fragments shown in these results is in accordance with that found by Nys and Rekker, and with the generally accepted variation. This is in contrast to the anomalous effects noted in the previous section.

Examination of the residuals of the regressions show that very few compounds are badly estimated, i.e. with errors approaching half of one log P unit. In all three regressions compounds 9, 76 and 77 are badly estimated, and in the analyses with sets A and B compounds 78 and 79 also fall into this category. These are all structures with two oxygen-containing groups, i.e. -COOH, -OH, or -O-. This may indicate the existence of some interaction between such groups. This type of interaction was noted by Hansch's group (IWASA *et al.*, 1965), and also by Nys and Rekker, and for this reason the proximity effect factor was introduced (NYS *et al.*, 1973). The estimated values from the analysis with set C features are included in Table 8.

The regression analyses discussed above were carried out at the 99.99% level. There seemed little point in attempting an analysis at any other significance level, since virtually all the coefficient values are of high statistical significance, and hence it would be expected that few would be omitted in such an analysis. Little would be gained in any event, since the feature sets are already sufficiently small to allow ready interpretation of the results.

In view of the good correlations and sensible interpretation of results in chemical terms achieved with these feature sets, and the consequent hope of the general usefulness of such sets, it was decided to write a computer program to allow automatic generation of structural features. Set C, with ring systems treated as whole units, was regarded as most suitable for straightforward feature derivation for WLN. Such an algorithm could also be the basis for a consideration of substituent position on rings. Structural features of the types in sets A and B, i.e. with rings fragmented, could more

conveniently be generated from connection tables, although production of such structural units from WLN is possible (OSINGA et al., 1974).

A computer program to derive structural features including ring systems, hydrocarbon fragments, nitrogen fragments, and simple functional groups was written, and tested on this data set. It gave a fragmentation indentical to the manual feature derivation of set C. An extended version of this program is fully described in the Appendix. It may be noted here that this program is intended to produce relatively simple structural features, as in set C here, and set A for the penicillin structures in the preceding section. The program is intended to deal with a wide range of structural types. It takes as input straightforward WLN's using the standard coding rules (SMITH et al., 1975) without multipliers or contractions.

Rates of Bromination for Benzene Derivatives

Cyclic structures comprise a large majority of known chemical substances (ADAMSON et al., 1973b). It is therefore of importance, if substructural analysis techniques are to be of general applicability, that they should be able to deal with cyclic structures, including heterocyclics and complex multiring systems. Such tecnhiques could be of particular value, if they could be used to investigate the effect on molecular properties of such factors as relative positions of heteroatoms, substituents, and ring fusion points.

It seemed probable that structural features suitable for such application could be very readily derived from WLN, and that this was one of the areas where the use of substructural analysis with this notation could be most valuable.

It was decided to initially investigate the use of WLN structural features with benzene derivatives, before going on to more complex situations. A detailed Hammett-type study of non-additivity of substituent effects on reaction kinetic data of benzene derivatives had been reported (DUBOIS et al., 1972). This was a particularly suitable data-set for the initial WLN investigation. It included a larger number of compounds, 44, and a wider range of property value, over 15 log. units, than is usual in such Hammett equation work. A large proportion of the structures, 15 out of 44, had three or more substituents on the ring. The property examined, electrophilic bromination on the aromatic ring, is well-known to be highly sensitive to substituent effects (STOCK et al., 1963, DE LA MARE et al., 1959). This enabled a clear assessment of the usefulness of structural features derived from WLN in accounting for effects due to positions

of substituents relative to the reaction-site, and relative to one another.

A good performance by the substructural analysis method here would be notable, since multisubstitution of this sort accounts for some of the poorer performances of the Hammett equation (EXNER, 1972c) and of some thermochemical property estimation schemes (HINE, 1975c).

The data-set used is listed in Table 13. The property value was taken to be the logarithm to base 10 of the reaction rate constant, log k.

A computer program was written to derive structural features from benzene derivatives: this is described in detail in the Appendix . The benzene ring was treated as a whole unit, common to all structures. All substituent groups were treated as whole units, and structural features representing relative position were derived by consideration of WLN locant pairs. The shortest path between ring locants was always considered, so that the conventional ortho, meta, and para representations of relative positions could be used.

Hydrogen atoms, in accordance with the convention in the majority of structure-property correlation techniques, were not considered explicitly as substituents in these analyses, nor in the other correlations of aromatic structures described in this thesis. Although hydrogen atoms are known to make a definite contribution to physicochemical properties (NYS et al., 1974, LEO et al., 1975), in cases where absence of any other substituent implies the presence of hydrogen its inclusion is not likely to be useful. A substructural

analysis study, closely related to work to be described in a later section of this thesis, has confirmed that under these circumstances inclusion of hydrogen does not affect the overall correlation, and adds little to the interpretation of the results (UFTON, 1976).

Each interaction was accounted for by a single term: e.g. a compound with Br meta to Me would have the term Br-m-Me assigned, rather than the two terms Br-m-Me and also Me-m-Br. Use of multiple terms would have unnecessarily increased the number of structural features, without affecting the correlation achieved.

This program was capable of producing structural features at varying levels of complexity (sets B, C and D below). For comparison the more generally applicable program discussed in a previous section was used to generate structural features which did not maintain substituents as whole units (set A below).

Four sets of structural features were derived and used in multiple regression analyses.

Set A: log k was assumed to be affected only by the type and number of small structural units present.

Set B: log k was assumed to be affected only by the type and number of substituents.

Set C: log k was assumed to be affected by the type and number of substituents, and by the positions of the substituents relative to the reaction site.

Set D: log k was assumed to be affected by type, number, and position relative to reaction site of all substituents, and also by the interaction between each pair of substituents.

Examples of structural feature derivation are given in Figure 13.

These sets of structural features were chosen as being the

most suitable for investigation of this data-set. Certain other sets were later applied, with a somewhat different rationale, as will be discussed below.

The results of the multiple regression analyses for the 4 sets of structural features, carried out at the 99.99% level, are summarised in Table 14. All give good correlations, significant at the 1% level.

Sets A and B differ in that set A substituents such as -OMe and $-NMe_2$ are fragmented into smaller units, while in set B such substituents are treated as whole features. For this particular data-set each feature in set B which is fragmented in set A corresponds uniquely to one feature in set A: i.e. $-CH_2CH_3$ gives $-CH_2-$, -OMe gives -O-, and $-NMe_2$ gives $-\overset{|}{N}-$. The analyses using the two sets therefore show identical correlations.

Set C gives a correlation statistically superior at the 1% level to that with set B, as assessed by the F-test. Set D gives a similarly improved correlation over set C. This indicates that positions of substituents relative both to the reaction site and to each other have an important effect on the variation in the observed reaction rate constants, as had been expected.

The regression coefficients and t statistics for the structural features of sets B, C and D from regression analyses carried out at the 99.99% level are shown in Tables 15, 16 and 17.

These results are largely in accordance with the generally accepted mechanism of substituent effects in this kind of reaction (DE LA MARE et al., 1959, STOCK et al., 1963). Me groups are activating, and OMe groups very strongly so, in the order of position relative to reaction site para > ortho >> meta: steric factors are likely to be at least partially responsible for the reduced ortho

effect, OH and NMe_2 groups are also very strongly activating, but comparison of relative positions is not possible, since the values were not available in this data-set. Halogen substituents are generally deactivating: F and Cl in the order m > o > p, and Br rather anomalously in the order m > p > o. Particular coefficients change sign in the various analyses, but the general trend remains constant. I is deactivating, but a value is available only for one position relative to the reaction site.

Of the terms representing substituent interaction, Me-OMe and Me-NMe_2 show a strongly deactivating effect, and Me-Me a weaker deactivating effect. These factors, ascribed to electronic interactions between activating, i.e. electron-releasing, substituents, have been noted (DUBOIS et al., 1972).

The multiple regression analysis with set D structural features was repeated, omitting variables insignificant at the 10% level. The overall result is shown in Table 14, and the regression coefficients etc. in Table 18. These results show clearly all the factors mentioned above, and enable a more rapid and convenient understanding than the corresponding 99.99% analysis, which includes a number of insignificant variables and high correlations.

It may be that with feature sets of this size and complexity, or greater, analysis omitting insignificant variables may become valuable. Analysis including as many variables as possible could still be worthwhile, in order to investigate possibly interesting trends among individually insignificant coefficients.

It will be noted that in this last analysis the regression constant, which would normally take the estimated value for the unsubstituted compound, was omitted from the regression, and the benzene ring, common to all structures, included. This is an artefact of

the regression program, and does not affect the overall correlation.

From the estimated values of log k from the regression with structural feature set C, eleven structures have a discrepancy between observed and estimated values greater than 0.5 log units. These are structures 1, 3, 4, 8, 10, 16, 35, 37, 39, 41 and 44. Apart from the unsubstituted parent compound these are all Me, OMe, or NEe_2 derivatives, indicating that the poor estimation is probably due to non-additivity caused by large substituent interactions.

The estimated values for the regression at the 99.99% level using set D structural features are listed in Table 13. Many of these are "perfectly estimated", since the structure contains a unique structural feature.

Property predictions were simulated using the "hold-one-out" method, i.e. omitting the structure under consideration from the regression analysis. Since the aim was to determine the maximum accuracy which can be obtained using such methods, structural feature set D, which gives a significantly better correlation than any other set, was used for the predictions. However, since this set is complex, and contains a number of unique features and perfectly correlated features, it is only possible to obtain predicted values for 26 out of the 44 compounds. The remaining compounds each contained structural features for which no value could be obtained when this structure was removed from the regression analysis.

It may be that this will prove to be a common problem with such complex structural feature sets, which might consequently be of less general use for property prediction than simpler sets.

Prediction of log k values were simulated by summing the regression coefficients for the structural features present in the compound, from the regression analysis excluding that structure, as shown for structure 14 in Figure 14, i.e. giving a predicted

value for log k of 5.04, cf. observed value of 4.95. It may be noted that the regression coefficientsdiffer somewhat from those in Table 17: this is due to the omission of structure 14 from the analysis.

The predicted values are listed in Table 13. Polysubstituted compounds are dealt with as adequately as simpler structures. The mean discrepancy between observed and predicted values is 0.22 log units, and the sum of squares ratio is 0.004.

Some criticism levelled at structure-property correlation applications of pattern recognition was based on the suggestion that the compositions of the data sets were such that seemingly good results could be obtained by trivial analysis procedures.

It did not seem likely that this could be the case with the analyses described here. All the correlations were of high statistical significance, and the use of feature sets incorporating a greater degree of "chemically sensible" complexity was accompanied by statistically significant improvements in correlation. However, it was thought worthwhile to reanalyse this data-set using less appropriate, though non-trivial, variables: poor results in this case would indicate that the good correlations obtained in the previous analyses were not simply artefacts of the method.

Log k values were correlated firstly with the number of substituents on the ring (regardless of their type), secondly with the number of different types of substituents (regardless of the total number of substituents), and thirdly with both of the variables together. The results of these analyses are shown in Table 19.

The correlations are very much worse than that with structural feature set B, statistically so at the 1% level. Individually the

correlations are significant at the 5% level, but not at the 1% level.

These poor results with inappropriate variables tend to confirm that the correlation procedure is non-trivial, and that the good correlations obtained are not an artefact of the data-set.

The analyses described above demonstrate clearly that sub-structural analysis procedures may be used to correlate structure with property for multisubstituted benzene derivatives, giving good correlations and chemically sensible results for a data set where positional and interactive effects of substituents are known to be of importance. The systematic derivation of features representing relative positions of substituents, reported here for the first time, may be of considerable importance in studies of this kind.

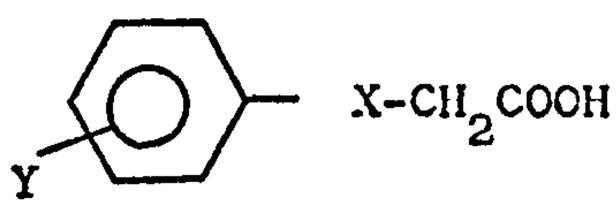
The success of these analyses, and the relative ease with which algorithmic feature generation may be carried out using WLN, suggested that this type of technique could be applicable to other properties of benzene derivatives, and to derivatives of other ring systems. These topics were examined in work to be discussed in following sections.

The work described in this section has been reported in part in a paper published in Journal of Chemical Information and Computer Sciences.

pk Values of Carboxylic Acid Derivatives of Benzene

The work described in this section had as its aim the application of the structural feature derivation procedure for cyclic structures described above to a data set of different characteristics. It also enabled the testing of an additional type of structural feature set.

The data used consisted of acid dissociation constants, pk values, for substituted β-phenylpropionic, phenoxy-, phenylthio-, phenylsulfinyl-, phenylsulfonyl-, phenylselenoacetic acids, i.e. with the structure shown below:



Y = various substituents

X = CH_2, O, S, SO, SO_2, Se

Property values were taken from two papers. In one case the values were measured at 20.0°C in aqueous solution (μ=0.10M KNO_3) (PETTIT et al., 1968). In the other case the values were measured in water at 25°C (PASTO et al., 1965): these values were made compatible by the necessary correction factor of 0.20 pk unit, as noted by Pettit et al. The data set used is listed in Table 20.

It will be noted that this data set is far from ideal for quantitative structure-property correlation. Although the number of structure, 98, is larger than in many such physical organic studies, no compound contains more than one substituent, in addition to the acid group. Thus in no case is more than one data point available as a measure of the effect of a given substituent in a particular position relative to a given acid group. Also the range of pk values is only just over 2.5 log units, considerably

smaller than for the reaction kinetics data set discussed above.

Five sets of structural features were used in these analyses:

A: breaking down the substituents into smaller units

B: keeping the substituents whole

C: including as structural features substituents and their relative positions (ortho, meta, and para)

D: including as structural features substituents and their relative positions (ortho, meta, and para) but treating the various acid groups, i.e. $-X-CH_2COOH$, as the same

E: including as structural features substituents and their co-occurrence, but not relative position.

Examples of structural feature derivation are given in Figure 15.

It will be seen that the use of structural feature set E is essentially equivalent to the Bocek-Kopecky additive model for disubstituted benzene derivatives (KOPECKY et al., 1965, BOCEK et al., 1967).

i.e. log activity = aX + bY +eX.eY + constant

The overall results of the regression analyses are shown in Table 21.

No analysis was possible using structural feature set C, because of the great number of variables, 113, involved.

The analysis using structural feature set B is statistically superior at the 1% level to the analyses with sets A and D. The correlation is not improved by the use of set E. The F values show that all the correlations are significant at the 1% level, with the exception of that using structural feature set D which is statistically insignificant.

It is not unexpected that the correlation with structural

feature set A is inferior to that with set B, since using set A structural features are derived without regard for their immediate environment. Thus in one of the examples in Figure 15 -S- is derived as a structural feature from both the acidic $-SCH_2COOH$ side-chain and the $-SCH_3$ group.

The very poor correlation with set D indicates that the major influence on pk is the nature of the ionising side-chain, rather than the nature of the other substituent on the ring, or the relative position of the two groups. It will be noted that in this case the introduction of extra parameters into the analysis has resulted in a very much worse correlation.

The regression coefficient values and t statistics for the structural features of set B are listed in Table 22. The predominant influence of the nature of the heteroatom in the acidic side-chain is demonstrated by the relative magnitudes of the coefficient values for the acidic groups and other substituents. A decrease in pk is seen, in the order $SO_2 > SO > O > S > Se > C$. Although the coefficients themselves are all highly statistically significant, there is no significant difference between any pair of coefficients. By comparison the coefficient values for other substituents are very small, though several are statistically significant. They are generally as would be expected from electronic factors (JAFFE, 1953, MCDANIEL et al., 1958): a certain inaccuracy is inevitable here, since position relative to the dissociating side-chain is not specified.

Examination of the estimated pk values, from the residuals of the regression with set B structural features listed in Table 20, shows that only 7 of the 98 values differ from the observed value by more than 0.1 pk units. These are structures 32, 38, 67, 43, 44, 71 and 84, i.e. phenylthio- ortho - OMe, ortho- SMe, meta-SMe, ortho-

COOH, and meta-NO_2, and phenylsulfinyl-meta-NO_2 acetic acids. This may indicate that some deviation from the normal substituent effects occurs for certain phenylthio derivatives: the ability of the thio group to transmit electronic effects is known to be greater than for the other groups considered here (PASTO et al., 1965). On the other hand Pasto et al. have noted the limited solubility in water of some phenylthio acids, which may lead to experimental error and hence these discrepancies.

The structural features from set E included in an analysis at the 10% significance level are listed in Table 23, together with their regression coefficient and t statistic values. Although three "cooccurrence" terms are included, none has a large coefficient value, and these results add nothing to the interpretation from the set B analysis.

The results of the analysis of this data set are less conclusive than those described in the previous section. The data set contained only disubstituted structures, and the compounds had been selected in order to obtain a Hammett plot for each of the subsets with different acidic groups. The more complex structural feature sets could therefore not be used.

The study of this data set is nonetheless of value. Firstly, data sets of this sort are likely to be common, particularly in areas where linear free energy relationship techniques are applied, and a failure of the substructural analysis techniques to deal with them would necessarily restrict their applicability. In the event, good correlations have been achieved, and the interpretation of the results is in accordance with those of workers using Hammett equation concepts.

Secondly, a number of useful points of technique may be noted. Simple structural features are adequate for good correlation

in this data set. The advantage of treating ring substituents as whole units is shown in this case by the significant improvement in correlation. The automatic derivation of "co-occurrence" structural features, corresponding to the Bocek-Kopecky additive model for disubstituted derivatives, from WLN has been demonstrated, although in this case their use did not bring a significant improvement in correlation. Finally, the very much worse correlation observed to result from an increase in the number of variables in set D, further demonstrates that the improvements in correlation brought about by the use of complex structural feature sets are non-trivial.

Heats of Formation for Unsaturated Aliphatics

The aim of the work described in this section was to develop methods for derivation of structural features to represent unsaturation in aliphatic compounds.

The data set used included gas-phase heats of formation for 70 alkenes and alkynes, taken from a review article (BENSON et al., 1969): this property has been measured with high accuracy for a relatively large number of unsaturated aliphatics. The structures and property values are listed in Table 24.

The structural features used to represent the unsaturated linkages were, for the most part, very similar to the augmented pair type of fragment, derived from connection tables (ADAMSON et al., 1976a). They included the multiple bond with its terminal atoms and their connections. Allenic linkages were treated as a single unit, including both multiple bonds and the central atom. In some cases, with small structures, these criteria resulted in the whole structure being treated as a single, unique structural feature: this is also the case with some thermochemical property estimation schemes (BENSON et al., 1969). Structural features of this sort have been regarded by some workers as the smallest units capable of giving consistent results for the estimation of the thermodynamic properties of unsaturated compounds (JANZ, 1967d).

Examples of structural feature derivation are given in Figure 16. The saturated parts of the structures were dealt with as described in the preceding section.

Details of the modifications to the computer program necessary for the derivation of these features are given in the Appendix. These procedures are applicable to unsaturated linkages containing heteroatoms, although compatible heat of formation data for such compounds was not available.

The procedures were able to detect the presence of conjugation, i.e. with unsaturated linkages separated by one single bond. This was a somewhat crude measure, since the type of unsaturation, and its environment, was not distinguished.

Some allowance was made for the cis type of interaction, by including an additional structural feature for each pair of non-hydrogen substituents in a cis relation about a double bond. Thermochemical additivity schemes often make use of this sort of correlation factor, though usually in a more elaborate form, with distinctions for particularly bulky groups, e.g. t-butyl (BENSON et al., 1969, JANZ, 1967c).

Examples of some structures with such cis interactions are shown in Figure 17, together with their Wiswesser notations. In some cases, e.g. structure 8, the cis/trans distinction included in the notation (SMITH et al., 1975) may be used straightforwardly. In other cases, because of the crude definition of the cis interaction used here, both cis and trans isomers are regarded as having a cis interaction, as in structures 22 and 23. Other structures, where cis and trans isomerism does not occur, may contain one or two cis interactions, as in structures 12 and 27 respectively.

Four sets of structural features were used in the analysis of this data set. These were:

set A - including only simple structural features, as described above

set B - as set A, including also cis interaction terms

set C - as set A, including also structural features representing conjugation

set D - as set A, including also structural features representing both cis interaction and conjugation

The overall results of the multiple regression analyses are

shown in Table 25. Structural feature sets B, C and D give correlations which are not statistically superior to that with set A. Examination of the residuals of the regressions showed that improvements in estimated values using sets B, C and D are not particularly marked for any structural type.

Thus, although the overall regressions results are statistically highly significant, the inclusion of the variables representing cis interactions and conjugation have little effect. Repeating the analysis with only cis and trans isomer pairs gave very much the same results, indicating that this is not a case of a significant effect in a small number of structures being lost in a larger data-set. Rather it may reflect the very simple nature of the representation of the isomerism, with no distinction between the type of cis groups. Similarly the relatively crude representation of conjugation probably explains the insignificant improvement brought about by the inclusion of this variable.

The structural features of set D are shown in Table 26, with the corresponding regression coefficients and t statistics. All the regression coefficients, with the exception of those for CH_3-, the cis interaction, and conjugation are statistically very highly significant. A number of coefficients however, as noted above, are based on a single observation.

These results show clearly the consistent negative effect of carbon branching on heat of formation.

It was not thought worthwhile for the purposes of this study to deal in a more exact fashion with cis interactions and the effect of conjugation, to try to obtain a significant improvement in correlation. This could probably best be done by a substructure search procedure, adapted for some particular application.

The work described in this section had demonstrated that sub-

structural analysis procedures based on WLN can deal adequately with multiply bonded compounds, and lead to highly significant correlations.

Heats of vaporization for diverse structures

The work described in this section had as its aim the application of the methods developed previously to a relatively large data set including compounds of various structural types. In particular the quality of simulated property predictions from correlations of data sets of varying size and composition were to be examined.

Heat of vaporization was the property correlated against structure in this study. This quantity has been accurately measured for a wide variety of structural types: thermochemical data of this type is amenable to computer processing (PEDLEY, 1976), so that automated data analysis of the kind described here could be of practical use in this area. This property is also of interest as it reflects inter-molecular forces, rather than intra-molecular effects.

The property values used, measured in kilocalories per gm. formula wt. at 25°C and 1 atmosphere pressure, were taken a standard compilation of thermochemical data (COX et al., 1970a).

The data set used is listed in Table 27. It comprises alkanes, alkenes, alcohols, ketones, benzenes and pyridines.

Heats of vaporization were used directly in the correlations, without conversion to logarithmic values, since this is conventional in the majority of additive schemes for thermochemical property estimation (COX et al., 1970b, JANZ, 1967c).

The procedures used for structural feature derivation were for the most part as described in preceding sections. Modifications were made to allow inclusion of pyridine structures with benzenes for correlation, to allow derivation of ring substituent interaction terms in an analysis including both cyclic and acyclic structures,

and to select a particular substructure as a feature: these points will be discussed further below,

A number of subsets of structures were examined separately: sets of data will be donoted as 'set (n)', where n is numeric, to avoid confusion with sets of structural features, denoted as 'set X', where X is upper case alphabetic.

It is convenient to deal firstly with the correlations, in three subsections: those data sets including only saturated aliphatic structures, those including unsaturated structures, and those including aromatic structures. The analyses of mixed data sets will then be described. Finally, simulated property predictions will be considered.

It should be noted that it is not possible to compare statistically the results of regression analyses of different data sets, i.e. different sets of values for the dependent variables.

Because of the relatively small number of variables generally involved these analyses were carried out for the most part at the 99.99% level.

Saturated Aliphatic Structures

Five subsets of data in this category were studied:

(1) 12 alcohols (structures 38-49)

(2) 11 ketones (structures 50-60)

(3) 23 alcohols and ketones (sets (1) and (2) combined)

(4) 37 alkanes (structures 1-37)

(5) 60 saturated aliphatics (sets (3) and (4) combined)

The structural features used for these correlations were the simple hydrocarbon fragments and functional groups described in earlier sections.

The overall results of the multiple regression analyses are given in Table 28. All the correlations are very highly significant, though, as noted above, statistical comparison of the results for different data sets is not possible.

The regression coefficients from the analyses of these sets of structures are for the most part individually statistically significant. They differ very little from the larger set, shown in Table 35, and are therefore not set out in full. Some discrepancies between individual values for different sets of structures may be expected, apart from the usual limitations on the reliability of coefficient values. For example, the incremental values for hydrocarbon groups are affected by the presence of ketonic groups (COX et al., 1970c). This may lead to an "averaging" effect, leading to lowered significance of coefficient values, with mixed sets of structures.

The use of these subsets of data for simulated property predictions will be described below.

Unsaturated Aliphatic Structures

Only one subset of data of this type was examined:

(6) 36 alkenes (structures 61-76)

This data set was analysed using structural features appropriate to unsaturated compounds, as described in an earlier section. The usefulness of the simple representation of cis substitution was again tested.

Two sets of structural features were used:

set A: including simple structural features

set B: as A, including also a term representing cis interaction

The overall regression results for these structural feature sets are shown in Table 29.

The use of set B gives no improvement in correlation over set A, indicating that the inclusion of the simple type of cis representation used here is not useful in accounting for the property variation, as was the case for the heats of formation data discussed above.

Set A gives a highly significant correlation. The regression coefficients and t statistics for this data set are shown in Table 30. It will be noted that carbon branching consistently reduces heat of vaporization: this will be discussed further below.

Aromatic Structures

Three subsets of data in this category were considered:

(7) 31 benzene derivatives (structures 97-127)

(8) 10 pyridine derivatives (structures 128-137)

(9) 41 aromatic structures (sets (7) and (8) combined)

The analyses of data set (7) were carried out using the types of structural features first developed for the correlation of reaction kinetics data for benzene derivatives, and described in an earlier section. Two sets of structural features were used:

set C: including type and number of substituents

set D: including type and number of substituents, plus positions of substituents relative to one another.

The overall results of the regression analyses are shown in Table 31. Structural feature set D gives a correlation superior at the 1% level to that with set C, indicating the importance of substituent interactions in this case.

In order to deal with derivatives of pyridine, the program for structural feature derivation was adapted to include the ring nitrogen atom as a substituent, as has been the procedure in some applications of Hammett analysis (EXNER, 1972g). Details of the modified program are given in the Appendix.

Data set (8) was then analysed using two sets of structural features:

set E: including number and type of substituent

set F: including number and type of substituent, plus relative positions of substituents

including the heteroatom as a substituent in each case.

The overall results of the multiple regression analyses are shown in Table 32. Despite the large discrepancy in the multiple correlation coefficients, 0.997 cf. 0.881, the set F analysis is

superior to set E only at the 5% level: a typical example of the inadequacy of this measure of correlation. This less significant improvement may well be accounted for by the very small number of structures, and small range of property values in the set.

Data subset (9), comprising all 41 benzene and pyridine structures, was analysed in some detail, in order to investigate the substituent interaction effects. Five sets of structural feature sets were used:

set G: including number and type of substituents

set H: as set G, and also including ortho substituent interactions

set I: as set H, and also including meta substituent interactions

set J: as set H, and also including para substituent interactions

set K: including number and type of substituents, and all ortho, meta, and para substituent interactions.

The overall results of these regression analyses are shown in Table 33. Structural feature set H gives a correlation better at the 1% level than set G, set I gives a correlation better at the 5% level than H, and set K gives a correlation not significantly better at the 10% level than set I. Set J gives a slightly worse correlation than set I, but the difference is not significant at 10%.

These results indicate that ortho interactions exert an important effect on heat of vaporization, meta interactions a less important effect, while the effect of para interactions is negligible. This important ortho effect is in accordance with the use of ortho correction factors in additive estimation schemes for thermochemical properties (BENSON et al., 1979) and for boiling points (KELLIE et al.,

1975).

The regression coefficients and t statistics for the set K structural features are listed in Table 34. These results confirm the relatively large effects of ortho interactions: in particular alkyl-alkyl interactions increasing the heat of vaporization and alkyl-hydroxy interactions reducing its value. Methyl substitution ortho to the ring nitrogen of pyridines decreases the otherwise positive effect of this substituent.

Diverse Structural Types

Two subsets of data for diverse structures were analysed. These were:

(10) 96 aliphatic structures (sets (5) and (6) combined)
i.e. alcohols, ketones, alkanes, and alkenes

(11) 137 diverse structures (sets (9) and (10) combined)
i.e. the total set of structures.

Data subset 10 was analysed using the straightforward structural feature set comprising hydrocarbon fragments, functional groups, and unsaturated units. The overall regression result was: no. of features = 11, no. included = 11, degress of freedom = 85, R => 0.999, r = 0.22, F = 7721.48 (range of property values = 11.59), a highly significant correlation.

The structural features in this analysis, together with their regression coefficients and t statistics are listed in Table 35. All the coefficient values are significant at the 1% level.

These results illustrate clearly the effect, noted above, of carbon branching in decreasing heat of vaporization, although, as is so frequently the case, the differences between pairs of coefficient values are not statistically significant. This trend is seen in both saturated fragments, $-CH_3 > -CH_2- > -\overset{|}{CH}- > -\overset{|}{\underset{|}{C}}-$, and unsaturated $CH_2=CH- > -CH=CH- > CH_2=\overset{|}{C}- > -CH=\overset{|}{C}- > -\overset{|}{C}=\overset{|}{C}-$. This is to be expected, since a greater degree of branching implies a decrease in molar volume and hence surface area, and consequently decreased non-polar intermolecular forces: the heat of vaporization is thereby lowered (Cox et al., 1970a).

The anticipated effect of the -OH and -C=O groups in increasing heat of vaporization is also clearly seen.

From the estimated heats of vaporization, i.e. the residuals of the regression, it is observed that for only four compounds is the discrepancy between observed and estimated values greater than 0.5 kcal./g.f.w. Three of these structures are CH_3-CO-CH_3, CH_3CH_2-CO-CH_3, and $CH_3(CH_2)_3CO(CH_2)_3CH_3$: this suggests that the widely-known anomalies in the incremental values for hydrocarbon groups in ketones, as discussed above, may be the only cause of non-additivity in this data set. The set does not, of course, contain any of the difunctionalities known to cause deviation from additivity for various thermodynamic and physicochemical properties (BENSON et al., 1969, NYS et al., 1973, IWASA et al., 1965).

The coefficient values from this analysis were used to further examine the adequacy with which this substructural analysis method accounts for effects due to chain branching. The differences in heat of vaporization between two pairs of structures, differing only in extent of carbon branching, were estimated: these values were then compared with the observed values, and those estimated by the Greenshield-Rossini and Laidler-Lovering structural contribution schemes (COX et al., 1970d).

The results are presented in Table 36. The substructural analysis technique estimates the values to a high degree of accuracy, in contrast to the other estimation procedures.

Data subset (11), i.e. the whole set of 137 structures, was analysed using two sets of structural features.

The feature derivation program was modified, so as to allow analysis of a data set including both cyclic and acyclic structures, giving features representing substituent interactions on rings.

Ring substituents are broken down into simpler units, as are acyclic structures: interaction terms are derived representing substituents as whole units. Full details of this modified program are given in the Appendix.

The feature sets used were:

set L: including as structural features hydrocarbon fragments, simple functionalities, unsaturated units, and whole ring systems.

set M: as set L, but including also terms representing substituent positions on rings.

Examples of structural feature derivation are given in Figure 18. The overall results of the regression analyses are shown in Table 37.

The correlation with set M structural features is superior at the 1% level to that with set L, re-emphasising the importance of relative positions of ring substituents.

The regression coefficients and t statistics for set M structural features from the analysis at the 99.99% level are listed in Table 38. The major effects noted above, i.e. those of polar groups, carbon branching, and intersubstituent interaction, are clearly shown here.

For comparison these values for set M structural features included in an analysis which excluded variables insignificant at the 10% level are listed in Table 39. So many variables have been excluded that the effects of carbon branching and substituent interaction cannot be identified clearly. This may suggest that analyses omitting variables whose effects are of low statistical significance may lead to a loss of potentially useful information, although they give highly reliable results and illustrate the most important

effects on the variation in the observed property values. For this reason the simulated property predictions carried out below all used analyses carried out at the 99.99% significance level.

Simulated Property Predictions

Simulated property predictions were carried out for the following structures:

a) n-heptane, 2,2 dimethyl pentane, toluene, chlorobenzene, pentan-2-one, and n-butanol, so that comparisons could be made with predicted values from other thermochemical estimation procedures (COX et al., 1970e).

b) cis-2-butene, so that the effect on prediction of including the cis structural feature could be tested.

c) 1,2,3 trimethylbenzene and 1-methyl,2-ethyl benzene, so that the effect on prediction on including substituent interaction terms could be tested.

Predictions were simulated by the hold-one-out procedure described above, using various applicable data subsets and/or structural feature sets in each case. It should be noted that reference to a particular data subset in this connection implies that subset minus the structure under consideration. The results of the simulated predictions are summarised in Tables 40 to 48.

n-heptane and 2,2 dimethylpentane are both well predicted from the alkane subset, from the total aliphatic subset, and from the total set including interaction terms; less well so from the total set without interaction terms.

Toluene is well predicted from the benzene and total aromatic subsets, provided interaction terms are included, poorly if they are not. It is reasonably well predicted from the total set, better if interaction terms are not included.

Chlorobenzene is well predicted from all the data subsets and structural feature sets tested.

Pentan-2-one is well predicted from those data subsets containing only ketones or only ketones and alcohols, poorly predicted by the other subsets. n-butanol is similar, though the disparity in performance is less marked.

Cis-2-butene was reasonably well predicted from data subset (6). Inclusion of the parameter representing cis interaction had very little effect on the prediction, as might have anticipated from its negligible effect on the overall correlation.

The predicted values using data set (11) were somewhat better than those with subset (16).

1,2,3-trimethyl benzene is reasonably well predicted by the subset of all aromatics and the total set, provided interaction terms are included.

1-methyl,2-ethyl benzene was chosen for property prediction testing, since it was the only structure in the set to contain one particular structure feature, Me-ortho-Et. When this compound is omitted from the set it is therefore impossible to calculate a value for this feature, and thus impossible to accurately predict the property value for this compound. This, as has been noted earlier, may well be a common problem, where property prediction is attempted, using complex structural feature sets.

Three alternative methods for alleviating this problem are possible: use of a simpler structural feature set, ignoring the outstanding structural feature, or approximating its value by that of a similar structural feature. In this case Me-ortho-Me was thought to be the most appropriate substitute. It will be seen from Table that the third course of action, using Me-ortho-Me, gives a reasonable prediction.

The simulated predictions for n-heptane, 2,2-dimethylpentane, toluene, chlorobenzene, pentan-2-one, and n-butanol were compared with predictions made for these compounds using other methods (COX et al., 1970e). The other methods were:

- Laidler-Lovering's method, the method of CH_2 increments, and Wright's method, all based on additive structural contributions.

- Chen's equation, considered by Cox and Pilcher to be the best of a group of methods using critical parameter data to predict heat of vaporization.

- Fishtine's equation, one of a number of methods using boiling point data to predict heat of vaporization, and recommended by Cox and Pilcher as the most generally effective prediction method.

The predicted values for the six compounds by these methods are summarised in Table 49, for comparison with predictions made using substructural analysis techniques. Two values are given for each compound from the simulated predictions described above: one derived from using the smallest relevant data subset, and one using the total data set.

It appears from these results that, although for each compound one of the other methods gives a better simulated prediction than substructural analysis, overall the whole set the methods described here are more consistently accurate.

This is, of course, far from being a comprehensive test of performance for the various methods. Also the different requirements for data input should be borne in mind. Thus the methods involving boiling points and critical parameters require that certain other properties of the compound under consideration should be known,

whereas substructural analysis, and the other techniques involving additive structural contributions, require heat of vaporization data for a number of compounds other than that under consideration.

It appears from these results that this type of substructural analysis is likely to give property predictions of at least comparable accuracy to other methods currently in use, for relatively accurate data of the kind discussed here.

Physico-chemical and Biological Properties of Benzimidazole Derivatives

The work described in this section involved the application of the type of structural feature sets originally devised for benzene derivatives to benzimidazole structures. It allowed the examination of several properties for a relatively large series of compounds.

The data sets used were pKa, aqueous solubility, and mammalian toxicity values for trifluoromethylbenzimidazoles, made available by Fisons Agrochemicals Division. In total 154 pKa values, 123 aqueous solubility (parts per million) values, and 129 rat oral LD_{50} (μmol) values were used: these data sets overlapped to a large extent. Logarithmic values were used for the correlation of solubility and LD_{50} values. The trifluoromethylbenzimidazole nucleus, with its numbering system, is shown in Figure 19, and the data sets are tabulated in Table 50.

These compounds have all been tested for herbicidal activity (BURTON et al., 1965) and are known to be uncouplers of oxidative phosphorylation (JONES et al., 1965). The uncoupling activities of some of these compounds have been studied, using substructural analysis techniques of the kind described here (UFTON, 1976). pKa values of compounds of this sort have been correlated with semi-empirical parameters (TOLLANAERE, 1973).

Since the position of nitrogen protonation is not known for these compounds, the structures were treated as symmetrical. The positions of substituents on the benzenoid ring can be related only to the nearer ring fusion point, and thereby to the nearer nitrogen atom, rather than to the -NH- group, considered to be important for biological activity (TOLLANAERE, 1973).

Four types of structural feature set were used for the

analysis of this set of compounds.

A - including type and number of substituents as structural features

B - including position of substituents, ortho or meta, relative to nearer ring fusion point

C - including only positions of substituents relative to one another. For monosubstituted structures, type of substituent is used.

D - including positions of substituents relative to the nearer ring fusion point and relative to one another.

Examples of structural feature derivation are given in Figure 20.

These structural feature sets are very similar to those used for benzene derivatives. Set C allows the investigation of inter-substituent interaction effects, regardless of position relative to ring fusion points.

It should be noted that, because of the different structures for which the three types of property value are available, the sets of structural features produced by each type of fragmentation will differ somewhat for the various properties.

The overall regression analysis results are summarised in Tables 51, 52 and 53. All these analyses are statistically significant at the 1% level. Although exact statistical comparison of the quality of correlations on different data sets is not possible, inspection of the correlation coefficients and F values indicates that the correlations for pKa data are superior to those for solubility, in turn superior to those for toxicity. This may reflect, at least in part, the relative accuracy of the measurement of the data.

pKa data regressions

The correlation with Type D structural features, i.e. including both substituent ring position and substituent interaction terms, was found to be significantly better at the 1% level than the correlations with Type A and Type B feature sets, and significantly better at the 5% level than that with the Type C feature set. This indicates the importance of both position and inter-substituent interactions in determining the effect of a substituent on pK value.

The regression coefficients of the Type D structural feature set are shown in Table 54. They are generally as would be expected from the known effects of substituents on heterocyclic pKs (ALBERT, 1963). The results may tentatively be further interpreted by assuming that substituents ortho to a ring fusion point exert a largely inductive effect, while for substituents meta to a ring fusion point mesomeric effects are of comparable importance. It must however be noted that a number of these coefficients are individually statistically insignificant, emphasising the need for discretion in interpreting the results in mechanistic terms.

Thus -OH and -OMe groups in the ortho position are base-weakening (-I,-M) and -Me groups are base-strengthening (+I,+M). There is little difference in the coefficient values for these substituents in the ortho compared with the meta position. This may indicate that a decrease in inductive effect in the meta compared with the ortho position, i.e. further from the N-containing ring, is compensated by an increased mesomeric effect in the meta position.

As expected ethyl and t-butyl groups are shown to be base-strengthening (+I,+M) and carboxylic acid and amide substituents base-

weakening (-I,-M). Positional comparisons cannot be made for these substituents, nor for several others, since the alternative substitution positions were not represented in the structures examined.

The -SH and -$NHCOCF_3$ substituents are perfectly correlated, and hence it is not possible to partition the observed base-weakening effect between these groups.

The NH_2 group is base-weakening, as would be expected due to the strong -I effect of the -NH_3^+ zwitterionic form (BURTON et al., 1965). Meta substitution appears to give a larger effect than ortho but the situation is complicated by the perfect correlation of the o-NH_2 group with the interaction term for NH_2-p-Cl. An interaction of this kind would be expected to reduce the -I effect (see below).

Halogen and -CF_3 substituents are base-weakening, with ortho substitution resulting in a greater effect than meta, presumably due to a weak +M effect for the meta substituent opposing the strong -I effect. The base-weakening effect for substitution both ortho and meta to a ring fusion varies in the order F < Cl < Br I < CF_3, in accordance with the previously observed effects of these substituents on pK (ALBERT, 1963, MACDANIEL et al., 1958).

The -SO_3H substituent gives a base-strengthening effect, though the unionised -SO_3H group usually exhibits a strong -I tendency. The situation is complicated by three perfectly correlated structural features for Cl-SO_3H interactions.

The -NHCOMe and -NHCOPh substituents both appear to be base-weakening. This is unexpected, since the -I effect for these groups is usually found to be weaker than +M.

The coefficients for substituent -OCOMe and -OCOEt and for -OCONHEt and -$OCONIIMe_2$ show differing effects on pK for pairs of

substituents which would be expected to have similar properties.

The regression coefficients for substituent interaction terms are nearly all small and statistically insignificant, with no obvious trends. However the statistical significance of the improvement obtained by including these values shows that this improvement is not due simply to the increase in the number of variables.

The strong base-strengthening interactions Cl-o-NH_2 and Cl-p-NH_2 (and probably Cl-p-NH_2, perfectly correlated with RF-o-NH_2) are likely to be due to mutual inhibition of electron-withdrawal between two -I groups (DUBOIS et al., 1972). Cl-p-OH can only occur in structure with the -OH group ortho to a ring fusion point, i.e. where it appears that the -I effect of -OH makes the major contribution to its effect on pK.

The base-weakening interaction Cl-o-OH can be explained as the inhibition of the strong +M effect of -OH, in those compounds where this is the predominant effect, by the weakly +M -Cl substituent.

It should be re-emphasised that since some individual coefficient values are not statistically significant, despite the good overall correlation, it is not to be expected that all regression coefficients and differences between pairs of coefficients will be reliable. Hence interpretation of the results should be based on trends rather than particular values, if the conclusions drawn are to be more than tentative.

Solubility data regressions

The correlations achieved using sets of structural features of Type B and Type D were not statistically superior at the 5% level

to that using Type A, the simplest set. This indicates that the greater part of the variation in the measured solubilities is accounted for simply by the number and type of substituents, with substituent positions and interactions of lesser importance. An analysis using Type C structural features could have added no further useful information.

The regression coefficients for the Type A structural feature set are shown in Table 55 and for the Type D set in Table 56.

The results from the analysis with Type A structural features are generally in accordance with expectations, with hydrophobic substituents, e.g. alkyls and halogens, decreasing solubility, and hydrophilic substituents, e.g. -OH and -COOH, increasing solubility. The -OCOMe and -OCOEt substituents appear to exert opposite effects, as was the case with pK. The statistical limitations on the reliability of coefficients, which may be individually insignificant, should again be borne in mind when considering seemingly anomalous values.

The interpretation of the results from the analysis with Type D structural features should be treated cautiously, since it is not superior, according to the most commonly applied statistical criterion, to the simpler Type A analysis. A definite trend is evident for those substituents where values for substitution both <u>ortho</u> and <u>meta</u> to the ring fusion point are available, in that a substituent tends to give a higher solubility in the <u>ortho</u> position than in the <u>meta</u> position. The $-NH_2$ groups is an exception.

Thus all the halogens and -OMe have a higher negative coefficient for <u>meta</u> cf. <u>ortho</u>, -OH has a higher positive coefficient for <u>ortho</u> cf. <u>meta</u>, and $-NO_2$, $-CF_3$, and Me have positive coefficients when <u>ortho</u> to a ring fusion point and negative coefficients when <u>meta</u>.

-OCOMe and -OCOEt now both have negative coefficients, with -OCOMe the more strongly negative. Both ortho and meta -F substituents show negative coefficient values, though in the Type A analysis F has a slightly positive coefficient. This sort of effect shows the possible dangers in placing too much emphasis on single coefficient values of low statistical significance.

No clear trends can be seen in the terms representing substituent interactions. The OH-p-OH interaction has a large and statistically significant negative effect on solubility, possibly due to a mesomeric interaction resulting in mutual inhibition of polarity (DUBOIS et al., 1972) The Cl-p-Cl and Cl-pNO_2 interactions also show negative effects, while OMe-o-OMe has a large positive effect on solubility.

Toxicity data regressions

The analyses using Type D structural feature sets are significantly better than those with Type A and Type B sets at the 5% level. This, together with the lack of improvement in correlation, as assessed by the residual error, from set A to set B, may indicate that relative positions relative to the benzimidazole ring in determining the substituents' effects on the measured LD_{50} values.

The detailed interpretation of the results is less clear than for the pK and solubility data sets, with fewer obvious trends among the regression coefficients. Many coefficients are small and statistically insignificant. The analysis using Type D structural features (Table 58) shows that -F, -Br, -Cl, -CF_3, -NO_2, and -CN substituents have a negative effect on measured LD_{50} values, i.e. increase

mammalian toxicity. -I, $-NH_2$, and $-NMe_2$ substituents have a positive effect on LD_{50}, while -Me, -OMe and OH groups have a negative effect when ortho to a ring fusion point and a positive effect when meta. SO_2X groups, with the exception of SO_2NH_2, have a positive effect, but comparison between different positions is generally lacking.

Interhalogen interaction terms appear to depend on relative position, with ortho pairs of halogens having a largely positive effect and meta pairs usually a negative effect. Halogen interactions with -Me, -OH, $-NO_2$, and -CN groups increase LD_{50}, with Cl-NO_2 interactions notably large and statistically significant. $-NO_2$ groups ortho and meta to one another similarly considerably increase LD_{50}, and a marked decrease in mammalian toxicity on the introduction of $-NO_2$ groups into chlorinated derivatives has been noted (BURTON et al., 1965). The interaction term for OMe-o-OMe has a large negative coefficient which may be compared with the positive effect of this feature on solubility (see above).

For comparison, the results of the analysis with Type B structural features are shown in Table 57. These support the summary of effects noted above.

The work described above demonstrated the applicability of substructural analysis techniques of the kind under consideration to multisubstituted heterocyclic systems. It indicated that good correlations could be obtained for both physicochemical and biological properties, suggesting that these methods could allow for a detailed analysis of a range of structure-property relationships for a series of compounds. This

could be useful if used in conjunction with the semi-empirical type of inter-property correlation: this has been investigation for part of this data set (UFTON, 1976).

A study of the structure-activity relationship for the uncoupling of oxidative phosphorylation of structures of this type was subsequently carried out using these methods (UFTON, 1976). This study gave statistically significant correlations, giving a further illustration of the whole range of applicability of this type of substructural analysis.

The results described above show that significant improvements in correlation are brought about for some properties by the inclusion of terms representing relative position of substituents. This suggests that this type of structural feature, originally devised for benzene derivatives, may be generally applicable to a variety of cyclic structures.

An alternative form of structural feature derivation was investigated for compounds of this sort (UFTON, 1976). Hydrogen atoms on the benzenoid ring were included explicitly as structural features, rather than the implicit consideration in the structural feature sets described above. These alternatives are also found in the various forms of Free-Wilson analysis (KUBINYI et al., 1976b). In the sub-structural analysis procedures it was found that inclusion of hydrogen had no effect on the overall correlation, as had been anticipated, and that the results, in terms of individual coefficient values, were less readily interpretable. This suggests that the most useful structural feature sets for aromatic structures may be those implicitly including hydrogen substituents, in accordance with the conventional chemical approach, and with suggestions regarding Free-Wilson methodologies (KUBINYI et al., 1976a).

The substructural analysis procedures described here have been used for simulated property predictions, using "training sets" of structures to allow prediction for all three properties (SAGGERS, 1976). The quality of the simulated predictions with type B structural feature sets, equivalent to Free-Wilson analysis, is in general superior to that with the simpler type A sets. Simulated predictions from type D sets are of high quality, but frequently are not possible because of the lack of coefficients for some of the necessary structural features, which do not occur in the "training set" of structures. This may well be a common problem with such complex structural feature sets, reducing their predictive usefulness.

pKa Values for Diverse Heterocyclic Systems

The work described here had as its aim the development of a substructural analysis technique which would allow for the treatment of data sets including a number of different types of heterocyclic system in a single analysis.

The technique devised was applicable to those ring systems, including fused systems, which may be treated as a derivative of a six-membered nucleus: only nitrogen heterocycles were considered. Heteroatoms and ring fusion points were treated as substituents, as in some applications of the Hammett equation (EXNER, 1972g). Other substituent groups were treated as whole units. Relative positions of substituents were represented as ortho, meta, or para, taking the shortest path between pairs of substituents, as in the structural feature sets for cyclic structures described in earlier sections.

An amount of approximation is involved here. Thus in the benzene derivative shown in Figure 21 the Me-ortho-He interactions 2-3 and 3-4 are not identical, because of the different position of the NH_2 group. A compromise is necessary between specifying structural features in sufficient detail and keeping the number of variables to an acceptable level. The good correlations achieved in the study of benzene reactivities described in an earlier section suggests that the level of approximation was not too great for that data set. However the situation is rather different for heterocycles. Thus, for example, it is evident that in the pyridine nucleus (Figure 21) there are three possible forms of meta interaction: 2-4 (equivalent to 4-6), 3-5 and 2-6. The 2-6 interaction is distinct in that a heteroatom is included between the meta position. In some of the analyses below, structural features were derived to distinguish between interaction terms involving substituents separated by heteroatoms and those involving only carbon atoms, and their usefulness

assessed.

In general, for each set of structures investigated, analyses were carried out on a number of sets of structural features of increasing complexity representing number and type of substituent only or position of substituents relative to heteroatoms, fusion points (if applicable), and/or other substituents.

A computer program (described fully in the Appendix) was written to derive these structural features from WLN. Details of the structural feature sets used for each set of structures investigated are given below.

These structural feature sets were tested by multiple regression analysis of pKa data for various heterocyclic systems. pKa values measured at 20°C in aqueous solution for 169 compounds were taken from standard compilations (ALBERT, 1963, ALBERT, 1971). The numbers of derivatives of each of eleven parent ring systems used are shown in Figure 22. Pyridine derivatives comprised approximately one third of the set, and pyrimidine derivatives made up another third. The remainder of the set consisted of derivatives of seven different ring systems. Just over 20% of the total set comprised derivatives of fused ring systems.

pKa was regarded as a suitable property value to test this method since it has been measured accurately for derivatives of a variety of heterocyclic ring systems. It has also been recognised as an important factor in determining some kinds of biological activity (FUJITA, 1966, TOLLENAERE, 1973).

The effects of substituents on pka values have been described in detail (ALBERT, 1963), and comparison of the results of the analyses with these documented effects enabled an assessment of the reliability

and usefulness of the technique.

The analyses were carried out on the whole set of structures, and on subsets of pyridines, pyrimidines, and the remaining diverse ring systems, to compare the correlations obtained on data sets of varied size and composition.

In some cases where a complex structural feature set produced a large number of variables, these analyses were carried out so as to include only those variables significant at the 10%, 5% or 1% level.

Pyridine Subset

The subset of 52 pyridine structures was analysed correlating pKa with several sets of structural features. These sets included:

A: number and type of substituents

B: position of each substituent relative to the heteroatom

C: as B, and additionally including relative positions of each pair of substituents

D: as C, making a distinction between meta substituent pairs in the 2 and 6 positions, separated by heteroatom, and other meta substituent pairs, separated by a carbon atom.

E: number and type of substituents, plus relative positions of each substituent pair as in D.

Examples of structural feature derivations are given in Figure 23.

The overall results of the regression analyses are shown in Table 59. The F values indicate that all the correlations are significant at the 1% level.

Comparison of the regression results by the F-test indicates that the improved correlations brought about by including firstly positions relative to the heteroatom, and secondly intersubstituent interactions are both statistically significant at the 1% level. Structural

feature set B, including intersubstituent relative positions but not including positions relative to the heteroatom, gives a correlation not significantly better than that with set A, including only number and type of substituent. Distinguishing between the two forms of meta interaction did not give a significant improvement in correlation, possibly due to the small number of structures affected. Values for both kinds of meta substituent were available for only two substituent pairs; Me-Me, occurring in 9 structures for which both meta interaction terms have negligible coefficient values, and Cl-NH_2 occurring in 2 structures for which the 2-6 meta interaction has a negligible value, and other meta interactions show a negative coefficient corresponding to about 0.6 of a pKa unit.

The best correlation for the pyridine subset data is clearly that using structural feature set C, i.e. indicating the position of each substituent relative to the ring nitrogen, and including the relative position of all substituent pairs. The structural features in this set, with their regression coefficients and t statistics are listed in Table 60.

These results are largely interpretable from the electronic properties of the substituent groups, although values were not available in the data set for all the structural features necessary for an exhaustive examination of trends in coefficient values.

Amino, methylamino and dimethylamino substituents are highly base-strengthening when ortho and para to theheteroatom, while the lesser effect of the methyl group does not appear to be position-dependent. Halogen and nitro substituents are base-weakening in all positions, while OMe and SMe substituents' effects are position-dependent, with strongly negative coefficients when ortho to the ring nitrogen, weakly negative when meta, and positive when para. This may be accounted for by the opposed inductive and resonance effects of these groups (ALBERT, 1963).

The most notable coefficient values for the terms representing intersubstituent interactions and those between amino, methylamino, dimethylamino and methyl groups. All these are base-weakening, probably representing inhibition of electron-release by similar substituents (DUBOIS et al., 1972).

Pyrimidine subset

The pyrimidine nucleus (Figure 24) appears to present more complex problems in deriving structural features useful in structure-property correlation than is the case with pyridine. The terms representing relative position of substituent and heteroatom, in the ortho case, reflect widely different physical situations because of the difference between the 2 and 4 positions, both of which are ortho to a heteroatom. This does not, in fact, affect the overall correlation, since the 4 position ortho interaction is perfectly correlated with the substituent-para-heteroatom term: thus any discrepancy between the two ortho terms will be correlated by the alteration of the "true" value for the para interaction. Substituent position relative to heteroatoms can, therefore, be accounted for equally well by using either explicit position on the ring, or substituent-heteroatom interaction terms. Reliable interpretation of the results is made easier by using explicit position, but at the expense of generality, i.e. of the applicability to deal with other classes of compound simultaneously.

Distinctions between 2-4 and 2-6 meta interactions, including a heteroatom, and 4-6 meta interactions, including a carbon atom, may be made as in the pyridine subset.

The structural feature sets used to analyse the pyrimidine data were very similar to those used for the pyridine structures. They included:

F: number and type of substituent

G: explicit substituent positions, i.e. 2, 4 (equivalent to 6), or 5

H: explicit substituent positions plus intersubstituent interaction terms

I: as set H, distinguishing the two forms of <u>meta</u> interaction

J: number and type of substituent, plus relative positions of substituent pairs as in set J.

Examples of structural feature derivation for pyrimidines are givenin Figure 24.

The overall results of the regression analyses are shown in Table 61. The F-values indicate that all the correlations are significant at the 1% level.

Comparison of the correlations by the F-test show that inclusion of substituent position, set G, and of relative position of substituents without distinction between the <u>meta</u> interactions, set H, does not significantly improve the correlation beyond that achieved with the structural features representing only number and type of substituent, set F. Only when intersubstituent interaction terms reflecting the two forms of <u>meta</u> interaction are included, i.e. structural feature set I, is the correlation shown to be improved at the 1% significance level. Set J gives a correlation which is not significantly better than that with the simple set F. These results suggest that the positions of substituents relative both to heteroatoms and to other substituents are of importance. Specification of the position of a substituent does not in itself improve the correlation, as it did with the pyridine subset. This may be interpreted as a measure of the non-additivity of the effects of substituents in the pyrimidine system. The importance of the distinction between the two possible <u>meta</u> relationships between substituents seems to further

indicate the importance of inter-substituent effects in this system.

It may be noted that structural feature set I produced 66 variables, i.e., in excess of the number of measured property values. Because of perfect correlation the number of features included was sufficiently small to allow a regression analysis, but with so large a number of variables the predictive usefulness of such a correlation is doubtful, as discussed below. In order to reduce the number of variables, the analysis was repeated in such a way that the regression program omitted variables insignificant at the 10% level. This eliminated approximately half the variables included at the 99.99% level. The structural features included in this 10% level analysis are listed in Table 62 together with their regression coefficients and t statistics: the overall result is shown in Table 61. These results again demonstrate the importance of intersubstituent interaction in this system: of the thirty structural features significant at the 10% level, half are substituent pair relative positions, and all except one of these represent *meta* interaction. The major trends in these results appears reasonable on the basis of known electronic factors (ROTH *et al.*, 1969). Amino derivative substituents in the 2 and 4 positions increase pKa markedly, while halogen and nitro groups reduce pKa, particularly in the 5 position. OMe and SMe substituents exert a positive effect on pKa in the 4 position and a negative effect in the 2 position. Of the substituent interaction terms amino substituents *meta* to one another show a negative effect on pKa, reduced when the substituents are separated by a heteroatom, and presumably indicating a mutual inhibition factor. Amino substituents *meta* to halogens, OMe and SMe show a strong interaction reducing pKa, except when the amino substituent is separated by a heteroatom from OMe, when a positive effect on pKa is observed.

When these results are compared with the coefficients for the same structural feature set, analysed at the 99.99% level, it appears that the same general conclusions may be drawn from each. The 10% level analysis, including only half the number of structural features, is less complex and hence easier to interpret; the 99.99% level analysis on the other hand allows for the observation of more trends in individually insignificant coefficients. The results of the two analyses are not in total accord: thus the Br-ortho-OMe structural feature which has a significant positive coefficient in the 10% level analysis appears to have a negligible effect in the analysis involving more variables. Some variation of this kind in particular coefficient values in different analyses is to be expected, and acts as a caution against placing undue emphasis on single coefficients in interpreting the results of such correlations.

Diverse subset

In order to deal with this subset of derivatives of nine ring systems the approximations introduced by the use of structural features representing the simplest form of relative position were accepted. To use, for example, explicit substituent positions or more accurately specified interaction terms as applied to particular ring systems would have introduced so many variables as to defeat the purpose of a generalised analysis.

Four sets of structural features were used, including:

K: number and type of substituents only (including heteroatoms and fused rings)

L: positions of substituents (including ring fusion points) relative to heteroatoms

M: as L, plus positions of substituents relative to ring fusion points

N: as M, plus positions of substituents in relation to one another

Examples of structural feature derivation are given in Figure 25.

The overall results of the regression analyses are given in Table 63. All the correlations are significant at the 1% level.

Comparison of the regression by the F-test shows that, compared with the analysis using set K structural features, no significant improvement is brought about by the use of structural feature set L, and an improvement only at the 10% level by the use of set M. The analysis with set N structural features brings about no further significant improvement.

In view of the known general similarity of substituent effects in a variety of nitrogen heterocyclic systems (ALBERT, 1963, ALBERT, 1971), this suggests that the approximations involved in the use of simple relative position terms for groups of diverse structures are too great to allow highly significantly improved calculations.

Total set

The subsets of pyridines, pyrimidines, and diverse structures were combined, giving a total of 169 structures. This set was analysed using the same types of structural features as for the subset of diverse structures, so as to allow for the different ring systems in this combined group of structures.

The sets of structural features used were as follows:

O: number and type of substituent only (including heteroatoms and fused rings)

P: positions of substituents (including ring fusion points) relative to heteroatoms

Q: as P, plus positions of substituents relative to ring fusion points

R: as Q, plus positions of substituents in relation to one another

The results of the regression analyses are summarised in Table 64. The correlations with structural feature sets P and Q are not significantly better than that with set O, and the analysis with set R is superior only at the 10% level (all the analyses being carried out so as to include as many features as possible). An analysis with set R, excluding structural features insignificant at the 10% level, was significantly better at 5% than the analysis using structural feature set O.

Structural features, with regression coefficients etc., for the analyses with set O structural features and with set R structural features at the 10% level are listed in Tables 65 and 66 respectively.

The results from the set O analysis reflect accurately, at a simple level, the effects of substituents upon pKa, which are largely as might be expected from the known electronic properties of these substituents (ALBERT, 1963). Thus ring nitrogen atoms have a negative coefficient, reflecting the base-weakening effect of multi-heteroatomic substitution. Amino, methylamino, and dimethylamino substituents are base-strengthening, as is the methyl group, to a lesser extent, while halogens and nitro substituents are strongly base-weakening. OMe and SMe substituents show a weak effect, due to an averaging out of their contributions at different positions observed in the more detailed analyses described above. The presence of a fused ring has a small overall effect upon pKa, which may again be partly due to an averaging effect.

The results listed in Table 66 show some aspects of the effects of substituents upon pKa in finer detail. Thus the effect of relative position of heteroatoms upon the base-weakening effect of the introduction

of more than one ring nitrogen is shown, as are the base-strengthening due to amino-type substituents ortho and para to the heteroatom. For the other substituent groups the position relative to ring nitrogen for their more reliably assessed effects are shown, for example the negative coefficients of the OMe and SMe groups in the ortho position, probably due to a large inductive effect. The intersubstituent interaction terms included, and therefore statistically significant for the whole set, are almost entirely those noted in the pyrimidine subset, emphasising the importance of substituent interaction the pyrimidine system. The representing ring fusion has only a relatively small negative coefficient, again showing the small effect of ring fusion on pKa, while the only structural feature included,representing interaction between substituent and fused ring,demonstrates the base-strengthening effect of the amino group ortho to a ring fusion.

Discussion of results

The work described above demonstrated that substructural analysis techniques can be used to produce structural features which allow detailed analysis of data sets containing heterocyclic compounds. It further indicated that such analyses can produce reliable and potentially useful results.

It has shown that a more detailed analysis is possible for a set of derivatives of a common parent ring system than for a set of diverse ring systems. An inverse relationship exists between the generality of application of a substructural analysis technique of this kind and the specificity of the structural features which may be derived. Although it is advantageous from this point of view to deal only with closely related structures at one time,it has been demonstrated that an analysis in considerable depth for derivatives of a number of ring systems is

possible, using simple procedures, although it may often be the case, as in the examples here, that highly significant improvements in correlation are not brought about, as discussed above. A more complex procedure for structural feature derivation, perhaps based on a minimum spanning tree algorithm, could be useful in making possible further generalisation of such analyses. Alternatively ring systems could be fragmented into individual rings, or still smaller units, by relatively simple procedures (WILLETT, 1976).

From the analyses on this data set, as with other analyses described in this thesis, it appears that in many cases the significantly better correlations are to be obtained by using the more detailed structural feature sets, in this case those including interaction terms. This raises the problem of the very large number of variables which may be included in such analyses. One potentially useful method of reducing the number of variables is to carry out the regression analysis in such a way that variables insignificant at a particular confidence level are omitted from the calculation. This has been demonstrated in the work above, using the 10% significance. If the aim of the analysis is the investigation, perhaps in a qualitative sense, of the factors involved in a structure-property relationship, it may be advantageous to use the analysis of highest statistical significance, even if this contains a very large number of variables. Useful information may be gained in this way, provided that the results are interpreted in terms of trends among the coefficients, which may individually be statistically insignificant. Analyses carried out with omission of less significant variables may be useful in that they concentrate on the more important factors.

If the aim of the analysis is quantitative prediction of unknown property values, the complex structural feature sets, even if giving

significant better correlations, may be of limited use, as noted in the preceding section. Because of the greater specificity of structural feature description in complex sets, it is likely that in many cases coefficient values for the features in a structure not included in the analysis will not be available, either because these structural features do not occur in any structure in the analysed set, or because they are perfectly correlated with other structural features. The extent of perfect correlation almost always increases with increasing complexity of structural feature sets. The simpler feature sets, with which this problem is less likely to arise, may be more useful for predictive purposes.

Boiling Points of Alicyclic Structures

The aim of the work described in this section was to determine whether good correlations, of potential usefulness, could be obtained by substructural analysis procedures for a property crucially dependent on the three-dimensional structure of molecules.

As noted in chapter 2, previous empirical correlation studies of this sort of property have relied largely upon calculated or experimentally determined molecular dimensions for derivation of appropriate variables. The aim here was rather to make use of features derived from representations of the structure diagram. Two possibilities then arose.

Firstly, the use of simple, non-stereochemical structural features could give a relatively crude account of three-dimensional structure. At the simplest level, atom and bond counts will give a rough indication of molecular size. It was thought possible that this sort of implicit representation of stereochemical factors could possibly give a reasonable correlation for some properties.

Secondly, the facilities in Wiswesser Line Notation for describing stereochemistry might enable the derivation of structural features explicitly representing aspects of three-dimensional molecular structure. Such features could be useful in correlation of particular molecular properties.

The property chosen for investigation was boiling point. This property has been measured for many organic substances, because of its importance for structure-determination, as noted earlier.

A recent study of the boiling points of several series of cyclic structures made use of structural features representing

stereochemical factors to correlate structure with boiling point (KELLIE et al, 1975). It was decided to reinvestigate a part of the data used in that study, comprising boiling points of methyl derivatives of cyclohexane and 1,3-dioxan. For liquids of this sort, boiling point is closely related to molar volume. This then seemed a suitable data set for testing both explicit and implicit representation of three-dimensional structure, by structural features derived from WLN: particularly since the findings of the original investigators, who used structural features representing the equatorial and axial nature of methyl substituents, were available for comparison.

The data set used in this work, comprising 29 derivatives of cyclohexane and 31 derivatives of 1,3 dioxan, is listed in Table 67.

These structures were coded in WLN, according to the tentative rules for describing cis-trans isomerism in such systems. Since all these compounds are known to exist in the chair form (KELLIE et al., 1975), and since only one type of substituent is present, it was relatively straightforward to devise an algorithm to derive the most stable conformation for each compound from its WLN, according to the principles of conformational analysis (ELIEL, 1962c). Structural features could then be generated to represent the equatorial and axial methyls, and their relative positions: geminal (gem), ortho, meta, and para. The correlations obtained using structural feature sets of this sort were compared with those using sets which did not include the equatorial-axial distinction, and therefore could only represent stereochemical factors implicitly.

With the derivatives of 1,3 dioxan there was the additional problem of representing substituent position relative to the heteroatoms. As will be seen from the structure of 1,3 dioxan, Figure 26, this situation is exactly analogous to the pyrimidines discussed in the

preceding section. There are two forms of meta interaction: 2,6 and 2,4, including a heteroatom, and 4,6 including a carbon atom, between the substituents.

It may be noted that these analyses also enabled the investigation of the usefulness of terms representing relative position of substituents for alicyclic systems, since the applications of such structural features described above were all concerned with aromatic rings.

Full details of the computer programs for generation of this type of structural feature set are given in the Appendix.

A total of eleven sets of structural features were used in the analysis of the relationship of boiling point with structure for cyclohexanes. These were devised to investigate the effects of equatorial and axial substitution, and relative position of substituents on the property. These sets were:

A - including on the number of methyl substituents

B - as A, also including the number of geminal interactions

C - as B, also including the number of ortho interactions

D - as C, also including the number of meta interactions

E - as D, also including the number of para interactions

F - including the numbers of equatorial and axial methyl substituents

G - as F, also including the numbers of geminal interactions

H - as G, also including the numbers of ortho interactions

I - as H, also including the numbers of meta interactions

J - as G, also including the number of para interactions

K - as I, also including the number of para interactions

Examples of structural feature derivation are given in Figures 27, 28 and 29.

The overall results of the regression analyses using these structural feature sets are shown in Table 68. All correlations are significant at the 1% level.

The statistical significance of the differences in overall correlation were assessed by the F-test. For those structural feature sets which made no distinction between equatorial and axial methyl groups, set B gave a correlation superior at the 10% level to that with set A, and set C superior at the 5% level to set B, while no

significant improvement was gained by the use of sets D or E.

Set F, making the equatorial-axial distinction, gives a worse correlation than set A, although the difference is not significant. However, when terms representing relative positions of substituents, are included, structural feature sets including the equatorial-axial distinction give significantly better correlation than the corresponding sets without this distinction. Thus, for example, sets H and K give correlations better at the 1% level than sets C and E respectively.

For the structural feature sets including the equatorial-axial distinction, set G gives a superior correlation to set F, and set H to set G, both at the 1% level. The correlations with set I and J are not significantly different to that with set H, while the correlation with set K is superior at the 5% level to that with set H.

These results suggest that the structural factors affecting boiling points to the greatest extent are equatorial-axial substitution; geminal substitution, and ortho interactions. This is in accordance with the findings of the original workers (KELLIE et al., 1975). Two other points, however, may be noted.

Firstly, a good correlation is obtained by considering solely number of methyls, indicating that in this case the implicit account of three-dimensional structure is reasonably effective.

Secondly, the inclusion of structural features representing meta and para interactions, not considered by the original authors (KELLIE et al., 1975), gives a correlation improved at the 5% level, when the equatorial-axial distinction is made.

The structural features of set K are listed in Table 69, together with their regression coefficients and t statistics. The contributions to boiling point of equatorial and axial methyls and geminal and ortho interactions here are very similar to those found by Kellie

and Riddell (KELLIE et al., 1975). It is notable that the effects of ortho interactions are opposed to those of meta and para interactions, indicative of the steric, rather than electronic, factors influencing this property.

In the analyses of the 1,3 dioxan derivatives, structural feature sets were derived so as to investigate the effects of substituent position relative to the heteroatoms. A total of thirteen structural feature sets were used:

set L - including only number of methyl substituents

set M - including number of equatorial and axial methyl substituents

set N - as M, including also the number of gem interactions

set O - as N, including also the number of ortho interactions

set P - including the positions of methyl substituents (2,4(or6), or 5)

set Q - as P, including also the numbers of gem, ortho, meta, and para interactions

set R - as Q, distinguishing between 2,4(or 2,6) and 4,6 meta interactions

set S - including the positions of equatorial and axial methyl substituents

set T - as S, including also the number of gem interactions

set U - as S, including also the positions of gem interactions

set V - as T, including also the numbers of ortho interactions

set W - as V, including also the numbers of meta and para interactions

set X - as W, distinguishing between types of meta interaction.

Examples of structural feature derivation are given in Figures 30 and 31.

The overall results of the regression analyses are listed in Table 70. All the correlations are statistically significant.

Much the same factors are seen in the comparison of the correlations with various structural feature sets as were noted with

the cyclohexanes. Introduction of the equatorial-axial distinction to feature sets including only number or position of methyl groups brings about no significant improvement, in sets M compared with L, and S compared with P. However for feature sets including substituent interactions, W compared with Q, inclusion of the equatorial-axial distinction brings about an improvement significant at the 1% level.

The inclusion of terms representing number and position of geminal interactions bring about no significant improvement; sets T and U compared with S, and N compared with M. Inclusion of ortho interactions improves the correlation at the 5% level, (set O compared with N) or 1% level (set V compared with T). Inclusion of meta and para interactions, when the equatorial-axial distinction is made, gives a correlation significantly improved at the 5% level: set W compared with V.

Inclusion of position of the methyl substituents relative to the heteroatoms brings about an improvement in correlation significant at 5%, set P compared with set L, or 1%, sets S and V compared with M and O respectively. This indicates that the position of substitution is of importance in determining the boiling point, as is to be expected, since the alignment of these dipolar molecules will be affected by the position of substitution (KELLIE et al., 1975).

Distinction between the two types of meta interaction, sets R and X compared with Q and W respectively, gives worse correlations, indicating that this distinction is inappropriate to this data set. This is a notable contrast to the situation with the pKa data discussed in the preceding section, and probably reflects the predominantly steric, rather than electronic, factors affecting boiling point.

The structural features of set W, together with their regression coefficients and t statistics, are listed in Table 71.

These values are broadly in line with those of the original investigators (KELLIE et al., 1975). Particularly noticable are the differences in the coefficient values for equatorial and axial substitution in the 2 and 4 positions. Some coefficient values for interaction terms are not statistically significant, but they follow for the most part the trends observed with cyclohexanes, i.e. ortho interactions positive, the remainder negative, geminal strongly so.

Finally these two data sets, cyclohexanes and dioxans, were combined to give a set of 60 structures, with a range of boiling point values of 192 degrees. An analysis was carried out using a set of 20 structural features including the equatorial-axial distinction and all substituent interactions. The ring oxygen atoms of the dioxans were treated as ring substituents to which the positions of the methyl groups were related. 19 structural features were included in this analysis, which gave a correlation with R = 0.990, r = 3.14, F = 103.69 (40 degrees of freedom). The structural features of this set, together with their regression coefficients and t statistics, are listed in Table 72. They are in general in accordance with the results described above, although for the reasons discussed for pyrimidine structures in the preceding section, detailed interpretation of coefficient values for positions relative to heteroatoms is difficult. This result shows the feasibility of studying diverse structural types with features representing three-dimensional structure.

The work discussed above demonstrated that substructural analysis techniques can give good correlations for a property crucially dependent upon three-dimensional structure. Reasonable results may be obtained using relatively crude structural features which reflect stereochemical factors implicitly. Significant improvements in correlation are brought about by the use of explicit representation of steric factors which, in this case, may be readily achieved using WLN. Significant improvements are also brought about by the use of terms representing relative position of substituents, indicating that such structural features may be usefully applied to a variety of alicyclic as well as aromatic structures.

Chapter Four, Part 2

Cluster Analyses

Classification, the grouping of entities according to their similarities, has for long been an important aspect of most sciences (SOKAL, 1974). This is certainly the case in chemistry, particularly organic chemistry. In this field,classifications based on structural ideas have been found useful from the earliest development of the subject (FISHER, 1973a, FISHER, 1973b). This is reflected in the very common practice of dividing the subject matter of organic chemistry texts and monographs into sections corresponding to structural types, usually according to the presence of functional grouping, ring systems etc.

Such classifications are almost always monothetic, i.e. based upon the presence or absence at each stage of a single attribute. Relatively little use has been made of more recently development methodologies of automatic classification, which rely upon a computerised analysis of similarity based upon a number of attributes, in this case structural features. As noted in an earlier chapter, classificatory procedures, both supervised and unsupervised, have been applied to problems of structure-property correlation.

One potentially useful procedure involves the algorithmic generation of structural features from a computer-readable structural representation, which may be subsequently input to an automatic classification procedure. The feasibility of this type of substructural analysis has been demonstrated, using a connection table structural representation, and using cluster analysis techniques (ADAMSON et al., 1973a, ADAMSON et al., 1975a, BUSH, 1976). The aim of the work described below was twofold: firstly to investigate the use of structural features derived from WLN for classification, and secondly to investigate various cluster analysis procedures of potential applicability to chemical structure classification.

A procedure of this sort has two possible applications. The first is information retrieval; providing an alternative to the usual "present/absent" substructure search. The second is structure-property correlation, where the classification is presumed to be indicative of attributes, i.e. property values, not included in the analysis. As noted in an earlier chapter, this application could be useful where accurate quantitative data is not available.

The evaluation of the success of classification techniques presents particular problems. A classification displays relationships between objects, and there are in general many possible relationships which could be considered, and many different ways of ordering the relationships in their assumed importance to the overall classification. Taking a group of chemical structures as an example, possible relationships between them could consider the size, i.e. number of atoms, diversity, i.e. number of types of atoms or groups, the presence/absence/number of any atoms, functional groups, or ring systems, the relative positions or the environments of any groups, or any combinations of these factors. There is no "correct" classification of any group of entities, and the value of any particular classification depends upon its usefulness in practice. The correlation between chemical structure and property achieved by means of classification has been used to assess the success of that classification (SNEATH, 1966, ADAMSON et al., 1973a, ADAMSON et al., 1975a, BUSH, 1976). Since, however, it appears that the use of regression analysis will generally yield more accurate results than classification where accurate data is available (BUSH, 1976), it may well be that the major applications of cluster analysis and related techniques will be for information retrieval, or for obtaining a qualitative overview of a data set. A formal evaluation of any technique would be likely to be either prohibitively time-consuming, or relatively meaningless in the absence of any practical

situation. It was therefore decided to evaluate the results of the work described below on the basis of an intuitive judgement of the adequacy of the classifications. This was thought to be sufficient to give a general impression of the success of the classificatory procedures.

Data Sets (Cluster Analysis)

The purpose of this work, as noted above, was an investigation, in general terms, of the usefulness of various cluster analysis methods with the type of structural features readily derived from WLN. It was decided that this would as well be achieved using "artificial" data sets, i.e. sets of structures chosen for this purpose, as with sets of "real" structures, i.e. from literature data, as used in the correlations described in earlier sections.

Four sets of structures were investigated. The first two, comprising relatively simple aliphatic structures and benzene derivatives respectively, were used to assess classifications based on simple WLN fragments.

A further set of benzene derivatives was used to examine classifications taking account of substituent positions.

Finally the set of 44 benzene derivatives, for which the correlation of structure with reaction kinetics data was discussed in an earlier section, was classified, in order to attempt clusterings based on structural features reflecting substituent positions relative to a specified position, in addition to relative intersubstituent positions.

For each of these data sets clustering was carried out using the techniques described above. The results are discussed below for each data-set in turn, and finally summarised.

It was found that in the majority of cases, the centroid and median clustering methods were unsatisfactory. Frequently, with these methods, the centre of a newly formed cluster lies within the boundaries of an existing cluster. Although mathematically valid, this situation

results in a meaningless dendrogram, with overlapping lines; so that interpretation of the classification in the way proposed here is not possible. These methods will not therefore be further discussed.

Structural features were derived algorithmically from WLN representation of structure, and the cluster analyses carried out using the CLUSTAN package (David Wishart, University of St. Andrews).

Aliphatic Structures (cluster analysis)

The structures are shown in Figure A1. There are 40 aliphatic structures, comprising straight-chain and branched hydrocarbons, primary and tertiary amines, ketones, and alcohols. Bifunctional and mixed functionality compounds are included.

The structural feature set used is shown in Figure A2.

The classifications are displayed as noted below.

Clustering method	Raw data	Standard data
Single link	AA	AF
Further neighbour	AB	AG
Group average	AC	AH
Ward	AD	AI
McQuitty	AE	AJ

Single link clustering gives a generally intuitively sensible classification, with both raw and standardised data, but the strong "chaining" effect, with structures joining the main group individually, prevents the emergence of well-defined clusters. This drawback is more pronounced with raw data.

The remaining four techniques using raw data give intuitively sensible classifications with well-defined clusters. The clusters are based on both functional groupings and carbon skeleton. Thus, in all cases, structures 11 (an alcohol) and 33 (containing both alcohol and ketone functions) are brought together at a very high similarity level.

These four clustering methods again give sensible, and potentially useful, classifications with standardised data. The well-defined clusters are here based largely on functional group. Certain differences in the detail of clustering may be noted: thus for example compound 40 (containing both primary and tertiary amine functions) is placed by

McQuitty's analysis as most similar to compound 23 (with two primary amine groups) and by the other three methods as most similar to nos. 37 and 38 (with two tertiary amine functions).

Benzene Derivatives (Cluster Analysis)

A set of 36 benzene derivatives was examined, varying from unsubstituted benzene to hexasubstituted. Substituents were Me, OMe, F, Cl, Br, NMe_2, and two types of substituent were present in some structures. The structures in the set are shown in Figure B1. The structural features used represented type and number of substituents.

The classifications are shown as noted below.

Clustering method	Raw data	Standard data
Single link	BA	BF
Furthest neighbour	BB	BG
Group average	BC	BH
Ward	BD	BI
McQuitty	BE	BJ

All the classifications with raw data show a greater degree of chaining than was the case with the aliphatic structures. The single link classification is virtually completely chained, and thereby valueless.

The remaining methods give apparently sensible classifications based on both type and number of substituent. Thus all four methods bring structures 5, 6, 7, 21, 22 and 23 (i.e. with several Me substituents) together as a cluster. Again differences of fine detail are apparent: the furthest neighbour algorithm brings together numbers 8, 9, 25, 26 and 27 (which all contain, inter alia, OMe groups), whereas McQuitty's method places 25 with the other Cl containing structures, 26 with the Br s etc. Group average and Ward's method give a classification intermediate between these two.

With the standardised data, the classifications obtained were based primarily on substituent type. Single link gave a somewhat chained, though still sensible clustering, while the other methods showed regular well-defined clusters. Furthest neighbour analysis differed somewhat from the other methods: for example compound number 21 (iCl, 4Me) was classed as most similar to the other mixed halomethyl structures by furthest neighbour, rather than most similar to the other chloro structures, as by the other methods.

Positional Isomer Benzenes (Cluster Analysis)

This set of structures, shown in Cl was devised so as to investigate classifications based on structural features representing substituent interaction. These structures include all the possible isomers of di- and tri- fluoro, chloro, bromo, and iodo benzenes.

Three sets of structural features were considered. Firstly, only structural features representing number and type of substituents, i.e. F, Cl, Br, and I were used. Secondly, features reflecting relative positions of substituents were included: this structural feature set is shown in CZ. Thirdly, relative positions of substituents, regardless of substituent type, were used: the feature set consisting of ortho, meta, and para interactions.

The classifications using the first structural feature set are set out as below:

Clustering method	Raw data	Standard data
Single link	CA	CF
Furthest neighbour	CB	CG
Group average	CC	CH
Ward	CD	CI
McQuitty	CE	CJ

It is particularly notable that all five clustering methods, with both raw and standardised data, show exactly the same pattern of clusters (though with different similarity levels). Classification is firstly by substituent type, and secondly by number of substituents. The anomalous position of structure 1 is presumably an artifact of the clustering procedures.

The classifications using the second structural feature set

are shown as below:

Clustering method	Raw data	Standard data
Single link	DA	DF
Furthest neighbour	DB	DG
Group average	DC	DH
Ward	DD	DI
McQuitty	DE	DJ

With raw data the five methods produce an identical clustering pattern. Classification is firstly by substituent type, secondly by substituent number, and thirdly by substituent position.

Variation in the classifications produced by the different clustering methods is observed with standard data. Group average, Ward's method, and McQuitty's method give classifications based firstly on substituent type, secondly on number and position of substituents. With single link and furthest neighbour, on the other hand, there is no initial substituent type division: rather similar substituent patterns are grouped together.

The classifications using structural features representing only relative position of substituents, regardless of substituent type are shown:

Clustering method	Raw data	Standard data
Single link	EA	EF
Furthest neighbour	EB	EG
Group average	EC	EH
Ward	ED	EI
McQuitty	EE	EJ

These classifications, with both raw and standardised variables, reflect only substitution pattern. The interrelations shown between substitution patterns differ between the various methods. Although no

method shows a clear-cut discrimination between di- and tri-substituted structures, a tendency to divide on the basis of number of substituents is seen with unstandardised variables. With standardisation the occurrence of ortho, meta, and para interactions are clearly dominating the clustering: thus the 1,4 and 1,2,4 substitution patterns (i.e. those including para substituents) are brought together.

Benzene Derivatives - Reaction Rate Set (Cluster Analysis)

The set of multisubstituted benzene derivatives, for which the correlation of structure against rate of bromination was described in an earlier chapter, were used in structure-based cluster analyses. The structures are listed in Table 13.

Three structural feature sets were used for the classifications:

a) number and type of substituent only

b) including positions of substituents relative to the reaction site

c) as b, including also relative positions of substituents

The analyses were performed using the group average method, Ward's method, and McQuitty's method, since these had given the most interesting clusterings with previous sets of structures. Both raw and standardised data were used.

For the structural feature set including only number and type of substituents the clusterings are shown thus:

Classification method	Raw data	Standard data
Group average	FA	FD
Ward's method	FB	FE
McQuitty's method	FC	FF

All these analyses show generally similar classifications, which reflect in an intuitively sensible manner the composition of the data-set. One major distinction between analyses using raw and standard data is apparent for all three clustering methods. Classifications with raw data make an initial division between structures with three or more substituents and those with less, whereas with standardised data no such breakdown is observed.

Dendrograms for cluster analyses using structural features representing substituent position relative to the reaction site as shown:

Classification method	Raw data	Standard data
Group average	FG	FI
Ward's method	FH	FJ

These classifications all appear reasonably sensible. There is relatively little difference from the corresponding classifications where substituent positions were not considered. The inter-relations of some sets of isomers, e.g. the methoxy-methyls (compound numbers 17-20 and 34-39), is noticable. With standardised data, where no marked division by number of substituents is seen, there is little structure in the classification.

The classifications obtained using structural features representing relative position of substituents as well as their position relative to the reaction site are shown:

Classification method	Raw data	Standard data
Group average	FK	FM
Ward's method	FL	FN

The classifications are in general a reasonable reflection of the data-sets. Ward's method gives more distinct clusters than does the group average method. With standardised data Ward's method gives a particularly good representation of interrelationship between isomers.

Summary of Classifications

The studies of automatic classifications of chemical structures using WLN fragments described above achieved their limited objective of demonstrating that such classifications can display relationships between structures in a manner in general accord with chemical intuition. These classifications were able to include relative positions of substituent groups.

Various clustering procedures were shown to yield different, but all potentially useful, classifications. Group average, Ward's method, and McQuitty's method gave particularly sensible classifications, with well-defined clusters, and are therefore possibly most useful for practical application. However the single link method, although it gave classifications of little practical value in several cases here, should not be ignored. It has the advantages of computational simplicity, and hence the potential for larger-scale application, and a firm mathematical basis, and may be less likely than other methods to impose an arbitrary form onto a data-set.

The use of standardised as well as raw data may also be advantageous, since it appears to tend to give classifications based on type, rather than number, of structural features.

Chapter Five

Conclusions

'Thús slowly, one by one, its quaint events were hammered out'

(Lewis Carroll)

Three aspects will be considered in concluding this thesis. Firstly, the work described will be summarised, and attention drawn to some of its more important features. Secondly, the potential for practical application of these methods will be discussed. Thirdly, some indications of possible further research in extending this work will be made.

Summary of Thesis

The work described in this thesis demonstrates the applicability to various sets of structure and property data of a form of substructural analysis using structural fragments corresponding to chemically significant groupings of atoms. Such fragments are particularly suitable for representing functional groupings, ring systems, ring substitution patterns, stereochemical features; and all these aspects have been investigated in the work described above.

Algorithmic procedures for automatic generation of structural fragments of this sort for standard WLN representation have been devised. Their implementation by relatively simple computer programs, capable of dealing with a wide variety of types of compound, has been described.

Two techniques of statistical analysis have been used in this work: multiple regression analysis and cluster analysis.

Multiple regressions, with some molecular property as independent variable and the occurrence of structural features of the sort noted above as dependent variables, have shown statistically significant correlations with a number of data sets, with widely differing structures and property values. In those cases where a direct comparison has been made, the use of relatively simple fragments givesvery similar correlations to those using alternative sorts of structural fragments, derived from connection tables. In some data sets, particularly with multisubstitution on ring systems, it has been shown that significant improvements in correlation are obtained by using structural features representing the relative position of substituent groups, heteroatoms etc. It has been shown that these improvements are not due simply to the increased number of variables, nor to fortuitous data set characteristics, but rather to the necessity to include

"interaction" terms for an adequate account of the variation in the data. Rationalisations, in physico-chemical terms, have been possible in most cases. It may be that the ability to deal with such potential non-additive effects, by automatic generation of terms to represent the many possible intra-molecular effects, will be a major contribution of this type of substructural analysis.

The use of cluster analysis with sets of chemical structures, using structural features of the sort described above, demonstrated that sensible and potentially useful classifications can be produced with this type of structural feature. The inclusion of relative position terms for ring substitution resulted in classifications reflecting these factors, and showing relationships between structures at a more detailed level than on the basis of occurrence of simpler structural features. This is a further example of the value of this type of structural fragment. Thus, by the choice of structural feature type, different, though all equally "correct", classifications of a set of structures can be produced. The flexibility of an automated substructural analysis procedure may be used to advantage in this kind of situation.

The availability of the CLUSTAN package enabled a comparison, for the first time, of different clustering methods with chemical structures. No single "best" method could be identified (although some were considered of little value): rather it was concluded that the use of a number of clustering methods on any set of structures could be advantageous, in highlighting different ways of viewing the relationships between structures.

The demonstrated usefulness of chemically significant structural features in the two types of analysis investigated here suggests that they might be also of use with other data analysis methods, e.g. discriminant analysis or principal components analysis. In general,

whatever statistical procedure is used;the substructural analysis methodology has the advantage of allowing a rapid, systematic and exhaustive generation of structural features at any level of detail which may be required. The use of fragments corresponding to chemically significant moieties has been shown to be of value in the types of analyses investigated here: it may well be that it will prove equally valuable with other statistical techniques. Fragments of this sort have, in all cases, the further advantage that an analysis may be planned, and the results subsequently interpreted, in conventional chemical terms.

This methodology also has considerable advantages as regards data-handling system requirements. The procedures discussed here use a WLN representation of structure, widely used in computer-based information systems, together with relatively simple programs for fragment generation. It is likely that only trivial amendments need be made to the existing software capabilities of many chemical information systems in order to implement a fragmentation method of this sort. Although the structural feature selection can be done by hand, as was done for the initial tests of the work described in this thesis, the ability to perform this task automatically gives the user a convenient, rapid, and flexible means of routinely dealing with large data sets.

One notable feature of the work described here is the variety of data examined by multiple regression analysis, including biological, physicochemical, and thermodynamic molecular properties. All the data-sets used consisted of relatively accurate, quantitative measurements, because of the necessity for this sort of data, in order to compare statistically results of regression analyses with different structural feature sets. There is of course no reason why fragmentation techniques as described here should not be used with semi-quantitative or qualitative data, with an appropriate analytical technique.

Potential Practical Application

In the discussion above it was suggested that the substructural analysis procedure under consideration here, i.e. a derivation of chemically significant fragments from a structure diagram or computer-readable equivalent followed by an appropriate form of data analysis, could have useful practical application. This will now be further considered, under two headings: applications in structure-property correlation, and applications in information retrieval. It may be useful to note that there can be no rigid distinction between these two. If the questions "which compounds contain these substructures" and "which compounds have this activity" are considered (as would usually be the case) as problems in information retrieval, then it seems sensible to consider the question "which substructures contribute most to this activity" as being in the same area. In this case however, it will be useful to make a loose distinction.

i) Structure-property correlation

The work in this thesis has shown that these techniques can be usefully applied to a range of molecular properties, and therefore their applicability in a number of fields must be considered.

The computerised estimation of thermochemical properties, based on group additivity schemes, is well suited to the application of automatic structure-handling techniques. As noted in the discussion in an earlier section, proposals to this effect have been made by a number of workers in this area. The multiple regression analyses of thermochemical data described in this thesis demonstrate, on a small scale, the applicability of a WLN-based substructural analysis procedure in this field. This method is currently being investigated by a research group active in this area (Personal communication from Dr. J. B. Pedley).

Another area of potential applicability is physical organic chemistry. The work with reaction rate data and pKa values described in earlier sections indicates the possible usefulness of substructural analysis in such problems. Two kinds of application could be envisaged: firstly as an aid to the analysis of large data sets as carried out by Wold for reaction kinetics data (WOLD et al., 1972), and secondly as a means of detailed, systematic treatment of intramoleculer interactions, as in the work in this thesis. The need for automation in this area is not so evident, largely because of the relatively simple compounds involved in mechanistic studies.

One particularly important potential application within the area of physicochemical properties could be the calculation of molecular properties used for semi-empirical calculations, such as partition coefficients and molar refractivities. Such properties are reasonably approximated by an additive model, and hence may be correlated

with structure by regression analysis, and unknown values predicted, as evidenced by the work of Nys and Rekker (NYS et al., 1974).

The application of substructural analysis to the correlation of partition coefficient with structure, described in an earlier section, illustrates how readily the technique fits in with existing practice. It should be noted that in this case, as in many others, the structural fragments most readily derived from WLN correspond very closely with those assigned manually on purely chemical considerations.

Automated data analysis procedures could be helpful here in two ways:

i) by making possible the rapid and convenient analysis of large amounts of computer-readable data, e.g. the Hansch data-base,

ii) by making possible the rapid and convenient analysis of a data-set using a variety of sets of structural features, thus investigating the effect of including terms representing intra-molecular interactions. This may be particularly useful in complex sets of structures with multi-substitution or multi-functionality.

The correlation of structure with biological activity, important in areas such as drug design, offers further opportunities for useful application of substructural analysis techniques. The situation here is more complicated than with the physical properties mentioned above, because of the possibility of several mechanisms of action involved in any biological property. Also the data available is frequently relatively crude, often semi-quantitative or qualitative.

It is likely that substructural analysis can most usefully

be used here as one part of an approach to these problems using a variety of methodologies. In particular the ability of the sort of methods discussed here to deal rapidly and flexibly with large amounts of structural information could make them well suited for an initial analysis of large and/or complex data-sets. Data analysis techniques such as cluster analysis, discriminant analysis, or principal components analysis ,could be used in conjunction with an automated fragmentation procedure to gain an initial crude understanding of the relation of structure to property, which could then be further investigated by more precise techniques. This "first look" analysis in substructural terms could then be used for physicochemical and mechanistic rationalisations. Subsequently a more precise quantitative correlation could be formulated, perhaps using physicochemical property variables. The contribution of substructural analysis in such cases would be to assist in the visualisation of likely forms of structure-property relationships, by systematically dealing in detail with large and complex data-sets. In particular they could help to draw attention to the effect of co-occurrence and relative position of structural features, factors which are hard to take account of in any large set of structures without the aid of such an automatic procedure.

Another useful contribution of substructural analysis ,together with the classification methods ,could be to divide a large data-set into sub-sets, on the basis of structural similarities. The various sub-sets could then be treated individually by more precise correlation procedures. This sort of classification could also be useful in distinguishing between different pharmacological activities or modes of action.

One interesting aspect of structure-biological property correlation is the potential use of multiple regression analysis with

structural features of the sort discussed here. With a large data-set of the kind of complex structures frequently encountered in this area, a large number of structural features is generally produced, particularly if detailed aspects of structure, e.g. ring substitution patterns, are dealt with. This will often lead to violation of the guidelines for the maximum desirable ratio of variables to observations for regression analysis. Under these circumstances it may be advisable to regard the regression approach as a "non-parametric" treatment, on a par with cluster analysis, discriminant analysis etc. i.e. a representation of a complex data-set. A value would be obtained representing the contribution to property for each fragment, which could be used for rationalisation in physicochemical terms and formulation of a more precise equation; but the conventional measures of significance for the regression could not be confidently applied. This could be a useful compromise between the desirability of explicitly accounting for the many detailed aspects of the chemical structures in order to directly determine the effect of such factors as substituent interactions, and the contending requirement for a relatively small set of independent variables for a statistically sound correlation equation.

ii) Information Retrieval

The work described in an earlier section has shown that substructural analysis, using chemically significant fragments and a cluster analysis procedure, can give classifications of chemical structure which appear sensible in chemical terms. Classifications using such fragments can include factors such as substitution patterns in displaying relationships between compounds.

Such classification proceduresscould find application within computer-based information systems as a complement to existing substructure search facilities. The major limitation to the method as described here is the limited number of compounds, probably a few hundred, able to be dealt with in one classification by currently available cluster analysis programs. This limitation is primarily due to the computer storage requirement for the similarity measures for each pair of structures which must all be stored during the classification. Until the introduction of classification programs capable of dealing simultaneously with a larger number of structures, it seems that automatic classification techniques will be restricted to small-scale applications.

One such application could lie in the ranking of output from structure searches. Thus in answer to a request for "compounds of similar structure to A", a conventional substructure search system can produce only a list of compounds with some structural features in common with A. An advance on this would be to calculate a measure of similarity between each compound and A, and to present the results in order of similarity, i.e. presumed relevance. This of course is not a classification procedure, since it involves only the first part of

such a procedure, the calculation of similarity.

Another application could be a "browsing" capability, i.e. investigation of the composition of small files, or sub-files of structures. This could have the aim, for example, of selecting a "typical" structure as representative of the set, for further study.

Future Research

Future investigations in this area of substructural analysis could be directed into three aspects: fragmentation procedures, statistical analysis techniques, and the operation of complete systems, perhaps emphasising efficiency of operation and analysis on a larger scale.

The usefulness of different types of structural feature set in analyses of this type should be further investigated, in order to decide whether any general principles can be identified as to the most useful types of structural features. Such work would need to examine a wider range of structural types and property values than was possible in the studies described in this thesis. In particular further comparisons of the "chemically significant" types of structural fragments readily derived from WLN with other types, e.g. from connection tables, could be of value. It would also be desirable to investigate the differences between the optimal types of fragments for structure-property correlation and for information retrieval.

With regard to the statistical analysis techniques, the work on cluster analysis described in this thesis has established that sensible classifications of chemical structures can be obtained using WLN-type fragments. More detailed study of this topic would be of value: to investigate with "real" data the potential value for both structure-property correlation and information retrieval.

Another area which could be well worth investigation is the use of WLN-type fragments with other data analysis techniques: notably discriminant analysis, principal components analysis, and the various mapping procedures. Of particular value would be comparative evaluations of a number of data analysis techniques on the same set of structure-

property data, in order to establish the extent to which the techniques yield usefully complementary information. This could result in a greater understanding of the value of these techniques, used alone or in conjunction.

Investigations of the "scaling-up" of these techniques, to deal with much larger numbers of structures than attempted in the work described here, could also be valuable. Apart from the computer systems aspects of such work, a further problem worth investigating is the optimal size of data-set for structure-property correlation. The ability to deal with large numbers of compounds simultaneously, if only to divide them into sub-sets on a systematic basis, is certainly valuable: but it does not necessarily follow that all analyses should be carried out on the largest scale possible. It is possible in some cases that dealing with sub-sets of data can allow separation of different mechanisms of action, and thus simplify the interpretation of results. Also slight discrepancies between fragment values in different sub-sets may be found: e.g. between CH_2 groups in ketones and hydrocarbons noted in the heats of vaporisation correlations. In such cases more precise correlations could be obtained on data sub-sets, with, however, the sacrifice of the generalised diverse-structure applicability, valuable for "lead-seeking". These points could well be examined, using realistic data-sets as far as possible.

In conclusion, the work described in this thesis has demonstrated that this form of substructural analysis can yield potentially useful results. Further work along the lines mentioned above will be valuable, but major advances with these techniques must await their application on a routine basis in operational environments.

APPENDIX

WLN FRAGMENTATION PROGRAMS

The computer programs written for the work described in this thesis derived structural features from WLN structural representation, for subsequent analyses by standard commercial statistical programs. The rationale for the type of structural features, and hence the type of fragmentation used, has been discussed in an earlier chapter.

As noted in that discussion, the fragmentation procedures used in this work, being relatively simple, could be emulated by many of the widely-known structure-handling systems. These programs were written mainly to assist the data-handling for the correlations described above, and thereby to demonstrate the simplicity and flexibility of the automated procedures needed in this area. No attempt was made to generalise the applicability,or optimise the efficiency, of these programs beyond the level required for this work.

The description to be given of these programs should therefore be taken as purely illustrative of some simple techniques useful in this area, rather than as a documentation of a fully tested and optimised working system. For this reason the description will be brief.

The programs were written in ICL COBOL, and originally run on an ICL 1907E computer. Subsequently they were transferred, with minimal alterations, to an ICL 1906S machine.

The programs all required less than 20K workds of core storage. A disc file was used for intermediate storage of data, the capacity required depending upon the number of compounds being processed,but generally not exceeding 10K words. No accurate assessments of run times were made,but the programs typically used less than 100 seconds of CPU time in dealing with over 100 structures, depending upon the complexity of the fragmentation procedure.

A program listing and outline flowchart are given for the algorithm which fragmented diverse structures.

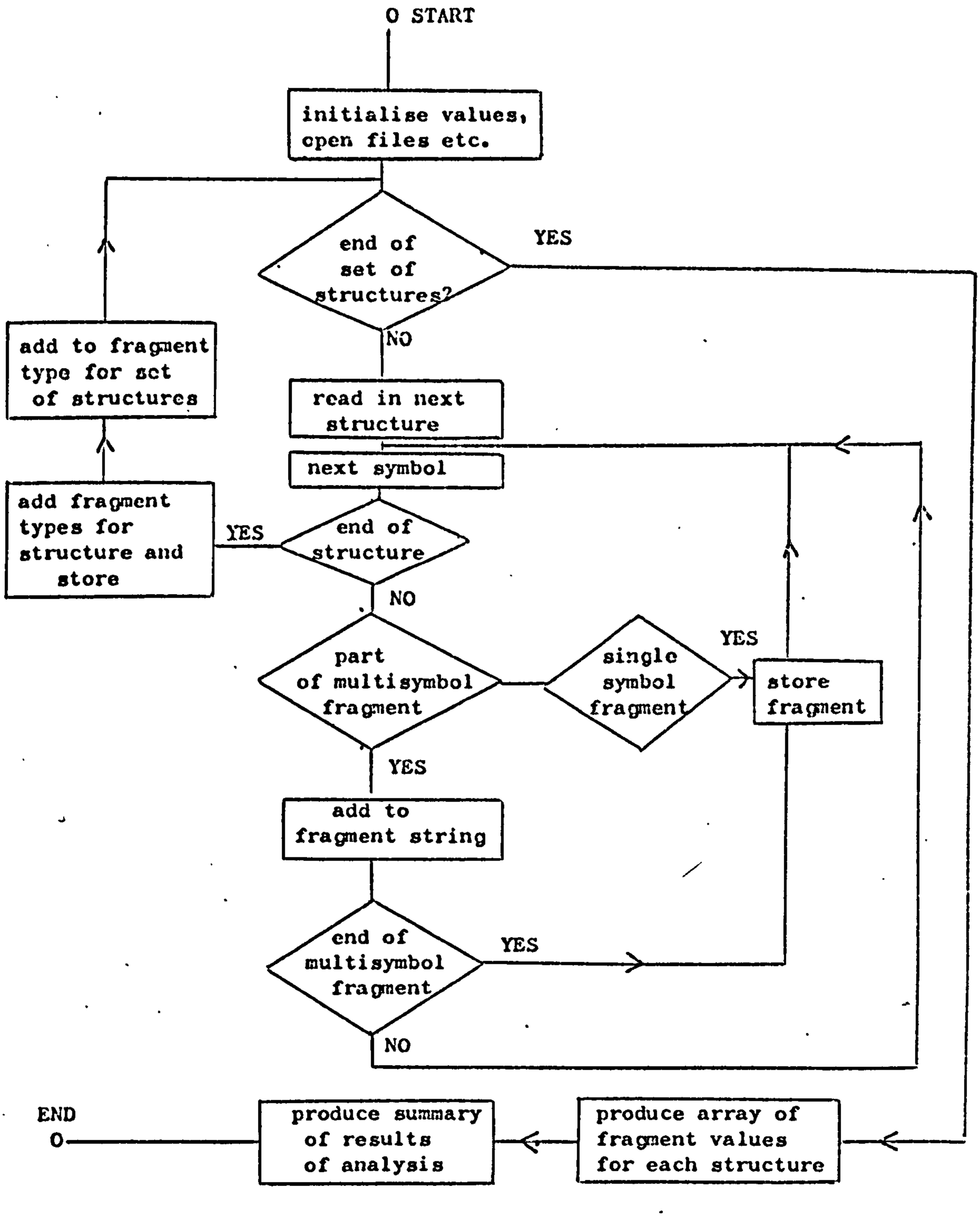
The fragmentation procedure was in this case very simple. Each WLN symbol was taken in turn and compared with a predetermined list of functionalities, unsaturated linkages, ring systems etc. represented by more than one WLN symbol. Numeric symbols, representing aliphatic chains, were fragmented into their component CH_2 and CH_3 groups. Apart from these cases, individual symbols from the WLN were taken as the fragments: this fragmentation procedure is therefore completely open-ended, i.e. capable of dealing with any structure, without the necessity of specifying all the fragments which could occur.

After the fragmentation of each structure the occurrence of each type of fragment is totalled, for each structure and then for the whole set. The fragment occurrence array for each structure is then output for subsequent statistical analysis.

In the algorithms for dealing with substitution on ring systems this type of fragmentation is not applied. Instead a table is set up with an entry for each substituent comprising its WLN locant and the symbols of the substituent group, after a canonicalisation procedure. This latter procedure corrects variations in the ordering of WLN symbols for any group: to take a simple example the COOH group may appear as QV or VQ. Fragments then consist of the substituent groups or heteroatoms, treated as whole units. Relative position terms are derived by considering locant pairs for each pair of substituents: thus for a benzene ring locants "BC" would indicate an ortho interaction, "CE" meta etc.

Structures, coded in WLN, and property values were input on punched cards, or as card-image on a disc-file. The output, arrays of fragment occurrence values for each structure, is produced, again on cards or card-image on disc, for subsequent input into a statistical analysis program.

OUTLINE FLOWCHART



REFERENCES

A.C.S. 1953
'Chemical Nomenclature', Advances in Chemistry Series No. 8
American Chemical Society, Washington, 1953

ADAMSON et al. 1973a
G. W. Adamson and J. A. Bush
'A method for the automatic classification of chemical structures'
Information Storage and Retrieval, 9, 561-568, (1973)

ADAMSON et al. 1973b
G. W. Adamson et al.
'Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part IV. Cyclic Fragments'
Journal of the Chemical Society Perkin I, 863-865,(1973)

ADAMSON et al. 1973c
G. W. Adamson et al.
'Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part V. More Detailed Cyclic Fragments'
Journal of the Chemical Society Perkin I, 2071-2076, (1973)

ADAMSON et al. 1974
G. W. Adamson and J. A. Bush
'Method for relating the structure and properties of chemical compounds'
Nature, 248, 406-408, (1974)

ADAMSON et al. 1975a
G. W. Adamson and J. A. Bush
'A Comparison of the Performance of some similarity and dissimilarity measures in the Automatic Classification of Chemical Structures'
Journal of Chemical Information and Computer Sciences, 15, 55-58, (1975)

ADAMSON et al. 1976a
G. W. Adamson and J. A. Bush
'The Evaluation of an Empirical Structure-Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics'
Journal of the Chemical Society Perkin I, 168-172, (1976)

ADLER et al. 1969
Yu. P. Adler and V. C. Stein
'Composition-Property Diagram Representation of Information for Computers'
Information Storage and Retrieval, 4, 329-332, (1969)

ALBERT 1963
A. Albert
'Ionization Constants', in A. R. Katritzky (ed.) "Physical Methods in Heterocyclic Chemistry", Vol. 1
Academic Press, London, 1963, ch. 1, p. 2

ALBERT 1971
A. Albert
'Ionization Constants' in A. R. Katritzky (ed.) "Physical Methods in Heterocyclic Chemistry", Vol. 3
Academic Press, London, 1971, ch. 1, p. 1

ALLEN et al. 1973
F. H. Allen et al.
'Cambridge Crystallographic Data Centre. II. Structural Data File'
Journal of Chemical Documentation, 13, 119-123, (1973)

ANDREWS 1972
P. R. Andrews
'Are Calculated Electron Populations Suitable Parameters for Multiple Regression Analyses of Biological Activity'
Journal of Medicinal Chemistry, 15, 1069-1072, (1972)

ARIENS 1971a
E. J. Ariens ed.
'Drug Design Vol. 1'
Academic Press, New York-London, 1971

ARIENS 1971b
E. J. Ariens
'A General Introduction to the field of Drug Design
ch. 1, p. 1 of ARIENS 1971a

ARIENS 1971c
E. J. Ariens
'Procedures Followed in Drug Design'
pp. 42-101 of ARIENS 1971b

ARIENS 1971d
E. J. Ariens
'Drug Action'
pp. 6-42 of ARIENS 1971b

ASH 1975
J. E. Ash
'Connection Tables and their Role in a System'
ch. 11, p. 156 of ASH et al., 1975

ASH et al. 1975
J. E. Ash and E. Hyde
'Chemical Information Systems'
Ellis Horwood, Chichester, 1975

AVIDON et al. 1974
V. V. Avidon and L. A. Leksina
'A descriptor Language for Analyzing the Similarity of the Chemical Structure of Organic Compounds'
Automatic Documentation and Mathematical Linguistics, 8, 61-65, (1974)
trans. from Nauchno-Tekhnicheskaya Informatsiya, Series 2, No. 3, 22-25, (1974)

BADGER 1969
G. M. Badger
'Aromatic Character and Aromaticity'
Cambridge University Press, Cambridge, 1969

BAKER et al. 1975
P. A. Baker, G. Palmer, P. W. L. Nichols
'The Wiswesser Line-Formula Notation'
ch. 9, p. 97 of ASH et al.,1975

BALENT et al. 1975
M. Z. Balent and J. M. Emberger
'A Unique Chemical Fragmentation System for Indexing Patent Literature'
Journal of Chemical Information and Computer Sciences, 15, 100-104, (1975)

BALL et al. 1970
G. H. Ball and D. J. Hall
'Some Implications of Interactive Graphic Computer Systems for Data Analysis and Statistics'
Technometrics, 12, 17-31, (1970)

BARTLETT 1965
P. D. Bartlett
'Non-classical Ions: Reprints and Commentary'
Bengamin, N.Y., 1965

BASSOW 1968
H. Bassow
'Construction and Use of Atomic and Molecular Models'
Pergamon, Oxford, 1968

BEASLEY et al. 1969
J. G. Beasley and W. P. Purcell
'An example of successful prediction of cholinesterase inhibitory potency from regression analysis'
Biochimica et Biophysica Acta, 178, 175-176, (1969)

BENSON et al. 1969
S. W. Benson et al.
'Additivity Rules for the estimation of thermochemical properties'
Chemical Reviews, 69, 279-324, (1969)

BERGMAN 1775
T. Bergman
'Disquisitio de attractionibus electivis', 1775
trans. 'Dissertation on Elective Attractions', London, 1785

BERGMAN 1784
T. Bergman
'Meditationes de Systemate Fossilium Naturali'
Nova Acta. Reg. Soc. Scient. Upsaliensis, 4, 63-128, (1974)

BERSOHN et al. 1976
M. Bersohn and A. Esack
'Computers and Organic Synthesis'
Chemical Reviews, 76, 269-282, (1976)
(and references therin)

BERZELIUS 1813
J. J. Berselius
Annals of Philosophy, 2, 359, (1813)
Annals of Philosophy, 3, 51,363, (1814)

BIAGI et al. 1972
G. L. Biagi, A. M. Barbaro, M. C. Guerra
'Partition Data of Chemotherapeutic and Steroid Agents Determined by Reversed-Phase Thin Layer Chromatography'
ch. 5, p. 61 in VAN VALKENBERG, 1972

BINGHAM et al. 1975
R. C. Bingham, M. J. S. Dewar and D. H. Lo
'Ground States of Molecules. XXV. MINDO/3. An Improved Version of the MINDO Semiempirical SCF-MO Method'
Journal of the American Chemical Society, 97, 1285-1293, (1975)

BIRD et al. 1967
A. E. Bird and A. C. Marshall
'Correlation of Serum Binding of Penicillins with Partition Coefficients'
Biochemical Pharmacology, 16, 2275-2290, (1967)

BLACKWOOD et al. 1975
J. E. Blackwood and P. A. Giles
'Chemical Abstracts Stereochemical Nomenclature of Organic Substances in the Ninth Collective Period'
Journal of Chemical Information and Computer Sciences, 15, 67-72, (1975)

BLAIR et al. 1974
J. Blair et al.
'Representation of the Constitutional and Steochemical Features of Chemical Systems in the Computer Assisted Design of Synthesis'
Tetrahedron, 30, 1845-1859, (1974)

BLOOM 1971c
B. M. Bloom
'The Rate of Contemporary Drug Discovery'
ch. 8, p. 176 of BLOOM et al., 1971a

BLOOM et al. 1971a
B. Bloom and G. E. Ullyot (eds.)
'Drug Discovery'
Advances in Chemistry Series 108, American Chemical Society, Washington, 1971

BLOOM et al. 1971b
ch. 1-4 of BLOOM et al., 1971a

BOCEK et al. 1964
K. Bocek et al.
'Chemical Structure and Biological Activity of p-Disubstituted Derivatives of Benzenes'
Experientia, 20, 667-668, (1964)

BOCEK et al. 1967
K. Bocek, J. Kopecky, M. Krivucova
'Chemical Structure and Biological Activity of o-Disubstituted Derivatives of Benzene'
Experientia, 23, 1038, (1967)

BOERHAAVE 1735
H. Boerhaave
'Elements of Chemistry'
trans. Dallowe London, 1735, vol. i, p. 20

BOLTON 1882
H. C. Bolton
'History of Chemical Notation'
Transactionsof the New York Academy of Sciences, 2, 102-106, (1882-3)

BOND et al. 1971
V. B. Bond et al.
'Interactive Searching of a Structure and Biological Activity File'
Journal of Chemical Documentation, 11, 168-170, (1971)

BOWMAN 1975
C. M. Bowman
'The Development of Chemical Information Systems'
ch. 2, p. 6 in ASH et al., 1975

BOWMAN et al. 1968
C. M. Bowman et al.
'A Chemically Oriented Information Storage and Retrieval System. II. Computer Generation of the Wiswesser Notations of Complex Polycyclic Structures'
Journal of Chemical Documentation, 8, 133-138, (1968)

BOWMAN et al. 1970
C. M. Bowman et al.
'A Chemically Oriented Information Storage and Retrieval System. III. Searching a Wiswesser Line Notation File'
Journal of Chemical Documentation, 10, 50-54, (1970)

BRASIE et al. 1965
W. C. Brasie and D. W. Kion
'Chemical Structure Coding'
Chemical Engineering Progress, 61, 102-108, (1965)

BROWN 1959
H. C. Brown
'Foundations of Structural Theory'
Journal of Chemical Education, 36, 104-110, (1959)

BROWN et al. 1976
H. D. Brown et al.
'The Computer-Based Chemical Structure Information System of Merck Sharp and Dohme Research Laboratories'
Journal of Chemical Information and Computer Sciences, 16, 5-10, (1976)

BRUGGER et al. 1976
W. E. Brugger, A. J. Stuper, P. C. Jurs
'Generation of Descriptors from Molecular Structures'
Journal of Chemical Information and Computer Sciences, 16, 105-110, (1976)

BRUICE et al. 1956
T. C. Bruice, N. Kharasch, R. J. Winzler
'A Correlation of Thyroxine-Like Activity and Chemical Structure'
Archives of Biochemistry and Biophysics, 62, 305-317, (1956)

BURSEY 1972
M. M. Bursey
'Interpretation of Mass Spectrometry Data through Linear Free Energy Relationships'
in CHAPMAN et al.,1972, ch. 10, p. 445

BURTON et al. 1965
D. E. Burton et al.
'2-Trifluoromethyl-benzimidazoles: A new class of herbicidal compounds'
Nature, 208, 1166-1169, (1965)

BUSH 1976
J. A. Bush
Doctoral thesis, University of Sheffield in preparation

BUSTARD 1974
T. M. Bustard
'Optimization of Alkyl Modifications by Fibonacci Search'
Journal of Medicinal Chemistry, 17, 777-778, (1974)

BYKOV 1962
G. V. Bykov
'The Origin of the Theory of Chemical Structure'
Journal of Chemical Education, 39, 220-224, (1962)

BYKOV 1965
G. V. Bykov
'Historical Sketch of the Electron Theories of Organic Chemistry'
Chymia, 10, 199-253, (1965)

CAHN 1974a
R. C. Cahn
'An Introduction to Chemical Nomenclature'
Butterworths, London, 1974

CAHN 1974b
R. S. Cahn
ch. 7, p. 118 of CAHN 1974b

CAMMARATA 1972
A. Cammarata
'Interrelationship of the Regression Models Used for Structure-Activity Studies'
Journal of Medicinal Chemistry, 15, 573-577, (1972)

CAMMARATA et al. 1970
A. Cammarata and S. J. Yau
'Predictability of correlations between in vitro tetracycline potencies and substituent indices'
Journal of Medicinal Chemistry, 13, 93-97, (1970)

CAMMARATA et al. 1972
A. Cammarata and K. S. Rogers
'The Interpretation of Drug Action through Linear Free Energy Relationships'
ch. O, p. 401 in CHAPMAN et al., 1972

CAMMARATA et al. 1976
A. Cammarata and G. K. Menon
'Pattern Recognition. Classification of Therapeutic Agents According to Pharmacophores'
Journal of Medicinal Chemistry, 19, 739-748, (1976)

CAMPEY et al. 1970
L. H. Campey, E. Hyde, A. R. H. Jackson
'Interconversion of Chemical Structure Systems'
Chemistry in Britain, 6, 427-430, (1970)

CANAS-RODRIGUEZ et al. 1972
A. Canas-Rodriquez and M. S. Tute
'Pitfalls in the Use of π Constants'
ch. 3, p. 41 in VAN VALKENBURG, 1972

CHAPMAN et al. 1972
N. B. Chapman and J. Shorter (eds.)
'Advances in Linear Free Energy Relationships'
Plenum, London, 1972

CHARTON 1971
M. Charton
'The Quantitative Treatment of the Ortho Effect'
Progress in Physical Organic Chemistry, 8, 235-315, (1971)

CHIPPERFIELD 1972
J. R. Chipperfield
'Linear Free Energy Relationships in Inorganic Chemistry'
ch. 7, p. 321 in CHAPMAN et al., 1972

CHU 1974
K. C. Chu
'Applications of Artificial Intelligence to Chemistry. Use of Pattern Recognition and Cluster Analysis to Determine the Pharmacological Activity of Some Organic Compounds'
Analytical Chemistry, 46, 1181-1187, (1974)

CHU et al. 1975
K.C. Chu et al.
'Pattern Recognition and Structure-Activity Relationship Studies. Computer-Assisted Prediction of Antitum or Activity in Structurally Diverse Drugs in an Experimental Mouse Brain Tumor System'
Journal of Medicinal Chemistry, 18, 539-545, (1975)

CLERC et al. 1973
J. T. Clerc, P. Naegeli, J. Seibl
'Artificial Intelligence'
Chimia, 27, 639, (1973)

CLINGING et al. 1974
R. Clinging and M. F. Lynch
'Production of Printed Indexes of Chemical Reactions. II. Analysis of reactions involving ring formation, cleavage and interconversion'
Journal of Chemical Documentation, 14, 69-71, (1974)

CONROW 1966
K. Conrow
'Computer Generation of Baeyer System Names of Saturated Bridged Bicyclic, Tricyclic, and Tetracyclic Hydrocarbons'
Journal of Chemical Documentation, 6, 206-213, (1966)

COOK 1974a
D. B. Cook
'Ab initio valence calculations in chemistry'
Butterworths, London, 1974

COOK 1974b
D. B. Cook
p. 13-15 of COOK, 1974a

COOK 1974c
D. B. Cook
p. 83-84 of COOK, 1974a

COOK 1974d
D. B. Cook
ch. 13, p. 209 of COOK, 1974

COOK 1974e
D. B. Cook
p. 3 of COOK, 1974a

COREY et al. 1972
E. J. Corey et al.
'Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structures by Interactive Computer Graphics'
Journal of the American Chemical Society, 94, 421-430, (1972)

COREY et al. 1972b
E. J. Corey et al.
'Techniques for perception by computer of synthetically significant structural features in complex molecules'
Journal of the American Chemical Society, 94, 431-439, (1972)

COREY et al. 1976
E. J. Corey and W. L. Jorgensen
'Computer-Assisted Synthetic Analysis. Synthetic Strategies Based on Appendages and the Use of Reconnective Transforms'
Journal of the American Chemical Society, 98, 189-203, (1976)

CORMACK 1971
R. M. Cormack
'A review of classification'
Journal of the Royal Statistical Society Series A, 134, 319-367, (1971)

COUSSE et al. 1973
H. Cousse, G. Mouzin, L. D. D'Hinterland
'Recherche d'une correlation par la methode des regressions multiples sur une serie de derives du pyridyl-2 methanol a activite spasmolytique'
Chimie Therapeutique, 8, 466-468, (1973)

COX et al. 1970a
J. D. Cox and G. Pilcher
'Thermochemistry of Organic and Organometallic Compounds'
Academic Press, London, 1970

COX et al. 1970b
J. D. Cox and G. Pilcher
pp. 118-124, 530-569 of COX et al., 1970a

COX et al. 1970c
J. D. Cox and G. Pilcher
p. 125 of COX et al., 1970a

COX et al. 1970d
J. D. Cox and G. Pilcher
pp. 122-124 of COX et al., 1970a

COX et al 1970e
J. D. Cox and G. Pilcher
pp. 124-125 of COX et al., 1970a

CRAM 1952
D. J. Cram
'Studies in Stereochemistry. V. Phenunium Sulphonate Ion-Pairs as Intermediates in the Intramolecular Rearrangements and Solvolysis Reactions that Occur in the 3-Phenyl-2-Butanol System'
Journal of the American Chemical Society, 74, 2129-2137, (1952)

CRAIG 1971a
P. N. Craig
'Interdependence between Physical Parameters and Selection of Substituent Groups for Correlation Studies'
Journal of Medicinal Chemisty, 14, 680-684, (1971)

CRAIG 1971b
P. N. Craig
'Comparison of Batch and Time-sharing Computer Runs for Correlating Structures and Bioactivity by the Hansch method'
Journal of Chemical Documentation, 11, 160-162, (1971)

CRAIG 1972a
P. N. Craig
'Comparison of the Hansch and Free-Wilson Approaches to Structure-Activity Correlation'
ch. 8, p. 115 of VAN VALKENBURG, 1972

CRAIG 1972b
P. N. Craig
'Structure-Activity Correlations of Antimalarial Compounds. 1. Free-Wilson Analysis of 2-Phenylquinoline-4-carbinols'
Journal of Medicinal Chemistry, 15, 144-149, (1972)

CRAIG 1975
P. N. Craig
'Structure/Property Correlations'
ch. 16, p. 259 of ASH et al., 1975

CRAIG et al. 1969
P. N. Craig and H. M. Ebert
'Eleven Years of Structure Retrieval Using the SK & F Fragment Code'
Journal of Chemical Documentation, 9, 141-146, (1969)

CRAMER et al. 1974
R. D. Cramer, G. Redl, C. E. Berkoff
'Substructural Analysis. A Novel Approach to the Problem of Drug Design',
Journal of Medicinal Chemistry, 17, 533-535, (1974)

CRISTENSEN et al. 1974
H. E. Cristensen, T. L. Luginbyhl, B. L. Caroll (eds.)
'The Toxic Substances List, 1974 edition'
National Institute for Occupational Safety and Health, Rockville, Maryland, Z0854, U.S.A., 1974

CRISTOFFERSON 1972
R. E. Crystofferson
'Ab initio calculations on large molecules'
Advances in Quantum Chemistry, 6, 333-393, (1972)

CROSLAND 1962a
M. P. Crosland.
Historical Studies in the Language of Chemistry
Heinemann, London, 1962

CROSLAND 1962b
M. P. Crosland
'Early Chemical Terminology'
part 2, p. 65 of CROSLAND, 1962a

CROSLAND 1962c
M. P. Crosland
'Chemical Symbolism'
part 4, p. 227 of CROSLAND, 1962a

CROSLAND 1962d
M. P. Crosland
p. 114-130 of CROSLAND, 1962b

CROSLAND 1962e
M. P. Crosland
p. 139-143 of CROSLAND, 1962f

CROSLAND 1962f
M. P. Crosland
'The Introduction of Systematic Nomenclature into Chemistry'
Part 3, p. 133 of CROSLAND, 1962a

CROSLAND 1962g
M. P. Crosland
'The Language of Organic Chemistry'
Part 5, p. 285 of CROSLAND, 1962a

CROSLAND 1962h
M. P. Crosland
pp. 338-354 of CROSLAND, 1962a

CROWE et al. 1973
J. E. Crowe et al.
'The Searching of Wiswesser Line Notations by means of a Character-Matching Serial Search'
Journal of Chemical Documentation, 13, 85-92, (1973)

CRUM BROWN 1864
A. Crum Brown
'On the Theory of Isomeric Compounds'
Transactions of the Royal Society of Edinburgh, 23, 707-719, (1864)

CRUM BROWN 1869
A. Crum Brown
'On Chemical Constitution, and its Relation to Physical and Physiological Properties'
Philosophical Magazine, 37, 395-400, (1869)

CRUM BROWN et al. 1868
A. Crum Brown and T. R. Fraser
'On the Connection between Chemical Constitution and Physiological Action. Part 1'
Transactions of the Royal Society of Edinburgh, 25, 151-203, (1868-69)

CURTIN 1954
D. Y. Curtin
Record Chem. Progr. (Kresge-Hooker Sci. Lib.), 15, 111, (1954)

DALTON 1810
J. Dalton
'New System of Chemical Philosophy'
London, 1810

DALTON 1840
J. Dalton
'On a new and easy method of analysing sugar'
p. 3-4 Manchester, 1840

DAMMERS 1975
H. F. Dammers
'Industrial Information Systems'
ch. 3, p. 13 of ASH et al., 1975

DAMMERS et al. 1968
H. F. Dammers and D. J. Polton
'Use of the IUPAC Notation in Computer Processing of Information on Chemical Structures'
Journal of Chemical Documentation, 8, 150-160, (1968)

DARVAS 1974
F. Darvas
'Application of the Sequential Simplex Method in Designing Drug Analogs'
Journal of Medicinal Chemistry, 17, 799-804, (1974)

DAVIDSON 1973
R. M. Davidson
'The Development of Stereochemistry (in the nineteenth century)'
B.A. Part II Thesis, University of Oxford, 1973

DAVIS 1973
S. S. Davis
'Use of substituent constants in structure-activity relations and the importance of the choice of standard state'
Journal of Pharmacy and Pharmacology, 25, 293-296, (1973)

DAVIS et al. 1974a
C. H. Davis and J. E. Rush
'Information Retrieval and Documentation in Chemistry'
Greenwood Press, Westport Connecticut, 1974

DAVIS et al. 1974b
ch. 8, p. 143 of DAVIS et al., 1974a

DAVIS et al. 1974c
C. H. Davis and J. E. Rush
ch. 9, p. 230 of DAVIS et al., 1974a

DAVIS et al. 1974d
S. S. Davis, T. Higuchi, J. H. Rytting
'Determination of Thermodynamics of Functional Groups in

Solutions of Drug Molecules'
Advances in Pharmaceutical Sciences, 4, 73-261, (1974)

DE LA MARE et al. 1959
P. B. D. de la Mare and J. H. Ridd
'Aromatic Substitution'
Butterworths, London, 1959

DEEMING 1986
S. N. Deeming
'On the use of Fibonacci searches in Structure-Activity Studies'
Journal of Medicinal Chemistry, 19, 977-978, (1976)

DEMILT 1951
C. deMilt.
'The Congress at Karlsruhe'
Journal of Chemical Education, 28, 421-425, (1951)

DEWAR 1866
J. Dewar
'On the Oxidation of Phenyl Alcohol and a Mechanical Arrangement adapted to illustrate Structure in the non-saturated hydrocarbons'
Proceedings of the Royal Society of Edinburgh, 6, 84-86, (1866-1867)

DEWAR 1969a
M. J. S. Dewar
'The Molecular Orbital Theory of Organic Chemistry'
McGraw-Hill, New York, 1969

DEWAR 1969b
M. J. S. Dewar
p. 73-74 of DEWAR 1969a

DEWAR et al.1962
M. J. S. Dewar and P. J. Grisdale
'Substituent Effects. IV. A Quantitative Theory'
Journal of the American Chemical Society, 84, 3548-3553, (1962)

DIERDORF et al. 1974
D. S. Dierdorf and B. R. Kowalski
'Three Dimensional Molecular Structure-Biological Activity Correlation by Pattern Recognition'
NTIS Report, AD 785 863

DINER et al. 1969
S. Diner et al.
'Localized Bond Orbitals and the Correlation Problem. III. Energy up to the Third Order in the Zero Differential Overlap Approximation. Application to o-electron systems'
Theoretica Chimica Acta, 15, 100-110, (1969)

DIRAC 1929
P. A. M. Dirac
'Quantum mechanics of Many Electron Systems'
Proceedings of the Royal Society, Series A, 123, 714-733, (1929)

DIXON et al. 1976
D. A. Dixon et al.
'Localized Molecular Orbitals for Polyatomic Molecules. IV. Large Boron Hydrides'
Journal of the American Chemical Society, 98, 2086-2096, (1976)

DONALDSON et al. 1974
N. Donaldson et al.
'Chemical Abstracts Index Names for Chemical Substances in the Ninth Collective Period (1972-1976)'
Journal of Chemical Documentation, 14, 3-15, (1974)

DOVING 1973
K. B. Doving
'Physiological data correlated with physical parameters'
in 'Transduction mechanisms in Chemoreception' European Chemoreception Research Organisation Symposium
Information Retrieval Limited, London, 1973

DUBOIS 1974
J. E. Dubois
'DARC System in Chemistry'
in WIPKE et al., 1974, p. 239-263

DUBOIS et al. 1972
J. E. Dubois et al.
'A Quantitative Study of Substituent Interactions in Aromatic Electrophilic Substitution. 1. Bromination of Polysubstituted Benzenes'
Journal of the American Chemical Society, 94, 6823-6828, (1972)

DUKE 1975a
B. J. Duke
'Electronic Calculations on Large Molecules'
in Chemical Society Specialist Periodical Reports, 'Theoretical Chemistry', Vol. 2, C.S. London, 1975, ch. 4, p. 159

DUKE 1975b
B. J. Duke
162-184 of DUKE 1975a

DUKE 1975c
B. J. Duke
184-191 of DUKE 1975a

DYSON 1944
G. M. Dyson
'A Notation for Organic Compounds'
Nature, 154, 114, (1944)

DYSON 1953
G. M. Dyson
'Development of Chemical Symbols and Their Relation to Nomenclature'
p. 99-105 of A.C.S., 1953

DYSON 1975
G. M. Dyson
'The Dyson-IUPAC Notation'
ch. 10, p. 130 of ASH et al., 1975

EAKIN 1975
D. R. Eakin
'The ICI CROSSBOW System'
ch. 14, p. 227 in ASH et al., 1975

EAKIN et al. 1974
D. R. Eakin and E. Hyde
'Evaluation of On-Line Techniques in a Sub-Structure Search System'
in WIPKE et al., 1974a, p. 1-30

EHRENSON et al. 1973
S. Ehrenson, R. T. C. Brownlee, R. W. Taft
'A Generalised Treatment of Substituent Effects in the Benzene Series. A Statistical Analysis by the Dual Substituent Parameter Equation'
Progress in Physical Organic Chemistry, 10, 1-80, (1973)

ELIEL 1962a
E. L. Eliel
'Stereochemistry of Carbon Compounds'
McGray-Hill, New York, 1962

ELIEL 1962b
E. L. Eliel
pp. 1-6, 16-25, 87-95 of ELIEL, 1962a

ENGLAND et al. 1971
W. England, L. S. Salmon, K. Ruedenberg
'Localised M.O.s: A Bridge between Chemical Intuition and Molecular Quantum Mechanics'
Topics in Current Chemistry, 23, 31-123, (1971)

ENGLAND et al. 1975
W. England, M. S. Gordon, K. Ruedenberg
'Localized Charge Distributions. VII. Transferable Localized Molecular Orbitals for Acyclic Hydrocarbons'
Theoretica Chimica Acta, 37, 177-216, (1975)

ENGLER et al. 1973
E. M. Engler, J. D. Andose, P. van R. Schleyer
'Critical Evaluation of Molecular Mechanics'
Journal of the American Chemical Society, 95, 8005-8025, (1973)

ESACK et al. 1974
A. Esack and M. Bersohn
'A Program for Rapid and Automatic Functional Group Recognition'
Journal of the Chemical Society, Perkin I, 4263-4270, (1974)

ESACK et al. 1975
A. Esack and M. Bersohn
'Computer Manipulation of Central Chirality'
Journal of the Chemical Society Perkin I, 1124-1129, (1975)

EVERITT 1974
B. Everitt
'Cluster Analysis'
Heinemann, London, 1974

EXNER 1966
O. Exner
'Additive physical properties. I. General Relationships and Problems of Statistical Nature'
Collection of Czechoslovak Chemical Communications, 31, 3222-3251, (1966)

EXNER 1967a
O. Exner
'Additive Physical Properties. II. Molar Volume as an Additive Property'
Collection of Czechoslovak Chemical Communications, 32; 1-23, (1967)

EXNER 1967b
O. Exner
'Additive Physical Properties. III. Re-examination of the Additive Character of Parachur'
Collection of Czechoslovak Chemical Communications, 32, 23-55, (1967)

EXNER 1972a
O. Exner
'The Hammett Equation- the Present Position'
in CHAPMAN et al., 1972, ch. 1, p. 1

EXNER 1972b
O. Exner
p. 2 of EXNER, 1972a

EXNER 1972c
O. Exner
p. 41-45 of EXNER, 1972a

EXNER 1972d
p. 4-19 of EXNER, 1972a

EXNER 1972e
O. Exner
p. 4 of EXNER, 1972a

EXNER 1972f
O. Exner
p. 46-52 of EXNER, 1972a

EXNER 1972g
O. Exner
pp. 44-46 of EXNER, 1972a

FARRELL et al. 1971
C. D. Farrell, A. R. Chauvenet, and D. A. Koniver
'Computer Generation of Wiswesser Line Notation'
Journal of Chemical Documentation, 11, 52-59, (1971)

FELDMAN 1973
A. Feldman
'A Chemical Teletype'
Journal of Chemical Documentation, 13, 53-56, (1973)

FELDMANN 1974
R. J. Feldmann
'Interactive Graphic Chemical Structure Searching'
pp. 55-81 in WIPKE et al., 1974a

FELDMANN et al. 1971
R. J. Feldmann and D. A. Koniver
'Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation'
Journal of Chemical Documentation, 11, 154-159, (1971)

FELDMANN et al. 1972a
R. J. Feldmann et al.
'An Application of Interactive Computing - A Chemical Information System'
Journal of Chemical Documentation, 12, 41-47, (1972)

FELDMANN et al. 1972b
R. J. Feldmann, S. R. Heller, C. R. T. Bacon
'An Interactive, Versatile, Three-Dimensional Display, Manipulation and Plotting System for Biomedical Research'
Journal of Chemical Documentation, 12, 234-237, (1972)

FERGUSON 1939
J. Ferguson
'The use of chemical potentials as indices of toxicity'
Proceedings of the Royal Society, Series B, 127, 387-404, (1939)

FIGEYS et al. 1975
H. P. Figeys et al.
'Approximate Charge Density Localization of Molecular Orbitals'
Theoretica Chimica Acta, 40, 253-261, (1975)

FINDLAY 1965a
A. Findlay
'A Hundred Years of Chemistry' 3rd edn.
Duckworth, London, 1965

FINDLAY 1965b
A. Findlay
ch. 7, p. 113 of FINDLAY, 1965a

FISANICK et al. 1975
W. Fisanick et al.
'Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files'
Journal of Chemical Information and Computer Sciences, 15, 73-84, (1975)

FISCHER 1891
E. Fischer
Berichte, 24, 2683, (1891)

FISHER 1973a
N. W. Fisher
'Organic Classification before Kekule'
Ambix, 20, 106-131, (1973)

FISHER 1973b
N. W. Fisher
'Organic Classification before Kekule Part II'
Ambix, 20, 209-233, (1973)

FOLEY 1972
D. H. Foley
'Considerations of Sample and Feature Size'
IEEE Transactions on Information Theory
IT-18, 618-626, (1972)

FREE et al. 1964
S. M. Free and J. W. Wilson
'A Mathematical Contribution to Structure-Activity Studies'
Journal of Medicinal Chemistry, 7, 395-399, (1964)

FUGMANN 1975
R. Fugmann
'The IDC System'
ch. 13, p. 195 of ASH et al., 1975

FUJITA 1966
T. Fujita
'The analysis of physiological activity of substituted phenols with substituent constants'
Journal of Medicinal Chemistry, 16, 791-796, (1973)

FUJITA et al. 1966
T. Fujita, J. Isawa, C. Hansch
'A New Substituent Constant, π , Derived from Partition Coefficient'
Journal of the American Chemical Society, 86, 5175-5180, (1964)

FUJITA et al. 1971
T. Fujita and T. Ban
'Structure-Activity Study of Phenethylamines as Substrates of Biosynthetic Enzymes of Sympathetic Transmitters'
Journal of Medicinal Chemistry, 14, 148-152, (1971)

FUKUNAGA et al. 1976
J. Y. Fukunaga, C. Hansch, E. E. Steller
'Inhibition of Dihydrofolate Reductase. Structure-Activity Correlations of Quinazolines'
Journal of Medicinal Chemistry, 19, 605-611, (1976)

FULLER et al. 1968
R. W. Fuller, M. M. Marsh, J. Mills
'Inhibition of Monoamine Oxidase by N-(Phenoxyethyl) cyclopropylamines. Correlation on Inhibition with Hammett Constants and Partition Coefficients'
Journal of Medicinal Chemistry, 11, 397-398, (1968)

GANELLIN et al. 1973
C. R. Ganellin et al.
'Conformation of Histamine Derivatives. I. Application of Molecular Orbital Calculations and Nuclear Magnetic Resonance Spectroscopy'
Journal of Medicinal Chemistry, 16, 610-620, (1973)

GEISS et al. 1975
F. Geiss et al.
'The Environmental Chemicals Data and Information Network of the European Communities (ECDIN):
Paper presented at the UNEP Workshop on an International Register of potentially toxic chemicals, Bithoven, 1975

GELBERG et al. 1962
A. Gelberg et al.
'A Program Retrieval of Organic Structure Information via Punched Cards
Journal of Chemical Documentation, 2, 7-11, (1962)

GEOFFROY 1718
St. F. Geoffroy
'Table des differents rapports observés en chimie entre differentes substances'
Memoires de l'Academie Royale des Sciences, Paris, 1718, p. 212

GERRATT 1974
J. Gerratt
'Valence Bond Theory'
in Chemical Society Specialist Periodical Report, 'Theoretical Chemistry', Vol. 1, p. 60-109, (1974)

GLUCK 1965
D. J. Gluck
'A Chemical Structure Storage and Search System Developed at Du Pont'
Journal of Chemical Documentation, 5, 43-51, (1965)

GOLEBIEWSKI et al. 1974
A. Golebiewski and A. Parcezewski
'Theoretical Conformational Analysis of Organic Molecules'
Chemical Reviews, 74, 519-530, (1974)

GOODFORD 1973
P. J. Goodford
'Prediction of Pharmacological Activity by the Method of Physicochemical Activity Relationships'
Advances in Pharmacology and Chemotherapy, 11, 51-97, (1973)

GOODFORD et al. 1973
P. J. Goodford et al.
'Predictions of the antimalarial activity of arylamidinoureas'
British Journal of Pharmacology, 48, 650-654, (1973)

GOTTARDI 1970
R. Gottardi
'A Modified Dot-Bond Structural Formula Font with Improved Stereochemical Notation Abilities'
Journal of Chemical Documentation, 10, 75-81, (1970)
(and reference therein)

COWER 1967
J. V. Gower
'A comparison of some methods of cluster analysis'
Biometrics, 23, 623-637, (1967)

GRANITO 1973
C. E. Granito
'CHEMTRAN and the interconversion of Chemical Substructure Systems'
Journal of Chemical Documentation, 13, 72-74, (1973)

GRANITO et al. 1965
C. E. Granito et al.
'Rapid Structure Searches via Permuted Chemical Line Notation. III. A Computer-Produced Index'
Journal of Chemical Documentation, 5, 229-233, (1965)

GRANITO et al. 1972
C. E. Granito, S. Roberts, G. W. Gibson
'The Conversion of Wiswesser Line Notation to Ring Codes. I. The Conversion of Ring Systems'
Journal of Chemical Documentation, 12, 190-196, (1972)

GRASSELLI 1973
J. G. Grasselli (ed.)
'CRC Atlas of Spectral Data and Physical Constants for Organic Compounds'
Chemical Rubber Company, Cleveland, Ohio, 1973

GREENWOOD et al. 1966
H. H. Greenwood and R. McWeeny
'Reactivity Indices in Conjugated Molecules. The Present Position'
Advances in Physical Organic Chemistry, 4, 73-145, (1966)

GUND et al. 1973
P. Gund, W. T. Wipke, R. Langridge
'Computer Searching of a Molecular Structure File for Pharmacophoric Patterns'
Proceedings of an International Conference on Computers in Chemical Research and Education, Zagreb, 1973

GUND et al. 1975
T. M. Gund et al.
'Computer Assisted Graph Theoretical Analysis of Complex Mechanistic Problems in Polycyclic Hydrocarbons. The Mechanism of Diamantane Formation from Various Pentacyclotetradecanes'
Journal of the American Chemical Society, 97, 743-751, (1975)

GUYTON 1782
L. B. Guyton de Morveau
'Mémoire sur les dénominations chimiques'
Observations sur las Physique, 19, 370-382, (1782)

HAHN 1975
F. E. Hahn
'Strategy and Tactics of Chemotherapeutic Drug Development'
Naturwissenschaften, 62, 449-458, (1975)

HALL 1973
G. G. Hall
'The Growth of Computational Quantum Chemistry from 1950 to 1971'
Chemical Society Reviews, 2, 21-28, (1973)

HALL et al. 1975
L. H. Hall, L. B. Kier, W. J. Murray
'Molecular Connectivity II: Relationship to Water Solubility and Boiling Point'
Journal of Pharmaceutical Sciences, 64, 1974-1977, (1975)

HAMMETT 1937
L. P. Hammett
'The Effect of Structure upon the reactions of organic compounds. Benzene derivatives'
Journal of the American Chemical Society, 59, 96-103, (1937)

HAMMETT 1940
L. P. Hammett
'Physical Organic Chemistry'
McGraw-Hill, New York, 1940

HAMMOND 1974
G. S. Hammond
'Information Management and Original Thought in Chemical Education'
Journal of Chemical Education, 51, 55-58, (1974)

HANSCH 1969
C. Hansch
'A Quantitative Approach to Biochemical Structure-Activity Relationships'
Accounts of Chemical Research, 2, 232-239, (1969)

HANSCH 1971a
C. Hansch
'Quantitative Structure-Activity Relationships in Drug Design'
in ARIENS, 1971a, ch. 2, p. 271

HANSCH 1971b
C. Hansch
p. 271-275 of HANSCH,1971a

HANSCH 1971c
C. Hansch
pp. 308-315 of HANSCH, 1971a

HANSCH 1971d
C. Hansch
pp. 290-296 of HANSCH, 1971a

HANSCH 1971e
C. Hansch
pp. 296-308 of HANSCH, 1971a

HANSCH 1971f
C. Hansch
pp. 315-335 of HANSCH, 1971a

HANSCH 1972
C. Hansch
'A Computerised Approach to Quantitative Biochemical Structure-Activity Relationships'
ch. 2, p. 20 in VAN VALKENBERG, 1972

HANSCH 1976
C. Hansch
'On the structure of medicinal chemistry'
Journal of Medicinal Chemistry, 19, 1-6, (1976)

HANSCH et al. 1962
C. Hansch et al.
'Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients'
Nature, 194, 178-180, (1962)

HANSCH et al. 1964a
C. Hansch and T. Fujita
'P-o- Analysis. A Method for the Correlation of Biological Activity and Chemical Structure'
Journal of the American Chemical Society', 86, 1616-1626, (1964)

HANSCH et al. 1972
C. Hansch and W. J. Dunn
'Linear Relationships between Lipophilic Character and Biological Activity of Drugs'
Journal of Pharmaceutical Sciences, 61, 1-19, (1972)

HANSCH et al. 1973a
C. Hansch and J. M. Clayton
'Lipophilic Character and Biological Activity of Drugs. II. The Parabolic Case'
Journal of Pharmaceutical Sciences, 62, 1-21, (1973)

HANSCH et al. 1973b
C. Hansch et al.
'"Aromatic" Substituent Constants for Structure-Activity Correlations'
Journal of Medicinal Chemistry, 16, 1207-1216, (1973)

HANSCH et al. 1973c
C. Hansch, S. H. Unger, A. B. Forsyth
'Strategy in Drug Design. Cluster Analysis as an aid in the selection of substituents'
Journal of Medicinal Chemistry, 16, 1217-1222, (1973)

HANSCH et al. 1974a
C. Hansch, A. Leo, D. Elkins
'Computerised Management of Structure-Activity Data. I. Multivariate Analysis of Biological Data'
Journal of Chemical Documentation, 14, 57-61, (1974)

HANSCH et al. 1974b
C. Hansch and M. Yoshimoto
'Structure-Activity Relationships in Immunochemistry. 2. Inhibition of Complement by Benzamidines'
Journal of Medicinal Chemistry, 17, 1160-1167, (1974)

HANSCH et al. 1975
C. Hansch, C. Silipo, E. E. Steller
'Formulation of De Novo Substituent Constants in Correlation Analysis: Inhibition of Dihydrofolate Reductase by 2,4-Diamino-5-(3,4-dichlorophenyl)-6-substituted Pyrimidines'
Journal of Pharmaceutical Sciences, 64, 1186-1191, (1975)

HASSENFRATZ et al. 1787
J. H. Hassenfratz and P. A. Adet
pp. 253-287 of LAVOISIER, 1787

HAYWARD et al. 1965
H. W. Hayward et al.
'Some Experiences with the Hayward Linear Notation System'
Journal of Chemical Documentation, 5, 183-190, (1965)

HELLER. 1974
S. R. Heller
'Computer Techniques for Interpreting Mass Spectrometry Data'
p. 175-202 in WIPKE et al., 1974a

HELLER et al. 1972
S. R. Heller and R. J. Feldmann
'An Interactive NMR Chemical Shift Search Program'
Journal of Chemical Education, 49, 291, (1972)

HELLER et al. 1973
S. R. Heller et al.
'A Conversational Mass Spectral Search System. IV. The Evolution of a System for the Retrieval of Mass Spectral Information'
Journal of Chemical Documentation, 13, 130-133, (1973)

HERMAN et al. 1973
F. Herman, A. D. McLean, R. K. Nesbet
'Computational Methods for Large Molecules and Localized States in Solids'
Plenum, New York, 1973

HERNDON 1972
W. C. Herndon
'Semi-empirical Molecular Orbital Calculations for Saturated Organic Compounds'
Progress in Physical Organic Chemistry, 9, 99-177, (1972)

HERNDON 1973
W. C. Herndon
'Resonance Energies of Aromatic Hydrocarbons. A Quantitative Test of Resonance Theory'
Journal of the American Chemical Society, 95, 2404-2406, (1973)

HESSE 1963
M. B. Hesse
'Models and Analogies in Science'
University of Notre Dame Press, 1966

HIGGINS 1789
W. Higgins
'A Comparative View of the Phlogistic and Anti-phlogistic Theories'
London, 1789

HIGUCHI et al. 1970
T. Higuchi and S. S. Davis
'Thermodynamic Analysis of Structure-Activity Relationships of Drugs. Prediction of Optimal Structures'
Journal of Pharmaceutical Sciences, 59, 1376-1383, (1970)

HILLER et al. 1973
S. A. Hiller et al.
'Cybernetic Methods of Drug Design. 1. Statement of the Problem - The Perceptron Approach'
Computers and Biomedical Research, 6, 411-421, (1973)

HINE 1962
J. Hine
'Physical Organic Chemistry'
McGraw-Hill, New York, 1962

HINE 1975a
J. Hine
'Structural Effects on Equilibria in Organic Chemistry'
Wiley, New York, 1975

HINE 1975b
J. Hine
'The Hammett and Taft Equations'
ch. 3, p. 55 of HINE, 1975a

HINE 1975c
J. Hine
pp. 4-13 of HINE, 1975a

HOTEL 1971
P. G. Hoel
'Introduction to Mathematical Statistics,'4th edition
Wiley, New York, 1971, pp. 348-356

HOFMANN 1865
A. W. Hofmann
'On the combining power of atoms'
Chemical News, 12, 176-179, 189, (1865)

HOLM et al. 1973
B. E. Holm, M. G. Howell, H. E. Kennedy, J. H. Kuney, J. E. Rush
'The Status of Chemical Information'
Journal of Chemical Documentation, 13, 171-183, (1973)

HOPFINGER et al. 1976
A. J. Hopfinger and R. D. Battershell
'Application of SCAP to Drug Design. 1. Prediction of Octanol-Water Partition Coefficients Using Solvent-Dependent Conformational Analysis'
Journal of Medicinal Chemistry, 19, 569-573, (1976)

HOYLAND 1969
J. R. Hoyland
'Semi empirical MO Theories: A Critique and a Review of Progress'
p. 31-81 of KIER, 1969

HUDSON et al. 1970
D. R. Hudson, G. E. Bass, W. P. Purcell
'Quantitative Structure-Activity Models. Some Conditions for Application and Statistical Interpretation'
Journal of Medicinal Chemistry, 13, 1184-1189, (1970)

HUMBER et al. 1975
L. G. Humber et al.
'"Predicted" compounds with "alleged" biological activities from analyses of structure-activity relationships'
IUPAC Information Bulletin, No. 49, 12-20, (1975)

HUMFFRAY et al. 1969
A. A. Humffray and J. J. Ryan
'Rate Correlations involving the Linear Combination of Substituent Parameters. Part IV. Base-catalysed Hydrolysis of Triethylphenoxysilanes'
Journal of the Chemical Society, B, 1138-1142, (1969)

HYDE 1975a
E. Hyde
'Chemical Structure Systems'
chapter 7, p. 86 of ASH et al., 1975

HYDE 1975b
R. M. Hyde
'Relationships between the Biological and Physicochemical Properties of Series of Compounds'
Journal of Medicinal Chemistry, 18, 231-233, (1975)

HYDE et al. 1967
E. Hyde et al.
'Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds'
Journal of Chemical Documentation, 7, 200-204, (1967)

HYDE et al. 1968
E. Hyde and L. Thomson
'Structure Display'
Journal of Chemical Documentation, 8, 138-146, (1968)

HYDE et al. 1975
E. Hyde and J. Ash
ch. 1 of ASH et al., 1975

I.C.L. 1971
Statistical Analysis Mark II
Applications Package, International Computers Ltd., Technical Publication 4301, London, 1971

IUPAC 1971a
IUPAC Nomenclature of Organic Chemistry 1969.
Sections A and B 3rd edn., Section C, 2nd edn.
Butterworths, London, 1971

IUPAC 1971b
IUPAC Nomenclature of Inorganic Chemistry, 2nd edn., Definitive Rules, 1970
Butterworths, London, 1971

IHE 1959
A. J. Ihde
'The Unraveling of Geometric Isomerism and Tautomerism'
Journal of Chemical Education, 36, 330-335, (1959)

INGOLD 1933
C. K. Ingold
'Significance of Tautomerism and of the Reactions of Aromatic Compounds in the Electronic Theory of Organic Reactions'
Journal of the Chemical Society, 1120-1127, (1933)

IOFFE et al. 1969
I. I. Ioffe et al.
'Prognosis of chemical reactions by statistical discrimination theory'
Doklady Akademii Nauk SSSR, 189, 1290-1293, (1969)
(English translation)

ISENHOUR et al. 1974
T. L. Isenhour, B. R. Kowalski, P. C. Jurs
'Application of Pattern Recognition to Chemistry'
CRC Reviews in Analytical Chemistry, 4, 1-44, (1974)

IWASA et al. 1965
J. Iwasa, T. Fujita, C. Hansch
'Substituent Constants for Aliphatic Functions Obtained from Partition Coefficients'
Journal of Medicinal Chemistry, 8, 150-153, (1965)

JACOBUS et al. 1970
D. P. Jacobus et al.
'Experience with the Mechanized Chemical and Biological Information Retrieval System'
Journal of Chemical Documentation, 10, 135-140, (1970)

JAFFE 1953
H. Jaffe
'A re-examination of the Hammet Equation'
Chemical Reviews, 53, 191-261, (1953)

JAGER et al. 1968
H. D. M. Jager, D. C. Maxwell, R. G. Ridley
'Information Retrieval and Future Developments at the Mass Spectrometry Data Centre'
Information Storage and Retrieval, 4, 133-137, (1968)

JANZ 1967a
G. Janz
'Thermodynamic Properties of Organic Compounds'
Academic Press, New York, 1967

JANZ 1967b
G. Janz
ch. 2 and 3, p. 13 of JANZ, 1967a

JANZ 1967c
G. Janz
ch. 4-6, p. 50 of JANZ, 1967a

JANZ 1967d
G. Janz
p. 67 of JANZ, 1967a

JARDINE et al. 1971a
N. Jardine and R. Sibson
'Mathematical Taxonomy'
Wiley, New York, 1971

JARDINE et al. 1971b
N. Jardine and R. Sibson
'Choice of methods for automatic classification'
Computer Journal, 14, 404-406, (1971)

JOCHELSON et al. 1968
N. Jochelson, C. M. Mohr, R. C. Reid
'The Automation of Structural Group Contribution Methods in the Estimation of Physical Properties'
Journal of Chemical Documentation, 8, 113-122, (1968)

JOHNSON 1973a
C. D. Johnson
'The Hammett Equation'
Cambridge University Press, Cambridge, 1973

JOHNSON 1973b
C. D. Johnson
ch. 4, p. 96 of JOHNSON, 1973a

JOHNSON 1973c
K. H. Johnson
'Scattered Wave Theory of the Chemical Bond'
Advances in Quantum Chemistry, 7, 143-185, (1973)

JOHNSON et al. 1973
K. H. Johnson, J. G. Norman, J. W. D. Connolly
'The SCF-X Scattered Wave Method'
p. 161-202 in HERMAN et al., 1973

JONES et al. 1965
O. T. G. Jones and W. A. Watson
'Activity of 2-trifluoromethylbenzimidazoles as uncouplers of oxidative phosphorylation'
Nature, 208, 1169-1170, (1965)

JURS 1970
P. C. Jurs
'Mass Spectral Feature Selection and Structural Correlation Using Computerised Learning Machines'
Analytical Chemistry, 42, 1633-1638, (1970)

JURS et al. 1975a
P. C. Jurs and T. L. Isenhour
'Chemical Applications of Pattern Recognition'
Wiley, New York , 1975

JURS et al. 1975b
P. C. Jurs and T. L. Isenhour
ch. 2, p. 9 and ch. 4, p. 31 of JURS et al., 1975a

JURS et al. 1975c
P. C. Jurs and T. L. Isenhour
ch. 3, p. 17 of JURS et al., 1975a

JURS et al. 1975d
P. C. Jurs and T. L. Isenhour
ch. 7, p. 105 of JURS et al., 1975a

JURS et al. 1975e
P. C. Jurs and T. L. Isenhour
ch. 5, p. 83 of JURS et al., 1975a

KAPOOR 1969
S. C. Kapoor
'Dumas and Organic Classification'
Ambix, 16, 1-65, (1969)

KARREMAN 1955
G. Karreman
'Topological Information Content and Chemical Reactions'
Bulletin of Mathematical Biophysics, 17, 279-285, (1955)

KATRITZKY et al. 1972
A. R. Katritzky and R. D. Topsom
'Linear Free Energy Relationships and Optical Spectroscopy'
in CHAPMAN et al., 1972, ch. 3, p. 119

KEKULE 1867a
F. A. Kekule
Zeitschriff Chem., 3, 217, (1867)

KEKULE 1867b
F. A. Kekule
Bull. Acad. roy. Belg., 24, 10, (1867)

KELLIE et al. 1975
G. M. Kellie and F. G. Riddell
'The von Auwers Boiling Point Rule. A New Approach'
Journal of the Chemical Society Perkin 1, 740-744, (1975)

KENNARD et al. 1972
O. Kennard, D. G. Watson, W. G. Twon
'Cambridge Crystallographic Data Centre. I. Bibliographic File'
Journal of Chemical Documentation, 12, 14-19, (1972)

KERMACK et al. 1922
W. O. Kermack and R. Robinson
'An explanation of the property of induced polarity of Atoms and an Interpretation of the Theory of Partial Valencies on an Electronic Basis'
Journal of the Chemical Society, 427-440, (1922)

KIER 1970
L. B. Ker (ed)
'Molecular Orbital Studies in Chemical Pharmacology'
Springer-Verlag, New York, 1970

KIER 1971a
L.B. Kier
'Molecular Orbital Theory in Drug Research'
Academic Press, New York-London, 1971

KIER 1971b
L.B. Kier
ch. 6, p. 68 of KIER, 1971a

KIER 1971c
L. B. Kier
ch. 8, p. 162 of KIER, 1971a

KIER 1972
L. B. Kier
'Molecular Orbital Studies of Biological Molecule Conformations'
ch. 15, p. 278 of VAN VALKENBERG, 1972

KIER et al. 1975a
L. B. Kier et al.
'Molecular Connectivity 1: Relationship to Nonspecific Local Anesthesia'
Journal of Pharmaceutical Sciences, 64, 1971-1974, (1975)

KIER et al. 1975b
L. B. Kier, W. J. Murray, L. H. Hall
'Molecular Connectivity. 4. Relationship to Biological Activitires'
Journal of Medicinal Chemistry, 18, 1272-1274, (1975)

KIHO 1971
Yu. K. Kiho
'The formalized definition of some conccpts of quantitative organic chemistry'

(Russian)

Organic Reactions (Tartu), 8, 429-444, (1971)

KIRSCH 1972
J. F. Kirsch
'Linear Free Energy Relationships in Enzymology'
ch. 8, p. 369 in CHAPMAN et al., 1972

KITAIGORODSKY 1973
A. I. Kitaigorodsky
'Molecular Crystals and Molecules'
Academic Press, New York-London, 1973
pp. 426-434

KLEIR et al. 1975
D. A. Kleir, D. A. Dixon, W. N. Lipscomb
'Localised Molecular Orbitals for Polyatomic Molecules. III. Monocyclic Aromatic Rings'
Theoretica Chimica Acta, 40, 33-45, (1975)

KLOPMAN et al. 1970
G. Klopman and B. O'Leary
'All-Valence Electrons S.C.F. Calculations'
Topics in Current Chemistry, 15, 445-534, (1970)

KOPECKY et al. 1965
J. Kopecky, K. Bocek, D. Vlachova
'Chemical Structure and Biological Activity of m- and p- Disubstituted Derivatives of Benzenes'
Nature, 207, 981, (1965)

KOPECKY et al. 1967
J. Kopecky and K. Bocek
'A Correlation between constants used in structure-activity relationships'
Experientia, 23, 125, (1967)

KOPPEL et al. 1972
I. A. Koppel and V. A. Palm
'The Influence of the Solvent on Organic Reactivity'
in CHAPMAN et al., 1972, ch. 5, p. 203

KOSKINEN et al. 1974
J. R. Koskinen and B. R. Kowalski
'Structure-reactivity correlations for organic molecules by pattern recognition'
NTIS report, AD-785 913, (1974)

KOSKINEN et al. 1975
J. R. Koskinen and B. R. Kowalski
'Interactive Pattern Recognition in the Chemical Laboratory'
Journal of Chemical Information and Computer Sciences, 15, 119-123 (1975)

KOSSEL 1916
W. Kossel
'Uber Molekulbildung uls Frage des Atombaus'
Annalen der Physik, 49, 229-362, (1916)

KOWALSKI 1974
B. R. Kowalski
'Pattern Recognition in Chemical Research'
Computers in Chemical and Biochemical Research, 2, 1-76, (1974)

KOWALSKI et al. 1969
B. R. Kowalski et al.
'Computerised Learning Machines Applied to Chemical Problems. Multicategory Pattern Classification by Least Squares'
Analytical Chemistry, 41, 695-700, (1969)

KOWALSKI et al. 1972a
B. R. Kowalski and C. F. Bender
'Pattern Recognition. A Powerful Approach to Interpreting Chemical Data'
Journal of the American Chemical Society, 94, 5632-5639, (1972)

KOWALSKI et al. 1972b
B. R. Kowalski and C. F. Bender
'The K-nearest neighbour classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation'
Analytical Chemistry, 44, 1405-1411, (1972)

KOWALSKI et al. 1973
B. R. Kowalski and C. F. Bender
'Pattern Recognition. II. Linear and Non-Linear Methods for Displaying Chemical Data'
Journal of the American Chemical Society, 95, 686-693, (1973)

KOWALSKI et al. 1974
B. R. Kowalski and C. F. Bender
'The Application of Pattern Recognition to Screening Potential Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test'
Journal of the American Chemical Society, 96, 916-918,)1974)

KRISHNAMURTHY et al. 1974
E. V. Krishnamurthy, P. V. Sankar, S. Krishnan
'ALWIN-Algorithmic Wiswesser Notation System for Organic Compounds'
Journal of Chemical Documentation, 14, 130-141, (1974)

KUBINYI 1976
H. Kubinyi
'Quantitative Structure-Activity Relationships. 2. A Mixed Approach, Based on Hansch and Free-Wilson Analysis'
Journal of Medicinal Chemistry, 19, 587-600, (1976)

KUBINYI et al. 1976a
H. Kubinyi and O. Kehrhahn
'Quantitative Structure-Activity Relatioships. 1. The Modified Free-Wilson Approach'
Journal of Medicinal Chemistry, 19, 578-586, (1976)

KUBINYI et al. 1976b
H. Kubinyi and O. Kehrhahn
'Quantitative Structure-Activity Relationships. 3. A Comparison of Different Free-Wilson Methods'
Journal of Medicinal Chemistry, 19, 1040-1049, (1976)

KUTTER et al. 1969
E. Kutter and C. Hansch
'Steric Parameters in Drug Design. Moncamine Oxidase Inhibitors and Antihistamines'
Journal of Medicinal Chemistry, 12, 647-652, (1969)

LANGMUIR 1921
I. Langmuir
'Types of Valence'
Science, 54, 59-66, (1921)

LARDER 1967
D. F. Larder
'Alexander Crum Brown and his Doctoral Thesis of 1861'
Ambix, 14, 112-132, (1967)

LAVOISIER et al. 1787
A. L. Lavoisier, L. B. Guyton de Morveau, C. L. Berthollet, and A. F. de Fourcroy
'Méthode de nomenclature chimique'
Cuthet, Paris, 1787

LEO et al. 1969
A. Leo, C. Hansch, C. Church
'Comparison of Parameters Currently Used in the Study of Structure-Activity Relationships'
Journal of Medicinal Chemistry, 12, 766-771, (1969)

LEO et al. 1971
A. Leo, C. Hansch, D. Elkins
'Partition Coefficients and Their Uses'
Chemical Reviews, 71, 525-616, (1971)

LEO et al. 1975
A Leo et al.
'Calculation of Hydrophobic Constant (Log P) from π and f Constants'
Journal of Medicinal Chemistry, 18, 865-868, (1975)

LEWIS 1916
G. N. Lewis
'The Atom and the Molecule'
Journal of the American Chemical Society, 38, 762-785, (1916)

LILJEFORS et al. 1976
T. Liljefors and N. L. Allinger
'Conformational Analysis. CXII. Conformations Energies, and Electronic Absorption Spectra of αβ-Unsaturated Aldehydes and Ketones'
Journal of the American Chemical Society, 98, 2745-2749, (1976)

LIN 1974
T. K. Lin
'Quantum Statistical Calculation for Correlation of Biological Activity and Chemical Structure. 1. Drug-Receptor Interaction'
Journal of Medicinal Chemistry, 17, 151-154, (1974)

LITTLE 1973
W. A. Little
'Molecular Modeling by Computer'
in HERMAN et al., 1973, p. 49-53

LOACH 1974
K. W. Loach
'A Numerical Identifier for the Chemical Elements Expressing Their Periodic Relationships'
Journal of Chemical Documentation, 14, 198-200, (1974)

LYNCH 1968a
M. F. Lynch
'Storage and retrieval of information on chemical structures by computer'
Endeavour, 27, 68-73, (1968)

LYNCH 1968b
M. F. Lynch
'Conversion of Connection Table Descriptions of Chemical Compounds into a Form of Wiswesser Notation'
Journal of Chemical Documentation, 8, 130-133, (1968)

LYNCH 1975a
M. F. Lynch
'Screening Large Chemical Files'
ch. 12, p. 177 in ASH et al., 1975

LYNCH 1975b
M. F. Lynch
'Final report to BLRDD on the project "Development and Assessment of an automatic system for analysing chemical reactions"'
PGSLIS, Sheffield University, 1975

LYNCH et al. 1971a
M. F. Lynch et al.
'Computer Handling of Chemical Structure Information'
Macdonald-Elsevier, London-New York, 1971

LYNCH et al. 1971b
ch. 2, p. 12 of LYNCH et al., 1971a

LYNCH et al. 1971c
M. F. Lynch et al.
ch. 4, p. 52 of LYNCH et al., 1971a

LYNCH et al. 1971d
M. F. Lynch et al.
p. 32-34 of LYNCH et al., 1971a

LYNCH et al. 1971e
M. F. Lynch et al.
ch. 3, p. 36 of LYNCH et al., 1971a

LYNCH et al. 1971f
M. F. Lynch et al.
ch. 5, p. 67 of LYNCH et al., 1971a

MACDANIEL et al. 1958
D. H. Macdaniel and H. C. Brown
'An Extended Table of Hammett Substituent Constants Based on the Ionization of Substituted Benzoic Acids'
Journal of Organic Chemistry, 23, 420-427, (1958)

MACKLE 1954
H. Mackel
'The Evolution of Valence Theory and Bond Symbolism'
Journal of Chemical Education, 31, 618-625, (1954)

MACQUER 1766
Dictionnaire de Chymie, Paris, 1766

MALKIN et al. 1925
T. Malkin and R. Robinson
'Phenyl Benzyl Diketone and Some Derivatives'
Journal of the Chemical Society, 369-377, (1925)

MARSHALL et al. 1974
G. R. Marshall, H. E. Busshard, R. A. Ellis
'Computer Modeling of Chemical Structures'
p. 203-237 in WIPKE et al., 1974a and references therein

MARTIN 1971
H. Martin (ed.)
'Pesticide Manual'
2nd ed., British Crop Protection Council, 1971

MARTIN et al. 1973a
Y. C. Martin and W. J. Dunn
'Examination of the Utility of the Topliss schemes for Analog Synthesis'
Journal of Medicinal Chemistry, 16, 578-579, (1973)

MARTIN et al. 1973b
Y. C. Martin, T. M. Bustard, K. R. Lynn
'Relationship between Physical Properties and Antimalarial Activities of 1,4-Naphthoquinones'
Journal of Medicinal Chemistry, 16, 1089-1093, (1973)

MARTIN et al. 1974
Y. C. Martin et al.
'Discriminant Analysis of the Relationship between Physical Properties and the Inhibition of Monoamine Oxidase by Aminonotetralins and Aminoindans'
Journal of Medicinal Chemistry, 17, 409-413, (1974)

MARTIN et al. 1976
Y. C. Martin and J. J. Hackbarth
'Theoretical Model-Based Equations for the Linear Free Energy Relationships of the Biological Activity of Ionizable Substances. 1. Equilibrium-controlled Potency'
Journal of Medicinal Chemistry, 19, 1033-1039, (1976)

MASINTER et al. 1974
L. M. Masinter et al.
'Applications of Artificial Intelligence for Chemical Inference. XII. Exhaustive Generation of Cyclic and Acyclic Isomers'
Journal of the American Chemical Society, 96, 7702-7714, (1974)

MASON 1976
S. F. Mason
'The Foundations of Classical Stereochemistry'
Topics in Stereochemistry, 9, 1-34, (1976)

MATHEWS 1975
R. J. Mathews
'A Comment on Structure-Activity Correlations, obtained using Pattern Recognition Methods'
Journal of the American Chemical Society, 97, 935-936, (1975)

MATTHEWS 1975
F. W. Matthews
'Organising Information for Retrieval'
ch. 4, p. 32 in ASH et al., 1975

MAZURS 1974
E. G. Mazurs
'Graphic Representations of the Periodic System During One Hundred Years'
University of Alabama Press, Alabama, 1974

McDANIEL et al. 1958
D. H. McDaniel and H. C. Brown
'An Extended Table of Hammett Substituent Constants Based on the Ionisation of Substituted Benzoic Acids'
Journal of Organic Chemistry, 23, 420-427, (1958)

McFARLAND 1971
J. W. McFarland
'On the understanding of drug potency'
Progress in Drug Research, 15, 123-146, (1971)

McGOWAN 1952
J. C. McGowan
'The Physical Toxicity of Chemicals. III. A Systematic Treatment of Physical Toxocity in Aqueous Solutions'
Journal of Applied Chemistry, 2, 651-658, (1952)

McQUITTY 1966
L. L. McQuitty
'Similarity analysis by reciprocal pairs for discrete and continuous data'
Educational and Psychological Measurement, 26, 825-831, (1966)

MEADOW et al. 1970
C. T. Meadow and H. R. Meadow
'Organization, Maintenance, and Search of Machine Files'
in Annual Review of Information Science and Technology, vol. 5, ch. 7, p. 170, (1970)

MEADOWS 1965
E. L. Meadows
'Estimating physical properties - The A.I.Ch.E. System'
Chemical Engineering Progress, 61, 93-95, (1965)

MENDENHALL 1974
D. M. Mendenhall
'Cost Comparison of Four Data Input Methods'
Journal of Chemical Documentation, 14, 109-111, (1974)

MEYER 1899
H. Meyer
'Zur Theorie der Alkoholnarkose'
Archiv fur Experimentelle Pathologie und Pharmakologie, 42, 109-118, (1899)

MEYER 1970
E. F. Meyer
'Towards an Automatic Three-Dimensional Display of Structural Data'
Journal of Chemical Documentation, 10, 85-86, (1970)

MEYER 1974
E. Meyer
'Topological Search for Classes of Compounds in Large Files'
p. 105-122 in WIPKE et al., 1974a

MORIGUCHI 1975
I. Moriguchi
'Quantitative Structure-Activity Studies. 1. Parameters relating to Hydrophobicity'
Chemical and Pharmaceutical Bulletin (Tokyo), 23, 247-257, (1975)

MUIR 1907
M. M. Pattison Muir
'A history of chemical theories and laws'
Wiley, New York, 1907, p. 189
reprinted 1975, Arno Press, New York

MULLEN 1967
J. M. Mullen
'Atom-by-Atom Typewriter Input for Computerised Storage and Retrieval of Chemical Structures'
Journal of Chemical Documentation, 7, 88-93, (1967)

MUNZ 1968
J. Munz
'The formal analysis of notation systems'
in J. P. Mitchell (ed.) 'Proceedings of the Wiswesser Line Notation meeting of the ACIDS program'
Report AD 665 397

MURRAY et al. 1975
W. J. Murray, L. H. Hall, L. B. Kier
'Molecular Connectivity III. Relationship to Partition Coefficients'
Journal of Pharmaceutical Sciences, 64, 1978-1981, (1975)

MURRAY et al. 1976
W. J. Murray, L. B. Kier, L. H. Hall
'Molecular Connectivity. 6. Examination of the Parabolic Relationship between Molecular Connectivity and Biological Activity'
Journal of Medicinal Chemistry, 19, 573-578, (1976)

MURRELL et al 1972
J. A. Murrell and A. J. Haryet
'Semi-empirical self-consistent-field molecular orbital theory of molecules'
Wiley, London, 1972

N.A.S. 1964
'Survey of Chemical Notation Systems'
National Academy of Sciences,
NAS-NRC Publication 1150, Washington, 1964

N.A.S. 1965
'Survey of European Non-Conventional Chemical Notational Systems'
National Academy of Sciences,
NAS-NRC Publication 1275, Washington, 1965

N.A.S. 1969
'Chemical Structure Information Handling - A Review of the Literature 1962-1968'
National Academy of Sciences,
NAS-NRC Publication, 1733, Washington, 1969

N.C.I. 1974
'Proceedings of the Symposium on Structure-Activity Techniques in the Treatment of Antitumor Drug Data'
(National Cancer Institute)
Cancer Chemotherapy Reports, Part 2, 4, 31-52, (1974)

NEELY et al. 1968
W. B. Neely and W. K. Whitney
'Statistical Analysis of Insecticidal Activity in a Series of Phosphoramidates'
Journal of Agricultural and Food Chemistry, 16, 571-573, (1968)

NEELY et al. 1970
W. B. Neely et al.
'Structure Activity Analysis of Some O,O-Dralkyl (p-Methylthio, p-methylsulfonyl) Phenyl Phosphates and Phosphorothioates prepared for their insecticidal activity'
Journal of Agricultural and Food Chemistry, 18, 45-49, (1970)

NEWMAN 1955
M. S. Newman
'A Notation for the Study of certain stereochemical problems'
Journal of Chemical Education, 32, 344-347, (1955)

NICHOLSON, 1795
Dictionary of Chemistry, London, 1795, Vol. , pp. 251-2.

NILSSON, 1965a
N. J. Nilsson
'Learning Machines'
McGraw-Hill, New York, 1965

NILSSON 1965b
N. J. Nilsson
ch. 3, p. 43 of NILSSON, 1965a

NILSSON 1965c
N. J. Nilsson
ch. 4, p. 65 of NILSSON, 1965a

NORRINGTON et al. 1975
F. E. Norrington et al.
'Physicochemical-Activity Relations in Practice.
1. A Rational and Self-Consistent Data Bank'
Journal of Medicinal Chemistry, 18, 604-607, (1975)

NUBLING et al. 1970
W. Nubling and W. Steidle
'The Dokumentationsring der Chemisch-pharmazeutischen Industrie;
Aims and Methods'
Angewandte Chemie International Edition, 9, 596-598, (1970)

NYS et al. 1973
G. G. Nys and R. F. Rekker
'Statistical Analysis of a Series of Partition Coefficients
with Special Reference to the Predictability of Folding of Drug
Molecules'
Chimie Therapeutique, 521-535, (1973)

NYS et al. 1974
G. G. Nys and R. F. Rekker
'The Concept of Hydrophobic Fragmental Constants (f-Values).
11. Extension of its Applicability to the Calculation of Lipophilicities of Aromatic and Heteroaromatic Structures'
Chimica Therapeutica, 9, 361-375, (1974)

O'LEARY et al. 1975
B. O'Leary, B. J. Duke, J. E. Eilers
'Utilization of Transferability in Molecular Orbital Theory'
Advances in Quantum Chemistry, 9, 1-67, (1975)

OHNACKER et al. 1970
G. Ohnacker and W. Kalbfleisch
'CCBF- A System for the Computer Processing of Chemical and
Biological Facts'
Angewandte Chemie International Edition, 9, 605-610, (1970)
(Permuted lists)

ORMEROD 1970
M. B. Ormerod
'The Architecture and Properties of Matter'
Arnold, London, 1970 and references therein

OSINGA et al. 1974
M. Osinga and A. A. Verrijn Stuart
'Documentation of Chemical Reactions. II. Analysis of the
Wiswesser Line Notation'
Journal of Chemical Documentation, 14, 194-198, (1974)

OSINGA et al. 1976
M. Osinga and A. A. Verijn Sturat
'Documentation of Chemical Reactions. III. Encoding of the Facets'
Journal of Chemical Information and Computer Sciences, 16, 165-171,
(1976)

OVERTON 1897
E. Overton
'Uber die osmotischen Eigenschaften der Zelle in iher Bedentung fur die Toxikologie und Pharmakologie'
Zeitschrift fur Physikalische Chemie, 22, 189-209, (1897)

OXMAN et al. 1976
M. A. Oxman et al.
'The Toxicology Data Bank'
Journal of Chemical Information and Computer Sciences, 16, 19-21, (1976)

PARKS et al. 1932
G. S. Parks and H. M. Huffman
'The Free Energies of Some Organic Compounds'
American Chemical Society Monograph No. 60
Chemical Catalogue Co., New York, 1932

PARTINGTON 1951
J. R. Partington
'An Advanced Theory of Physical Chemistry'
Vol. 2, pp. 17-28, 283-303
Vol. 4, pp. 42-66
Longmans Green and Co., London, 1951

PARTINGTON 1962
J. R. Partington
'A History of Chemistry'
Vol. 3, MacMillan, London, 1962

PARTINGTON 1964a
J. R. Partington
'A History of Chemistry'
Vol. 4, MacMillan, London, 1964

PARTINGTON 1964b
J. R. Partington
ch. 11, p. 337 of PARTINGTON, 1964a

PARTINGTON 1964c
J. R. Partington
ch. 14, p. 432 of PARTINGTON, 1964a

PARTINGTON 1964d
J. R. Partington
pp. 918-922 of PARTINGTON, 1964a

PAULING 1939
L. C. Pauling
'The nature of the chemical bond and the structure of molecules and crystals' First edition
Cornell University Press, Ithaca, New York, 1939

PEARSON 1972
R. G. Pearson
'The Influence of the Reagent on Organic Reactivity'
in CHAPMAN et al., 1972, ch. 6, p. 281

PEDLEY 1976
J. B. Pedley (ed.)
'Computer Analysis of Thermochemical Data (CATCH Tables)'
School of Molecular Sciences, University of Sussex, Brighton

PERKINS et al. 1975
W. J. Perkins and B. J. Hammond
'Computer-aided thought in biomedical research'
Nature, 256, 171-175, (1975)

PERRIN 1974
C. L. Perrin
'Testing of Computer-Assisted Methods for Classification of Pharmacological Activity'
Science, 183, 551-552, (1974)

PETERSEN 1970
Q. R. Petersen
'Some Reflections on the Use and Abuse of Molecular Models'
Journal of Chemical Education, 47, 24-29, (1970)
(and references therein)

PHARMA-DOK 1972
Pharma-Dokumentationsring e.V. and Derwent Publications Ltd.
'RINGDOC literature documentation. Instruction Bulletin No. 5. Ring Code - General and Chemical Code'
4th edition, April, 1972

PITZER 1940
K. S.Pitzer
'Chemical Equilibrium, Free Energies, and Heat Contents for Gaseous Hydrocarbons'
Chemical Reviews, 27, 39-57, (1940)

PLATT 1960
J. R. Platt
'The Need for better Macromolecular Models'
Science, 131, 1309-1310, (1960)

POPLE et al. 1970
J. A. Pople and D. L. Beveridge
'Approximate Molecular Orbital Theory'
McGraw-Hill, New York, 1970

PORT et al. 1974
G. N. J. Port and B. Pullman
'An ab initio SCF Molecular Orbital Study on the Conformation of Serotonin and Bufotenine'
Theoretica Chimica Acta, 33, 275-278, (1974)

PORTER 1965
G. Porter
'The Chemical Bond since Frankland'
Proceedings of the Royal Institution of Great Britain, 40, 384-397, (1965)

PULLMAN 1972
A. Pullman
'Quantum Biochemistry at the All- or Quasi-all-Electrons Level'
Topics in Current Chemistry, 31, 45-103, (1972)

PULLMAN et al. 1973
A. Pullman and G. N. J. Port
'An ab initio SCF Molecular Orbital Study of Acetylcholine'
Theoretica Chimica Acta, 32, 77-79, (1973)

PULLMAN 1976
B. Pullman (ed.)
'Quantum Mechanics of Molecular Conformations'
Wiley, London, 1976

PURCELL et al. 1970
W. P. Purcell and J. M. Clayton
'Quantitative Structure-Activity Relationships and Molecular Orbitals in Medicinal Chemistry'
p. 145-155 of KIER, 1970

PURCELL et al. 1971
W. P. Purcell and J. M. Clayton
'Physicochemical Approaches to Drug Design'
ch. 5, p. 123 of BLOOM et al., 1971a

PURCELL et al. 1973a
W. P. Purcell, G. E. Bass, J. M. Clayton
'Strategy of Drug Design: A Guide to Biological Activity'
Wiley, New York, 1973

PURCELL et al. 1973b
W. P. Purcell, G. E. Bass, J. M. Clayton
ch. 1, p. 1 of PURCELL et al., 1973a

PURCELL et al. 1973c
W. P. Purcell, G. E. Bass, J. M. Clayton
ch. 2, 3, 4, p. 21 of PURCELL et al., 1973a

PURCELL et al. 1973d
W. P. Purcell, G. E. Bass, J. M. Clayton
ch. 5, 6, p. 89 of PURCELL et al., 1973a

QUAM et al. 1934
G. N. Quam and M. B. Quam
'Types of Graphic Classifications of the Elements'
Journal of Chemical Education, 11, 27, 217, 288, (1934)
(and references therein)

RANDIC 1974
M. Randic
'On the recognition of identical graphs representing molecular topology'
Journal of Chemical Physics, 60, 3920-3928, (1974)

RANDIC 1975
M. Randic
'On Characterization of Molecular Branching'
Journal of the American Chemical Society, 97, 6609-6615, (1975)

RANKIN et al. 1971
K. Rankin and S. J. Tauber
'Linguistics as a Basis for Abalyzing Chemical Structure Diagrams'
Journal of Chemical Documentation, 11, 139-141, (1971)

RASHEVSKY 1955
N. Rashevsky
'Life, Information Theory, and Topology'
Bulletin of Mathematical Biophysics, 17, 229-235, (1955)

REDL et al. 1974
G. Redl, R. D. Cramer, C. E. Berkoff
'Quantitative Drug Design'
Chemical Society Reviews, 3, 273-292, (1974)

REID et al. 1966
R. Reid and T. K. Sherwood
'The Properties of Gases and Liquids (Their Estimation and Correlation)'
2nd ed., McGraw Hill, New York, 1966

RICHARDS et al. 1974
W. G. Richards et al.
'Bibliography of Ab Initio Molecular Wave Functions. Supplement for 1970-73'
Clarendon Press, Oxford, 1974

RICHARDSON 1901
G. M. Richardson
'The foundations of stereochemistry'
American Book Company, New York, 1901

RITCHIE et al. 1964
C. D. Ritchie and W. F. Sayer
'An examination of Structure-Reactivity Relationships'
Progress in Physical Organic Chemistry, 2, 323-400, (1964)

ROBINSON 1932
R. Robinson
'Two lectures on an outline of an electrochemical (electronis) theory of the course of organic reactions'
Institute of Chemistry, London, 1932

ROBINSON 1974
F. A. Robinson
'Therapeutic Innovation: the end or a new beginning'
Chemistry in Britain, 10, 129-136, (1974)

ROMANEC 1974
M. J. Romanec
'Code for Chemical Ring Compounds with Application of Fusion Lines, Suited for Calculation of Their Physical Properties'
Journal of Chemical Documentation, 14, 49-52, (1974)

ROSANOFF 1906
M. A. Rosanoff
'On Fischer's Classification of Stereo-Isomers'
Journal of the American Chemical Society, 28, 114-121, (1906)

ROSSLER et al. 1970
S. Rossler and A. Kolb
'The GREMAS System, an Integral Part of the IDC System for Chemical Documentation'
Journal of Chemical Documentation, 10, 128-134, (1970)

ROTH et al. 1969
B. Roth and J. Z. Strelitz
'The Protonation of 2,4 Diaminopyrimidines. 1. Dissociation Constants and Substituent Effects'
Journal of Organic Chemistry, 34, 821-836, (1969)

RUEDENBERG 1973
K. Ruedenberg
'Description of molecular in terms of localized orbitals'
p. 149-156 in HERMAN et al., 1973

RUSSELL 1963a
C. A. Russell
'The Electrochemical Theory of Berzelius. Part I. Origins of the Theory'
Annals of Science, 19, 117-126, (1963)

RUSSELL 1963b
C. A. Russell
'The Electrochemical Theory of Berzelius. Part II. An Electrochemical View of Matter'
Annals of Science, 19, 127-145, (1963)

RUSSELL 1971a
C. A. Russell
'The History of Valency'
Leicester University Press, 1971

RUSSELL 1971b
C. A. Russell
ch. 2, p. 21 of RUSSELL, 1971a

RUSSELL 1971c
C. A. Russell
ch. 3, p. 44 of RUSSELL, 1971a

RUSSELL 1971d
C. A. Russell
ch. 6, p. 108 of RUSSELL 1971a

RUSSELL 1971e
ch. 7, p. 137 of RUSSELL, 1971a

RUSSELL 1971f
C. A. Russell
ch. 8, p. 142 of RUSSELL, 1971a

RUSSELL 1971g
C. A. Russell
ch. 5,p. 92 of RUSSELL, 1971a

RUSSELL 1971h
C. A. Russell
ch. 4, p. 81 of RUSSELL, 1971a

RUSSELL 1971i
C. A. Russell
ch. 11, p. 224 and ch. 12, p. 242 of RUSSELL, 1971a

RUSSELL 1971j
C. A. Russell
ch. 13, p. 313 of RUSSELL, 1971a

SAGGERS 1974
D. T. Saggers
'The Application of the Computer to a Pesticide Screening Programme'
Pesticide Science, 5, 341-352, (1974)

SANDERSON 1962
R. T. Sanderson
'Teaching Chemistry with Models'
Van Nostrand, New York-London, 1962

SANKAR et al. 1974
P. V. Sankar, E. V. Krishamurthy, S. Krishran
'Representation of Stereoisomers in ALWIN'
Journal of Chemical Documentation, 14, 141-146, (1974)

SANTORA et al. 1975
N. J. Santora and K. Auyany
'Non-computer Approach to Structure-Activity Study. An Expanded Fibonacci Search Applied to Structurally Diverse Types of Compounds'
Journal of Medicinal Chemistry, 18, 959-963, (1975)

SAVORY 1976
C. G. Savory
'A survey of crystal and molecular models'
Education in Chemistry, 13, 136-139, (1976)

SCHERAGA 1968
H. A. Scheraga
'Calculations of Conformations of Polypeptides'
Advances in Physical Organic Chemistry, 6, 103-184, (1968)

SCHIFFMAN 1975
S. S. Schiffman
'Physicochemical Correlates of Olfactory Quality'
Science, 185, 112-117, (1975)

SCHNAARE 1971
R. L. Schnarre
'Electronic Aspects of Drug Action'
ch. 5, p. 404 of ARIENS, 1971a

SCHRODINGER 1928
E. Schrodinger
'Collected papers on wave mechanics'
Translated from the second German edition
Blackie and Son, London-Glasgow, 1928

SCHULTZ 1974
J. L. Schultz
'Handling Chemical Information in the Du Pont Central Report Index'
Journal of Chemical Documentation, 14, 171-179, (1974)

SHORTER 1972a
J. Shorter
'The Separation of Polar, Steric, and Resonance Effects by the Use of Linear Free Energy Relationships'
in CHAPMAN et al., 1972
ch. 2, p. 71

SHORTER 1972b
J. Shorter
p. 103-110 of SHORTER, 1972a

SHORTER 1973
J. Shorter
'Correlation Analysis in organic chemistry'
Clarendon Press, Oxford, 1973
Appendix p. 103

SIDGWICK 1933
N. V. Sidgwick
'Some Physical Properties of the covalent link in chemistry'
Cornell University Press, Ithaca, New York, 1933
ch. 4, p. 95

SIEGAL 1956
S. Siefal
'Non-parametric statistics for the behavioural sciences'
McGraw-Hill, New York, 1956
p. 19

SLIK 1963
J. A. Silk
'A Linear Notation for Organic Compounds'
Journal of Chemical Documentation, 3, 189-195, (1963)

SILVERMAN et al. 1974
M. Silverman and P. R. Lee
'Pills, Profits and Politics'
University of California Press, Berkeley and Los Angeles, 1974

SINGER et al. 1967
J. A. Singer and W. P. Purcell

'Relationships among Current Quantitative Structure-Activity Models'
Journal of Medicinal Chemistry, 10, 1000-1002, (1967)

SJOSTROM et al. 1974
M. Sjostrom and S. Wold
'Statistical Analysis of the Hammett Equation. II. A unified inductive sigma scale'
Chemica Scripta, 6, 114-121, (1974)

SKOLNIK et al. 1964
H. Skolnik and A. Clow
'A Notation System for Indexing Pesticides'
Journal of Chemical Documentation, 4, 221-227, (1964)

SMEATON 1954
W. A. Smeaton
'The Contributions of P. J. Macquer, T. O. Bergman, and L. B. Guyton de Morveau to the reform of chemical nomenclature'
Annals of Science, 10, 87-106, (1954)

SMITH 1975
D. H. Smith
'The Scope of Structural Isomerism'
Journal of Chemical Information and Computer Sciences, 15, 203-207, (1975)

SMITH et al. 1975
E. G. Smith and P. A. Baker
'The Wiswesser Line-Formula Chemical Notation (WLN)'
Third Edition
Chemical Information Management Inc., Cherry Hill, New Jersey, 1975

SNEATH 1957
P. H. A. Sneath
'The Application of computers to taxonomy'
Journal of General Microbiology', 17, 201-226, (1957)

SNEATH 1966
P. H. A. Sneath
'Relations between Chemical Structure and Biological Activity in Peptides'
Journal of Theoretical Biology, 12, 157-195, (1966)

SNEATH et al. 1973
P. H. A. Sneath and R. R. Sokal
'Numerical Taxonomy'
W. H. Freeman, San Francisco, 1973

SNEDECOR et al. 1967a
G. W. Snedecor and W. G. Cochran
'Statistical Methods', 6th edition
Iowa State University Press, Ames Iowa, 1967

SNEDECOR et al. 1967b
W. G. Snedecor and G. W. Cochran
p. 141 of SNEDECOR et al., 1967a

SOKAL 1974
R. R. Sokal
'Classification: Purposes, Principles, Progress, Prospects'
Science, 185, 1115-1123, (1974)

SOKAL et al. 1958
R. R. Sokal and C. D. Michener
'A Statistical Method for Evaluating Systematic Relationships'
University of Kansas Science Bulletin, 38, 1409-1438, (1958)

SOLOVEICHIK et al. 1967
S. Soloveichik and H. Krakauer
ch. 2, p. 3 of 'Werner Centennial'
G. B. Kaufmann (ed.), Advances in Chemistry Series No. 62
American Chemical Society, Washington, 1967

SOLTZBERG et al. 1976
L. J. Soltzberg and C. L. Wilkins
'Computer Recognition of Activity Class fromMolecular Transforms'
Journal of the American Chemical Society, 98, 4006, (1976)

SPINKS 1973
A. Spinks
'The changing role of chemistry in product innovation'
Chemistry and Industry, 885-891, (1973)

STOCK et al. 1963
L. M. Stock and H. C. Brown
'A Quantitative Treatment of Directive Effects in Aromatic Substitution'
Advances in Physical Organic Chemistry, 1, 35-154, (1963)

STOCKTON et al. 1974
F. G. Stockton and R. L. Merritt
'The Shell Chemical Typewriter File System'
Journal of Chemical Documentation, 14, 166-170, (1974)

STONHAM et al. 1975
T. J. Stonham and M. A. Shaw
'Automatic Classification of Mass Spectra by Means of Digital Learning Nets - Existence of Characteristic Features of Chemical Class in Mass Spectra'
Pattern Recognition, 7, 235-241, (1975)

STREITWIESER 1961
A. Streitwieser
'Molecular Orbital Theory for organic chemists'
Wiley, New York, 1971

SNEDECOR et al. 1967b
W. G. Snedecor and G. W. Cochran
p. 141 of SNEDECOR et al., 1967a

SOKAL 1974
R. R. Sokal
'Classification: Purposes, Principles, Progress, Prospects'
Science, 185, 1115-1123, (1974)

SOKAL et al. 1958
R. R. Sokal and C. D. Michener
'A Statistical Method for Evaluating Systematic Relationships'
University of Kansas Science Bulletin, 38, 1409-1438, (1958)

SOLOVEICHIK et al. 1967
S. Soloveichik and H. Krakauer
ch. 2, p. 3 of 'Werner Centennial'
G. B. Kaufmann (ed.), Advances in Chemistry Series No. 62
American Chemical Society, Washington, 1967

SOLTZBERG et al. 1976
L. J. Soltzberg and C. L. Wilkins
'Computer Recognition of Activity Class from Molecular Transforms'
Journal of the American Chemical Society, 98, 4006, (1976)

SPINKS 1973
A. Spinks
'The changing role of chemistry in product innovation'
Chemistry and Industry, 885-891, (1973)

STOCK et al. 1963
L. M. Stock and H. C. Brown
'A Quantitative Treatment of Directive Effects in Aromatic Substitution'
Advances in Physical Organic Chemistry, 1, 35-154, (1963)

STOCKTON et al. 1974
F. G. Stockton and R. L. Merritt
'The Shell Chemical Typewriter File System'
Journal of Chemical Documentation, 14, 166-170, (1974)

STONHAM et al. 1975
T. J. Stonham and M. A. Shaw
'Automatic Classification of Mass Spectra by Means of Digital Learning Nets - Existence of Characteristic Features of Chemical Class in Mass Spectra'
Pattern Recognition, 7, 235-241, (1975)

STREITWIESER 1961
A. Streitwieser
'Molecular Orbital Theory for organic chemists'
Wiley, New York, 1971

STULL et al. 1969
D. R. Stull, E. F. Westrum, G. C. Sinke
'The Chemical Thermodynamics of Organic Compounds'
Wiley, New York, 1969, ch. 6, p. 140

STUPER et al. 1975
A. J. Stuper and P. C. Jurs
'Classification of Psychotropic Drugs as Sedatives or Tranquilizers Using Pattern Recognition Techniques'
Journal of the American Chemical Society, 97, 182-187, (1975)

STUPER et al. 1976
A. J. Stuper and P. C. Jurs
'ADAPT: A Computer System for Automated Data Analysis using Pattern Recognition Techniques'
Journal of Chemical Information and Computer Sciences, 16, 99-105, (1976)

SUGDEN 1930
S. Sugden
'The Parachor and Valency'
Routledge, London, 1930

SWAIN et al. 1968
C. G. Swain and E. C. Lupton
'Field and Resonance Components of Substituent Effects'
Journal of the American Chemical Society, 90, 4328-4337, (1968)

TAFT 1952
R. W. Taft
'Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters'
Journal of the American Chemical Society, 74, 3120-3127, (1952)

TAFT et al. 1959
R. W. Taft and I. C. Lewis
'Evaluation of Resonance Effects on Reactivity by Application of the Linear Inductive Energy Relationship. V. Concerning a σ_R Scale of Resonance Effects'
Journal of the American Chemical Society, 81, 5343-5352, (1959)

TATE 1967
F. A. Tate
'Handling Chemical Compounds in Information Systems'
in 'Annual Review of Information Science and Technology'
ed. C. A. Cuadra, 2, 285-309, (1967)

TAUBER et al. 1972
S. J. Tauber and K. Rankin
'Valid Structure Diagrams and Chemical Gibberish'
Journal of Chemical Documentation, 12, 30-34, (1972)

TAYLOR 1976a
F. Sherwood Taylor
'The Alchemists'
Paladin, London, 1976

TAYLOR 1976b
F. Sherwood Taylor
ch. 5, p. 51 of TAYLOR, 1976a

TEELING-SMITH 1967
G. Teeling-Smith (ed.)
'Innovation and the Balance of Payments - tje Experience in the Pharmaceutical Industry'
Office of Health Economics, London, 1967

THIELE 1899
J. Thiele
'Theorie der ungesättigten und aromatischen'
Liebig's Annalen, 306, 87-142, (1899)

THOMSON 1914
J. J. Thomson
'The Forces between Atoms and Chemical Affinity'
Philosophical Magazine, 27, 757-789, (1914)

THOMSON 1936
R. C. Thomson
'Dictionary of Assyrian Chemistry and Geology'
Oxford, 1936

THOMSON et al. 1967
L. H. Thomson, E. Hyde, F. W. Matthews
'Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation'
Journal of Chemical Documentation, 7, 204-209, (1967)

TING et al. 1973
K. H. Ting et al.
'Applications of Artificial Intelligence: Relationships between Mass Spectra and Pharmacological Activity of Drugs'
Science, 180, 417-420, (1973)

TOLLANAERE 1973
J. P. Tollanaere
'Structure-activity relationships of three groups of uncouplers of oxidative phosphorylation: Salicylanilides, 2-trifluoremethyl-benzimidazoles, and phenols'
Journal of Medicinal Chemistry, 16, 791-796, (1973)

TOPLISS 1972
J. G. Topliss
'Utilization of Operational Schemes for Analog Synthesis in Drug Design'
Journal of Medicinal Chemistry, 15, 1006-1011, (1972)

TOPLISS et al. 1972
J. G. Topliss and R. J. Costello
'Chance Correlation in Structure-Activity Studies using Multiple Regression Analysis'
Journal of Medicinal Chemistry, 15, 1066-1068, (1972)

TOPLISS et al. 1975
J. G. Topliss and Y. C. Martin
'Utilization of Operational Schemes for Analog Synthesis in Drug Design'
ch. 1, p. 1 in E. J. Ariens (ed.)
'Drug Design', Vo. 5
Academic Press, London-New York, 1975

TRIBBLE et al. 1972
M. T. Tribble and J. G. Traynham
'Linear Correlations of Substituent Effects in ^{1}H, ^{19}F, and ^{13}C Nuclear Magnetic Resonance Spectroscopy'
in CHAPMAN et al., 1972, ch. 4, p. 143

TUKEY et al. 1966
J. W. Tukey and M. B. Wilk
'Data Analysis and Statistics: An Expository Overview'
AFIPS Conference Proceedings, Vol. 29, Fall Joint Computer Conference, 1966

TUTE 1971
M. S. Tute
'Principles and Practice of Hansch Analysis'
Advances in Drug Research, 6, 1-77, (1971)

UFTON 1976
J. Ufton
M.Sc. dissertation, PGSLIS, Sheffield University, 1976

UNGER et al. 1973
S. H. Unger, C. Hansch
'On Model Building in Structure-Activity Relationships. A re-examination of Adrenergic Blocking Activity of β-Halo-β-arylalkylamines'
Journal of Medicinal Chemistry, 16, 745-749, (1973)

VALLS 1974
J. Valls
'Reaction Documentation'
p. 83-103 of WIPKE et al., 1974a

VALLS et al. 1975
J. Valls and O. Schier
'Chemical Reaction Indexing'
ch. 15, p. 243 of ASH et al., 1975

VAN RIJSBERGEN 1975
C. J. van Rijsbergen
'Information Retrieval'
Butterworths, London, 1975

VAN VALKENBERG 1972
W. van Valkenberg (ed.)
'Biological Correlations - The Hansch Approach'
Advances in Chemistry Series No. 114
American Chemical Society, Washington, 1972

VAN DER STOUW et al. 1974
G. G. Vander Stouw, P. M. Elliott, A. C. Isenberg
'Automated Conversion of Chemical Substance names to Atom-Bond Connection Tables'
Journal of Chemical Documentation, 14, 183-193, (1974)

VERKADE 1953
P. E. Verkade
'Organic Chemical Nomenclature' Past, Present and Future'
p. 75-82 of A.C.S., 1953

VERLOOP 1972
A. Verloop
'The Use of Linear Free Energy Parameters and Other Experimental Constants in Structure-Activity Studies'
ch. 2, p. 133 in E. J. Ariens (ed.) 'Drug Design'
Vol. III, Academic Press, London-New York, 1972

VILLARREAL et al. 1976
J. Villarreal et al.
'CRYSRC: A Generalised Chemical Information System Applied to a Structural Data File'
Journal of Chemical Information and Computer Sciences, 16, 220-225, (1975)

VLEDUTS 1963
G. E. Vleduts
'Concerning one system of classification and codification of organic reactions'
Information Storage and Retrieval, 1, 117-146, (1963)

VOGEL 1948
A. I. Vogel
'Physical and Chemical Constitution. Part XXIII'
Journal of the Chemical Society, 1833-1855, (1948)

VOGEL et al. 1950
A. I. Vogel et al.
'Bond Refractivities and Bond Parachors'
Chemistry and Industry, 358, (1950)

VON NIESSEN 1973
W. von Niessen
'A Theory of Molecules in Molecules. II. The Theory and its application to the Molecules Be-Be, Li_2-Li_2 and to the Internal Rotation in C_2H_6'
Theoretica Chimica Acta., 31, 111-135, (1973)

VON NIESSEN 1974
W. von Niessen
'Localized Molecular Orbitals for Aromatic Molecules. Mono- and Disubstituted Benzenes'
Theoretica Chimica Acta., 33, 185-200, (1974)

VON NIESSEN 1975
W. von Niessen
'On a Definition of Bond Energies Based on Localized Molecular Orbitals'
Theoretica Chimica Acta., 38, 9-20, (1975)

WALTON 1969
A. Walton
'The Use of Models in Stereochemistry'
Progress in Stereochemistry, 4, 335-375, (1969)

WARD 1963
J. H. Ward
'Hierarchical Grouping to Optimise an Objective Function'
Journal of the American Statistical Association, 58, 236-244, (1963)

WEINER et al. 1973
M. L. Weiner and P. H. Weiner
'A study of Structure-Activity Relationships of a Series of Diphenylaminopropanols by Factor Analysis'
Journal of Medicinal Chemistry, 16, 655-667, (1973)

WEINSTEIN et al. 1971
H. Weinstein, R. Pauncz, M. Cohen
'Localised Molecular Orbitals'
Advances in Atomic and Molecular Physics, 7, 97-140, (1971)

WERNER 1893
A. Werner
'Beitrag zur Konstitution anorganischer verbindungen'
Zeitschrift fur anorganische und allgemeine chemie, 3, 267-330, (1893)

WHELAND 1944
G. W. Wheland
'The Theory of Resonance and its application to organic chemistry'
Wiley, New York, Chapman and Hall, London, 1944

WILLETT 1976
P. Willett
M.Sc. dissertation, PGSLIS, Sheffield University, 1976

WILLIAMS 1974
M. E. Williams
'Use of Machine-Readable Data Bases'
in Annual Review of Information Science and Technology, Vol. 9, ch. 7, p. 221, (1974)

WILLIAMS et al. 1971
W. T. Williams et al.
'Controversy concerning the criteria for taxonometric strategies'
Computer Journal, 162-165, (1971)

WILLIAMS et al. 1976
S. G. Williams and F. E. Norrington
'Determination of Positonal Weighting Factors for the Swain and Lupton Substituent Constants $\mathcal{F}$ and R'
Journal of the American Chemistry Society, 98, 508-516, (1976)

WILLIAMSON 1852
A. W. Williamson
'On the constitution of salts'
Journal of the Chemical Society, 4 (old style), 350-355, (1852)

WINDERLICH 1953
R. Winderlich
'History of the Chemical Sign Language'
Journal of Chemical Education, 30, 58-62, (1953)

WIPKE 1974
W. T. Wipke
'Computer-Assisted Three-Dimensional Synthetic Analysis'
in WIPKE et al. 1974a, p. 147-174

WIPKE et al. 1974a
W. T. Wipke et al. (eds.)
'Computer Representation and Manipulation of Chemical Information'
NATO Advanced Study Institute, Noordwijkerhout, Netherlands, 1973
Wikey - New York, 1974

WIPKE et al. 1974b
W. T. Wipke and T. M. Dyott
'Simulation and Evaluation of Chemical Synthesis. Computer Representation and Manipulation of Stereochemistry'
Journal of the American Chemical Society, 96, 4825-34, (1974)

WISWESSER 1952
W. J. Wiswecser
'The Wiswesser Line-Formula Notation'
Chemical and Engineering News, 30, 3523-3526, (1952)

WISWESSER 1968
W. J. Wiswesser
'107 Years of Line-Formula Notations (1861-1968)'
Journal of Chemical Documentation, 8, 146-150, (1968)

WISWESSER 1975
W. J. Wiswesser
'Historical Development of Chemical Notations'
ch. 8, p. 92 of ASH et al., 1975

WISWESSER et al. 1973
W. J. Wiswesser, C. L. Crum, K. J. Windlinx, R. A. Creager
'Some Chemical Notation Co-operative Activities', 13, 74-77, (1973)

WISWESSER et al. 1974
W. J. Wiswesser, R. A. Creager, K. J. Windlinx
'A Chemical-Biological Data Base for Herbicidal Information'
Report AD/A-003 254, 1974

WOHL 1971
A. J. Wohl
'A Molecular Orbital Approach to Quantitative Drug Design'
ch. 4, p. 381 of ARIENS, 1971a

WOLD 1974
S. Wold
'A Theoretical Foundation of Extrathermodynamic Relationships (Linear Free Energy Relationships)'
Chemica Scripta, 5, 97-106, (1974)

WOLD et al. 1972
S. Wold and M. Sjostrum
'Statistical Analysis of the Hammett Equation. I. Methods and Model Calculations'
Chemica Scripta, 2, 49-55, (1972)

WOODRUFF et al. 1975
H. B. Woodruff, S. R. Lowry, T. L. Isenhour
'A Text Search System Using Boolean Strategies for the Identification of Infrared Spectra'
Journal of Chemical Information and Computer Sciences, 15, 207-212, (1975)

WOOTON et al. 1975
R. Wootoon et al.
'Physicochemical-Activity Relationships in Practice. 2. Rational Selection of Benzenoid Substituents'
Journal of Medicinal Chemistry, 18, 607-613, (1975)

YUKAWA et al. 1959
Y. Yukawa and Y. Tsuno
'Resonance Effects in Hammett Relationship. II. Sigma Constants in Electrophilic Reactions and their Intercorrelation'
Bulletin of the Chemical Society of Japan, 32, 965-971, (1959)

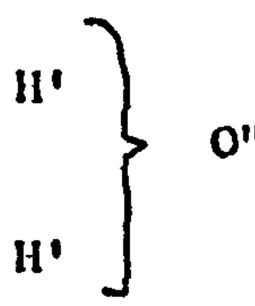
YUKAWA et al. 1966
Y. Yukawa, Y. Tsuno and M. Sawada
'Resonance Effect in Hammett Relationship. IV. Linear Free Energy Relationship based on the Normal Substituent Constants'
Bulletin of the Chemical Society of Japan, 39, 2274-2286, (1966)

ZAHRADNIK et al. 1960
R. Zahradnik and M. Chvapil
'Study of the relationship between the magnitude of biological activity and the structure of aliphatic compounds'
Experientia, 16, 511-512, (1960)

ZWOLINSKI et al. 1972a
B. J. Zwolinski and J. Chao
'Critically Evaluated Tables of Thermodynamic Data'
in M. T. P. International Review of Science, Physical Chemistry Series One, Vol. 10, H. A. Skinner (ed.)
Butterworths, London, 1972

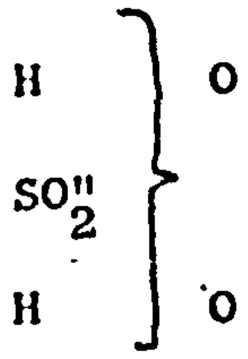
ZWOLINSKI et al. 1972b
B. J. Zwolinski and J. Chao
'Use of Electronic Computer for Handling Thermodynamic Data'
pp. 116-118 of ZWOLINSKI et al., 1972a

FIGURES



water

(Odling)

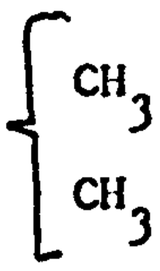


sulphuric
acid

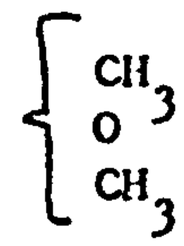
(Kekule)

Figure 1

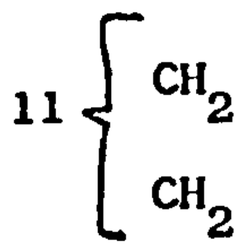
Superscript Dashes



ethane



diethyl ether

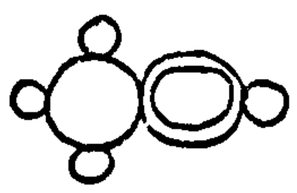


ethylene

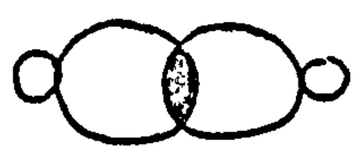
(dashes represent multiple bonding)

Figure 2

Brackets (Frankland)

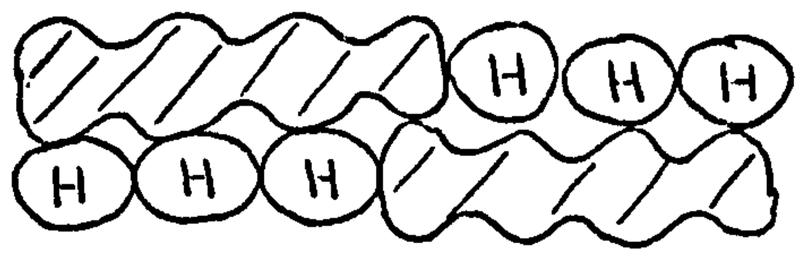


methanol

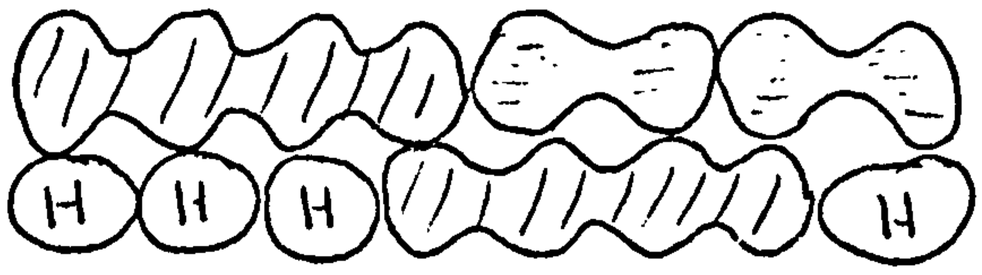


acetylene
(intersection indicates
multiple bonding)

Loschmidt



ethane

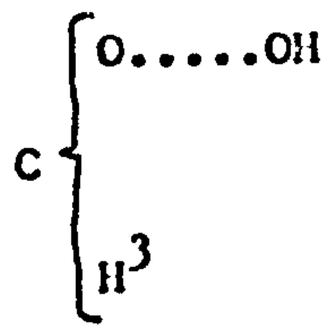


acetic acid

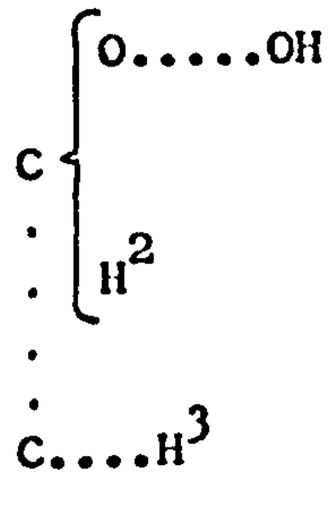
Kekule

(size of units representing atoms depends on combining power
- each vertical contact represents one bond)

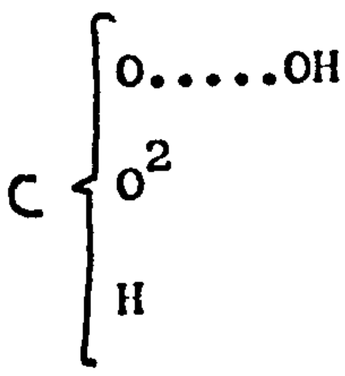
Figure 3 Touching and Intersecting Circles



methyl alcohol



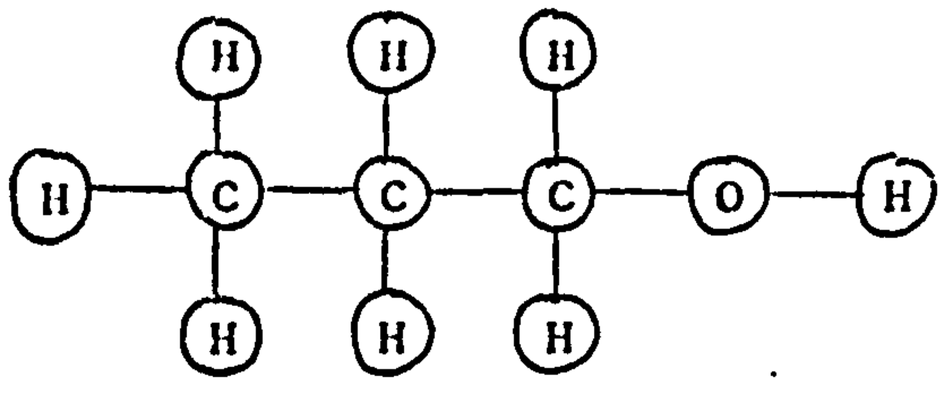
ethyl alcohol



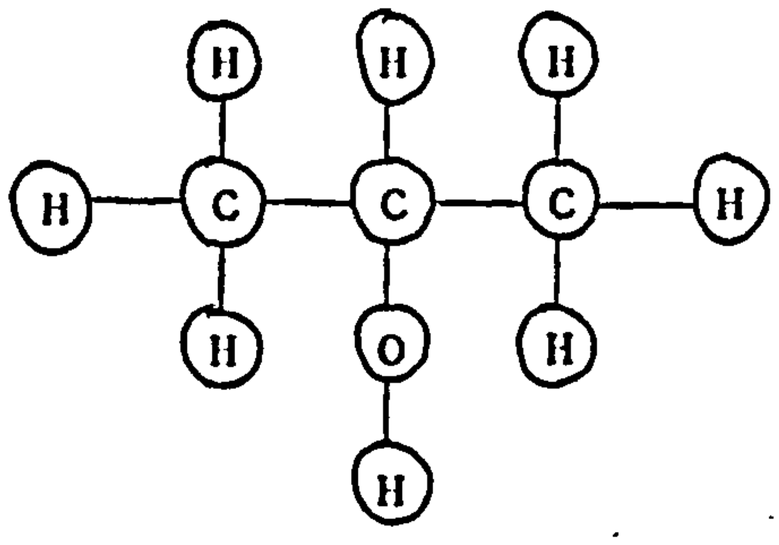
formic acid

- "doubled" oxygen atoms are due to the use of atomic weight 8 for oxygen

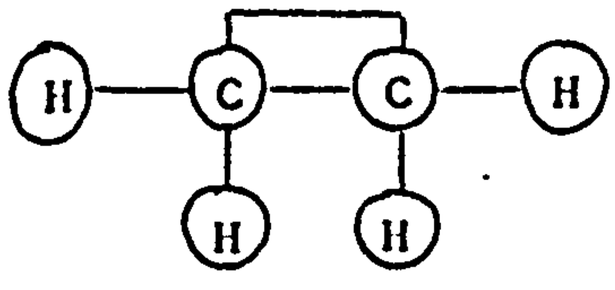
Figure 4 Couper's formulae



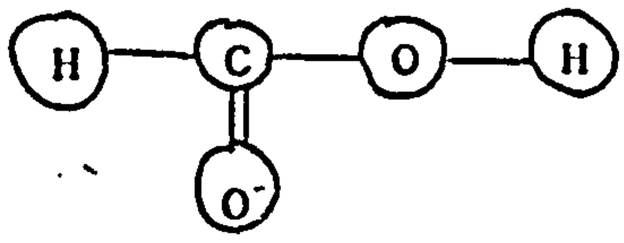
propan-1-ol



propan-2-ol

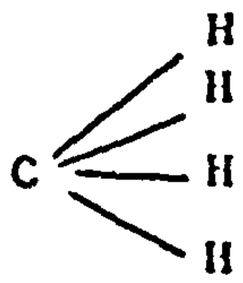


ethylene

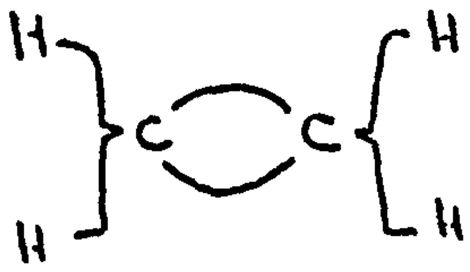


formic acid

Figure 5 Crum Brown's formulae

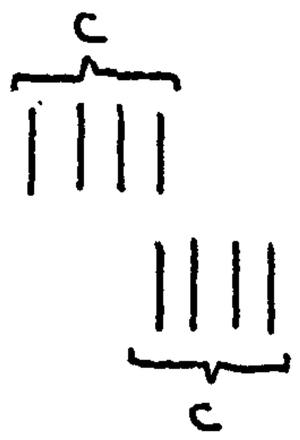


methane

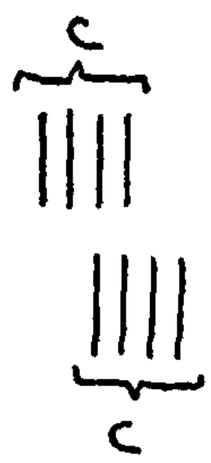


ethylene

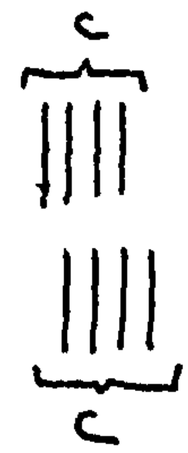
Lothar Meyer



ethane



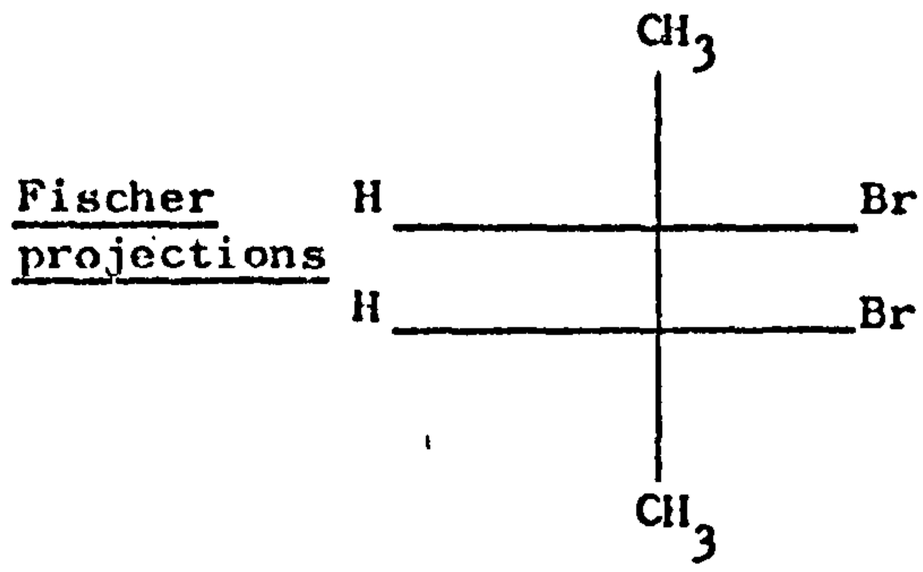
ethene

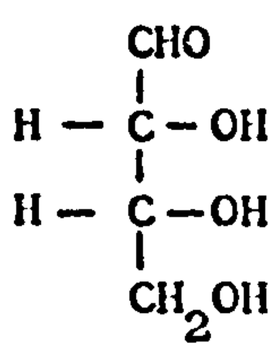


ethyne

Wilbrand

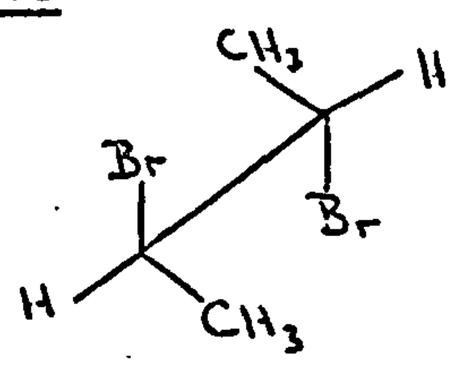
Figure 6

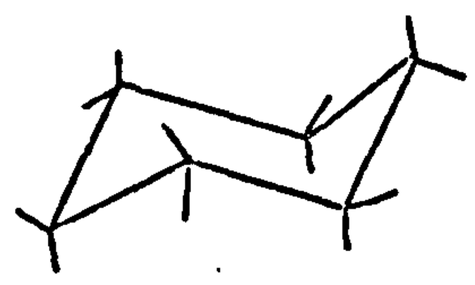




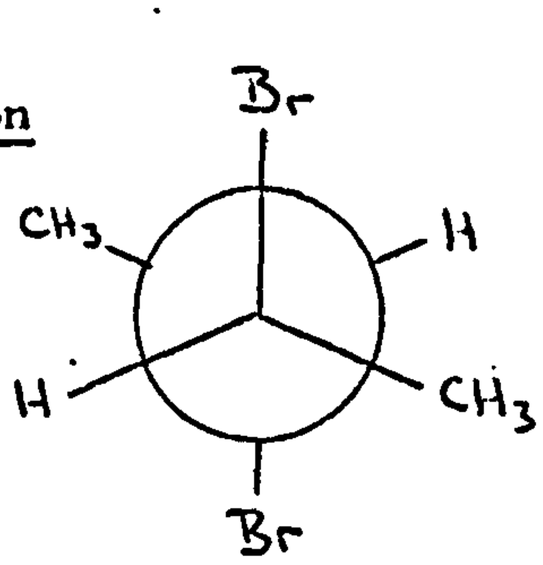
(abbreviated Fischer)

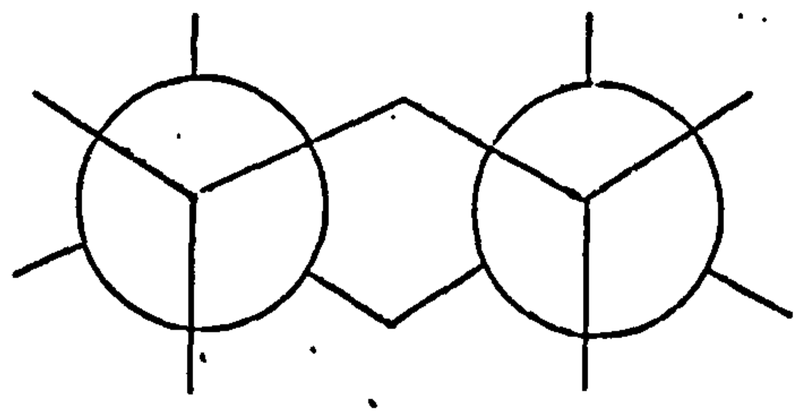
perspective



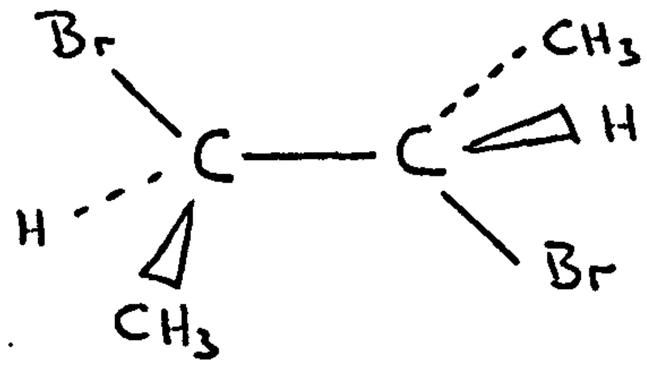


Newman projection





wedge/dot



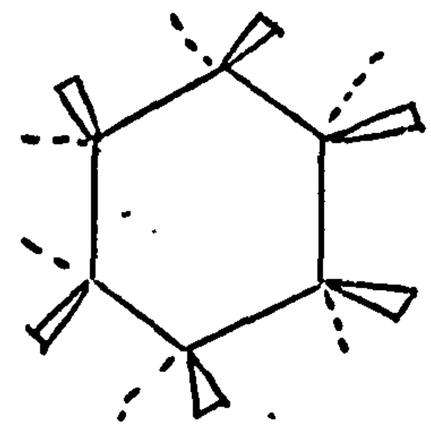
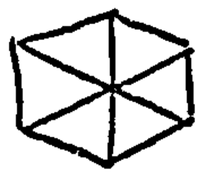


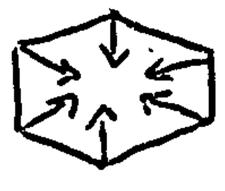
Figure 7

benzene formulae

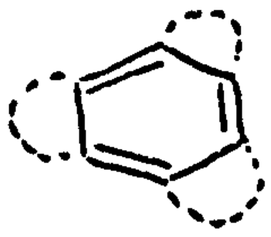
Kekule



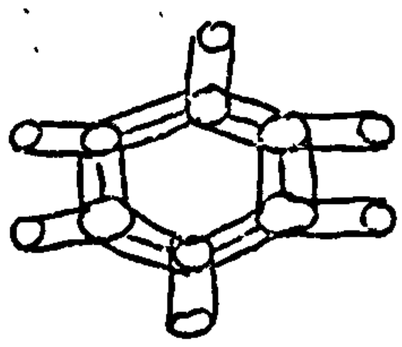
Claus



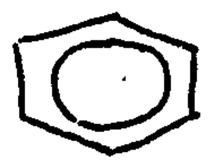
Armstrong-Baeyer-Lother Meyer



Thiele

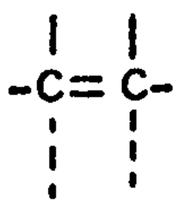


Thomson

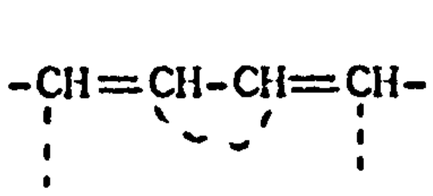


modern

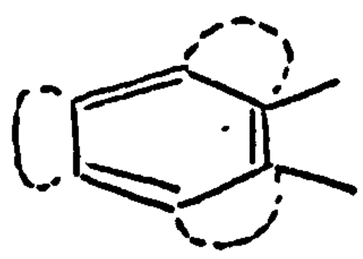
Thielè'sformulae



double bond



conjugated system

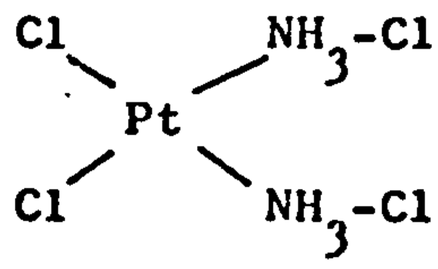


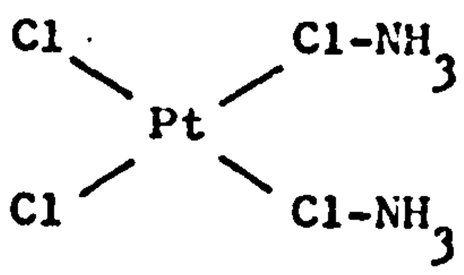
aromatic system

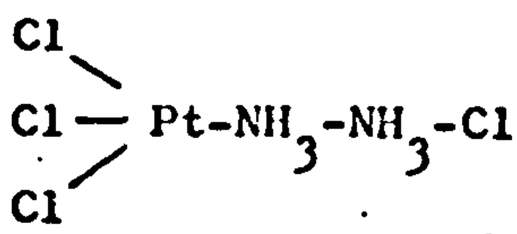
Figure 8

composition $Pt(NH_3)_2Cl_4$

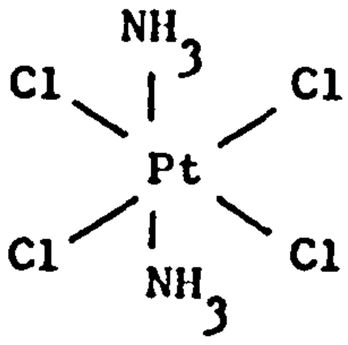
previously suggested structures







Werner's structures



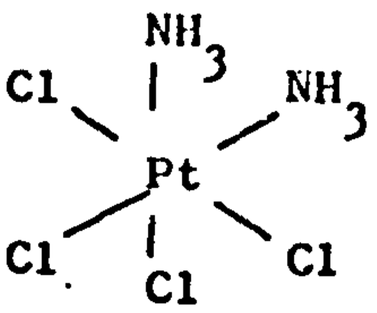
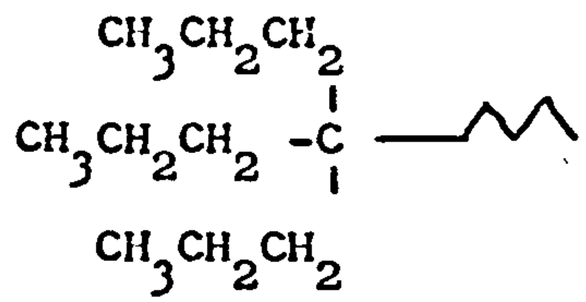
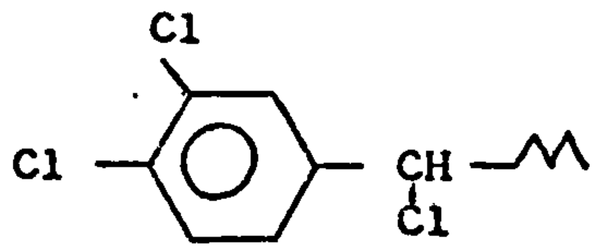
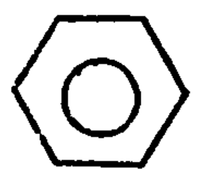
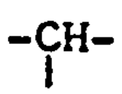
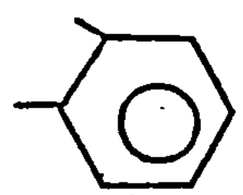
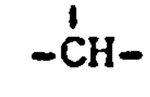
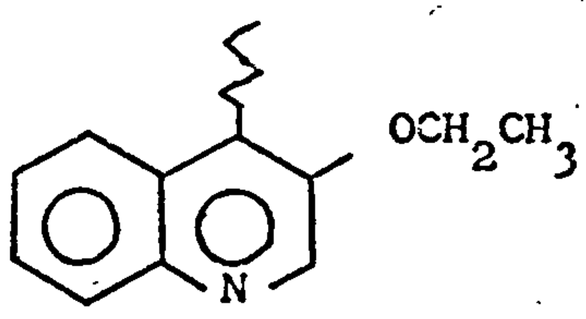
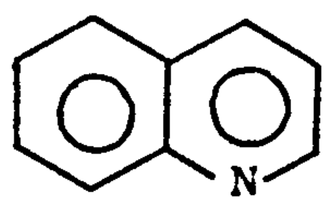
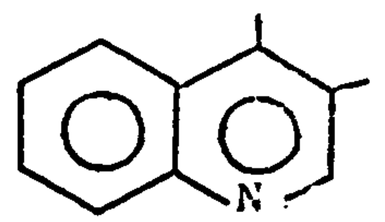


Figure 9

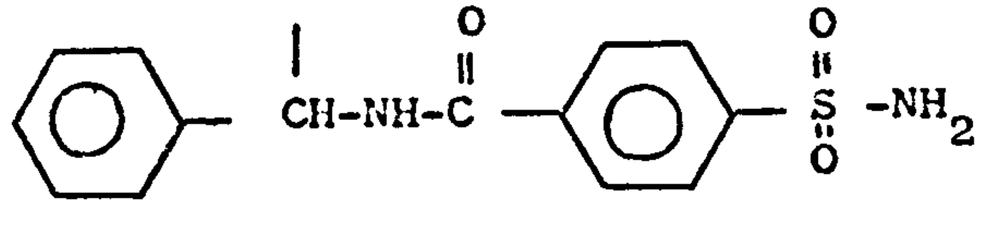
Figure 10

Structural Features

Structure	Set A	Set B
	one three CH_3- six $-CH_2-$	as A
	one  one  three Cl	one  one  one Cl (chain) two Cl (ring)
	one  one -O- one $-CH_2-$ one CH_3	one  one -O- (ring) one $-CH_2-$ one CH_3

330

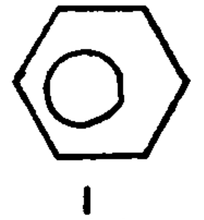
Figure 11



Set A

two	
one	-CH-
one	NHCO
one	SO_2NH_2

Set B

one	
one	
one	-CH-
one	NHCO
one	SO_2NH_2

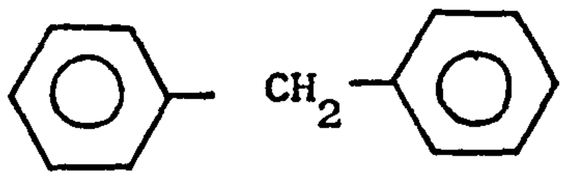
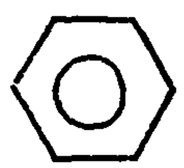
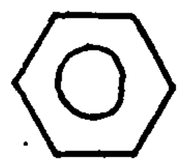
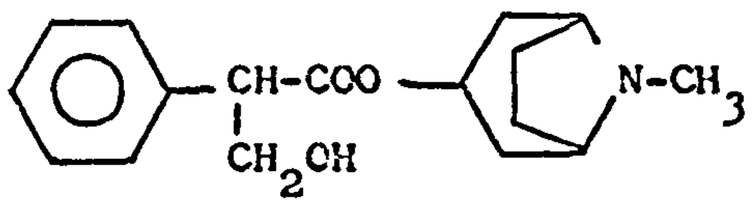
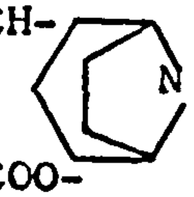
 - indicates common penicillin nucleus

Penicillin serum binding

Examples of structural feature derivation

Figure 12

Structural Features

Structure	Set A[a]	Set B[b]	Set C[c]
$CH_3-COO-CH_3$	two $-CH_3$ one $-COO-$	two $-CH_3$ one $-COO-$	as A
	two  one $-CH_2-$	two  one $-CH_2-$	as A
	one $-CH_3$ five $-CH_2-$ four $-CH-$ one $-N-$ one $-COO-$ one $-OH$ one 	one $-CH_3$ one $-CH_2-$ four $*CH_2*$ one $-CH-$ two $*CH*$ one $*CH*$ one $*N*$ one $-COO-$ one $-OH$ one 	one $-CH_3$ one $-CH_2-$ one $-CH-$ one  one $-COO-$ one $-OH$ one 

Notes

[a] - indicates any bond

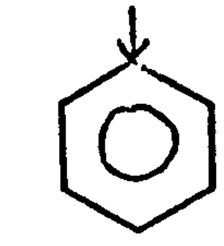
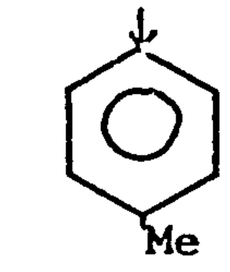
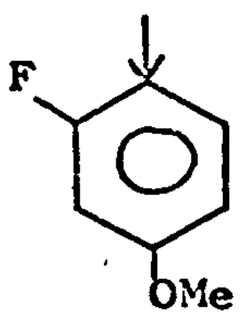
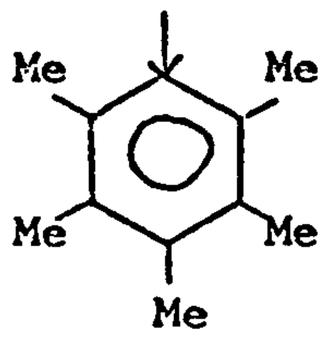
[b] - indicates chain bond, *indicates ring bond .

[c] - indicates any bond

Diverse Structures Partition Coefficients

Examples of Structural Feature Derivation

Structural Features

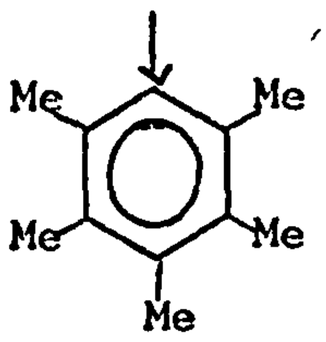
Structure	Set A	Set B	Set C	Set D
	one benzene ring	as A	as A	as A
	one benzene ring one Me	as A	one benzene ring one Me-p-RS	as C
	one benzene ring one F- one -O- one Me	one benzene ring one F- one -OMe	one benzene ring one F-o-RS one OMe-p-RS	one benzene ring one F-o-RS one OMe-p-RS one F-m-OMe
	one benzene ring five Me	as A	one benzene ring two Me-o-RS two Me-m-RS one Me-p-RS	one benzene ring two Me-o-RS two Me-m-RS one Me-p-RS four Me-o-Me four Me-m-Me two Me-p-Me

Notes: ↓ and RS indicate the reaction site

Benzene derivatives reaction kinetics

Examples of structural feature derivation

Figure 13



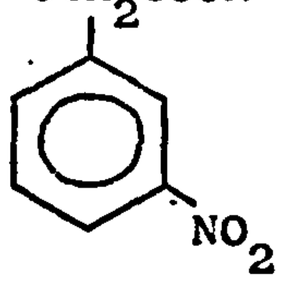
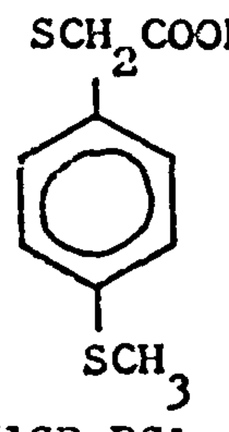
Structural Features	Regression coefficient x multiplicity
two Me-o-RS	3.165 x 2
two Me-m-RS	1.298 x 2
one Me-p-RS	3.869 x 1
Me-o-Me	-0.155 x 4
Me-m-Me	-0.254 x 4
Me-p-Me	-0.201 x 2
regression constant	-5.178 x 1
Summation	5.039

Notes: ↓ and RS indicate reaction site

Benzene Derivatives Reaction Kinetics

Simulated Property Prediction for Structure 14

Figure 14

Structure and WLN	Structural Feature Set A	B	C	D	E
 WNR C01VQ	one -O- one $-CH_2-$ one -COOH one $-NO_2$	one OCH_2COOH one NO_2	one OCH_2COOH one NO_3 one NO_2-meta-OCH_2COOH	one ACID GROUP one NO_2 one NO_2-meta-ACID GROUP	one OCH_2COOH one NO_2 one NO_2-with-OCH_2COOH
 QV1SR DS1	two -S- one $-CH_2-$ one -COOH one $-CH_3$	one SCH_2COOH one SCH_3	one SCH_2COOH one SCH_3 one SCH_3-para-SCH_2COOH	one ACID GROUP one SCH_3 one SCH_3-para-ACID GROUP	one SCH_2COOH one SCH_3 one SCH_3-with-SCH_2COOH

Note: the benzene ring is common to all structures and is not included here

Benzene acid pK values

Examples of structural feature derivation

Figure 15

Structure	Structural Features	
$CH_2=CH_2$	one	$CH_2=CH_2$ [a]
$CH_3CH=CH_2$	one	$-CH=CH_2$
	one	$-CH_3$
$CH_3-\overset{CH_3}{\underset{}{C}}=CH_2$	one	$-\overset{}{\underset{}{C}}=CH_2$
	two	$-CH_3$
$CH_3CH_2CH=\overset{CH_3}{\underset{}{C}}-CH_3$	one	$-CH=\overset{}{\underset{}{C}}-$
	three	$-CH_3$
	one	$-CH_2-$
$CH_3C\equiv C-CH_3$	one	$-C\equiv C-$
	two	$-CH_3$
$CH_3-CH=C=CH_2$	one	$-CH=C=CH_2$
	one	$-CH_3$
$CH_2=CH-CH=CH_2$	two	$-CH=CH_2$
		(CONJUGATED)

Note [a] unique feature

Figure 16

Unsaturated Aliphatics Heats of Formation

Examples of structural feature derivation

Examples of CIS interactions

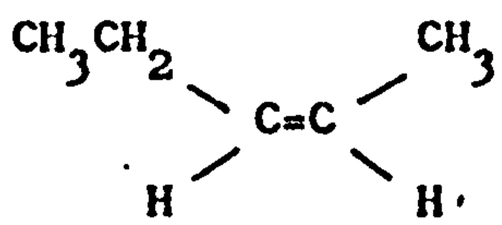
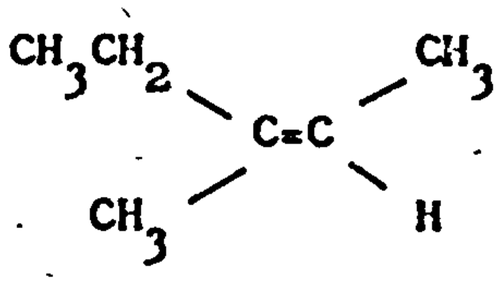
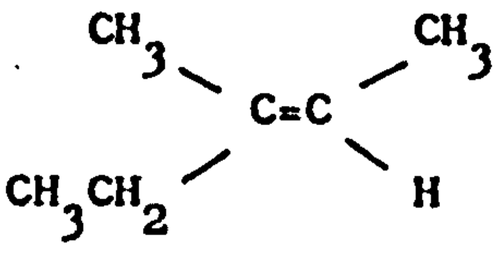
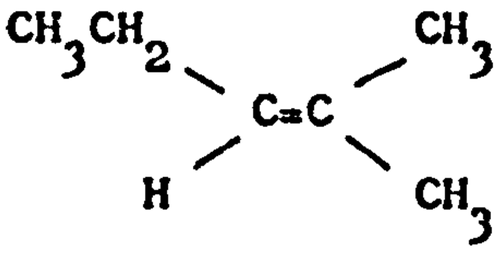
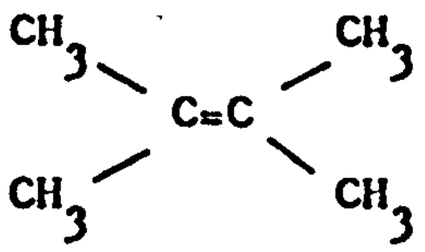
Structure Number	Structure	WLN
8.		3U2 - C
22.		2Y1 & U2 - C
23.		2Y1 & U2 - T
12.		2UY1 & 1
27.		1Y1 & UY1 & 1

Figure 17

Unsaturated Aliphatics Heats of Formation

Examples of CIS interactions

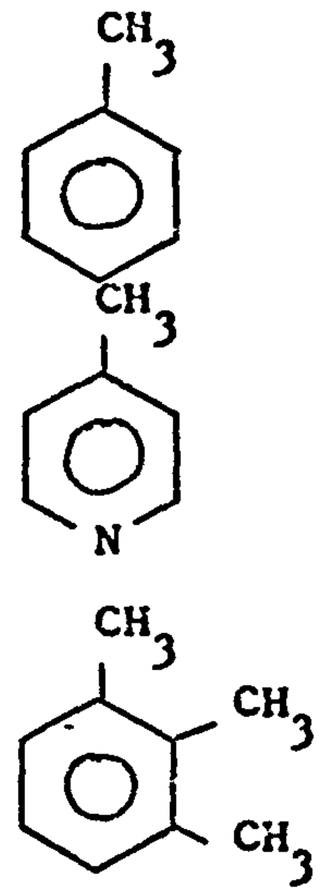
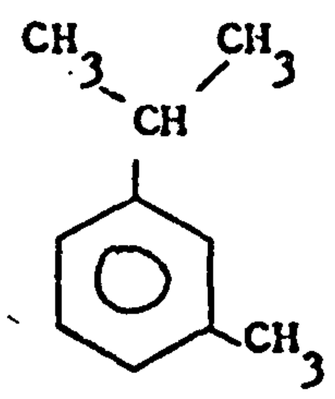
Structure	Structural Features	
	Set L	Set M
	one benzene ring one $-CH_3$	as L
	one pyridine ring one $-CH_3$	one pyridine ring one $-CH_3$ one CH_3-para-ring N
	one benzene ring three $-CH_3$	one benzene ring three $-CH_3$ two CH_3-ortho-CH_3 one CH_3-meta-CH_3
	one benzene ring one -CH- three $-CH_3$	one benzene ring one -CH- three $-CH_3$ one CH_3-meta- $CH(CH_3)_2$

Figure 18 Diverse Structures Heats of Vaporization

Examples of Structural Feature Derivation for data subset (10)

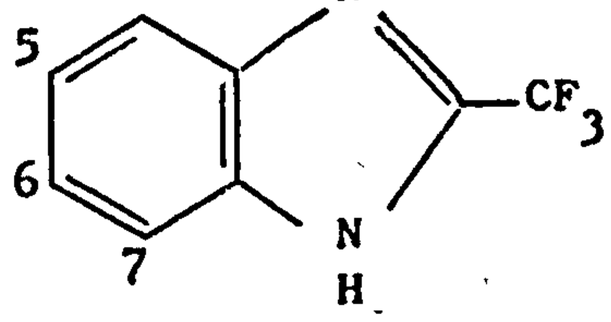


Figure 19

Trifluoromethylbenzimidazole nucleus

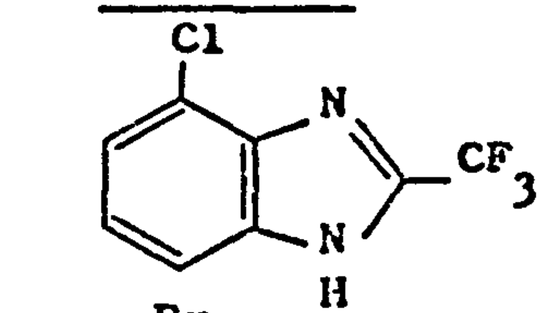
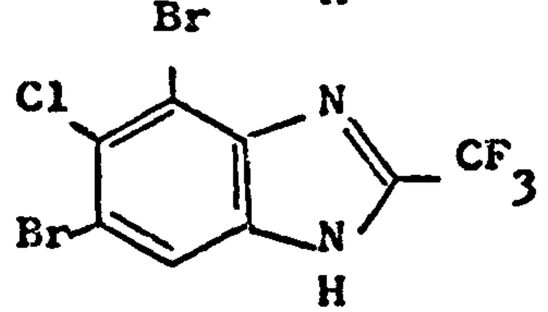
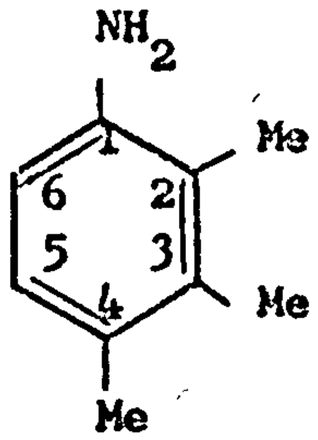
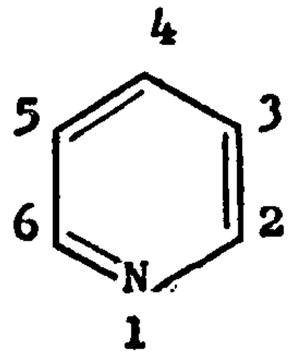
Structure	Structural Features			
	Set A	Set B	Set C	Set D
	one Cl	one Cl-o-RF	as A	as B
	two Br one Cl	one Br-o-RF one Br-m-RF one Cl-m-RF	two Br one Cl two Cl-o-Br one Br-m-Br	one Br-o-RF one Br-m-RF one Cl-m-RF two Cl-o-Br one Br-m-Br

Figure 20

Examples of structural feature derivation for benzimidazole derivatives



Benzene derivative example



Pyridine nucleus

Figure 21

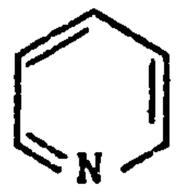
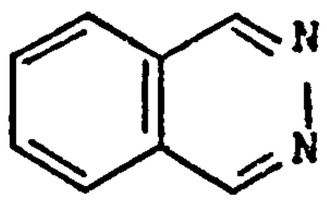
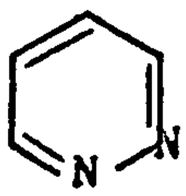
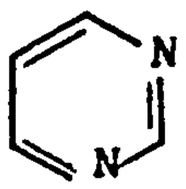
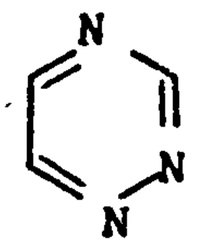
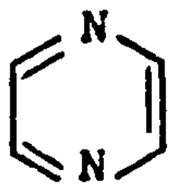
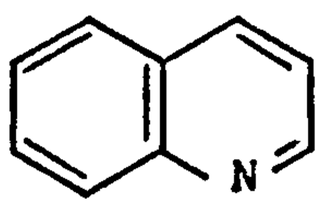
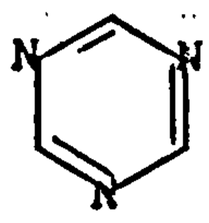
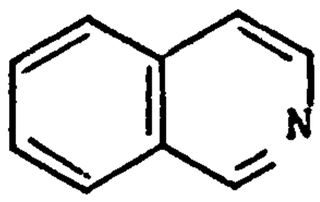
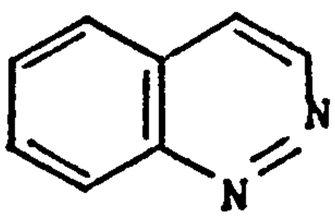
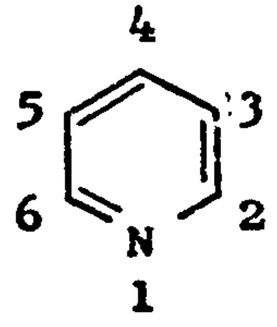
Parent ring system	No. of derivatives	Parent ring system	No. of derivatives
 pyridine	52	 phthalazine	5
 pyridazine	10	 quinoxaline	5
 pyrimidine	60	 1,2,4 triazine	2
 pyrazine	6		
 quinoline	13	 1,3,5 triczine	3
 isoquinoline	8		
 cinnoline	5		

Figure 22

Types and numbers of heterocyclic systems in analysis of pKa data



pyridine nucleus

2 - 6 meta interactions

decnated as meta*

Structural Feature Sets

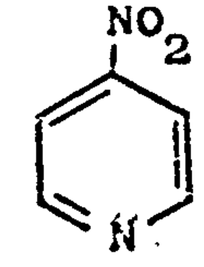
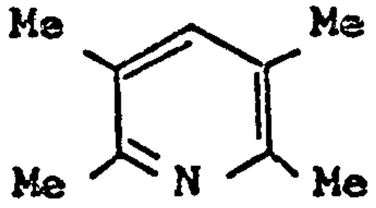
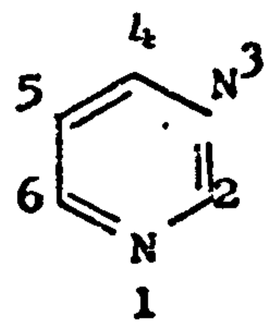
Structure and WLN	A=	B	C	D	E
 T6NJ DNW	one NO_2	one NO_2 - para - RING N	one NO_2 - para - RING N	one NO_2 - para - RING N	one NO_2
	four Me	two Me - ortho - RING N	two Me - ortho - RING N	two Me - ortho - RING N	four Me two Me - ortho -
		two Me - meta - RING N	two Me - meta - RING N	two Me - meta - RING N	one Me - meta* - one Me - meta -
 T6NJ B1 C1 E1 F1			two Me - ortho - Me two Me - meta - Me two Me - para - Me	two Me - ortho - Me one Me - meta* - Me one Me - meta - Me two Me - para - Me	two Me - para -

Figure 23

pKa values for heterocyclic structures

Examples of structural feature derivation for pyridines



- 4 and 6 positions are equivalent and denoted as 4

- meta* denotes 2 - 4 and 2 - 6 meta interactions

Structural Feature Sets

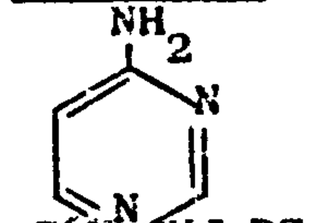
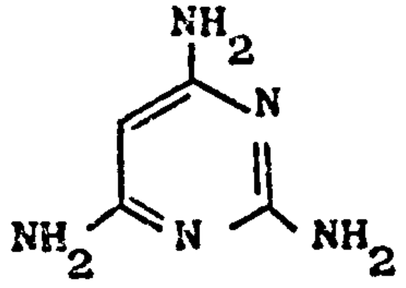
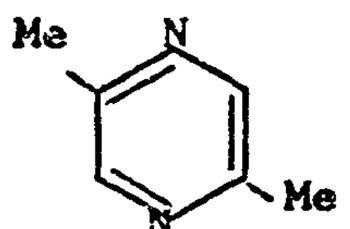
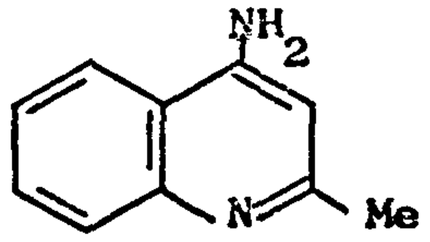
Structure and WLN					
 T6N CNJ DZ	one NH_2	one 4 - NH_2	one 4 - NH_2	one 4 - NH_2	one NH_2
 T6N CNJ BZ DZ FZ	three NH_2	one 2 - NH_2 two 4 - NH_2	one 2 - NH_2 two 4 - NH_2 three NH_2 - meta - NH_2	one 2 - NH_2 two 4 - NH_2 one NH_2 - meta - NH_2 two NH_2 - meta* - NH_2	three NH_2 one NH_2 - meta - NH_2 two NH_2 - meta* - NH_2

Figure 24

pKa values for heterocyclic structures

Examples of structural feature derivation for pyrimidines

Structure and WLN	K	L	M	N
 T6N DNJ B1 E1		one RING N-para-RING N	one RING N-para-RING N	one RING N-para-RING N
	two RING N	two Me-ortho-RING N	two Me-ortho-RING N	two Me-ortho-RING N
	two Me	two Me-meta-RING N	two Me-meta-RING N	two Me-meta-RING N
				one Me-para-Me
 T66 BMJ C1 EZ	one RING N	one Me-ortho RING N	one Me-ortho-RING N	one Me-ortho RING N
	one FUSED RING	one NH_2-para-RING N	one NH_2-para-RING N	one NH_2-para-RING N
	one NH_2	one RINGFUSION-ortho-RING N	one RINGFUSION-ortho-RING N	one RINGFUSION
	one Me	one RINGFUSION-meta-RING N	one RINGFUSION-meta-RING N	-ortho-RING N
			one Me-meta-RINGFUSION	one RINGFUSION-meta-RING N
			one Me-para-RINGFUSION	one Me-Meta-
			one NH_2-ortho-RINGFUSION	RINGFUSION
			one NH_2-meta-RINGFUSION	one me-para-RINGFUSION
				one NH_2-ortho-RINGFUSION
				one NH_2-meta-RINGFUSION
				one Me-meta-NH_2

Structural Feature Sets

Figure 25

pKa values for heterocyclic structures
Examples of structural feature derivation for diverse structures

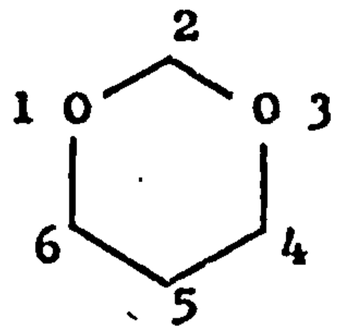
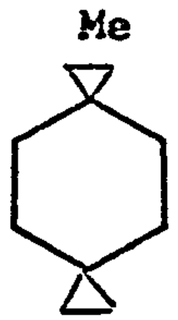


Figure 26

Alicyclic Structure Boiling Points

Structure of 1,3 dioxan

Structural Features

Structure	Set A	Set B	Set C	Set D	Set E
	two Me	as A	as A	as A	two Me one Me-para-Me
Set F	**Set G**	**Set H**	**Set I**	**Set J**	**Set K**
two Eq Me	as F	as F	as F	two eq Me one eq Me-para-eq Me	as J

Notes

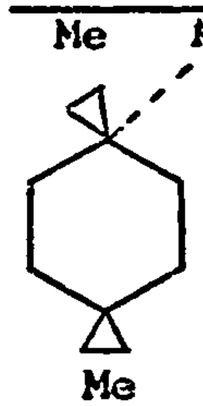
eq = equatorial, ax = axial

the ring system is not included, as it is common to all structures

Figure 27

Structural Feature Derivation for Cyclohexanes

Structural Features

Structure	Set A	Set B	Set C	Set D	Set E
	three Me	three Me one Me-gem-Me	as B	as B	three Me one Me-gem-Me two Me-para-Me

Set F	Set G	Set H	Set I	Set J	Set K
two eq Me one ax Me	two eq Me one ax Me one eq Me-gem-ax Me	as G	as G	two eq Me one ax Me one eq Me-gem-ax Me one eq Me-para-eq Me one ax Me-para-eq Me	as J

Notes

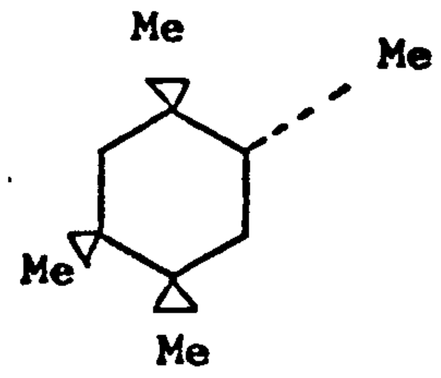
eq = equatorial, ax = axial

the ring system is not included as it is common to all structures

Figure 28

Structural Feature Derivation for Cyclohexanes

Structural Features

Structure	Set A	Set B	Set C	Set D	Set E
	four Me	as A	four Me two Me-ortho-Me	four Me two Me-ortho-Me two Me-meta-Me	four Me two Me-ortho-Me two Me-meta-Me two Me-para-Me

Set F	Set G	Set H	Set I	Set J	Set K
three eq Me one ax Me	as F	three eq Me one ax Me one eq Me-ortho-ax Me one eq Me-ortho-eq Me one eq Me-meta-ax Me one eq Me-meta-eq Me	three eq Me one ax Me one eq Me-ortho-ax Me one eq Me-ortho-eq Me one eq Me-meta-ax Me one eq Me-meta-eq Me	three eq Me one ax Me one eq Me-ortho-ax Me one eq Me-ortho-eq Me one eq Me-para-ax Me one eq Me-para-eq Me	three eq Me one ax Me one eq Me-ortho-ax Me one eq Me-ortho-eq Me one eq Me-meta-ax Me one eq Me-meta-eq Me one eq Me-para-ax Me one eq Me-para-eq Me

Notes

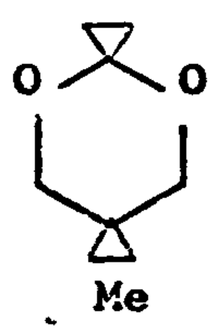
eq = equatorial, ax = axial

the ring system is not included as it is common to all structures

Figure 29

Structural Feature Derivation for Cyclohexanes

Structural Features

Structure	Set L	Set M	Set N	Set O	Set P
Me  Me	two Me	two eq Me	as M	as N	one 2-Me one 5-Me
	Set Q	**Set R**	**Set S**	**Set T**	**Set U**
	one 2-Me one 5-Me one Me-para-Me	as Q	one 2-eq Me one 5-eq Me	as S	as S
	Set V	**Set W**	**Set X**		
	as S	one 2-eq Me one 5-eq Me one eq Me-para-eq Me	as S		

Notes

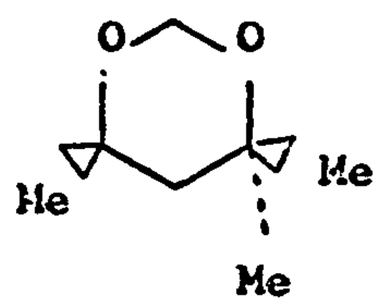
eq = equatorial

the ring system is not included as it is common to all structures

Figure 30

Structural Feature Derivation for 1,3 Dioxans

Structural Features

Structure	Set L	Set M	Set N	Set O	Set P
	three Me	two eq Me one ax Me	two eq Me one ax Me one Me-gem-Me	as N	three 4-Me

Set Q	Set R	Set S	Set T	Set U
three 4-Me one Me-gem-Me two Me-meta-Me	three 4-Me one Me-gem-Me two Me-(4,6)-meta-Me	two 4-eq Me one 4-ax Me	two 4-eq Me one 4-ax Me one eq Me-gem-ax Me	two 4-eq Me one 4-ax Me one 4-gem

Set V	Set W	Set X
as T	two 4-eq Me one 4-ax Me one eq Me-gem-ax Me one eq Me-meta-eq Me one eq Me-meta-ax Me	two 4-eq Me one 4-ax Me one eq Me-gem-ax Me one eq Me-(4,6)meta-eq Me one eq Me-(4,6)meta-ax Me

Notes

eq = equatorial, ax = axial, gem = geminal

ring system is not included as it is common to all structures

Figure 31

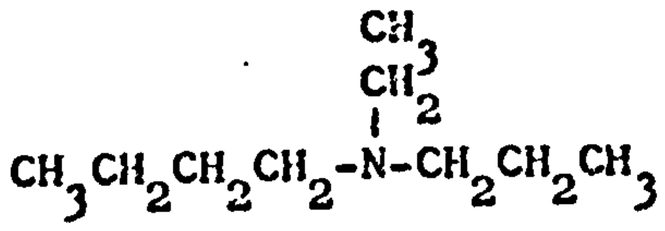
Structural Feature Derivation for 1,3-Dioxans

FIGURES (CLUSTER ANALYSIS)

Classification Aliphatics Set

1. CH_3-CH_3
2. $CH_3CH_2CH_3$
3. $CH_3CH_2CH_2CH_3$
4. $(CH_3)_2CH-CH_3$
5. $CH_3-CH_2-CH(CH_3)-CH_3$
6. $CH_3-CH_2-CH(CH_3)-CH_3$
7. $CH_3-CH(CH_3)-CH(CH_3)-CH_3$
8. $CH_3CH_2-C(CH_3)_3$
9. CH_3OH
10. CH_3CH_2OH
11. $CH_3CH_2CH_2OH$
12. $(CH_3)_2CH-OH$
13. $CH_3CH_2-CH(CH_3)-OH$
14. $CH_3-CH(CH_3)-CH_2OH$
15. $HO-CH_2CH_2-OH$
16. $CH_3-(CH_2)-OH$
17. CH_3-NH_2
18. $CH_3CH_2NH_2$
19. $CH_3CH_2CH_2NH_2$
20. $CH_3-CH(CH_3)-NH_2$
21. $CH_3CH_2-CH(CH_3)-NH_2$
22. $CH_3-CH(CH_3)-CH(CH_3)-NH_2$
23. $H_2N-CH_2CH_2-NH_2$
24. $H_2N-CH(CH_3)-CH(CH_3)-NH_2$
25. $CH_3-C(=O)-CH_3$
26. $CH_3CH_2-C(=O)-CH_3$
27. $CH_3CH_2CH_2-C(=O)-CH_3$
28. $CH_3CH_2-C(=O)-CH_2CH_3$
29. $CH_3-CH(CH_3)-C(=O)-CH_2CH_3$
30. $CH_3CH(CH_3)-C(=O)-CH(CH_3)-CH_3$
31. $CH_3-C(=O)-CH_2-C(=O)-CH_3$
32. $CH_3CH_2-C(=O)-CH_2-C(=O)-CH_3$
33. $HO-CH_2CH_2-C(=O)-CH_3$
34. $CH_3-N(CH_3)-CH_3$
35. $CH_3CH_2-N(CH_3)-CH(CH_3)-CH_2-CH_3$
36. $CH_3-CH_2-N(CH_3)-CH_2CH_3$
37. $CH_3-N(CH_3)-CH_2-CH(CH_3)-CH_3$
38. $CH_3-N(CH_3)-CH_2CH_2-N(CH_3)-CH_3$

A1

39.


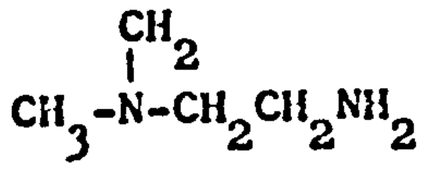
40.


Figure A2

1. $-CH_3-$
2. $-CH_2-$
3. $-\overset{|}{C}H-$
4. $-OH$
5. $-\overset{\overset{O}{\|}}{C}-$
6. $-NH_2$
7. $-\overset{|}{N}-$

Classification of aliphatics

Structural feature set

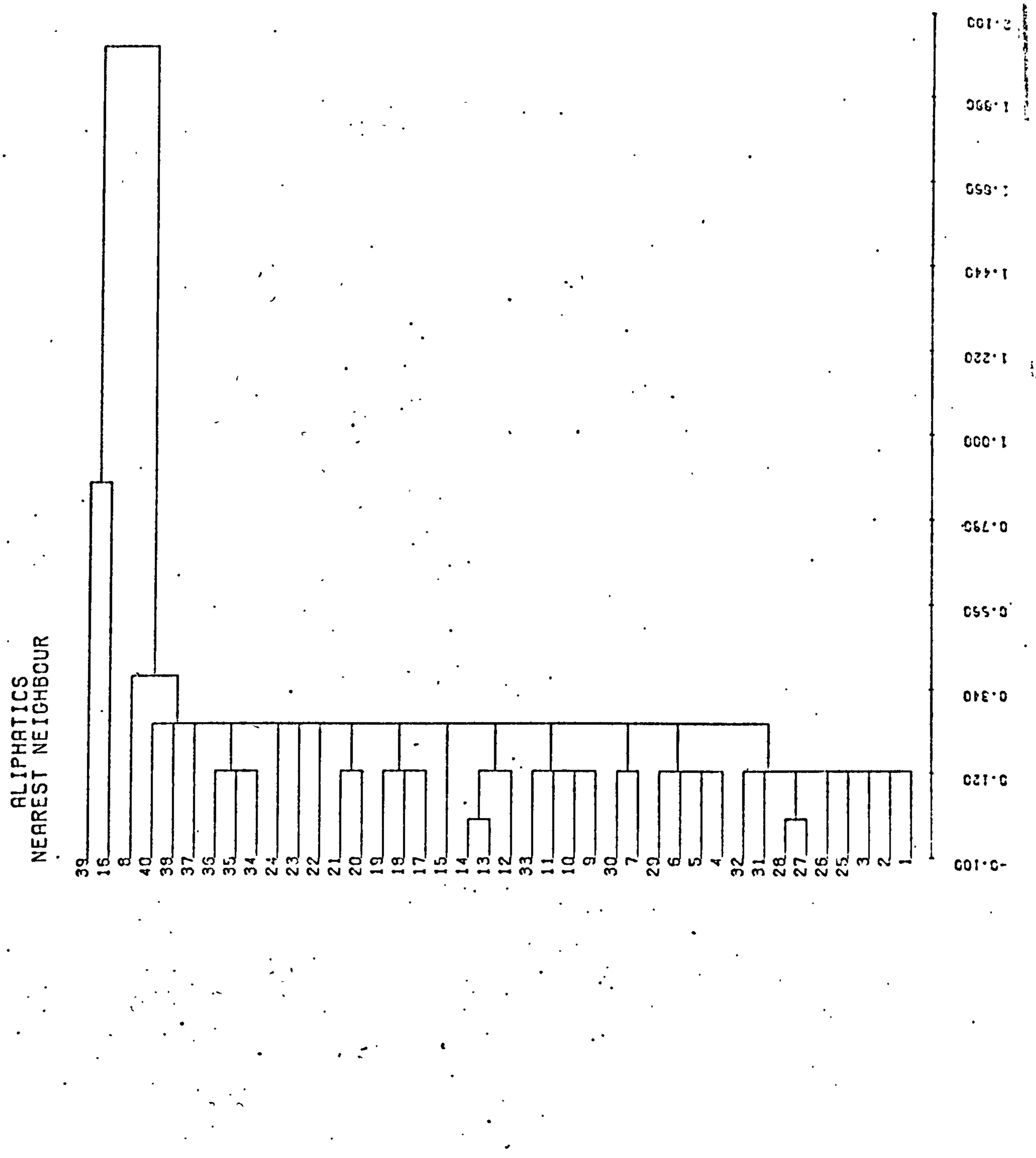


Figure AA

CALCOMP OUTPUT FOR :LILGWA.CLUSTESTB
FROM FILE PRODUCED ON 4FEB77 AT 20.59.23

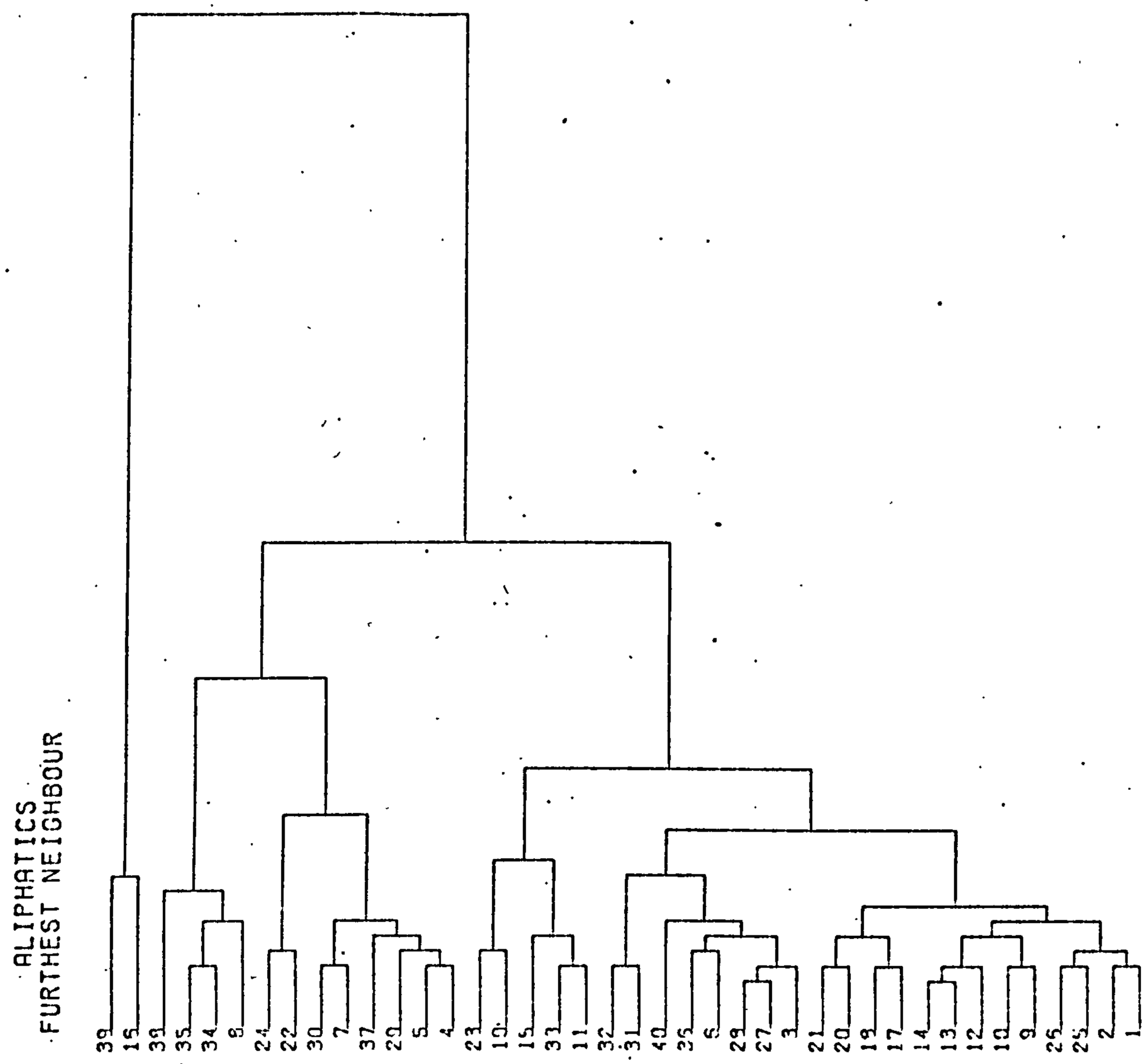


Figure AB

CALCOMP OUTPUT FOR FILE :LILGWA.CLUSIESIB PRODUCED ON 5FEB77 AT 09.45.00

CUT HERE

84

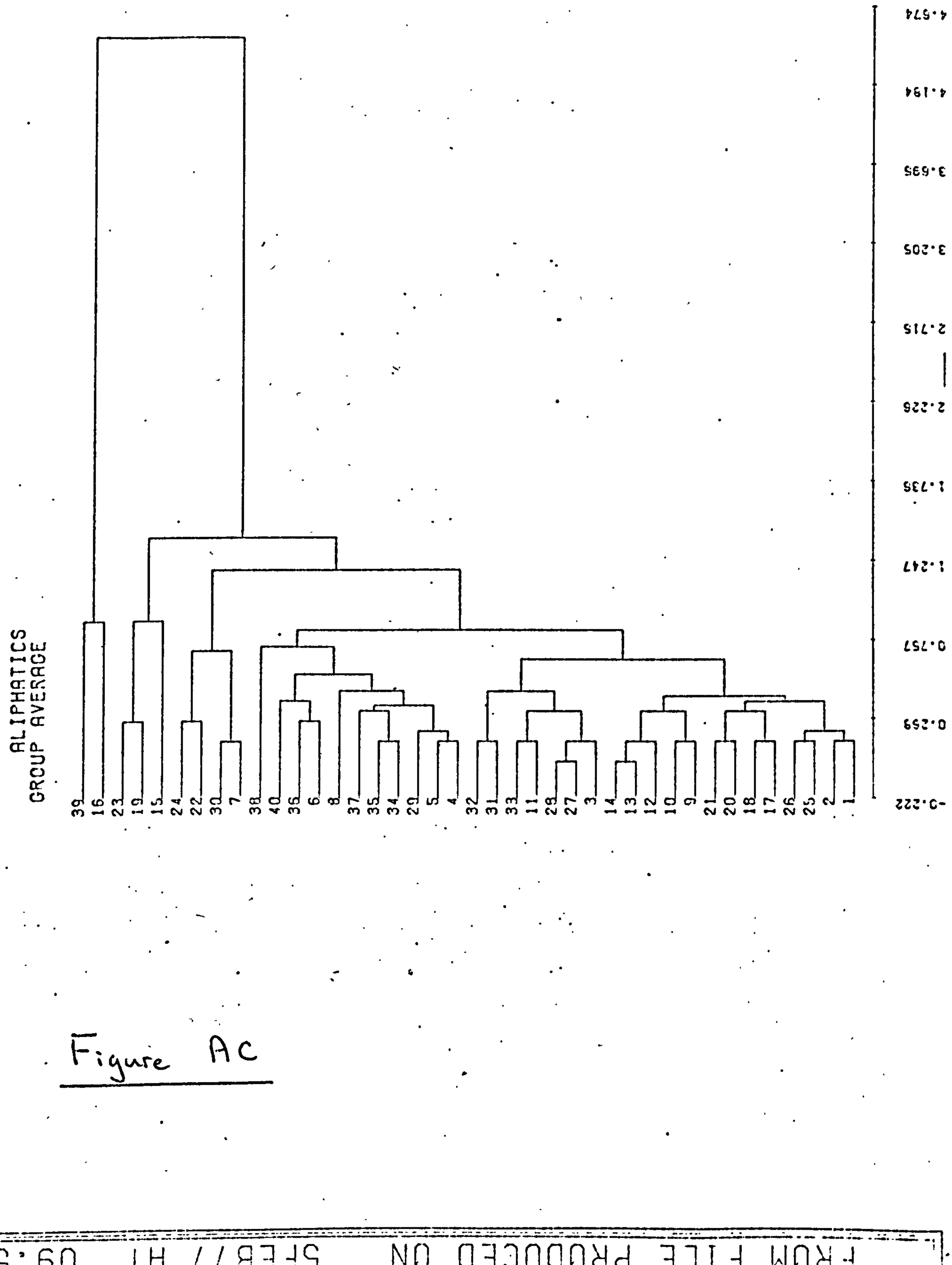


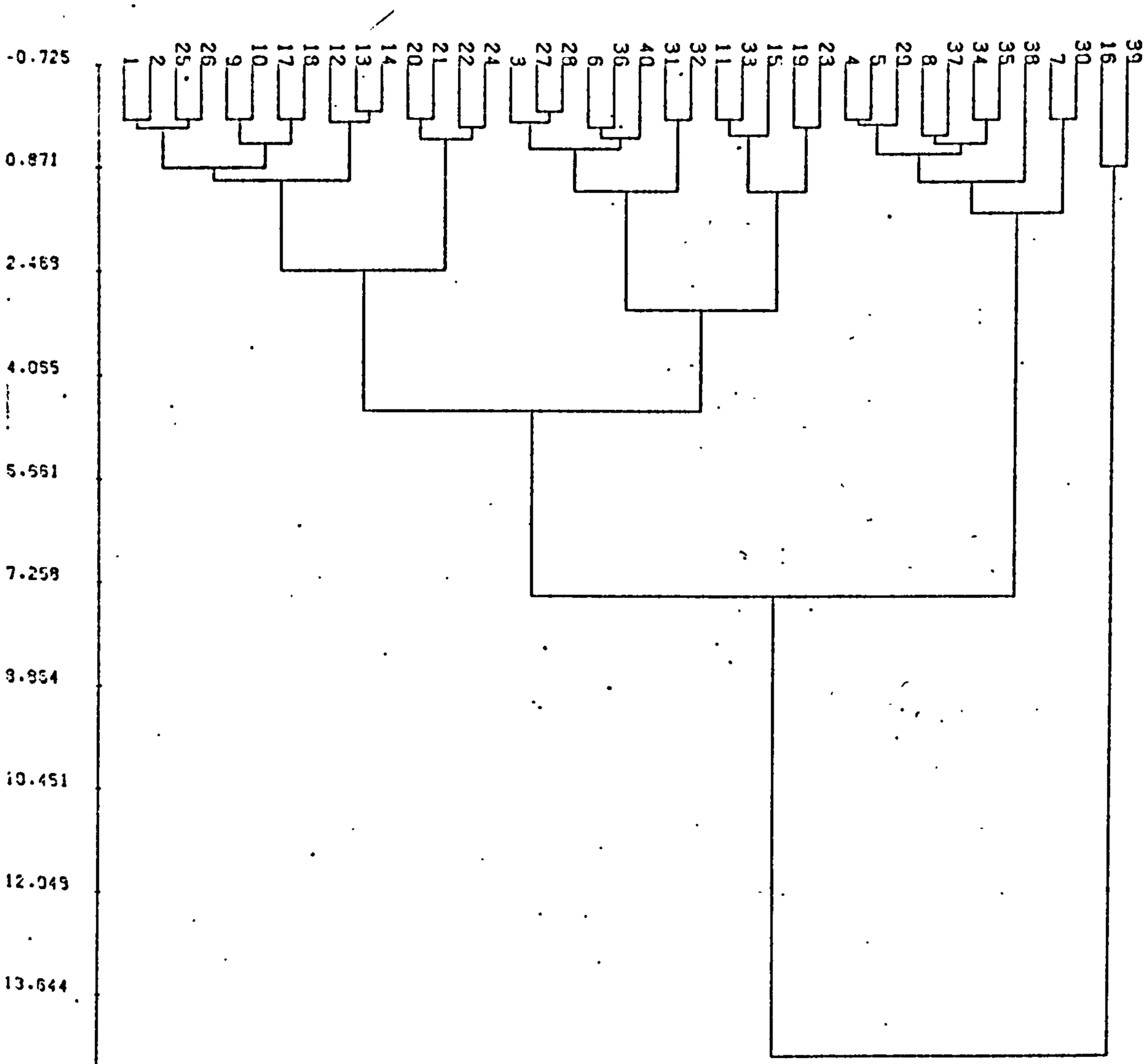
Figure AC

CALCOMP OUTPUT FOR ;LILGWA,CLUSTESTB
FROM FILE PRODUCED ON 5FEB77 AT 09.57.46

AC

CALCOMP OUTPUT FOR :LI1GWA.CLUSTESTB
FROM FILE PRODUCED ON 5FEB77 AT 11.21.15

Figure A1b



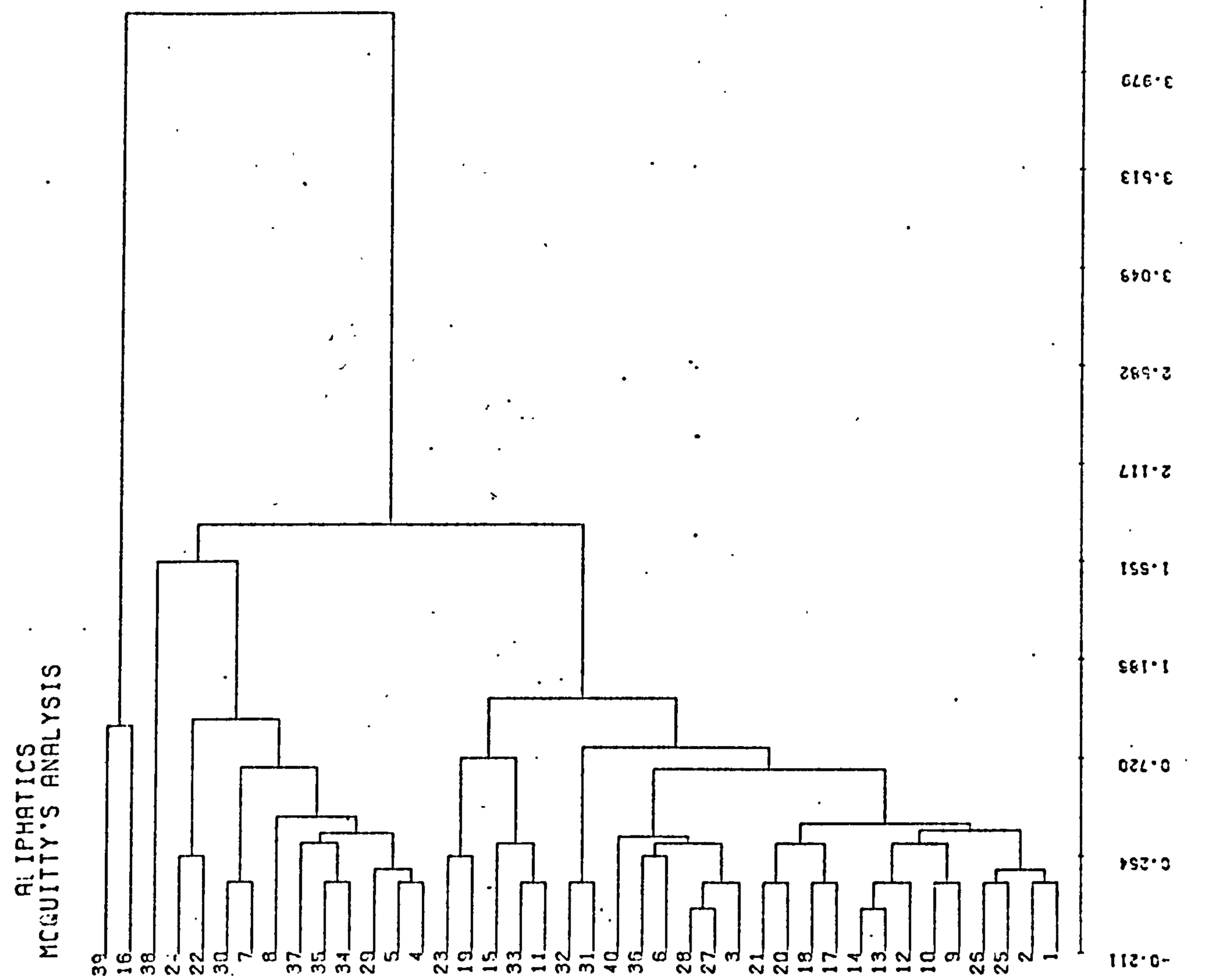


Figure AE

CUT HERE

CALCOMP OUTPUT FOR :LIIGWA.CLUSTESTB
FROM FILE PRODUCED ON 5FEB77 AT 16.24.05

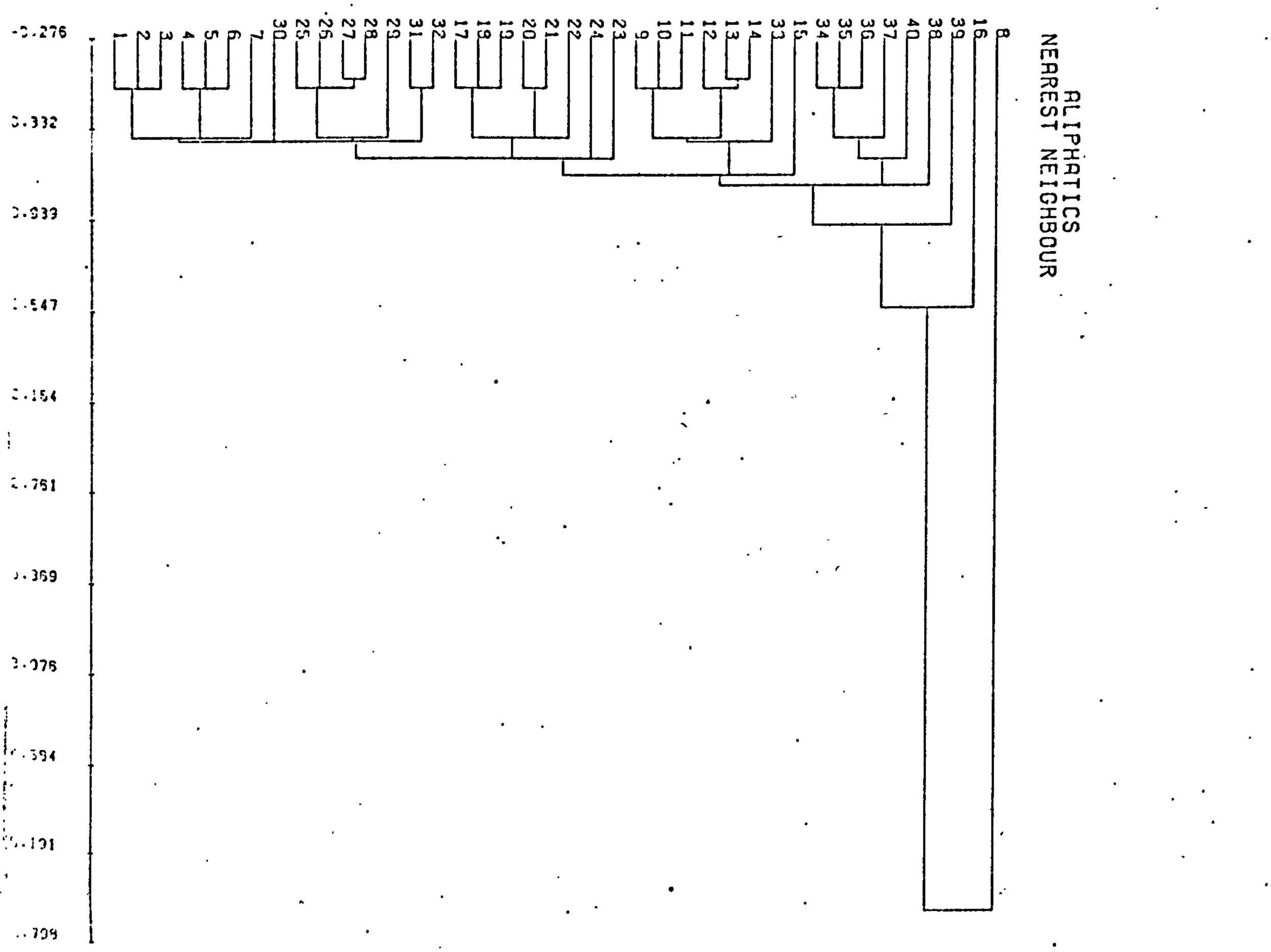


Figure A1

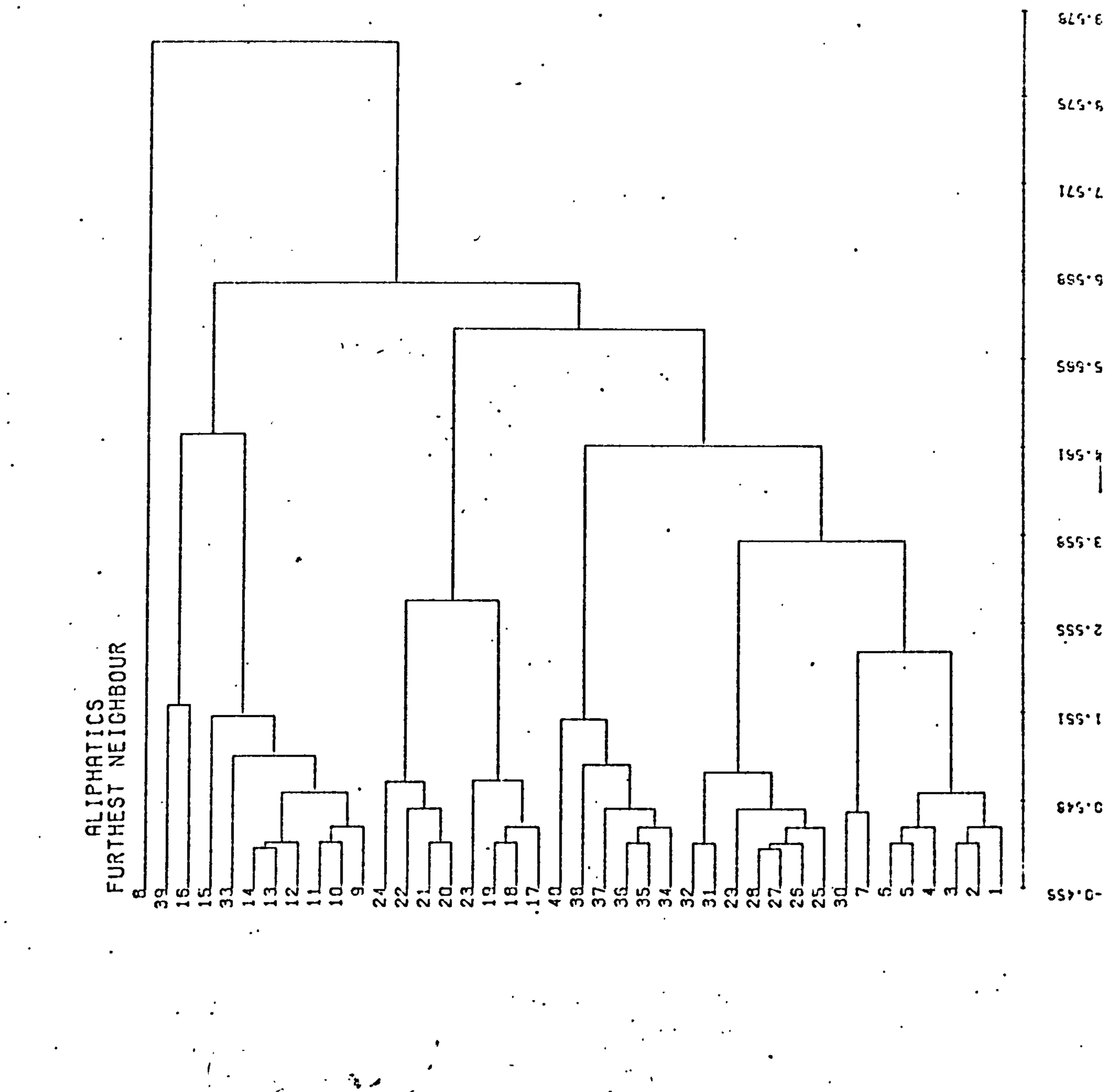


Figure AG

CALCOMP OUTPUT FOR :LLIGWA.CLUSTESTB
FROM FILE PRODUCED ON 7FEB77 AT 18.50.42

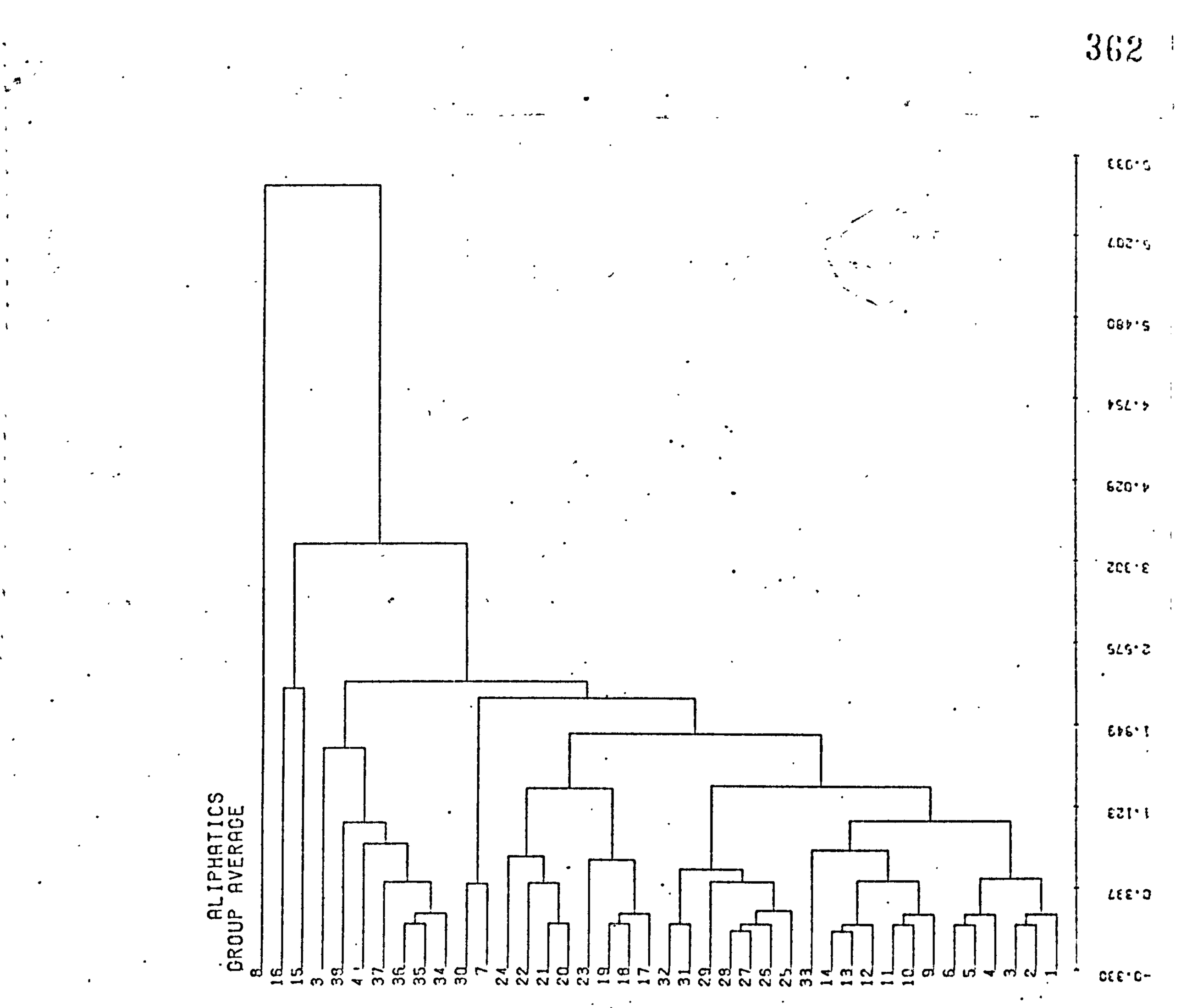


Figure AH

CALCOMP OUTPUT FOR ;LLIGWA.CLUSTESTB
FROM FILE PRODUCED ON 8FEB77 AT 18.08.39

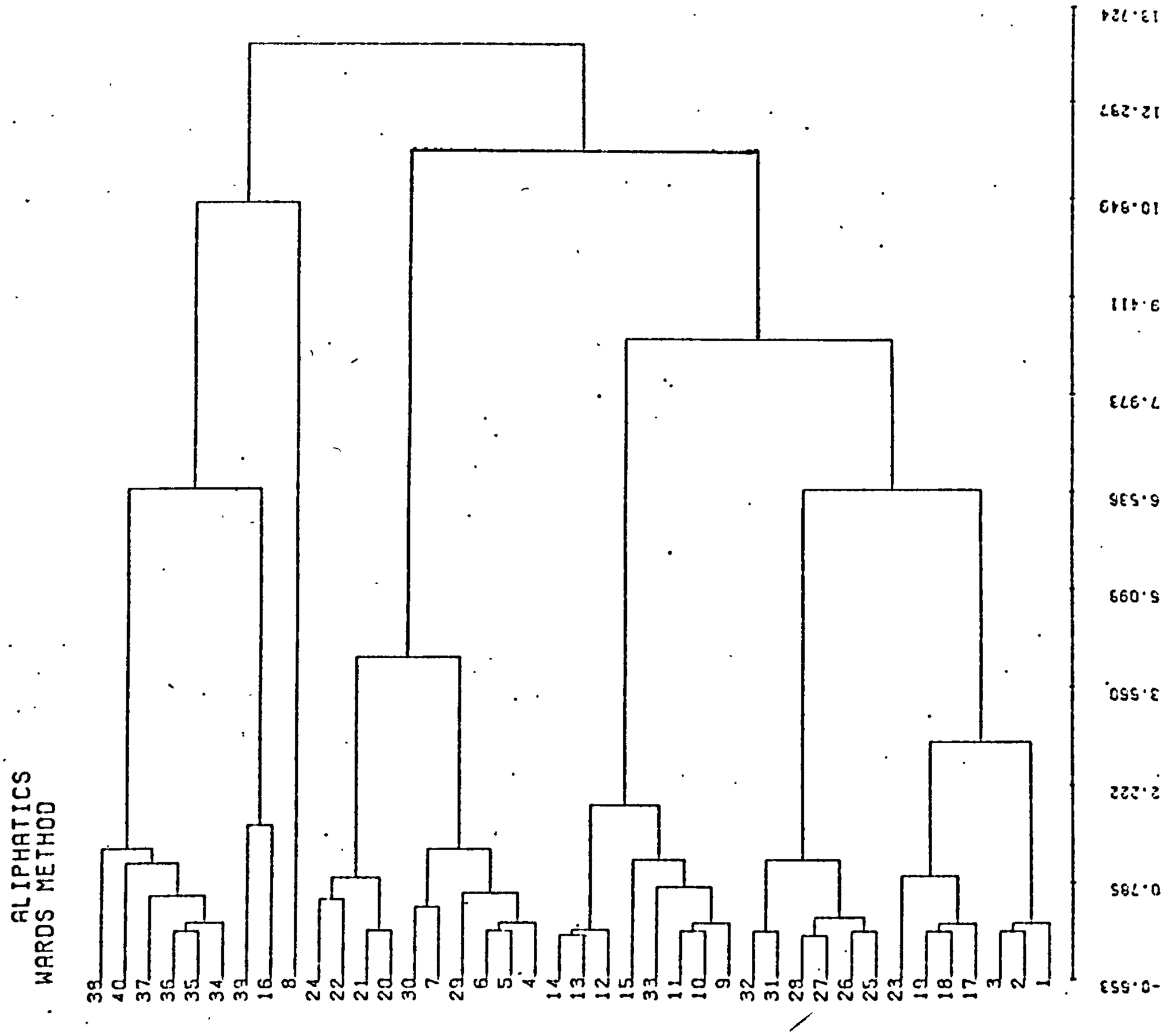


Figure A I

CALCOMP OUTPU: FOR :LLIGWA.CLUSTESTB
FROM FILE PRODUCED ON FEB7 1: 1: 1:

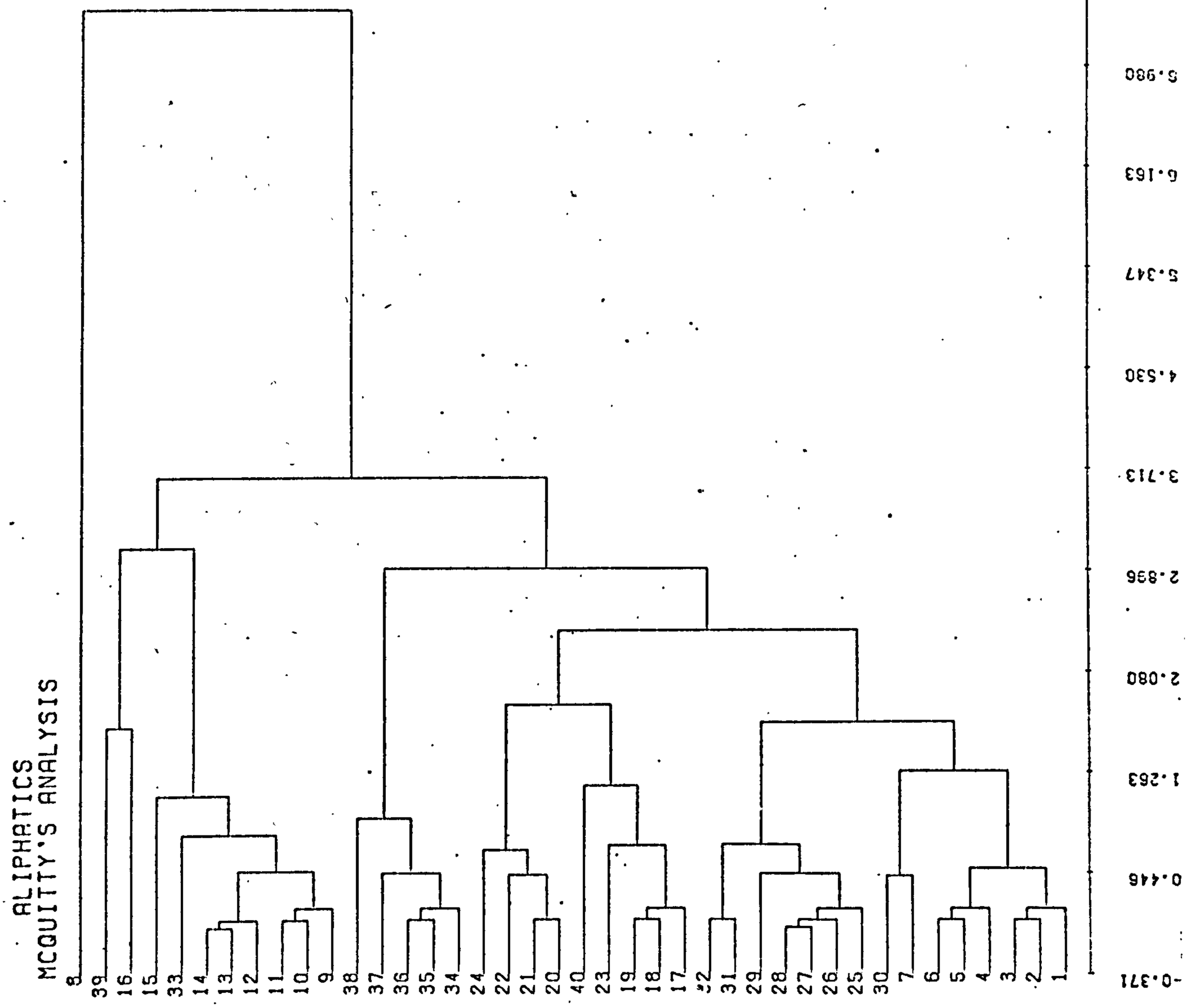
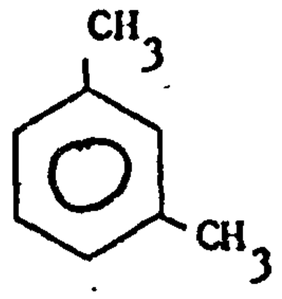


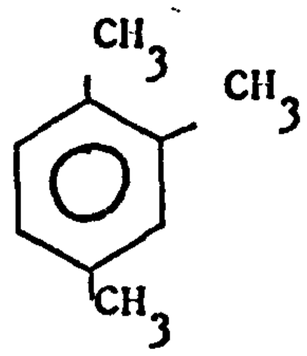
Figure A5

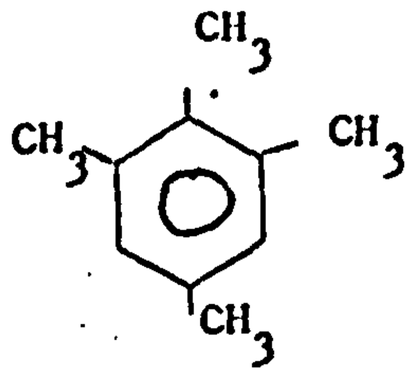
Figure B1

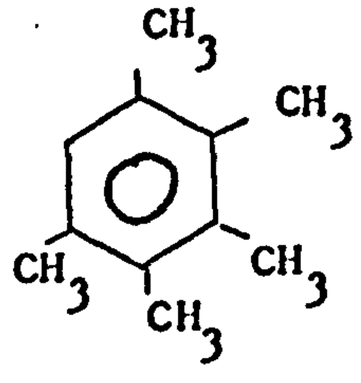
1. 

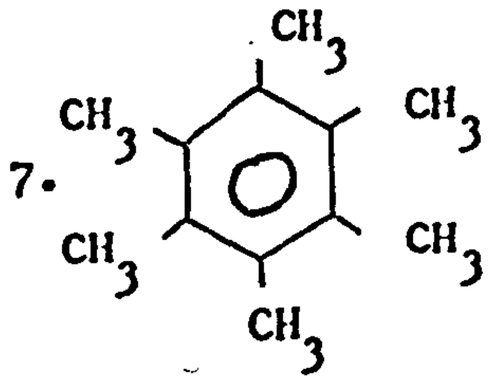
2. 

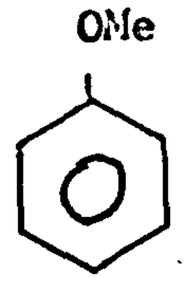
3. 

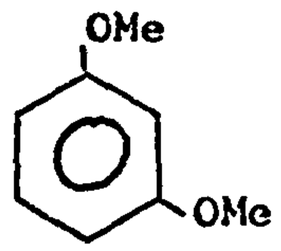
4. 

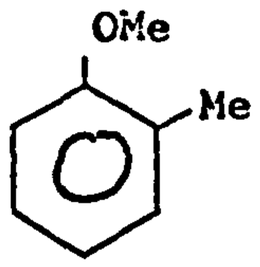
5. 

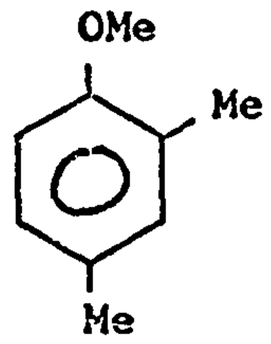
6. 

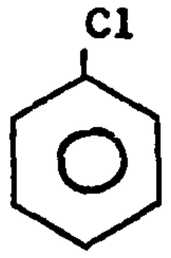
7. 

8. 

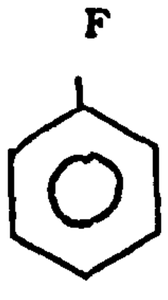
9. 

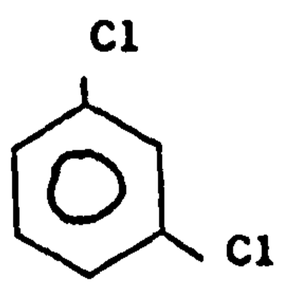
10. 

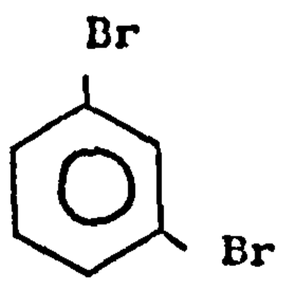
11. 

12. 

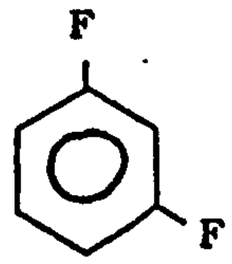
13. 

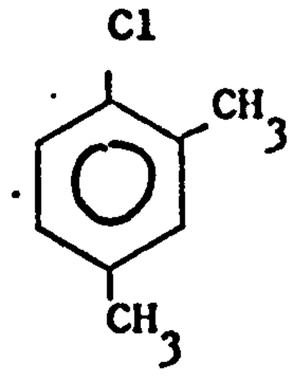
14. 

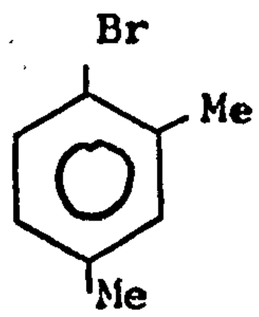
15. 

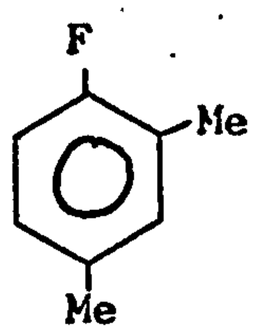
16. 

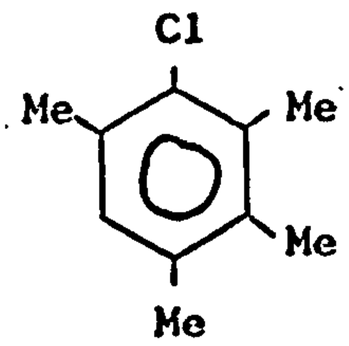
Classification - Benzene Set

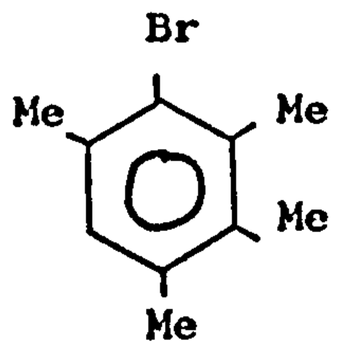
17. 

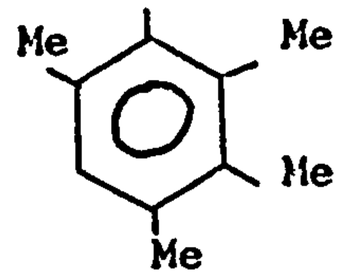
18. 

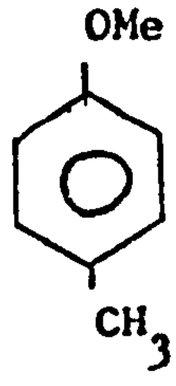
19. 

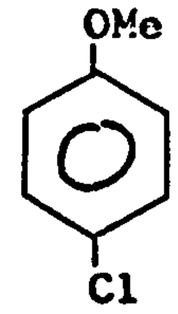
20. 

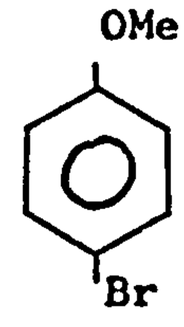
21. 

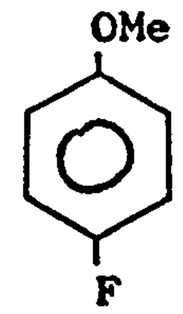
22. 

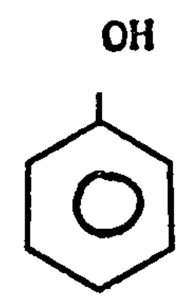
23. 

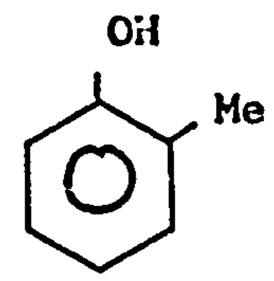
24. 

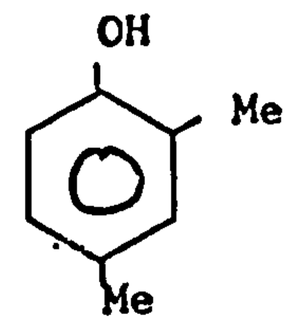
25. 

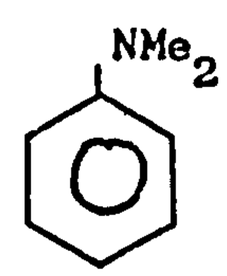
26. 

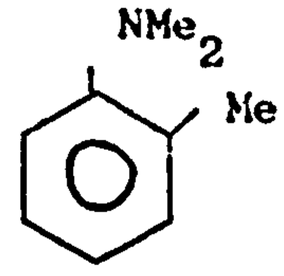
27. 

28. 

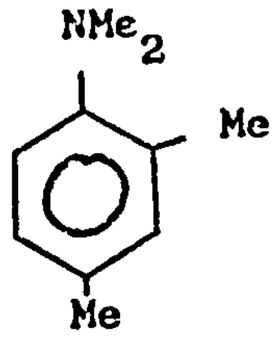
29. 

30. 

31. 

32. 

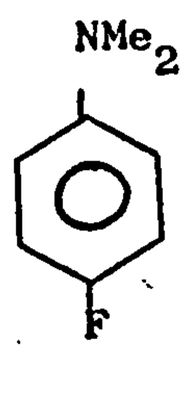
33.



34.

35.

36.



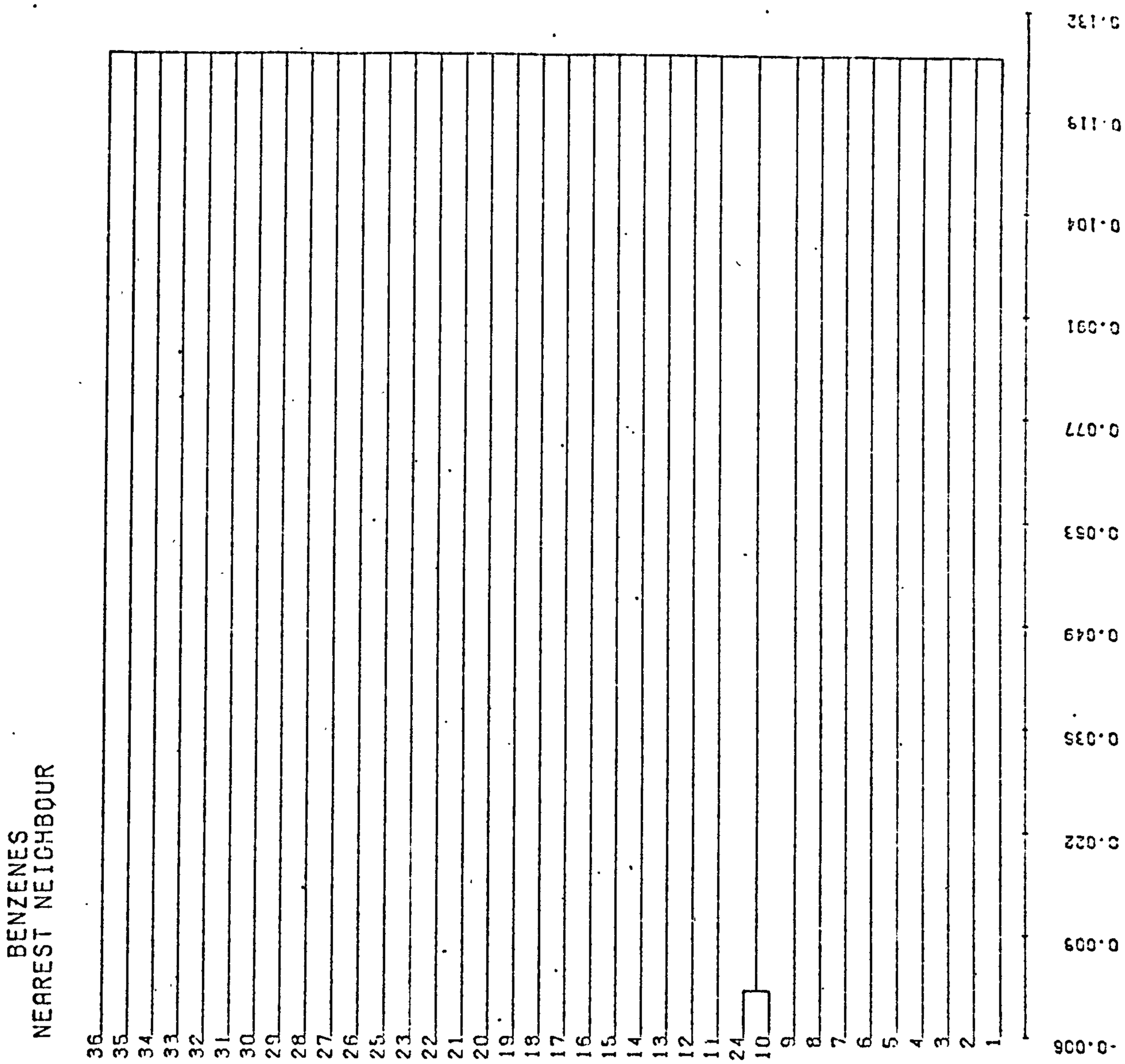


Figure 8A

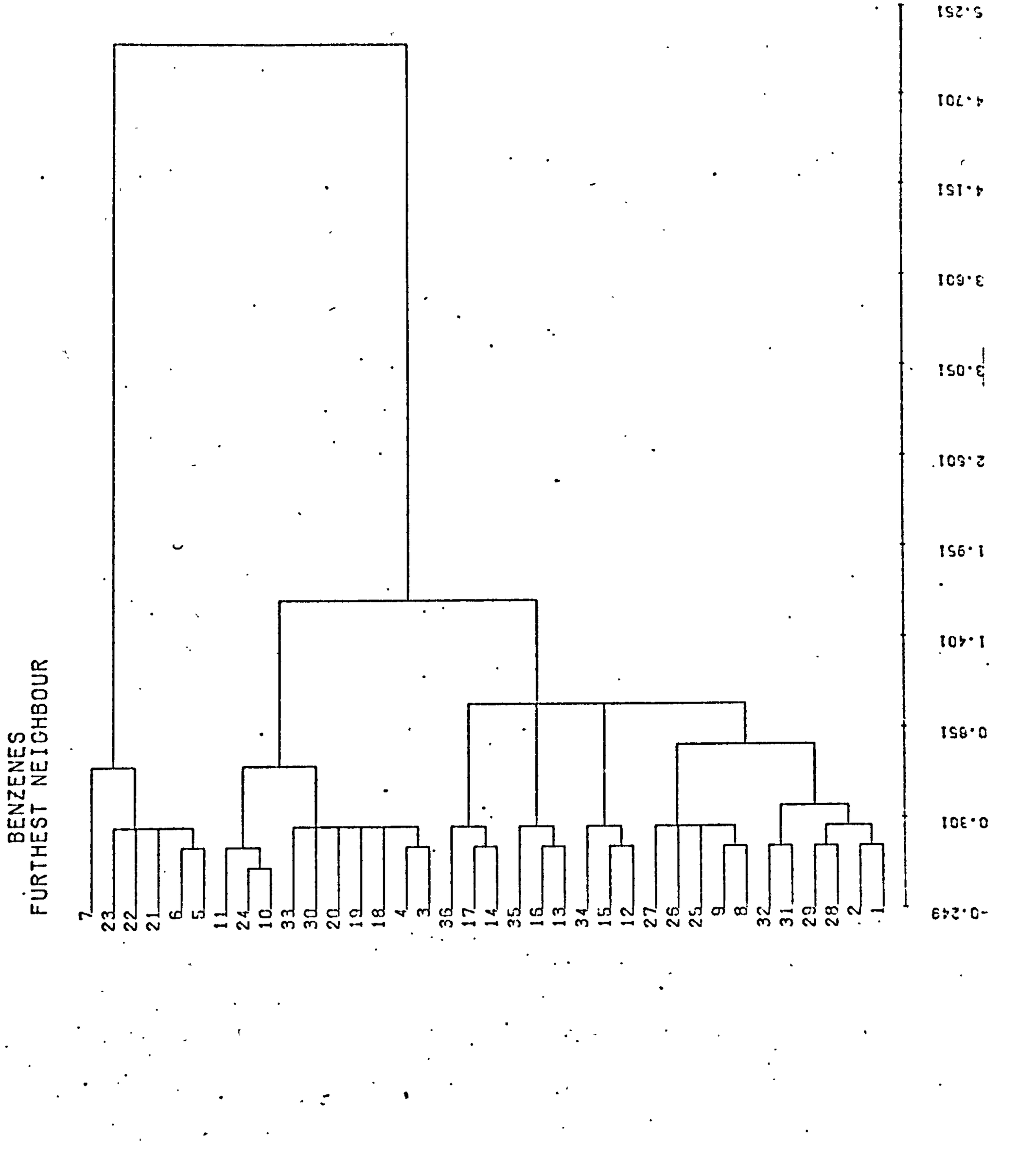


Figure BB

BB

CALCOMP OUTPUT FOR :LILGWA.CLUSTESTD
FROM FILE PRODUCED ON 5FEB77 AT 09.57.17

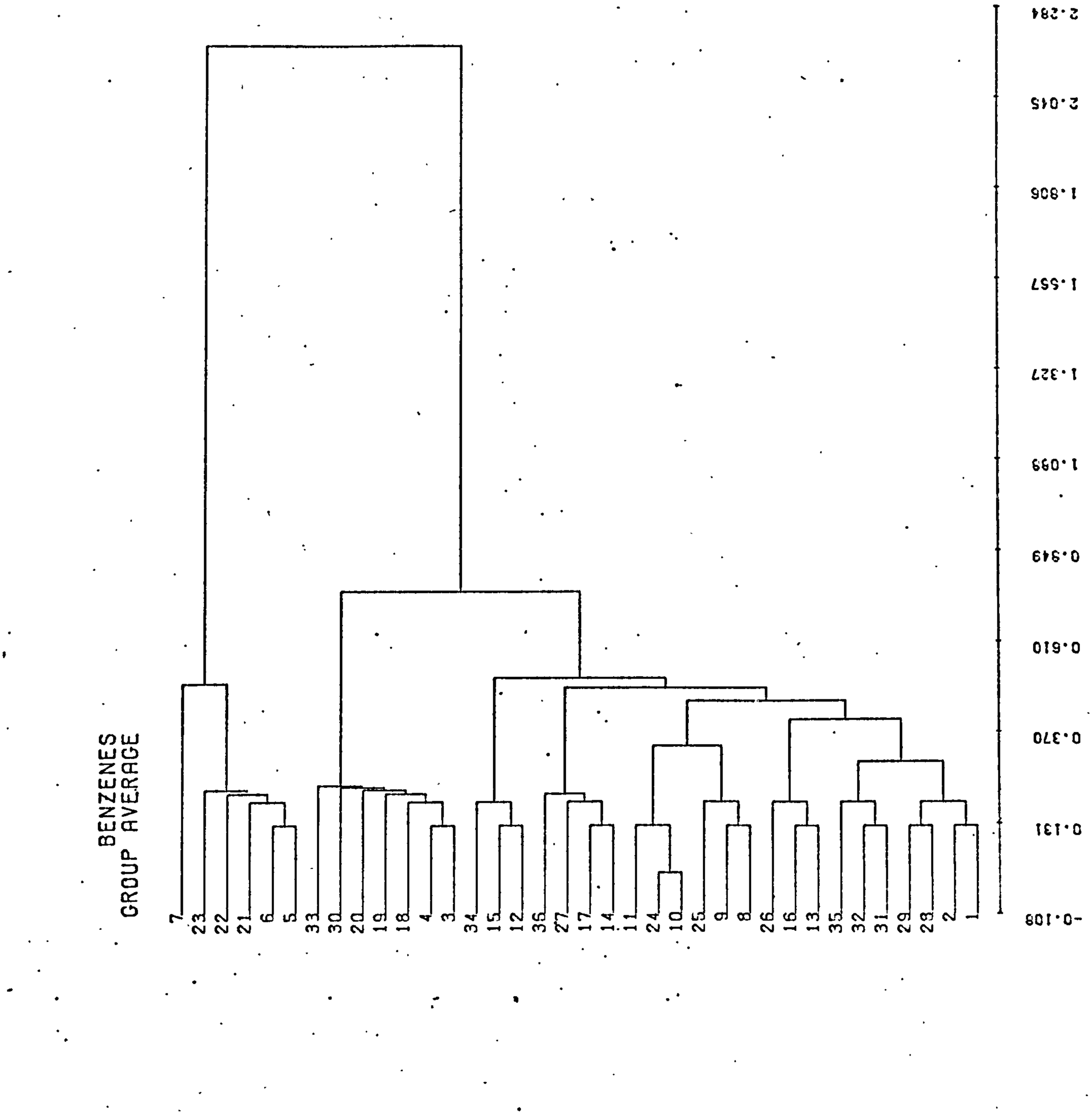


Figure BC

BC

CALCOMP OUTPUT FOR :LILGWA.CLUSTESTD
FROM FILE PRODUCED ON 5FEB77 AT 10.17.42

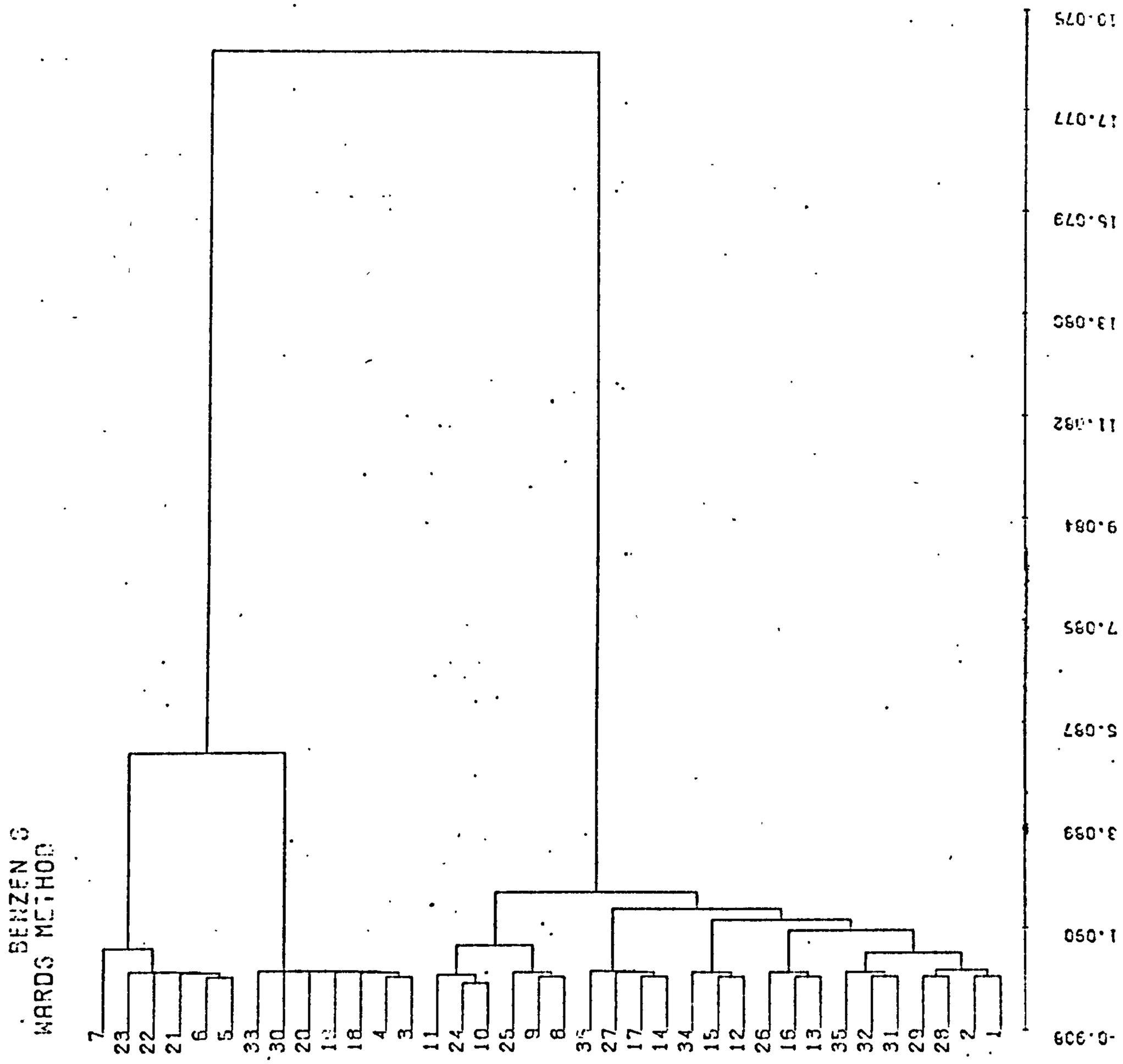
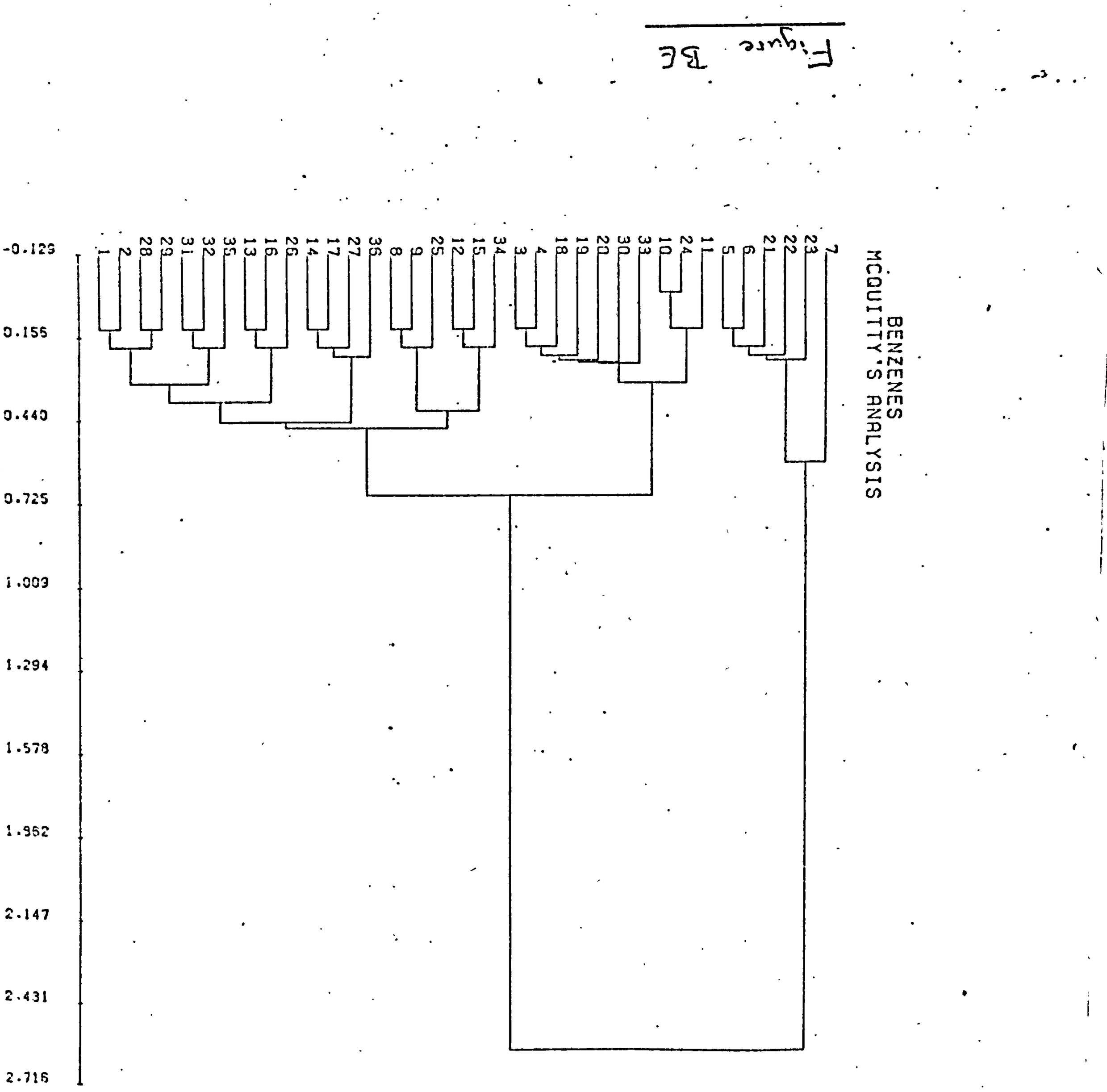


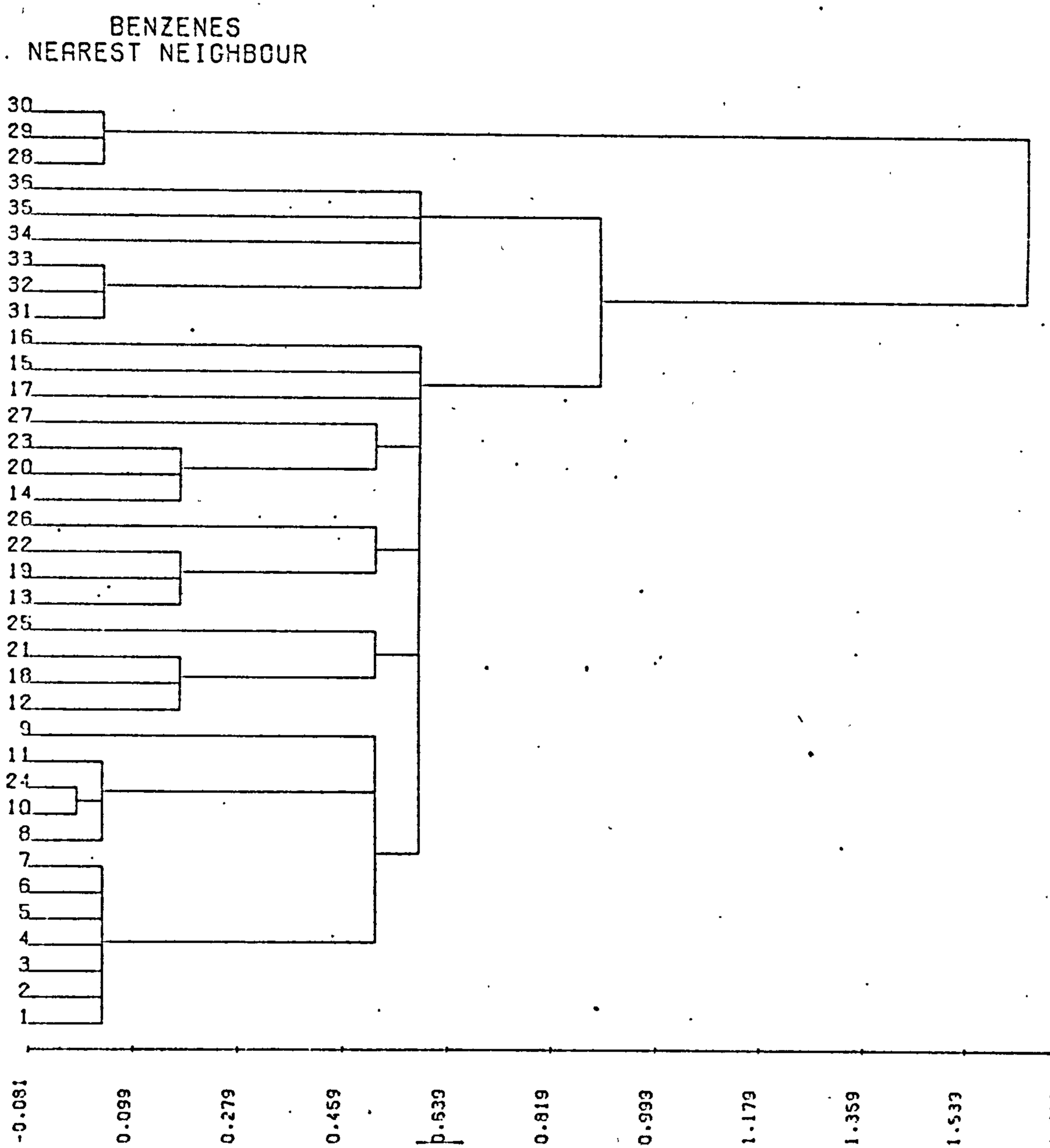
Figure BD

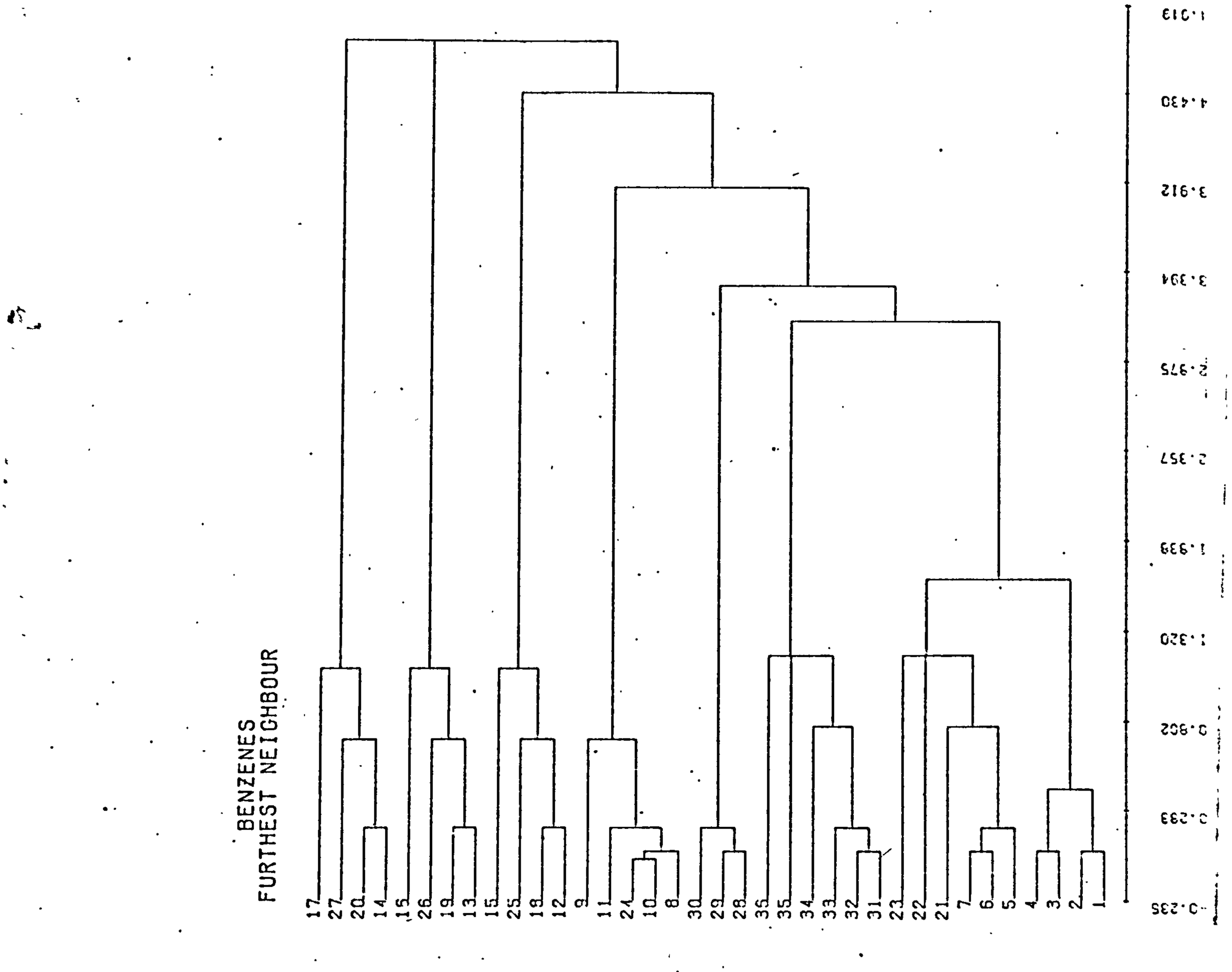
CALCOMP OUTPUT FOR .LI1GWA.CLUSTESTD
FROM FILE PRODUCED ON 5FEB77 AT 12.25.11



CALCOMP OUTPUT FOR
FROM FILE PRODUCED ON 5FEB77 AT 14.18.

Figure BF



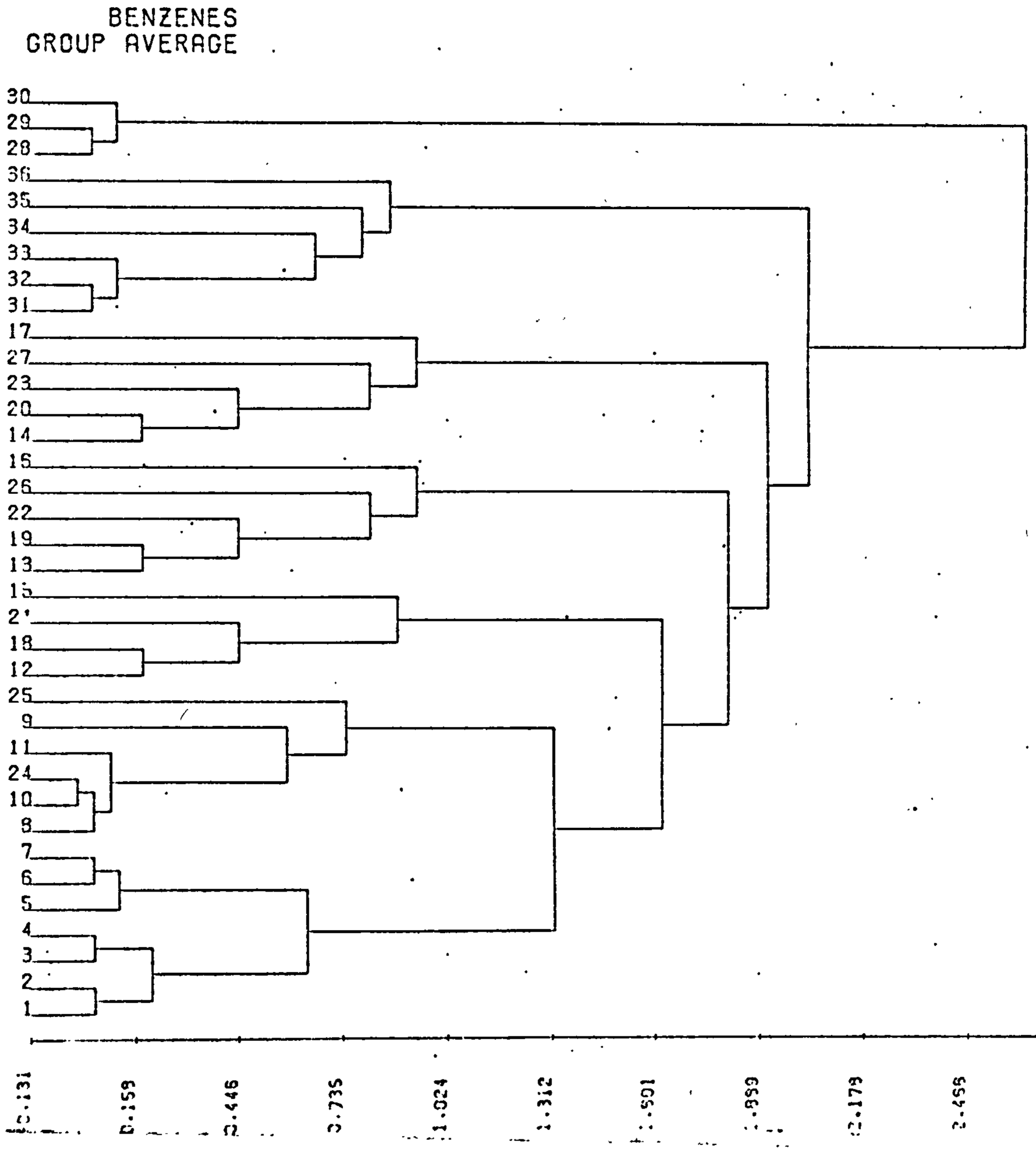


BG 5

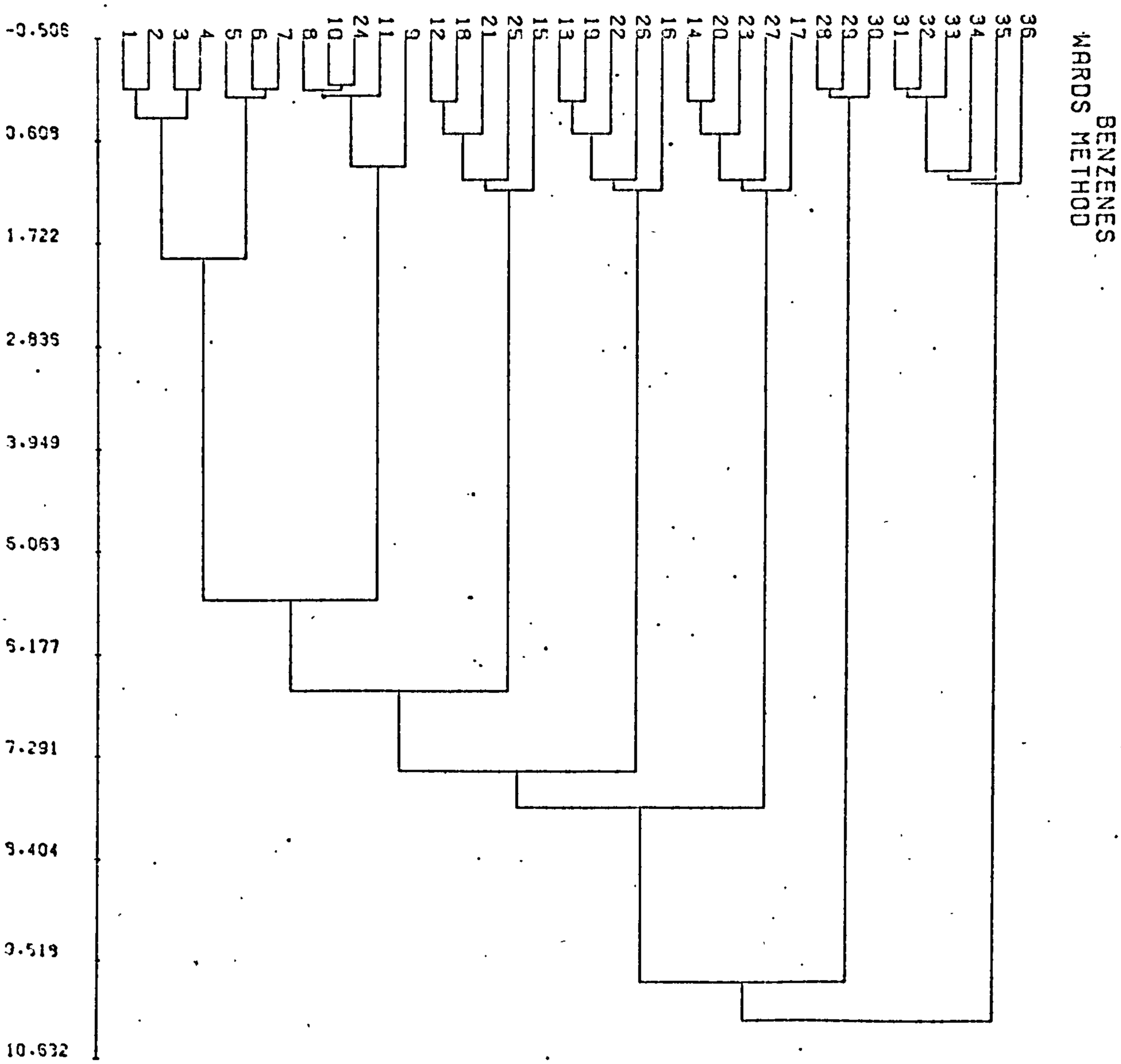
Figure BG

COMP OUTPUT FOR LIGNA 8:LUSTESTD 18.:.52
FROM FILE PRODUC:D 0:77 AT

Figure BH



CALCOMP OUTPUT FOR :LI1GWA.CLUSTESTD
FROM FILE PRODUCED ON 5FEB77 AT 15.15.34



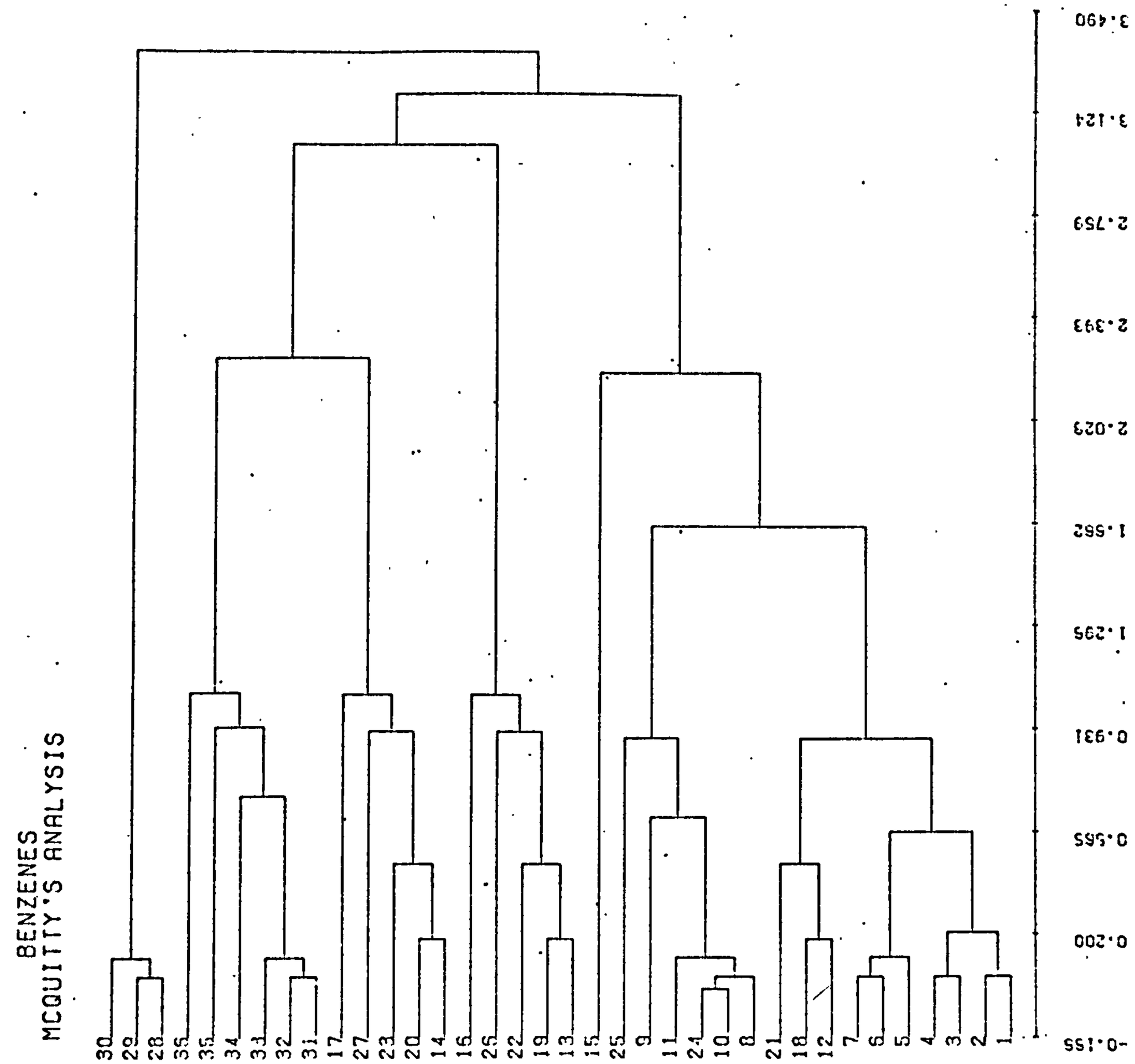
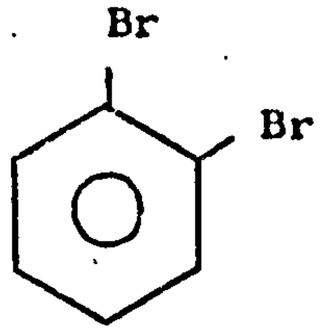


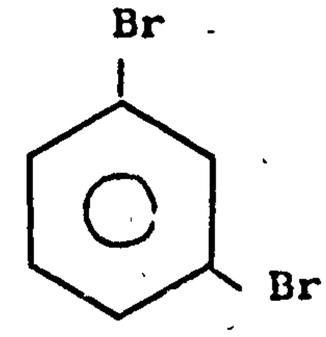
Figure BJ

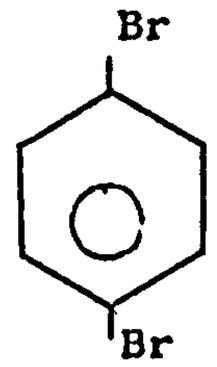
CALCOMP OUTPUT FOR FILE :LIIGWA.CLUSTESTD FROM FILE PRODUCED ON 9FEB77 AT 13.58.2

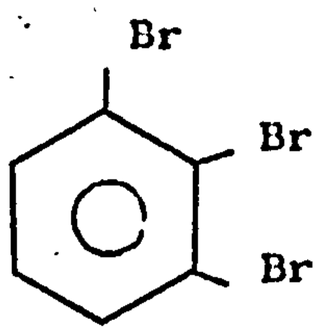
BJ

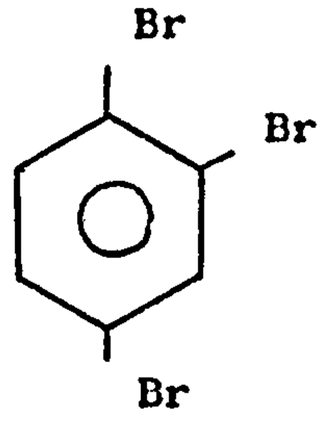
Figure C1

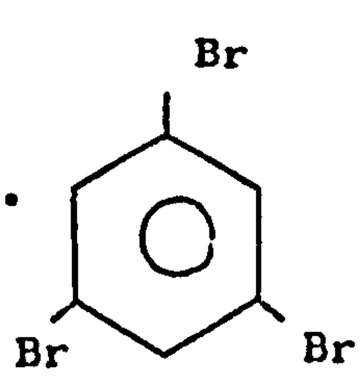
1.


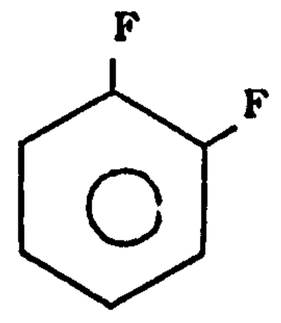
2.


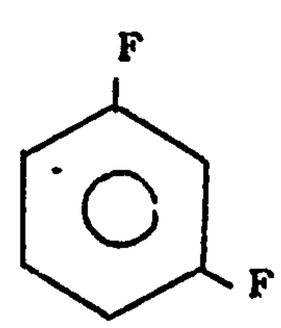
3.


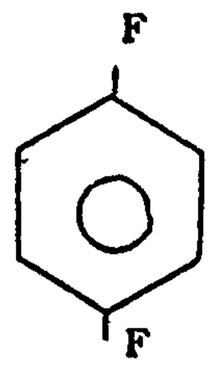
4.


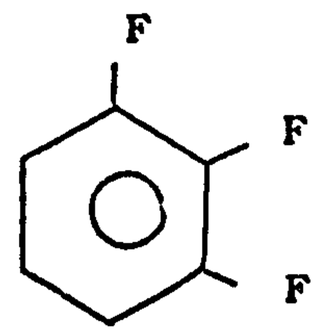
5.


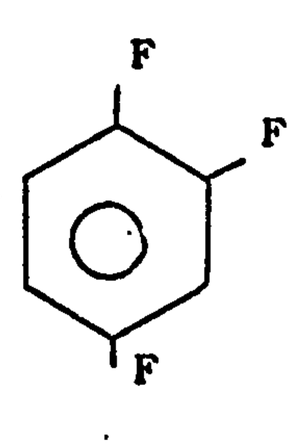
6.


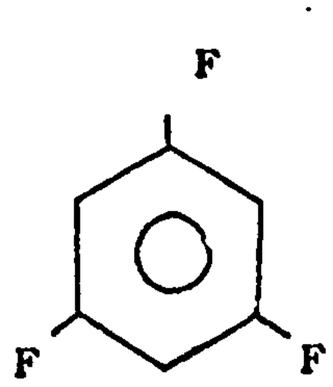
7.


8.


9.


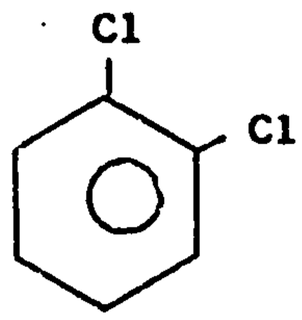
10.


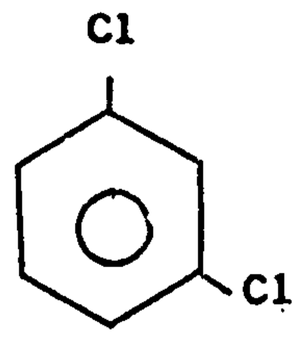
11.


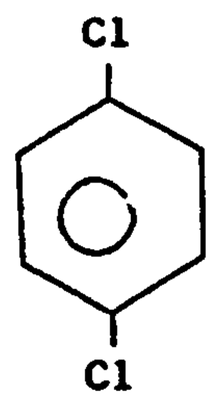
12.


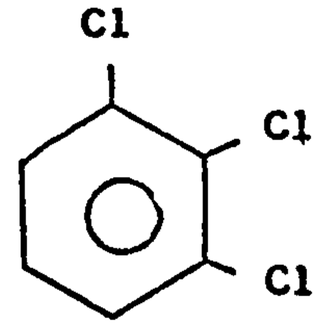
Classification - positional isomer

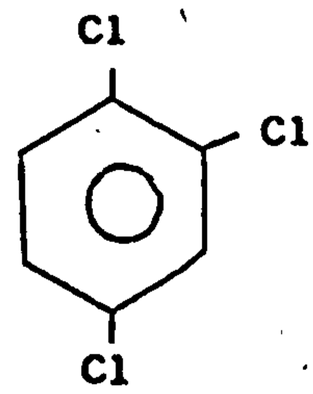
benzenes set

13. 

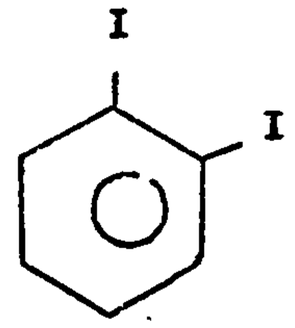
14. 

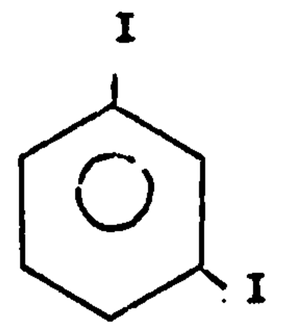
15. 

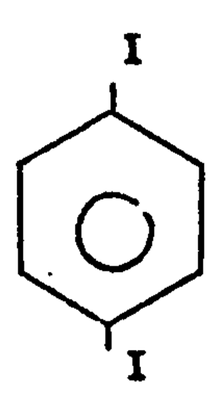
16. 

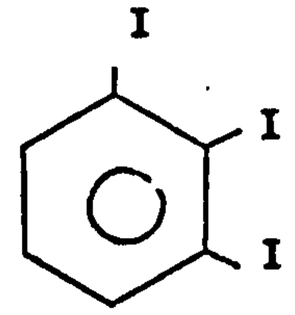
17. 

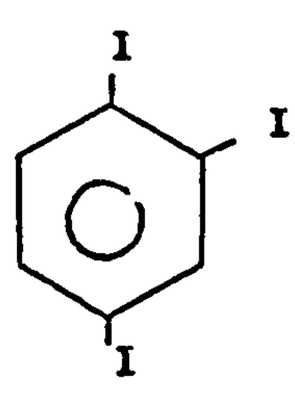
18.

19. 

20. 

21. 

22. 

23. 

24.

positional isomer

benzenes

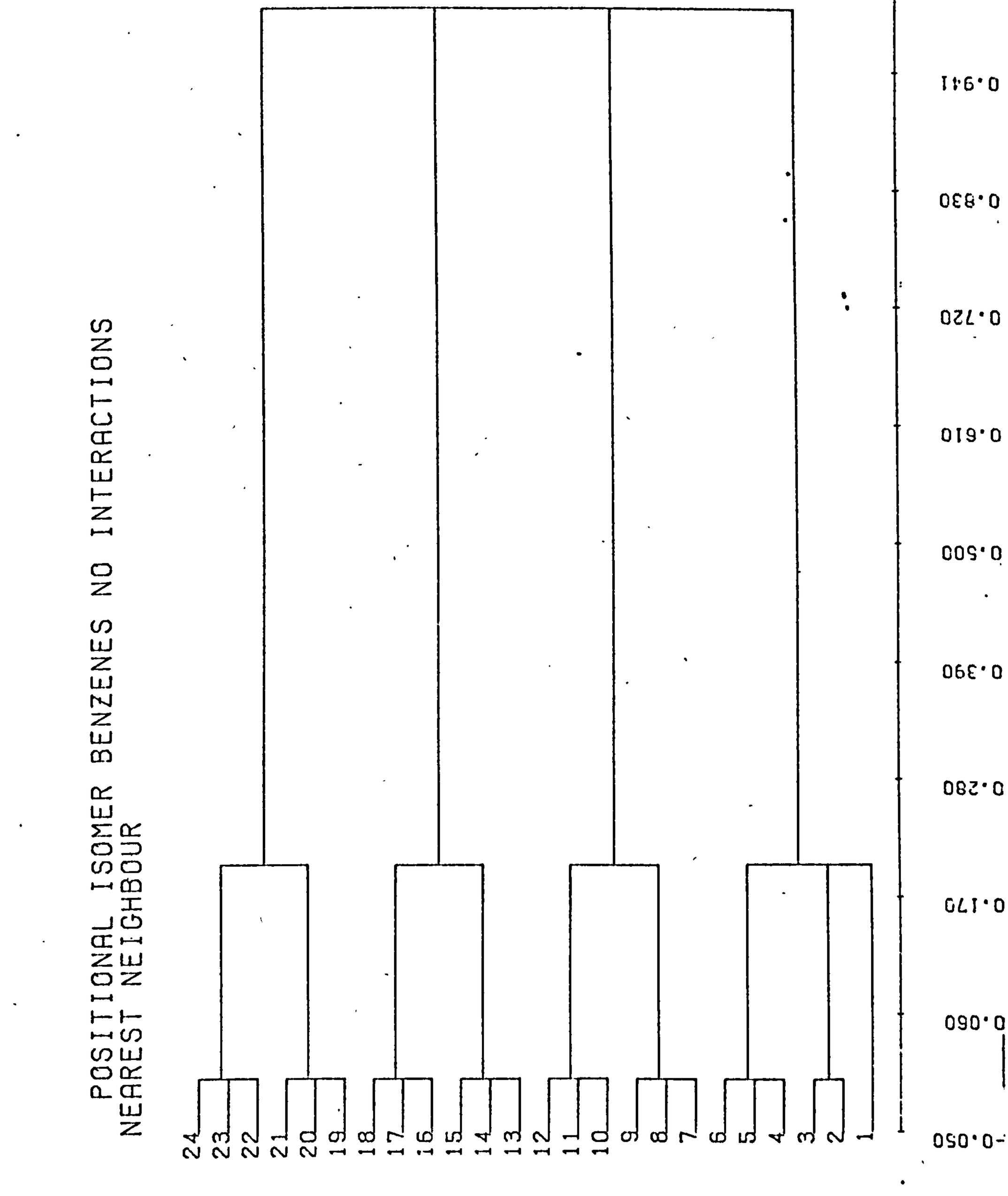

POSITIONAL ISOMER BENZENES NO INTERACTIONS
NEAREST NEIGHBOUR
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
1.050
0.941
0.830
0.720
0.610
0.500
0.390
0.280
0.170
0.060
-0.050

Figure CA

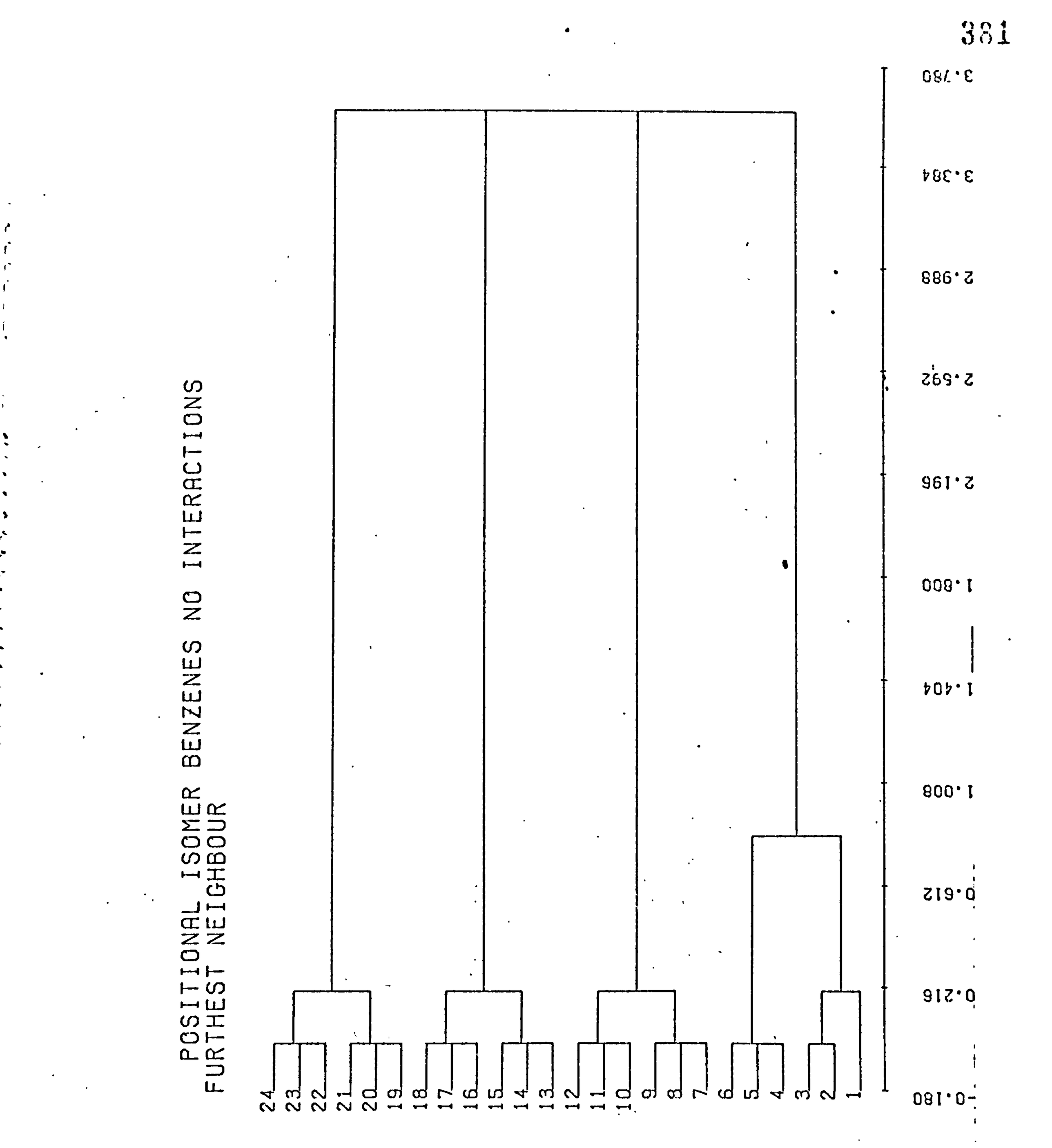

POSITIONAL ISOMER BENZENES NO INTERACTIONS
FURTHEST NEIGHBOUR
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
3.780
3.384
2.988
2.592
2.196
1.800
1.404
1.008
0.612
0.216
0.180

Figure CB

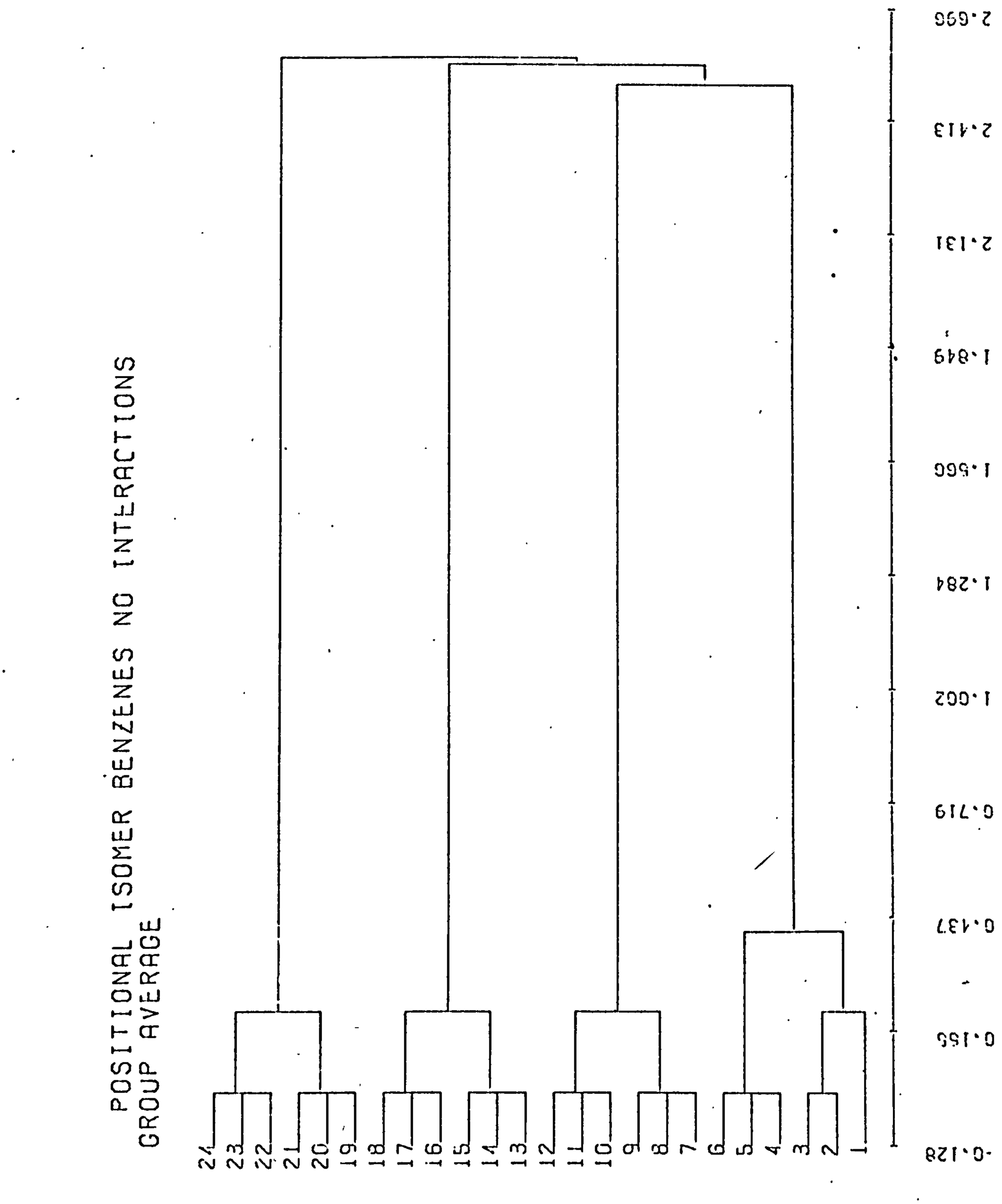

POSITIONAL ISOMER BENZENE NO INTERACTIONS
GROUP AVERAGE
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
2.656
2.413
2.131
1.849
1.566
1.284
1.002
0.719
0.437
0.155
-0.128

Figure CC

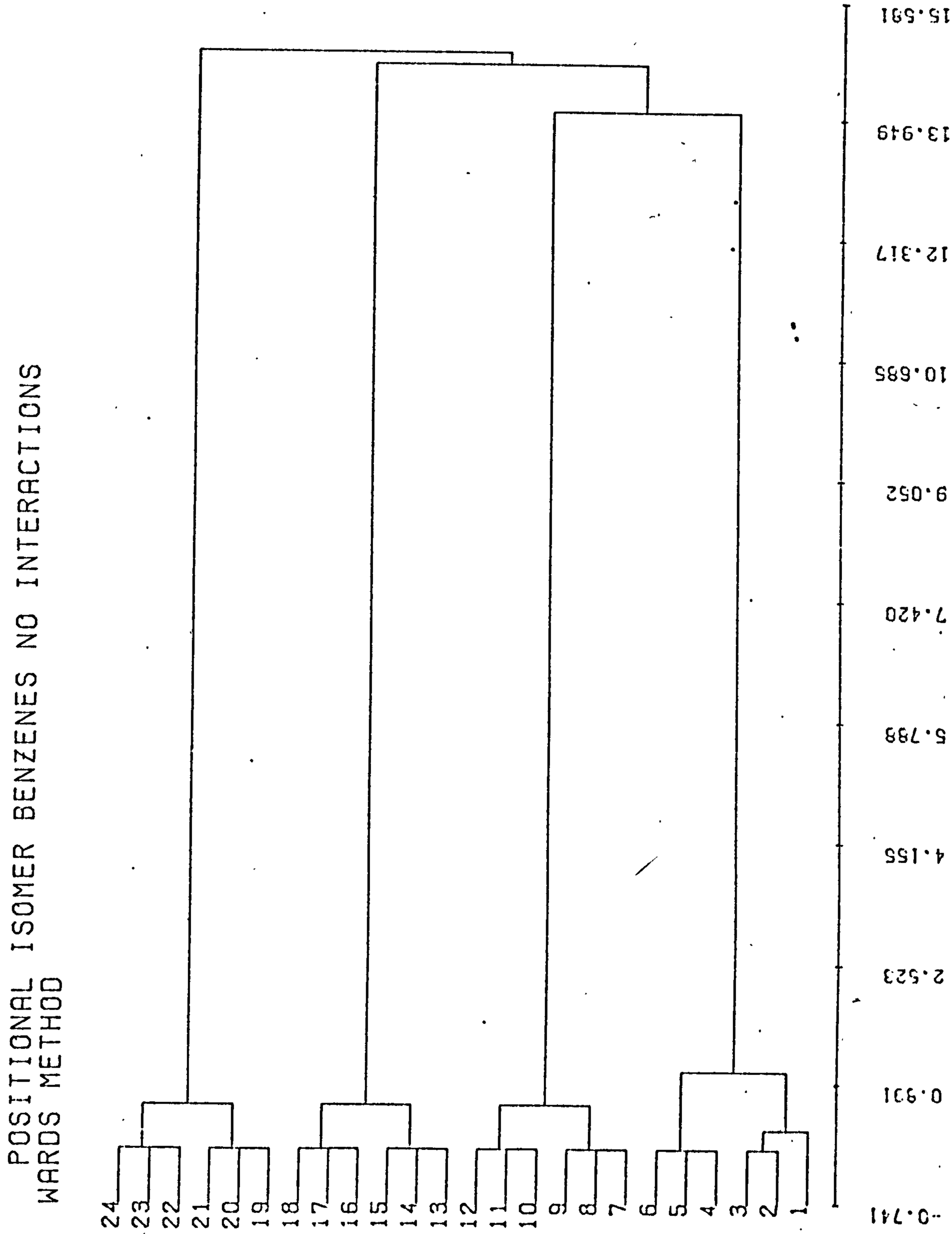

POSITIONAL ISOMER BENZENES NO INTERACTIONS
WARDS METHOD
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
15.581
13.949
12.317
10.685
9.052
7.420
5.788
4.155
2.523
0.891
-0.741

Figure CD

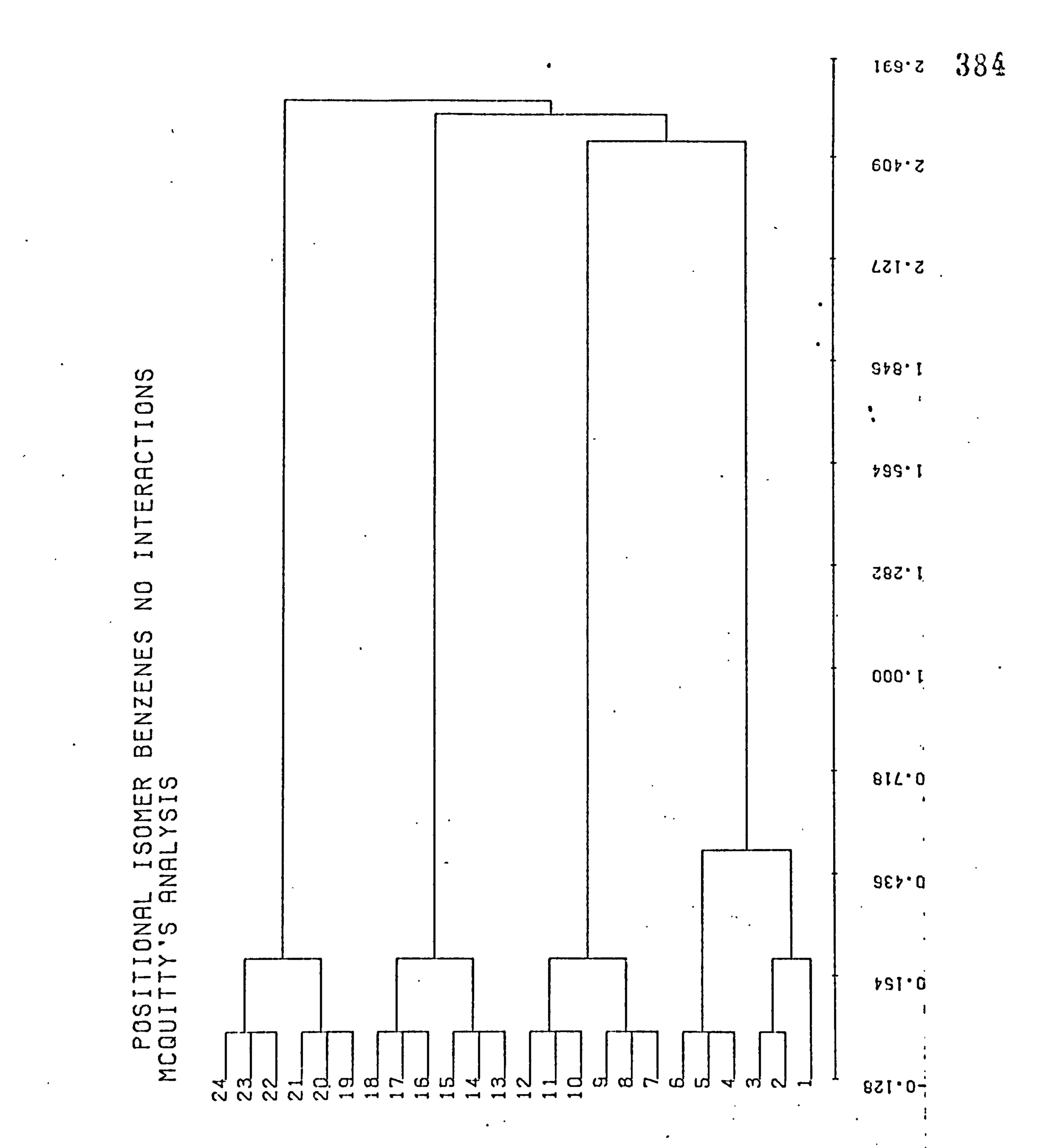

POSITIONAL ISOMER BENZENES NO INTERACTIONS
MCQUITTY'S ANALYSIS
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
2.631
2.409
2.127
1.845
1.564
1.282
1.000
0.718
0.436
0.154
-0.128

Figure CE

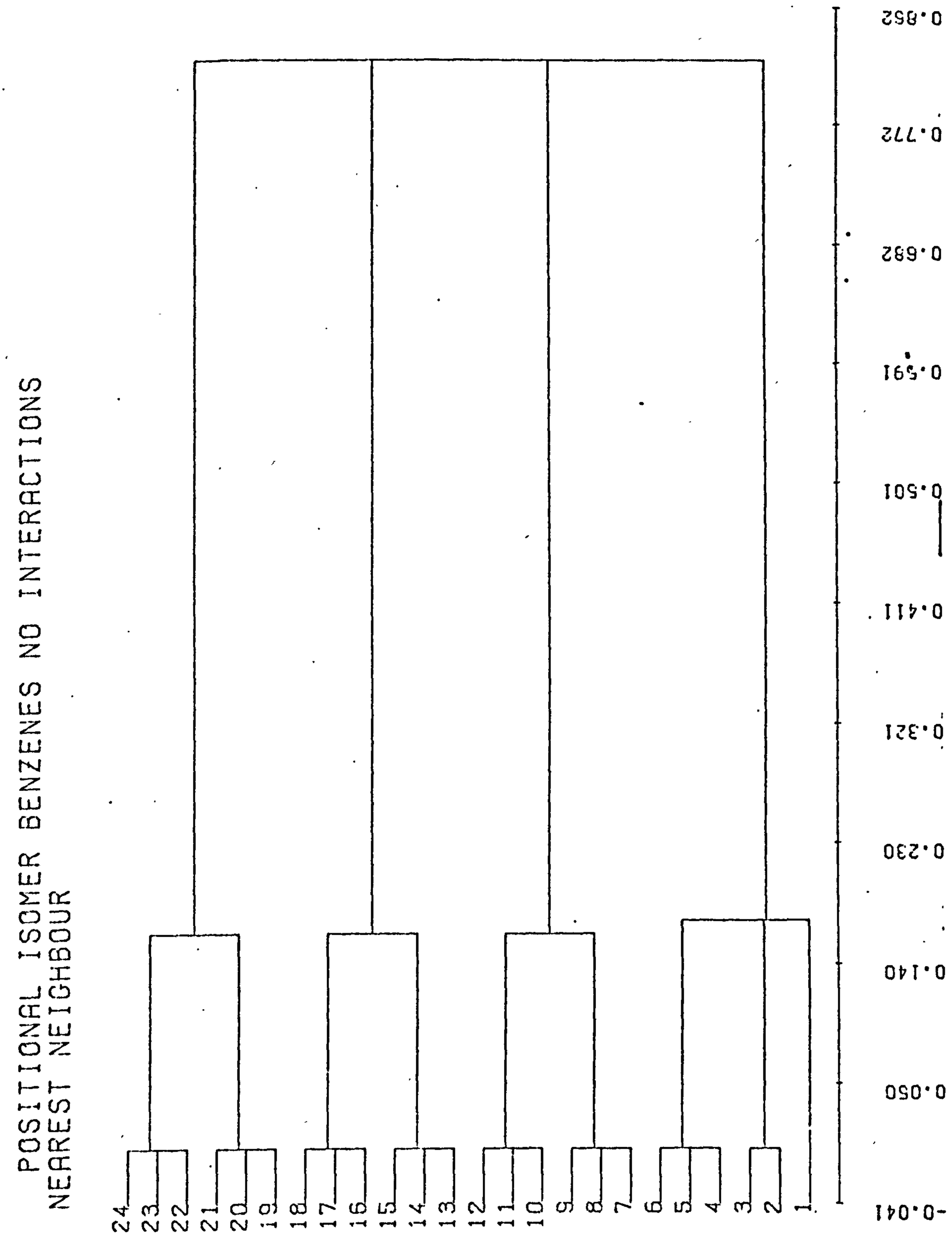

POSITIONAL ISOMER BENZENES NO INTERACTIONS
NEAREST NEIGHBOUR
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0.862
0.772
0.682
0.591
0.501
0.411
0.321
0.230
0.140
0.050
-0.041

Figure CF

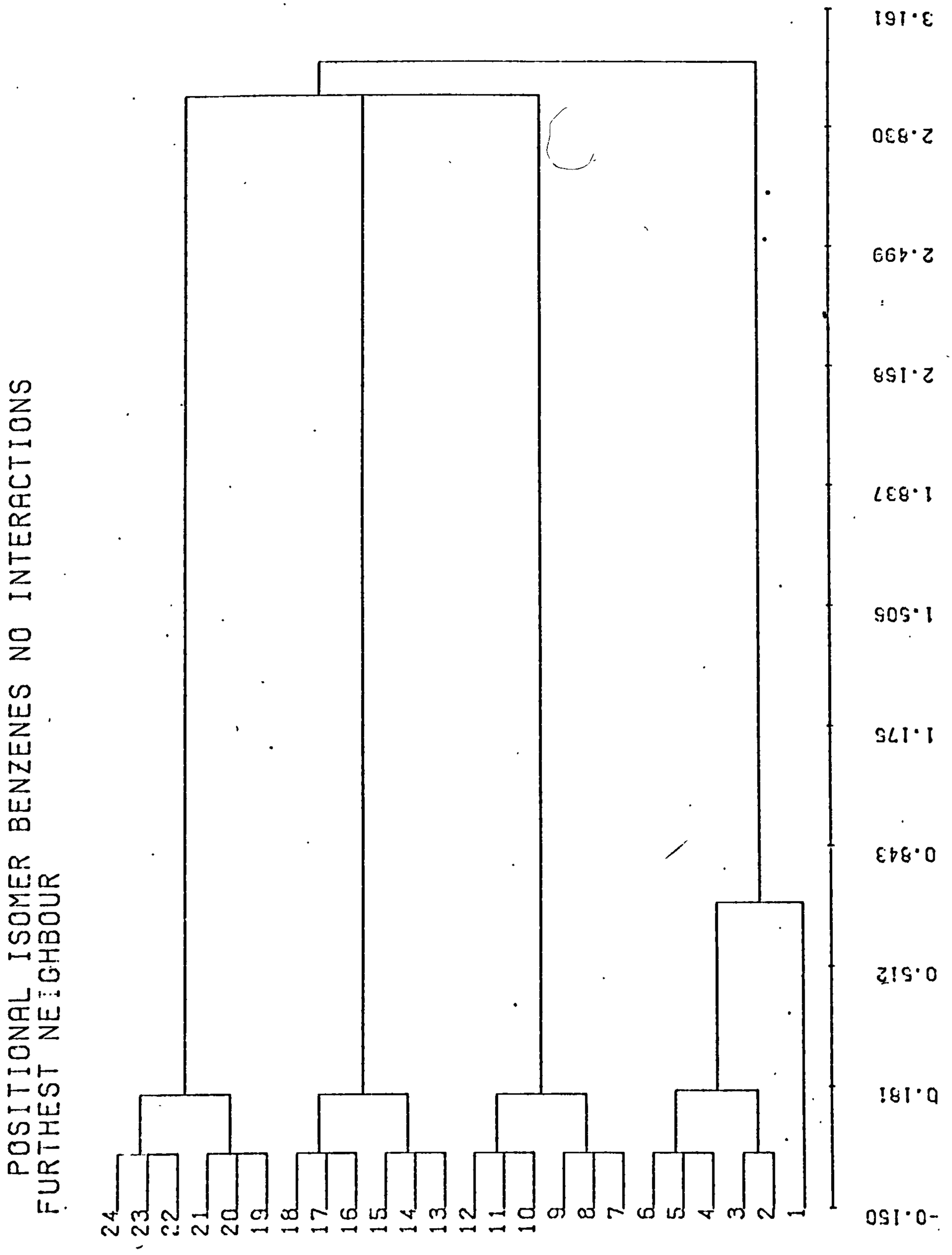

POSITIONAL ISOMER BENZENES NO INTERACTIONS
FURTHEST NEIGHBOUR
3.161
2.830
2.499
2.168
1.837
1.505
1.175
0.843
0.512
0.181
-0.150
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Figure CG

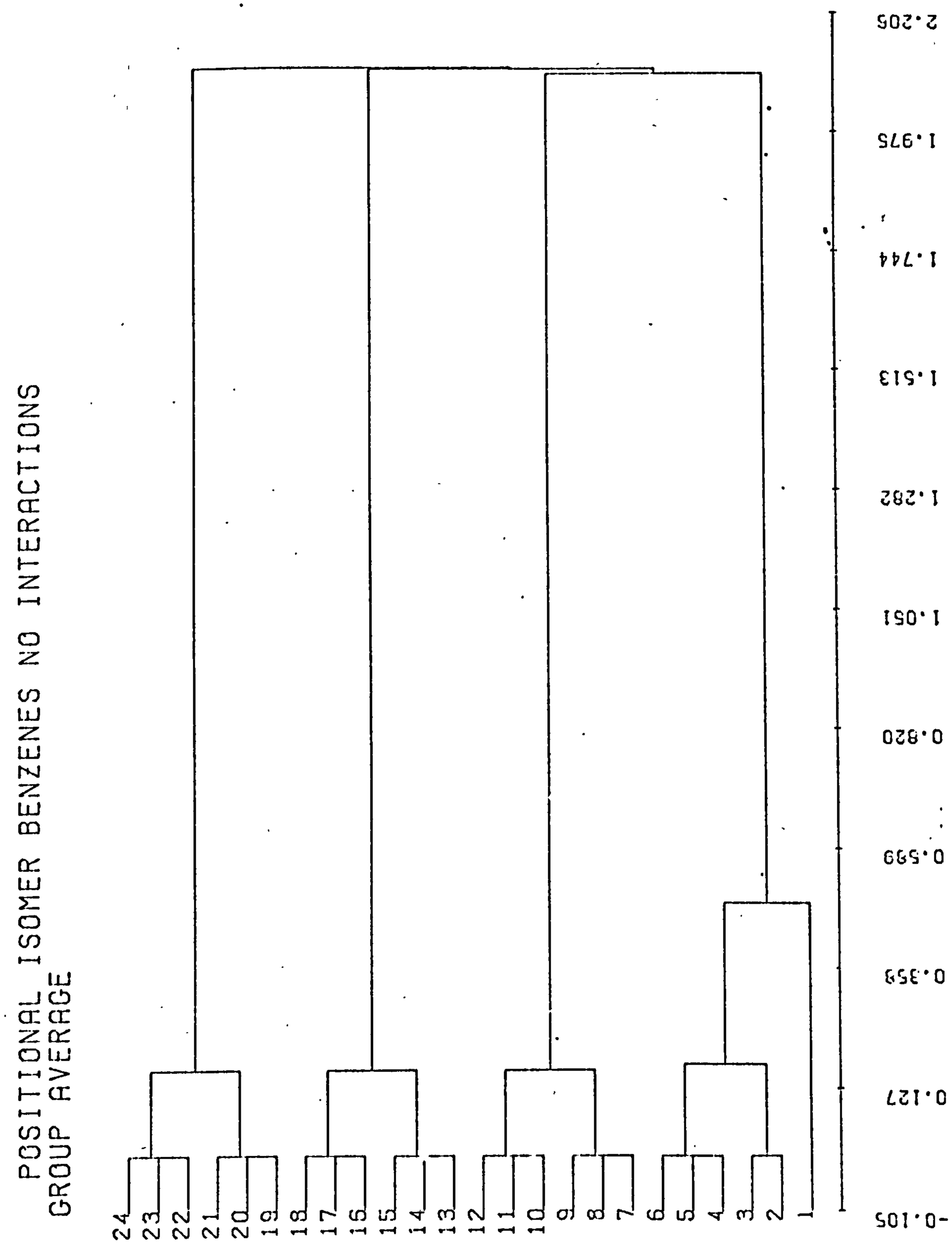

POSITIONAL ISOMER BENZENES NO INTERACTIONS
GROUP AVERAGE
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
2.205
1.975
1.744
1.513
1.282
1.051
0.820
0.589
0.359
0.127
-0.105

Figure CH

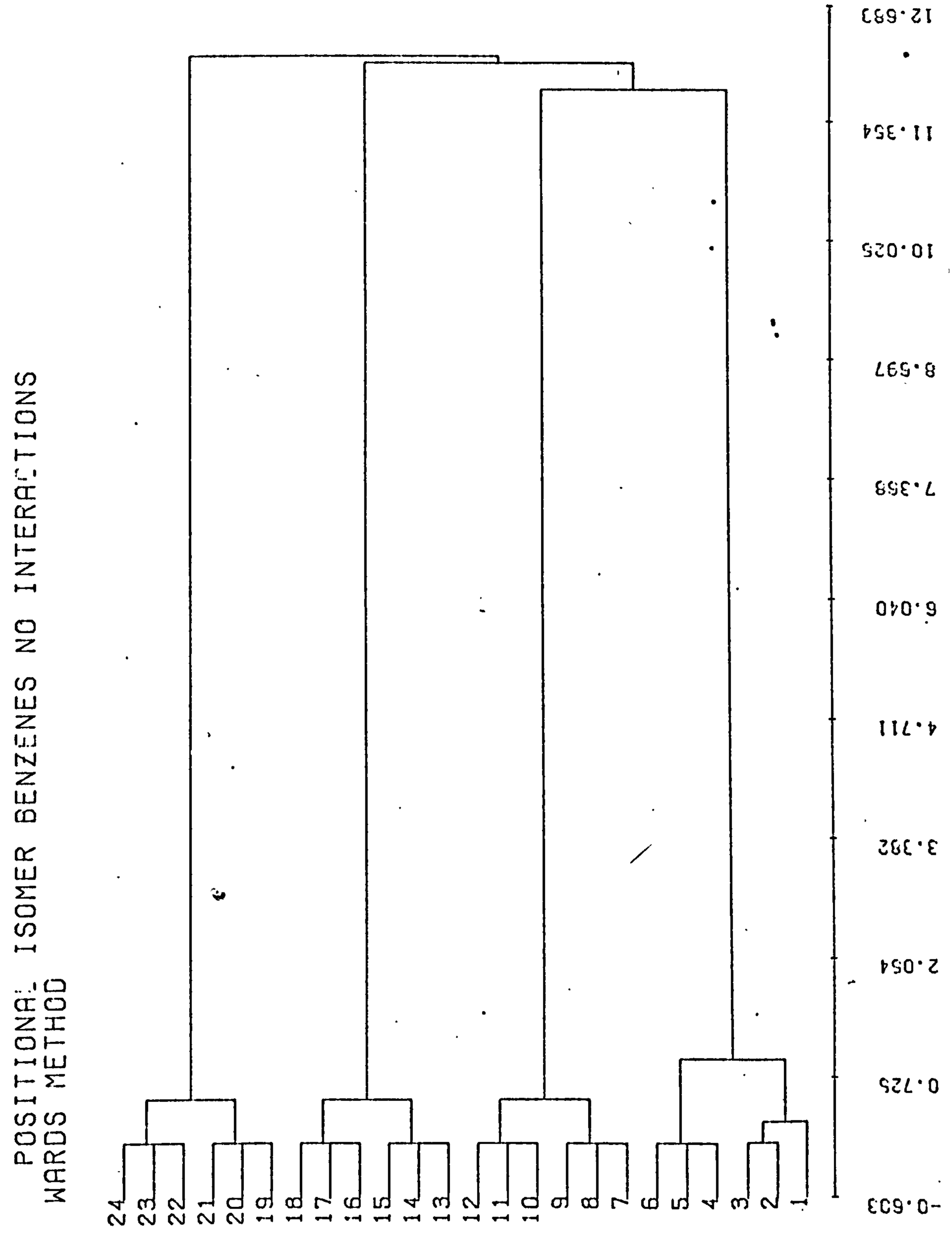

POSITIONAL ISOMER BENZENES NO INTERACTIONS
WARDS METHOD
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
12.683
11.354
10.025
8.697
7.368
6.040
4.711
3.382
2.054
0.725
-0.603

Figure CI

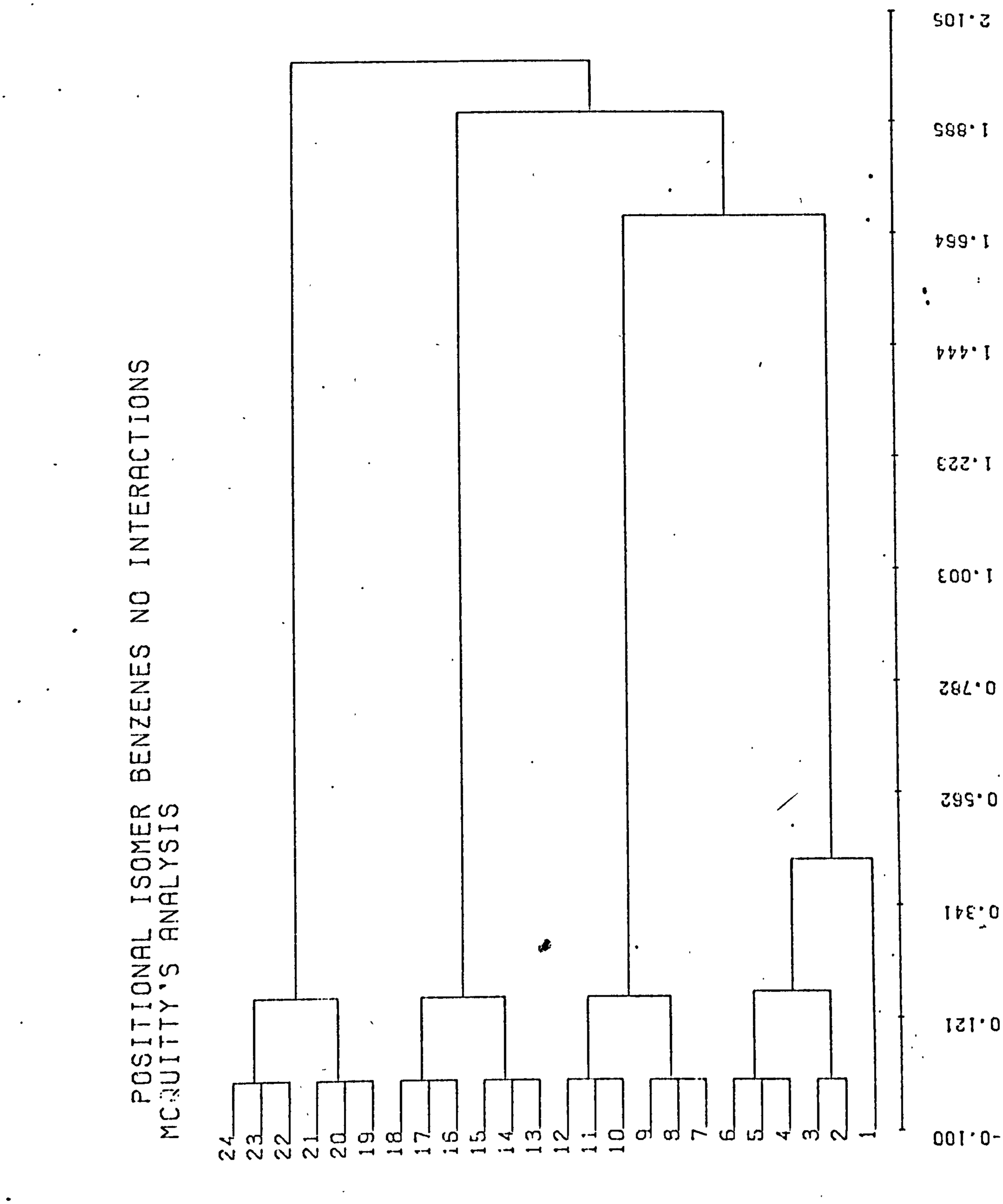

POSITIONAL ISOMER BENZENES
MCQUITTY'S ANALYSIS
NO INTERACTIONS
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
2.105
1.885
1.654
1.444
1.223
1.003
0.782
0.562
0.341
0.121
-0.100

Figure CJ

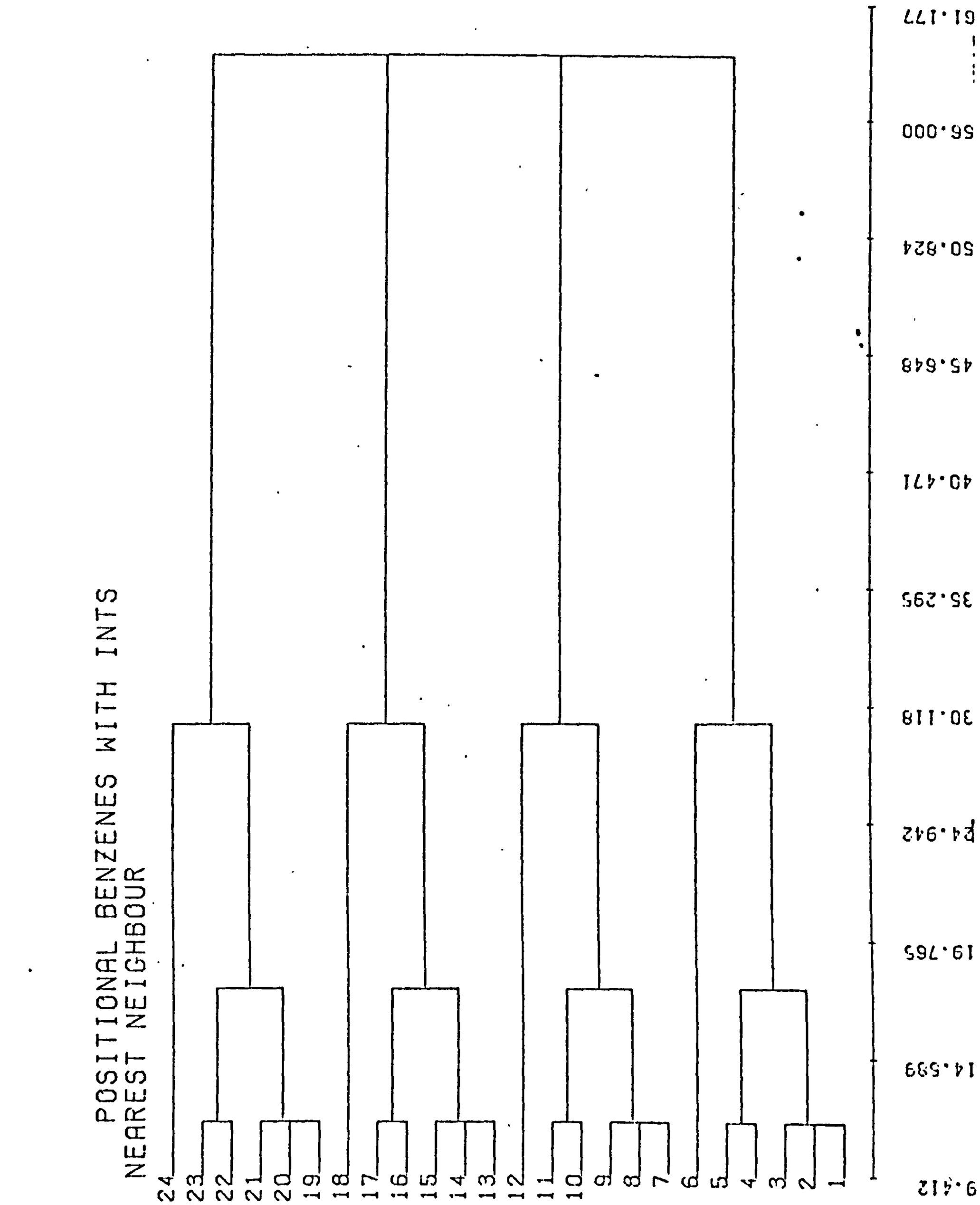

POSITIONAL BENZENES WITH I
NEAREST NEIGHBOUR
INTS
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
61.177
56.000
50.824
45.648
40.471
35.295
30.118
24.942
19.765
14.589
9.412

Figure DA

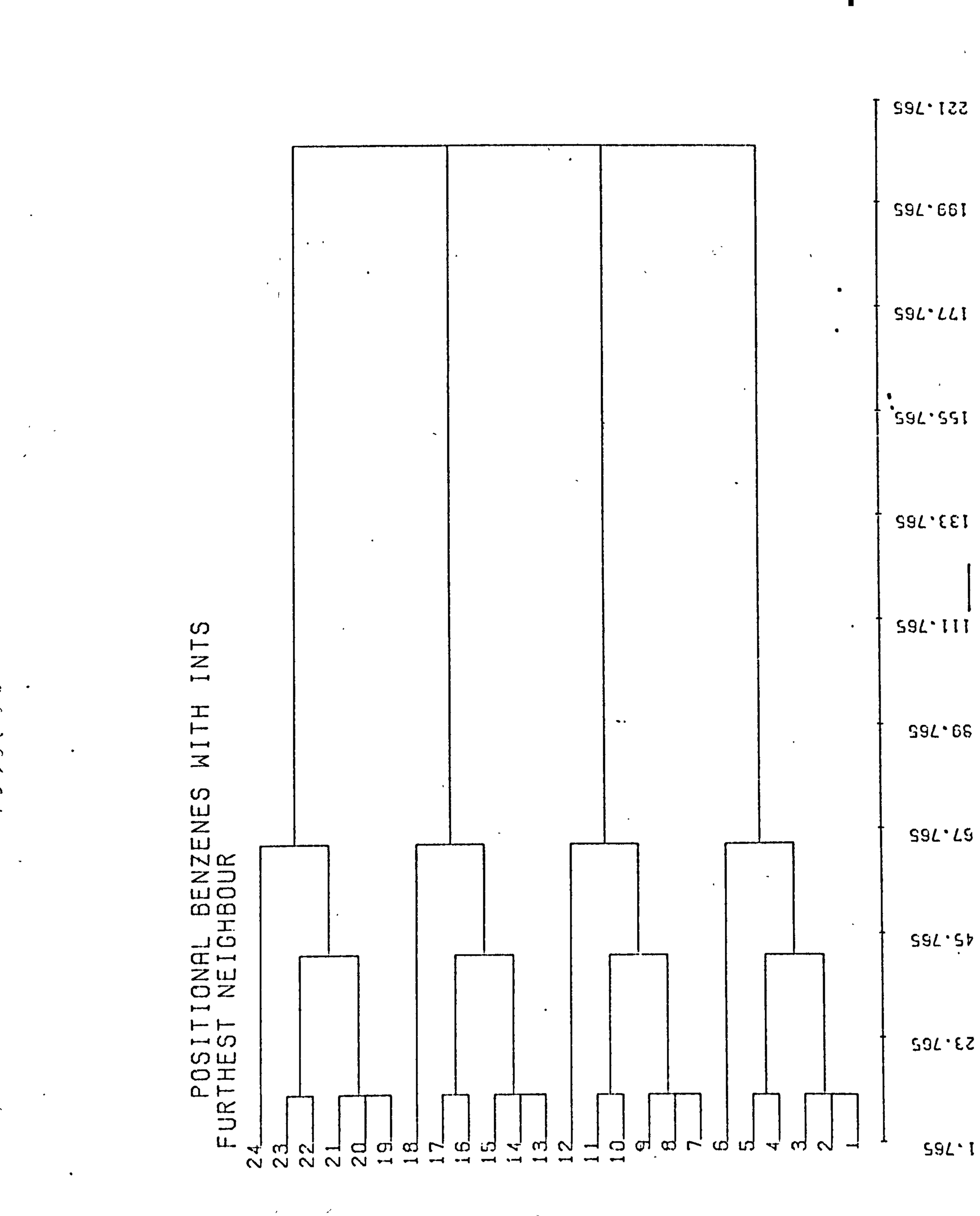

POSITIONAL BENZENES WITH INTS
FURTHEST NEIGHBOUR
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
221.765
199.765
177.765
155.765
133.765
111.765
99.765
67.765
45.765
23.765
1.765

Figure DB

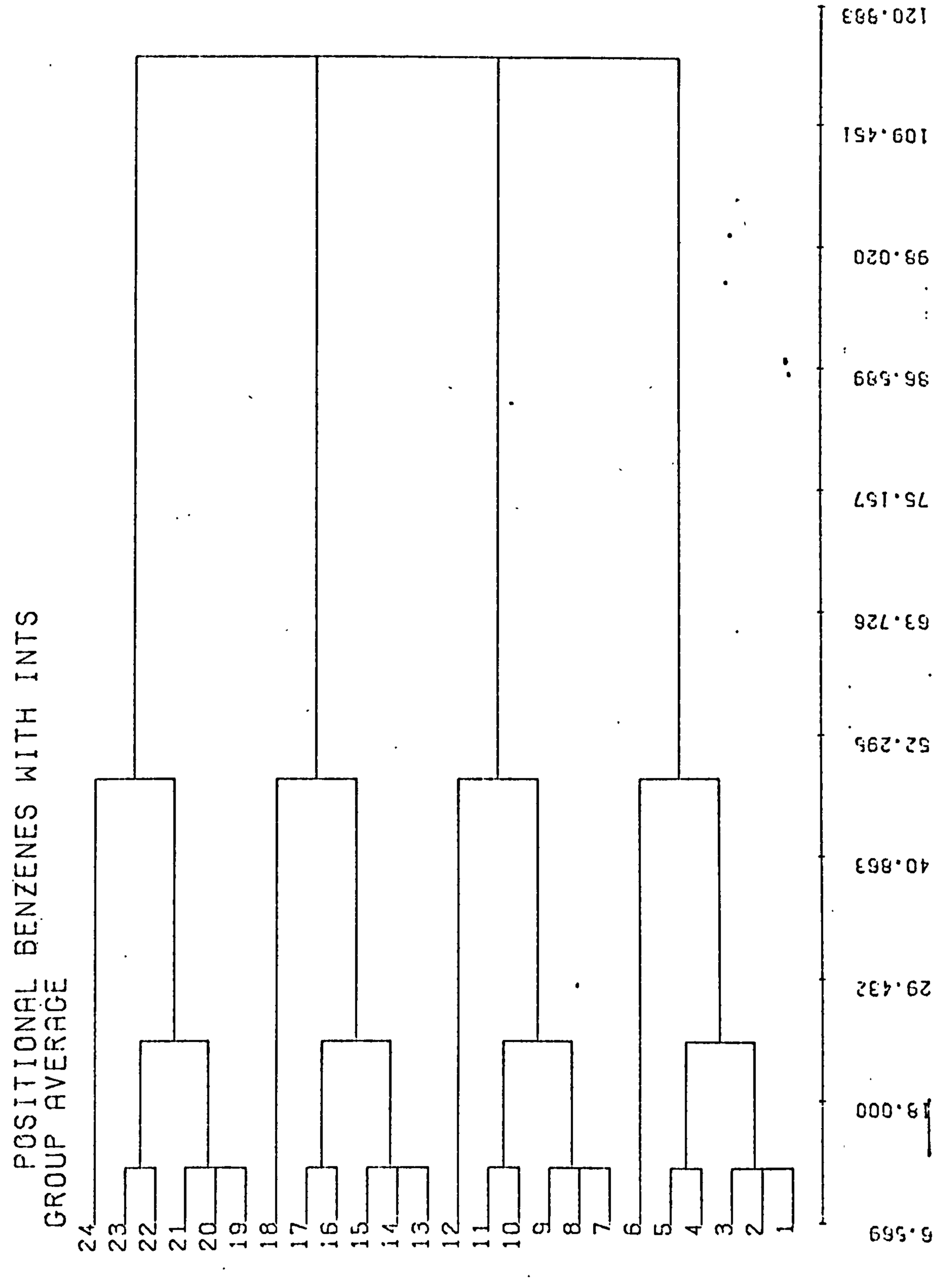

POSITIONAL BENZENES WITH INTS
GROUP AVERAGE
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
120.983
109.451
98.020
86.589
75.157
63.726
52.295
40.863
29.432
19.000
5.569

Figure DC

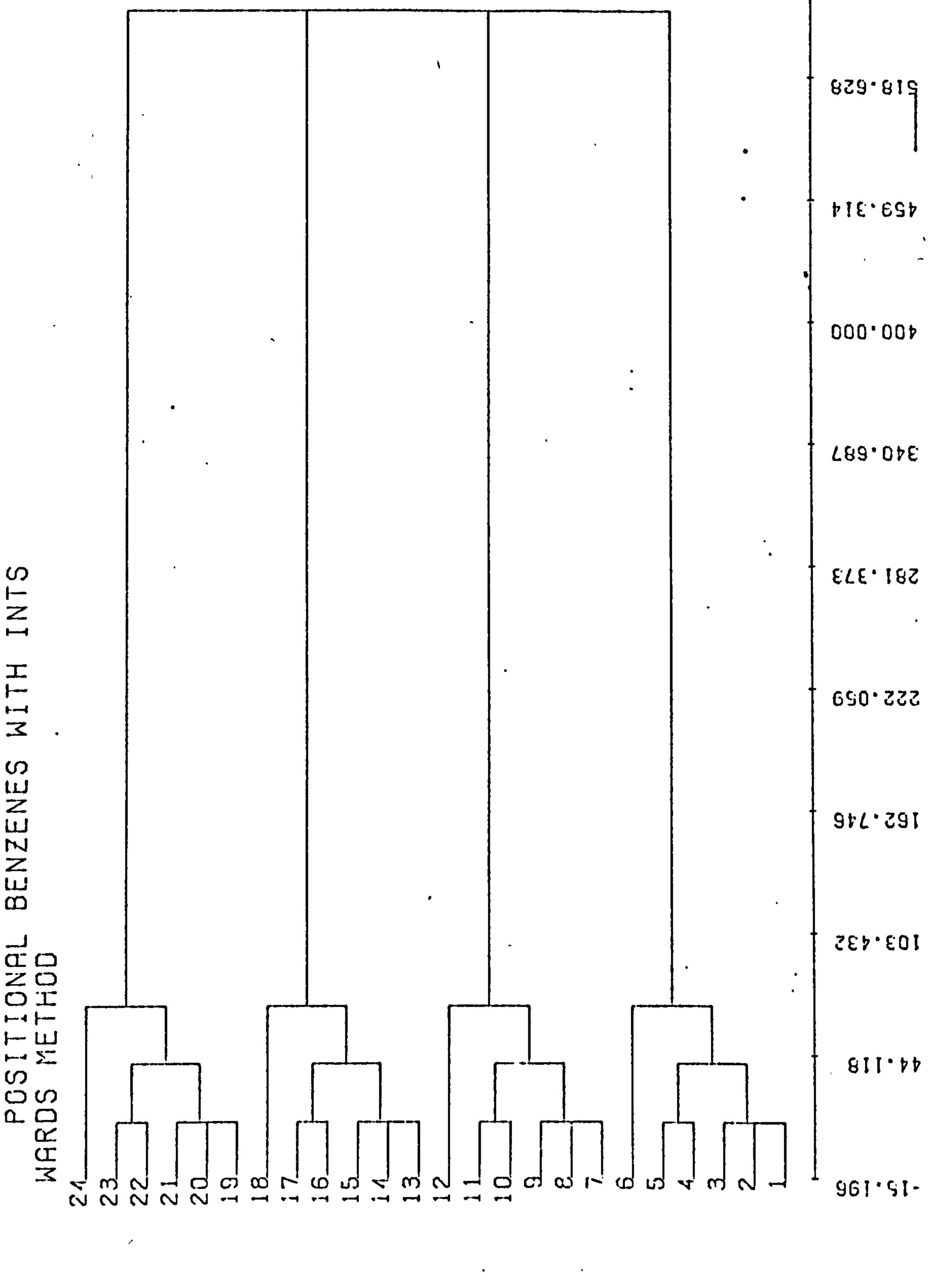

POSITIONAL BENZENES WITH INTS
WARDS METHOD
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
577.942
518.628
459.314
400.000
340.687
281.373
222.059
162.746
103.432
44.118
-15.196

Figure DD

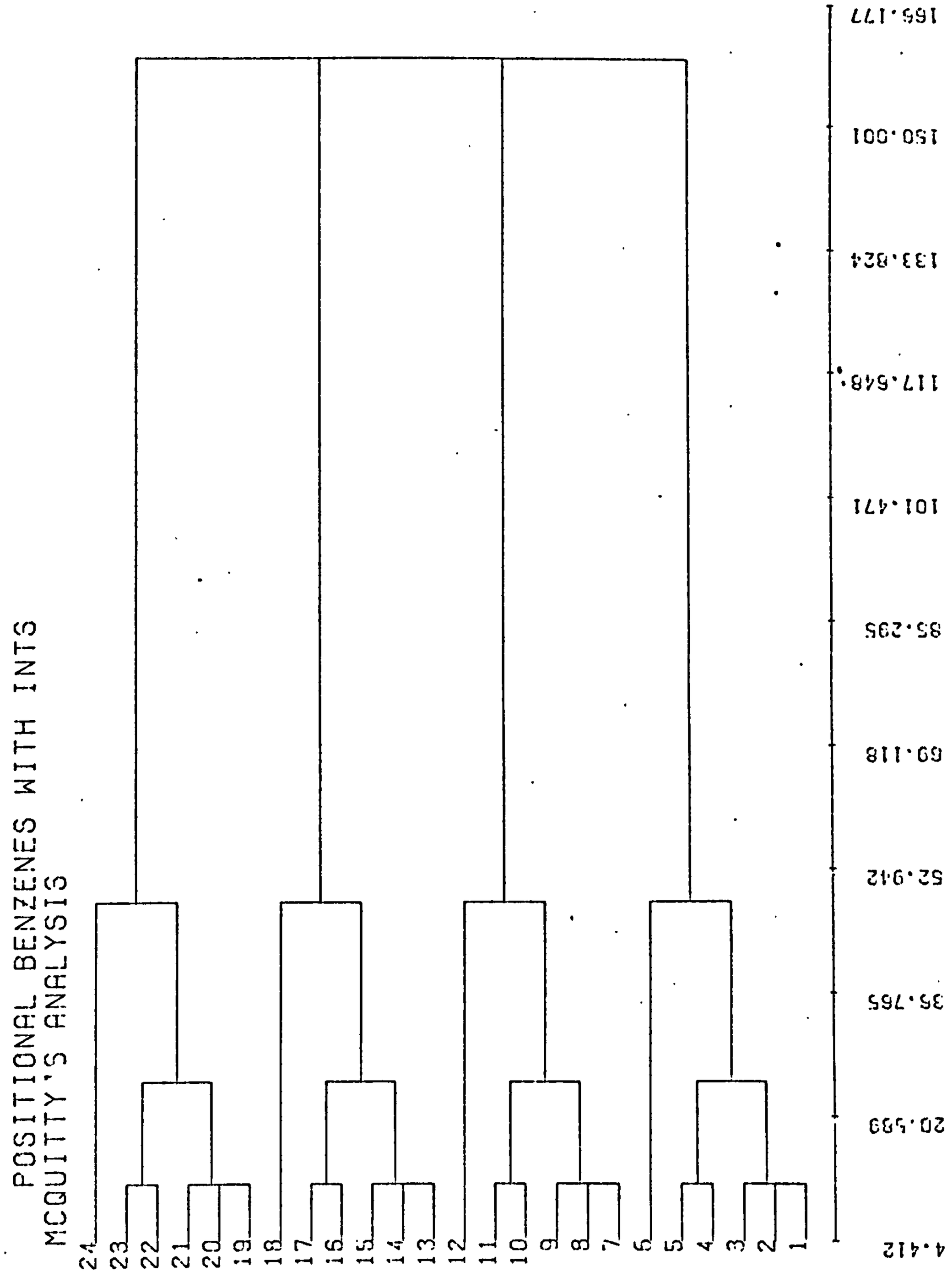

POSITIONAL BENZENES WITH INTS
MCQUITY'S ANALYSIS
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
165.177
150.001
133.824
117.648
101.471
85.295
69.118
52.942
35.765
20.589
4.412

Figure DE

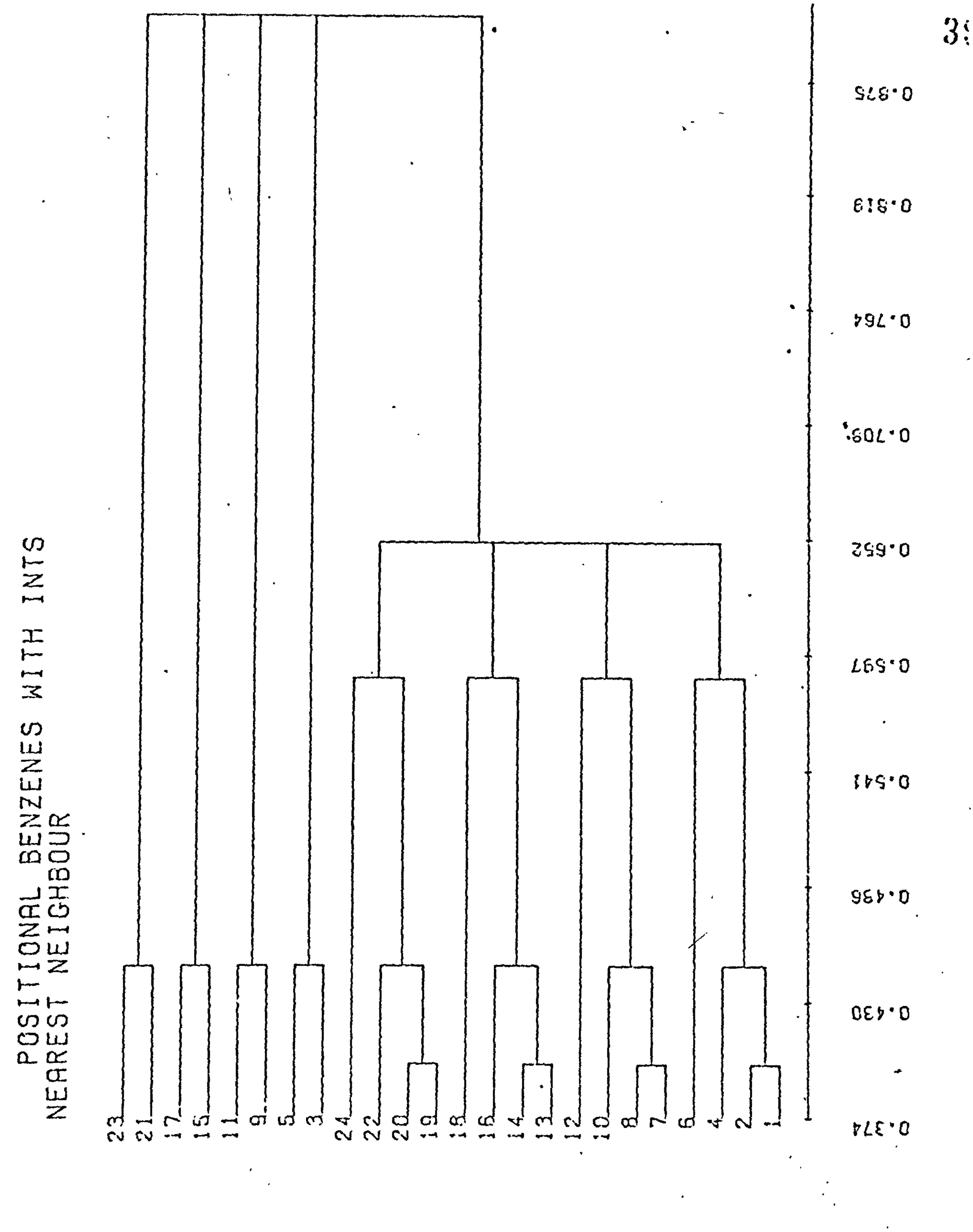

POSITIONAL BENZENES WITH INTS
NEAREST NEIGHBOUR
23
21
17
15
11
9
5
3
24
22
20
19
18
16
14
13
12
10
8
7
6
4
2
1
0.374
0.430
0.436
0.541
0.597
0.652
0.709
0.764
0.819
0.875

Figure DF

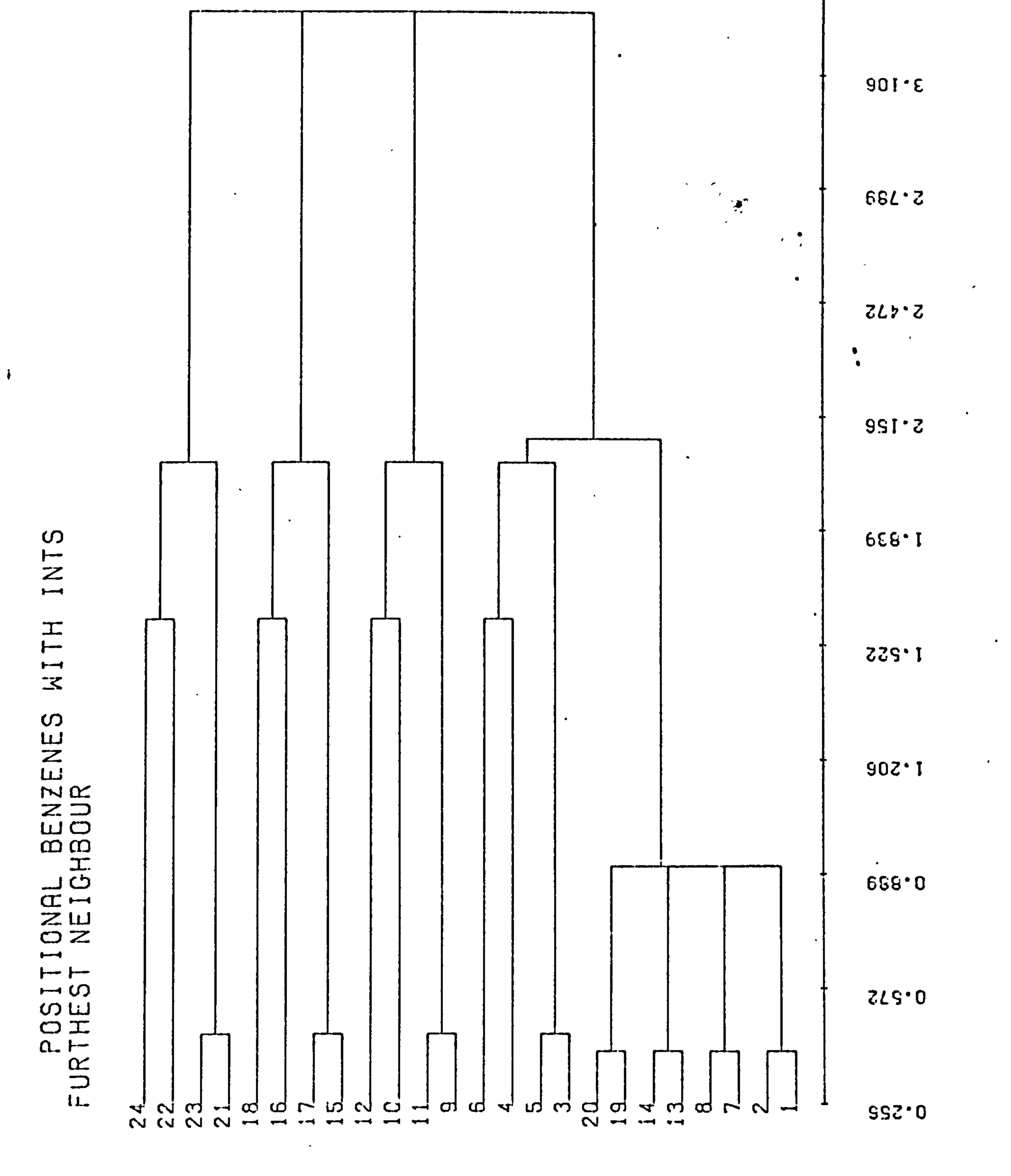

POSITIONAL BENZENE WITH INTS
FURTHEST NEIGHBOUR
24
22
23
21
18
16
17
15
12
10
11
9
6
4
5
3
20
19
14
13
8
7
2
1
3.422
3.106
2.799
2.472
2.156
1.839
1.522
1.206
0.889
0.572
0.255

Figure DG

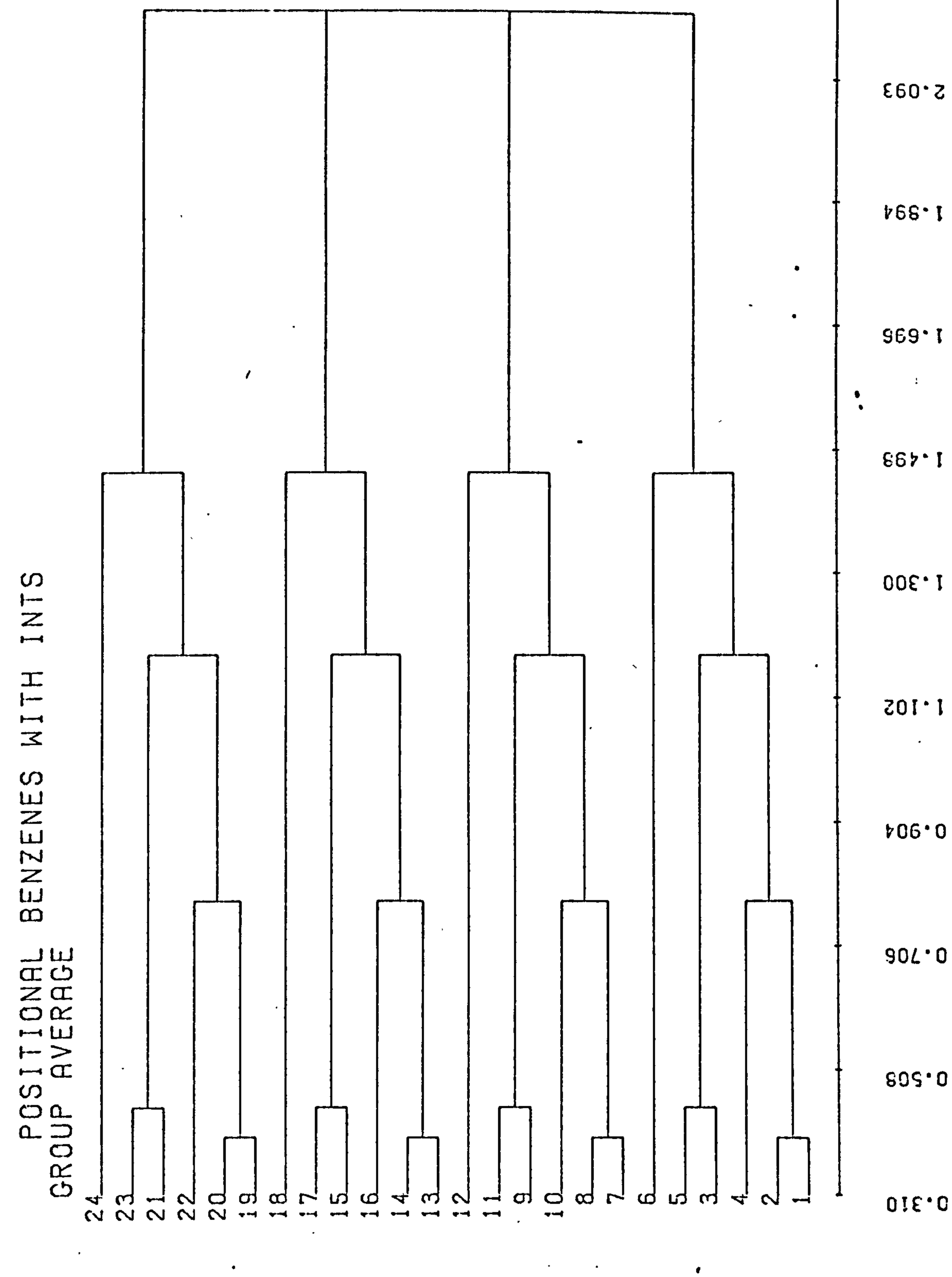

POSITIONAL BENZENES WITH INTS
GROUP AVERAGE
24
23
21
22
20
19
18
17
15
16
14
13
12
11
9
10
8
7
6
5
3
4
2
1
2.291
2.093
1.894
1.695
1.493
1.300
1.102
0.904
0.706
0.508
0.310

Figure DH

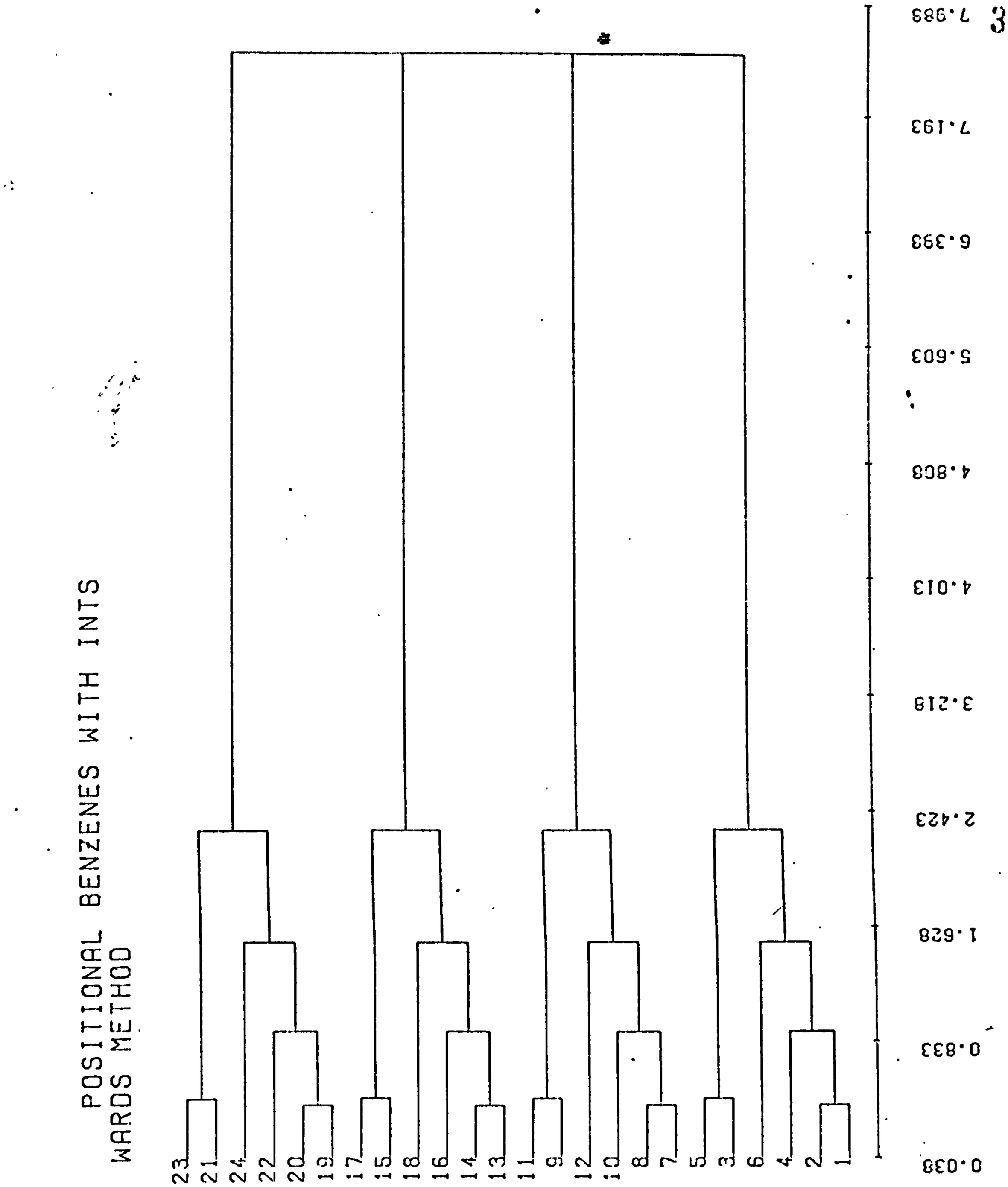

POSITIONAL BENZENES WITH INTS
WARDS METHOD
23
21
24
22
20
19
17
15
18
16
14
13
11
9
12
10
8
7
5
3
6
4
2
1
0.038
0.833
1.628
2.423
3.218
4.013
4.808
5.603
6.398
7.193
7.989

Figure DI

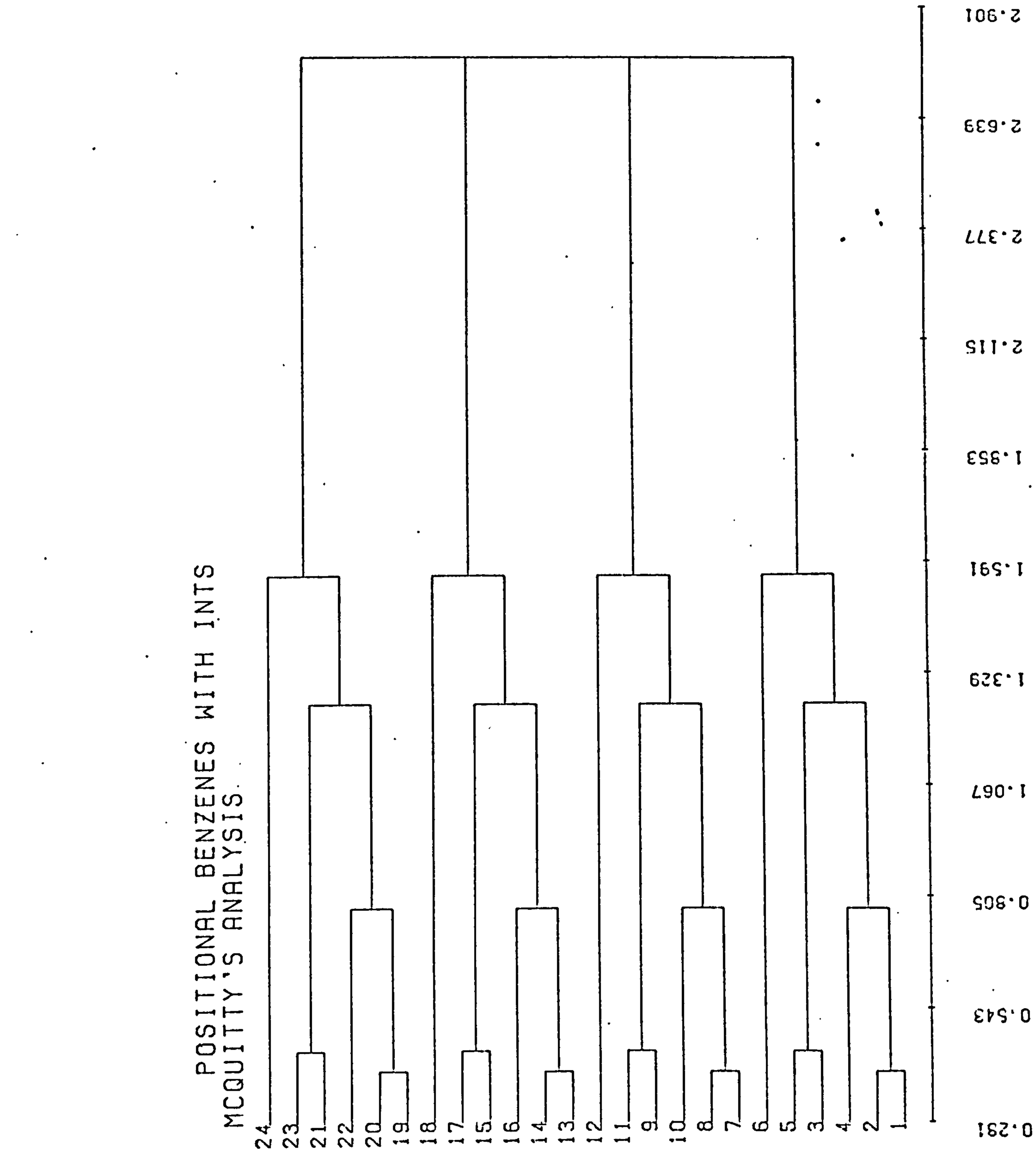

POSITIONAL ISOMERS OF DIMETHYLBENZENES WITH INTS
MCQUITTY'S ANALYSIS.
24
23
21
22
20
19
18
17
15
16
14
13
12
11
9
10
8
7
6
5
3
4
2
1
2.901
2.639
2.377
2.115
1.853
1.591
1.329
1.067
0.805
0.543
0.281

Figure DJ

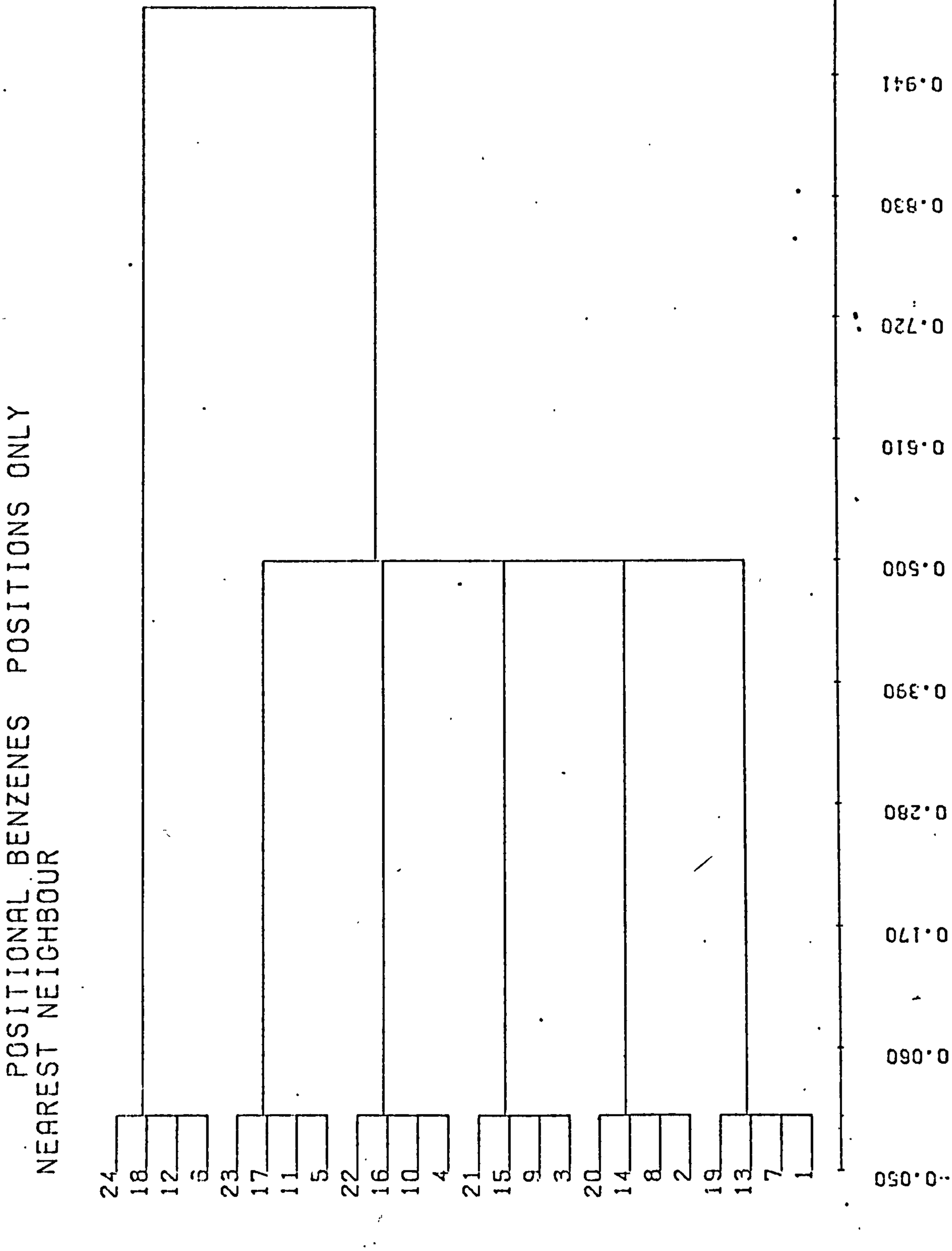

POSITIONAL BENZENES PATTERNS ONLY
NEAREST NEIGHBOUR
24 18 12 6 23 17 11 5 22 16 10 4 21 15 9 3 20 14 8 2 19 13 7 1
1.050
0.941
0.830
0.720
0.610
0.500
0.390
0.280
0.170
0.060
-0.050

Figure EA

EA

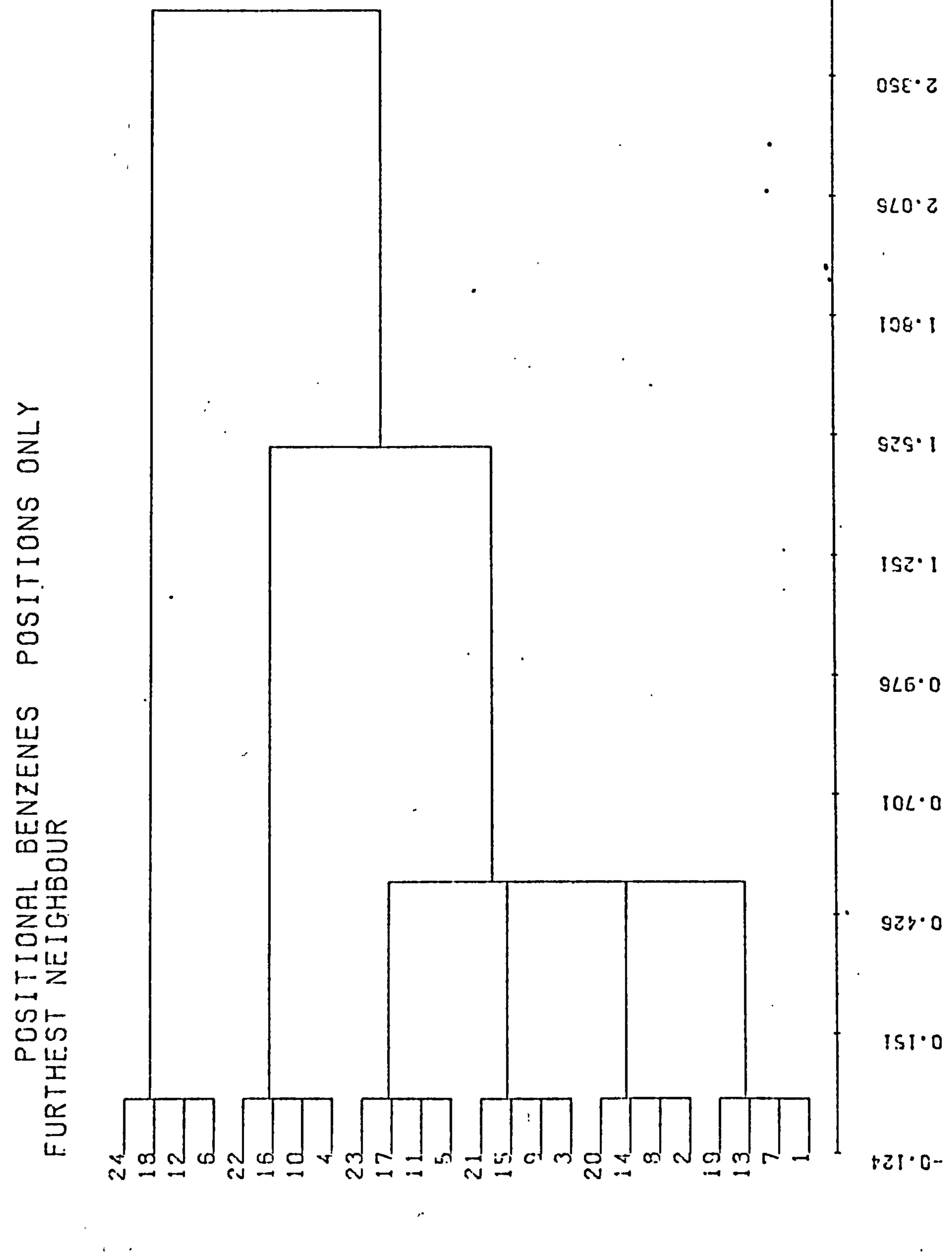

POSITIONS ONLY
POSITIONAL BENZENES
FURTHEST NEIGHBOUR
24 18 12 6 22 16 10 4 23 17 11 5 21 15 9 3 20 14 8 2 19 13 7 1
2.625
2.350
2.075
1.801
1.525
1.251
0.976
0.701
0.426
0.151
-0.124

Figure 3B

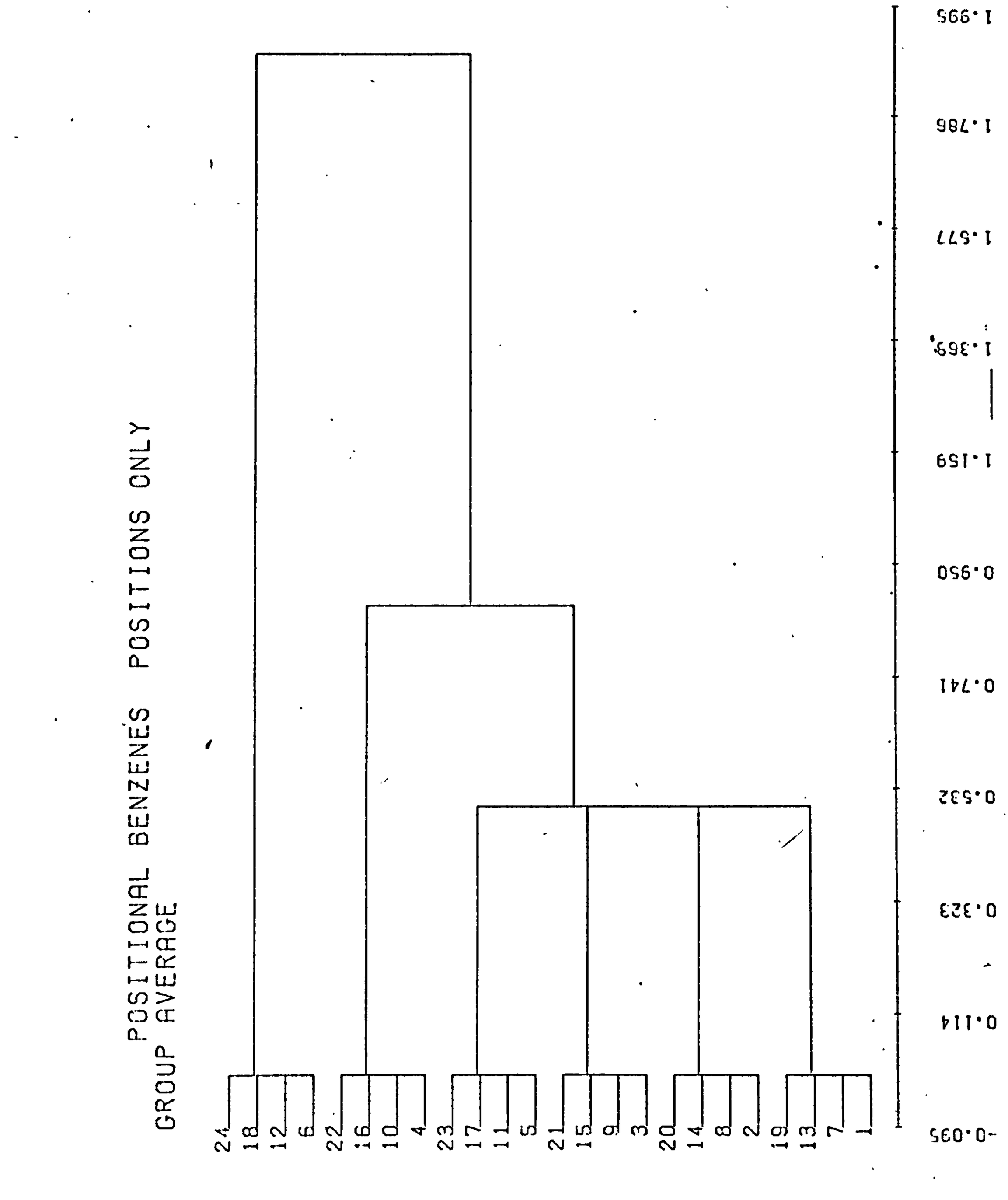

POSITIONAL BENZENES PATTERNS ONLY
GROUP AVERAGE
24
18
12
6
22
16
10
4
23
17
11
5
21
15
9
3
20
14
8
2
19
13
7
1
1.995
1.786
1.577
1.369
1.159
0.950
0.741
0.532
0.323
0.114
-0.095

Figure EC

EC

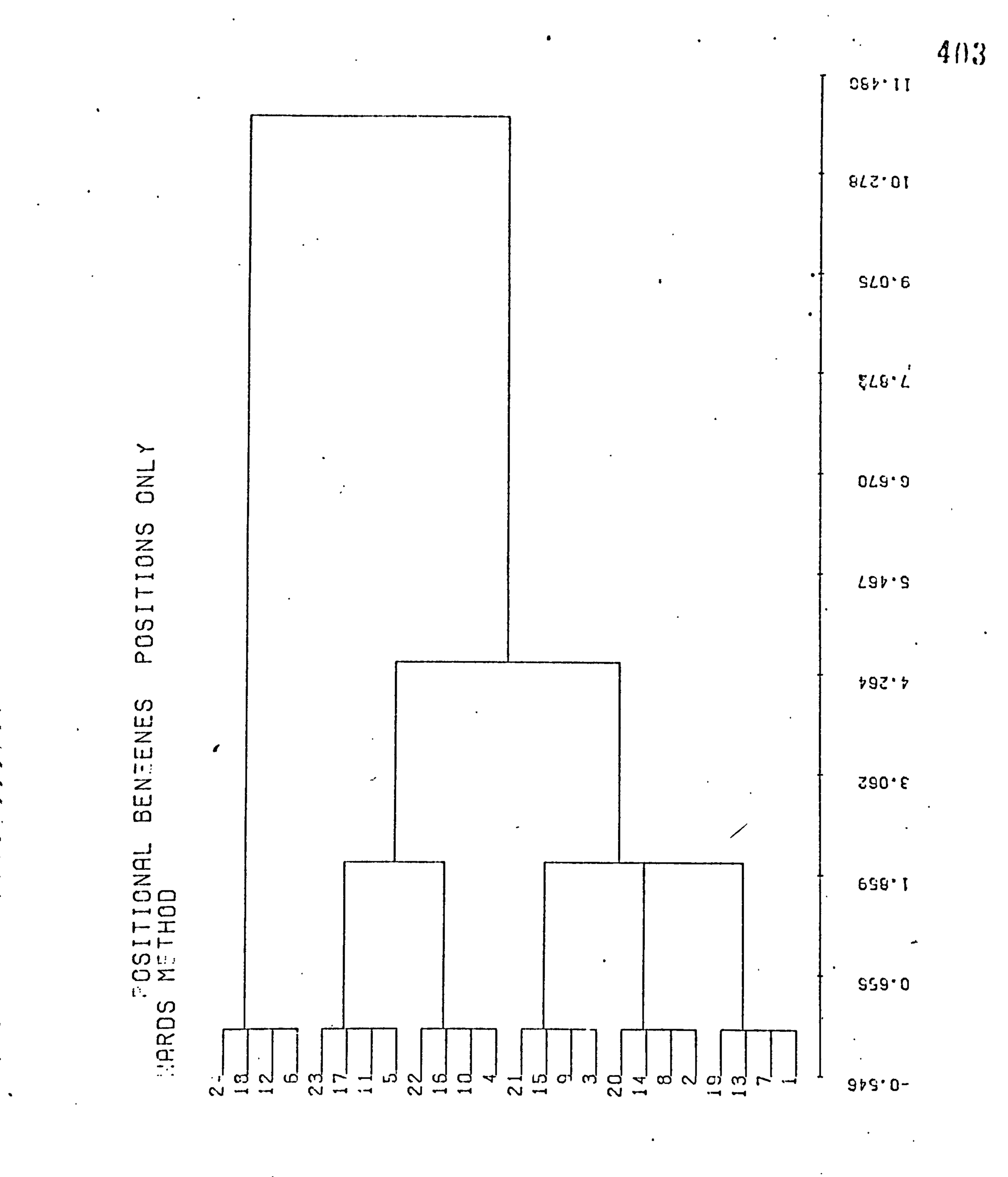

POSITIONAL BENZENES POSITIONS ONLY
WARDS METHOD
11.480
10.278
9.075
7.873
6.670
5.467
4.264
3.062
1.859
0.655
-0.546
2
13
12
6
23
17
11
5
22
16
10
4
21
15
9
3
20
14
8
2
19
13
7
1

Figure ED

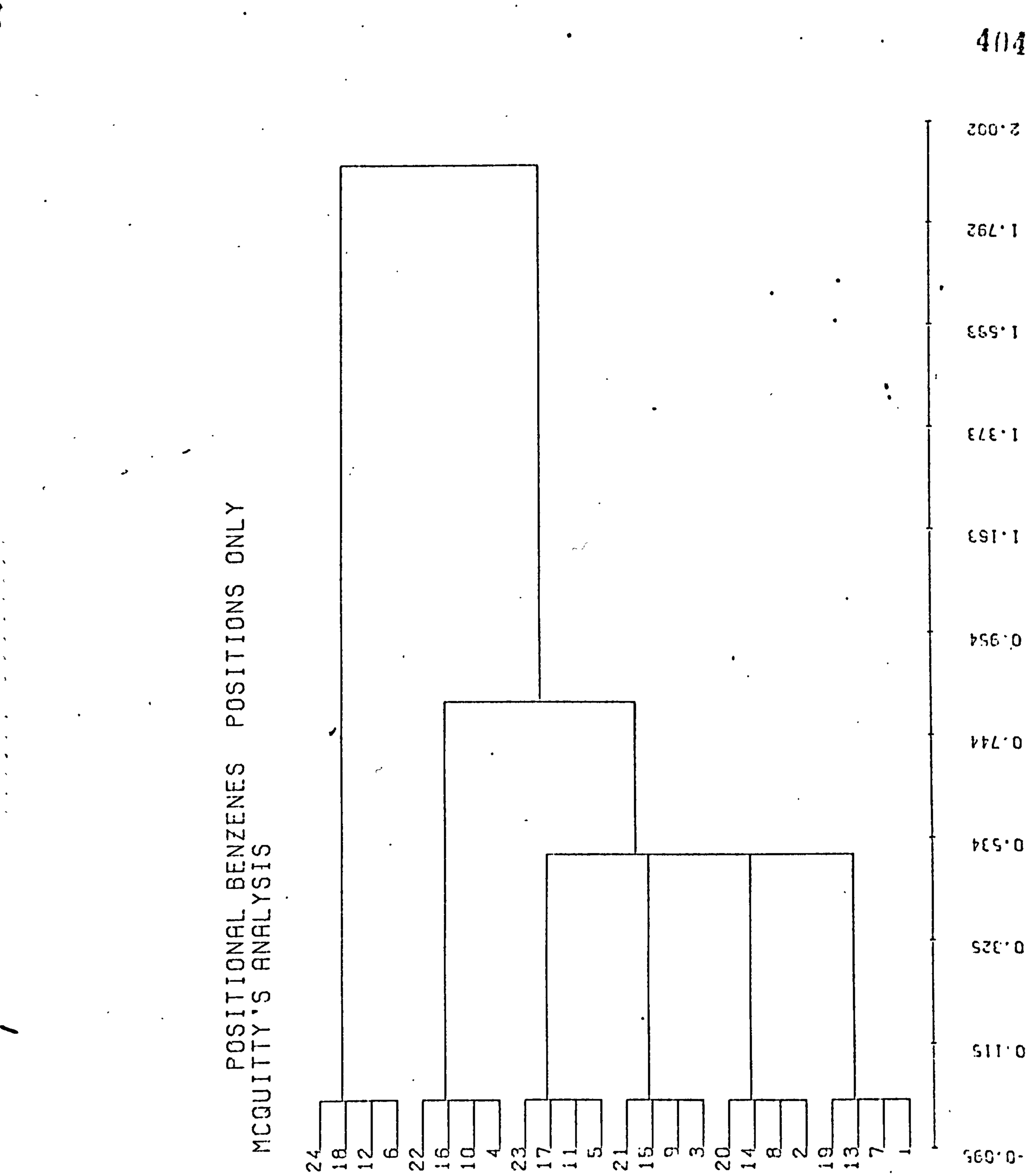

POSITIONAL BENZENES POSITIONS ONLY
MCQUITY'S ANALYSIS
24 18 12 6 22 16 10 4 23 17 11 5 21 15 9 3 20 14 8 2 19 13 7 1
2.002
1.792
1.583
1.373
1.153
0.954
0.744
0.534
0.325
0.115
0.000

Figure EE

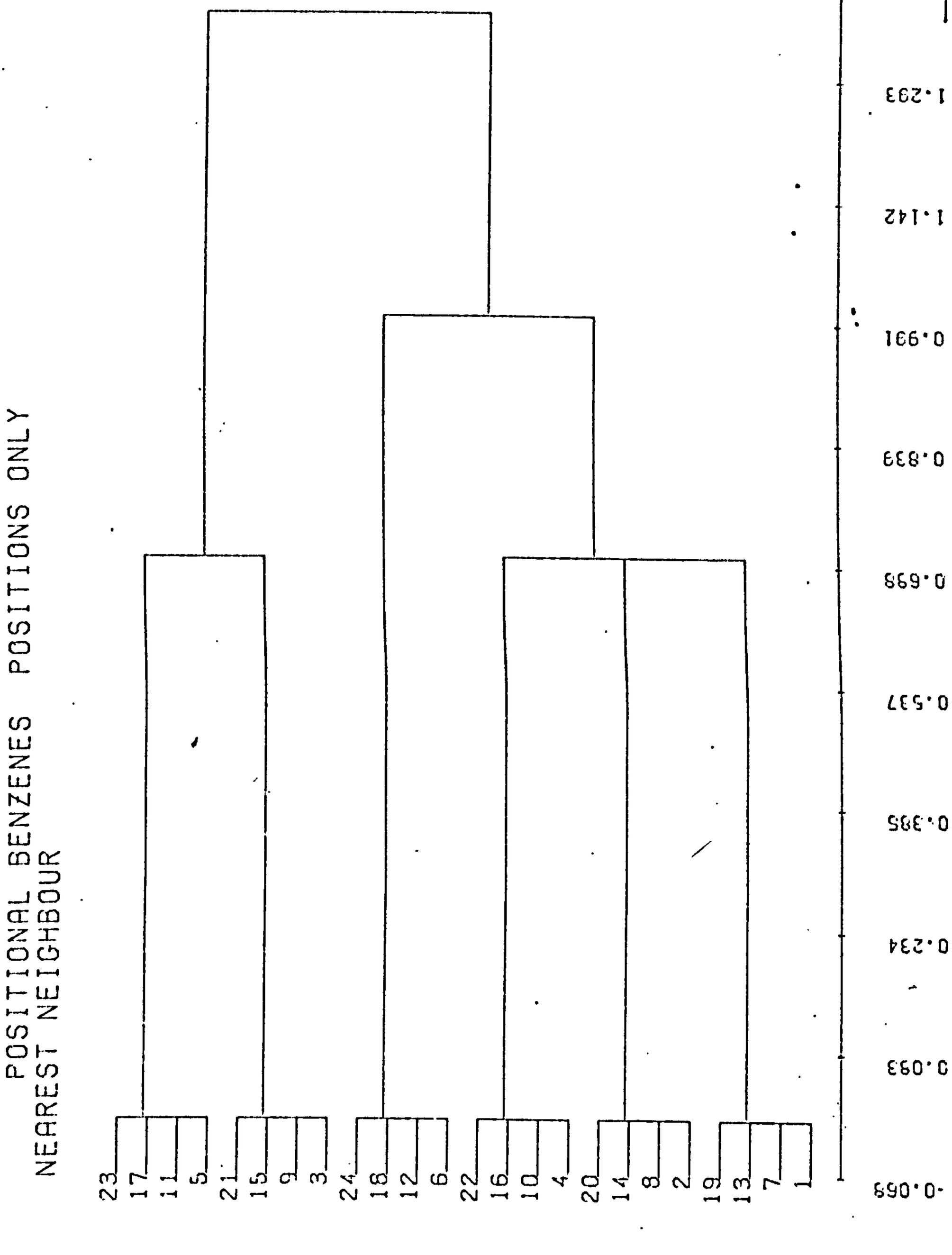

POSITIONAL BENZENES
NEAREST NEIGHBOUR
23 17 11 5 21 15 9 3 24 18 12 6 22 16 10 4 20 14 8 2 19 13 7 1
1.444
1.293
1.142
0.991
0.839
0.688
0.537
0.385
0.234
0.083
-0.069

Figure EF

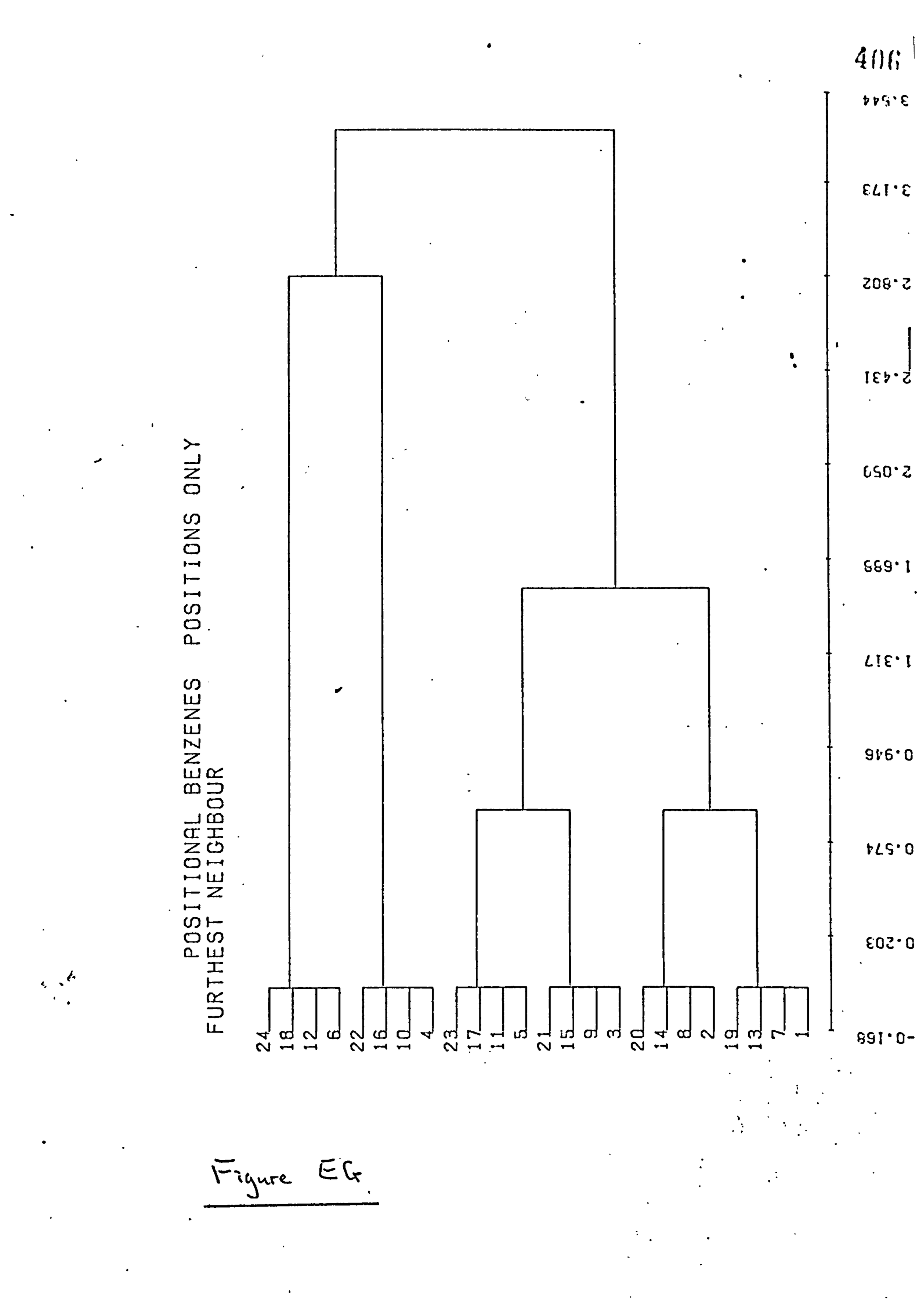

POSITIONAL BENZENES POSITIONS ONLY
FURTHEST NEIGHBOUR
24
18
12
6
22
16
10
4
23
17
11
5
21
15
9
3
20
14
8
2
19
13
7
1
3.544
3.173
2.802
2.431
2.059
1.689
1.317
0.946
0.574
0.203
-0.168

Figure EG

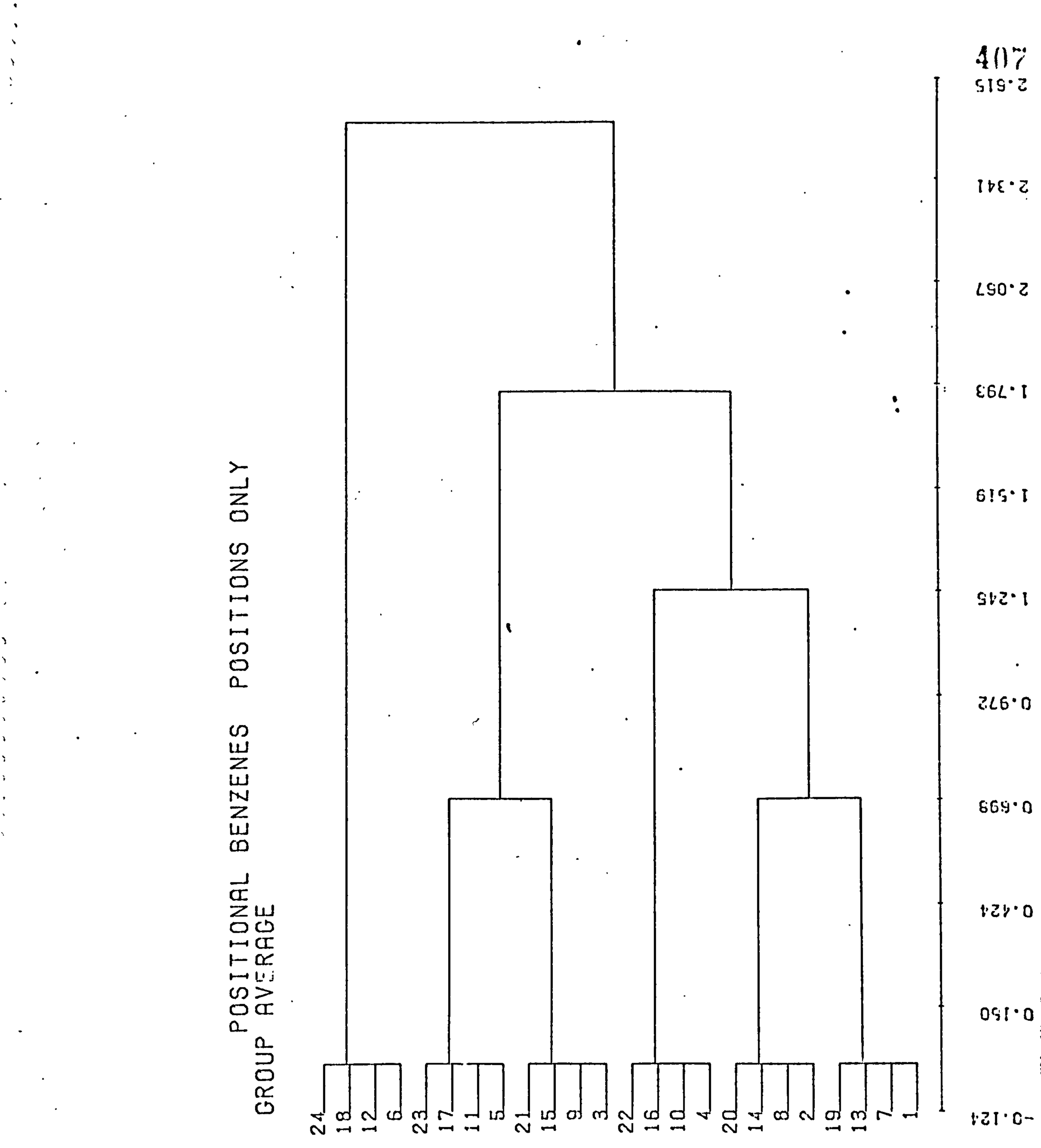

POSITIONAL BENZENES
GROUP AVERAGE
POSITIONS ONLY
24
18
12
6
23
17
11
5
21
15
9
3
22
16
10
4
20
14
8
2
19
13
7
1
2.515
2.341
2.057
1.793
1.519
1.245
0.972
0.699
0.424
0.150
-0.124

Figure EH

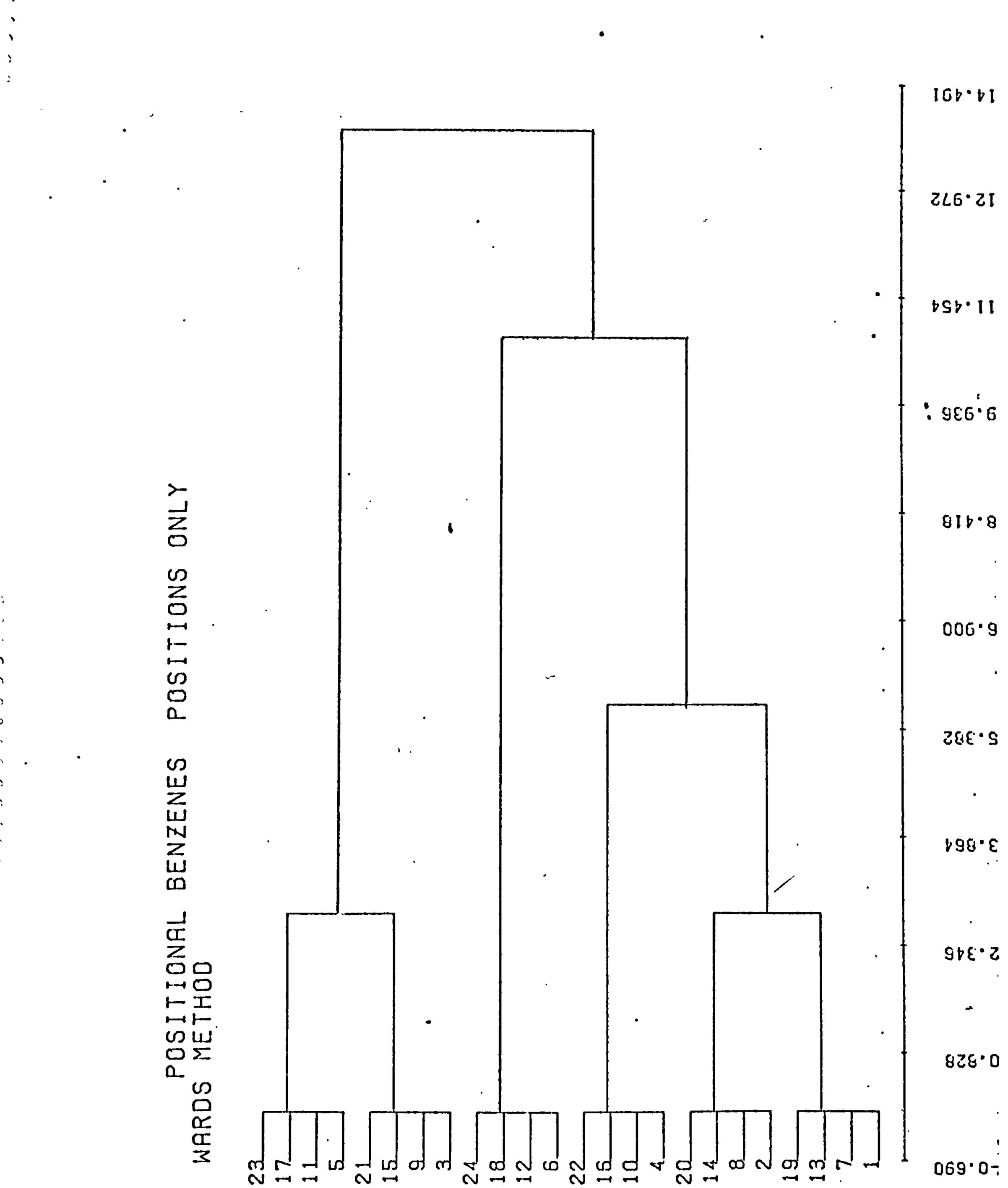

POSITIONAL BENZENES POSITIONS ONLY
WARDS METHOD
23
17
11
5
21
15
9
3
24
18
12
6
22
16
10
4
20
14
8
2
19
13
7
1
0.690
0.828
2.345
3.864
5.382
6.900
8.418
9.936
11.454
12.972
14.491

Figure EI

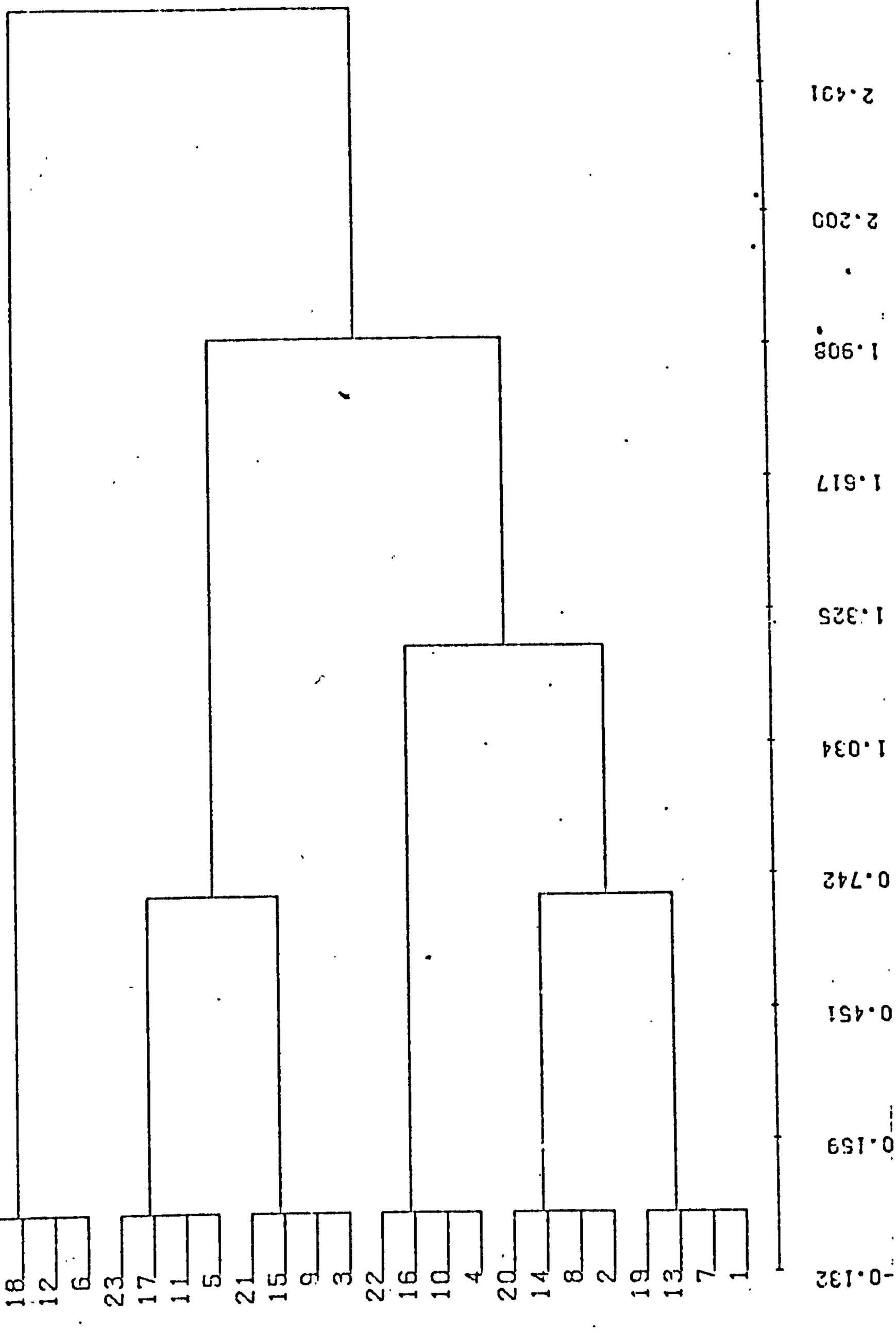

POSITIONAL BENZENES POSITIONS ONLY
MCQUITTY'S ANALYSIS
24
18
12
6
23
17
11
5
21
15
9
3
22
16
10
4
20
14
8
2
19
13
7
1
2.793
2.431
2.200
1.908
1.617
1.325
1.034
0.742
0.451
0.159
-0.132

Figure EJ

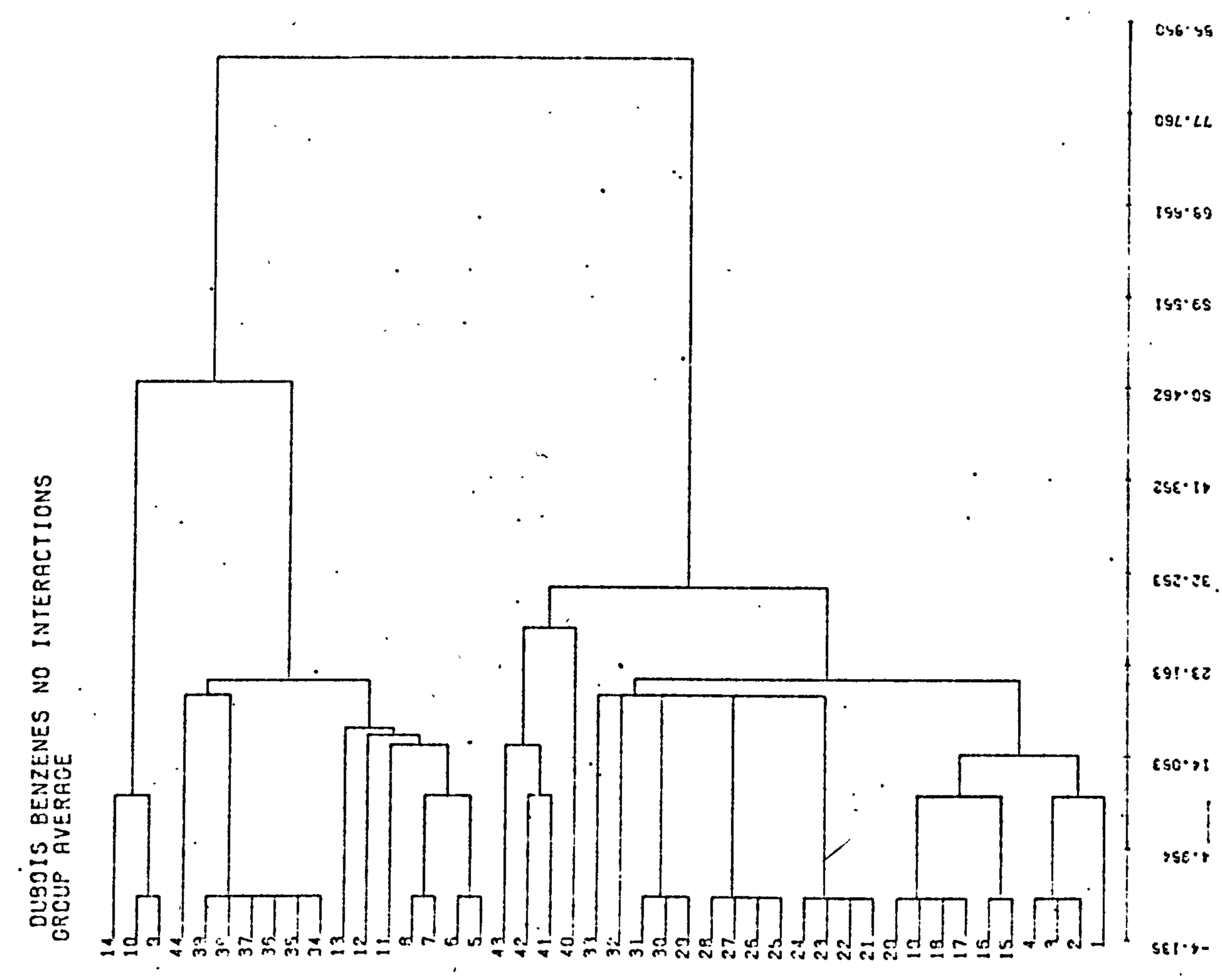


Figure FA

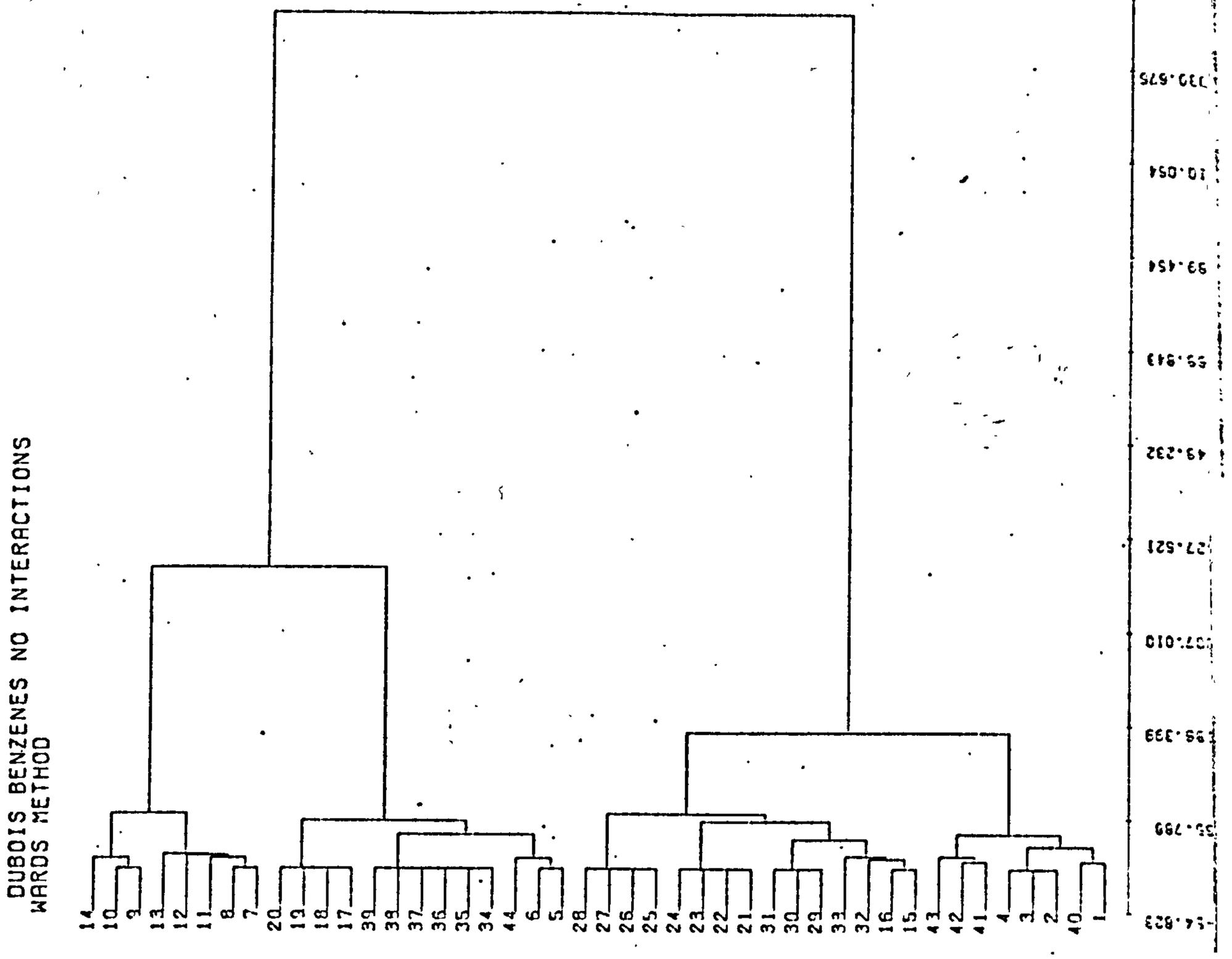


Figure 7B

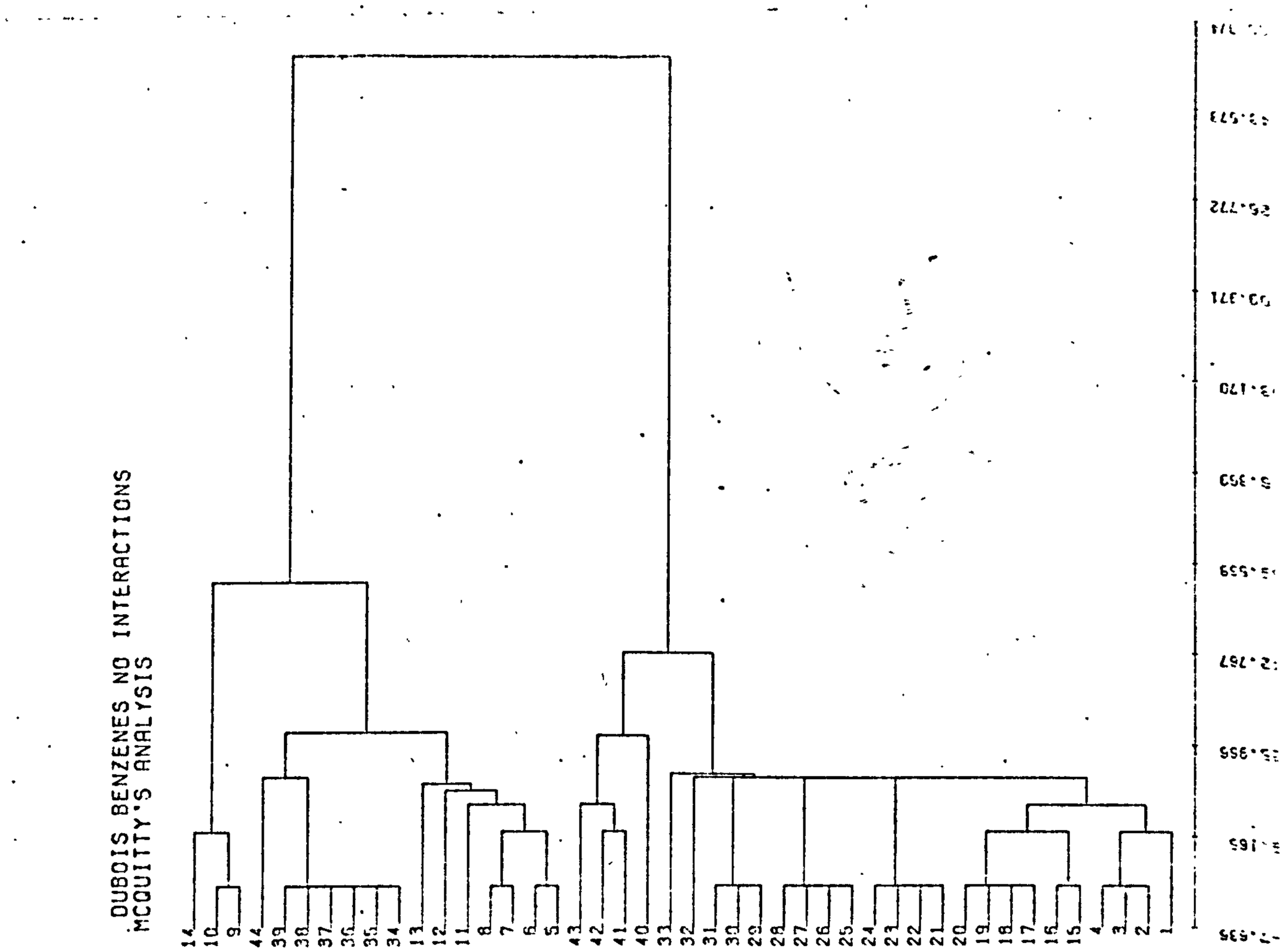


Figure FC

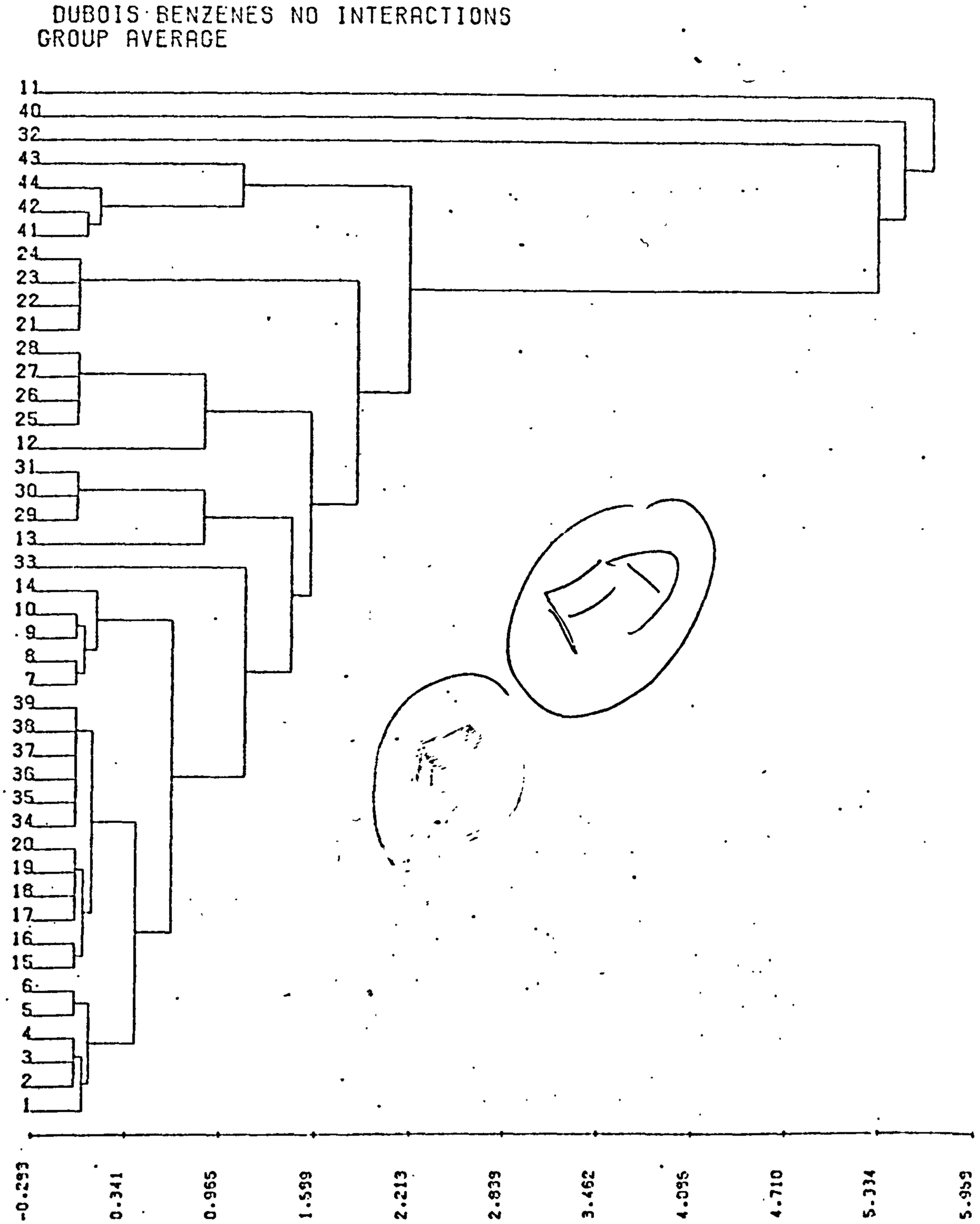


Figure 11D

DUBOIS BENZENES NO INTERACTIONS
WARDS METHOD

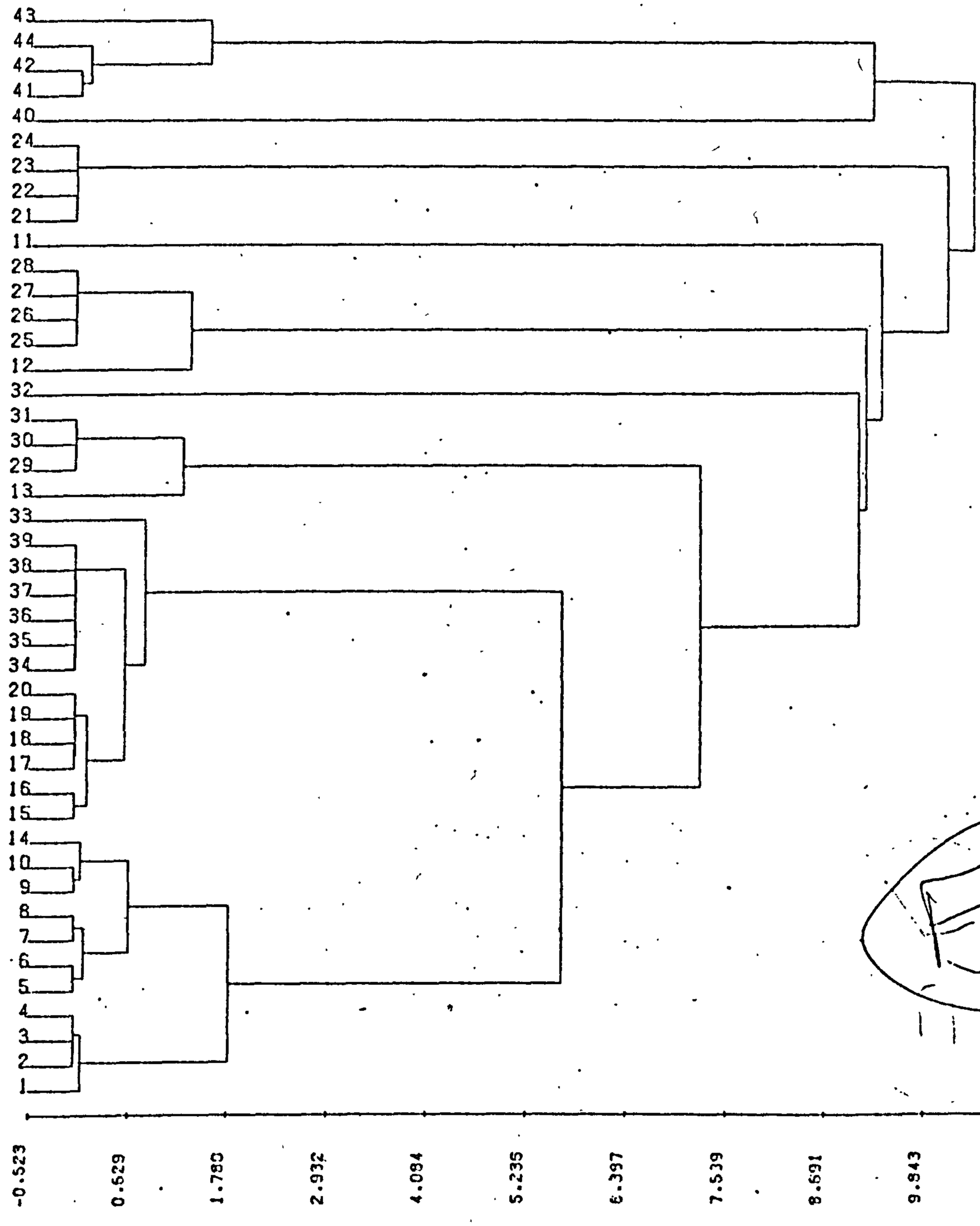


Figure FE

CALIFORNIA COMPUTER PRODUCTS, INC. ANAHEIM, CALIFORNIA CHART NO

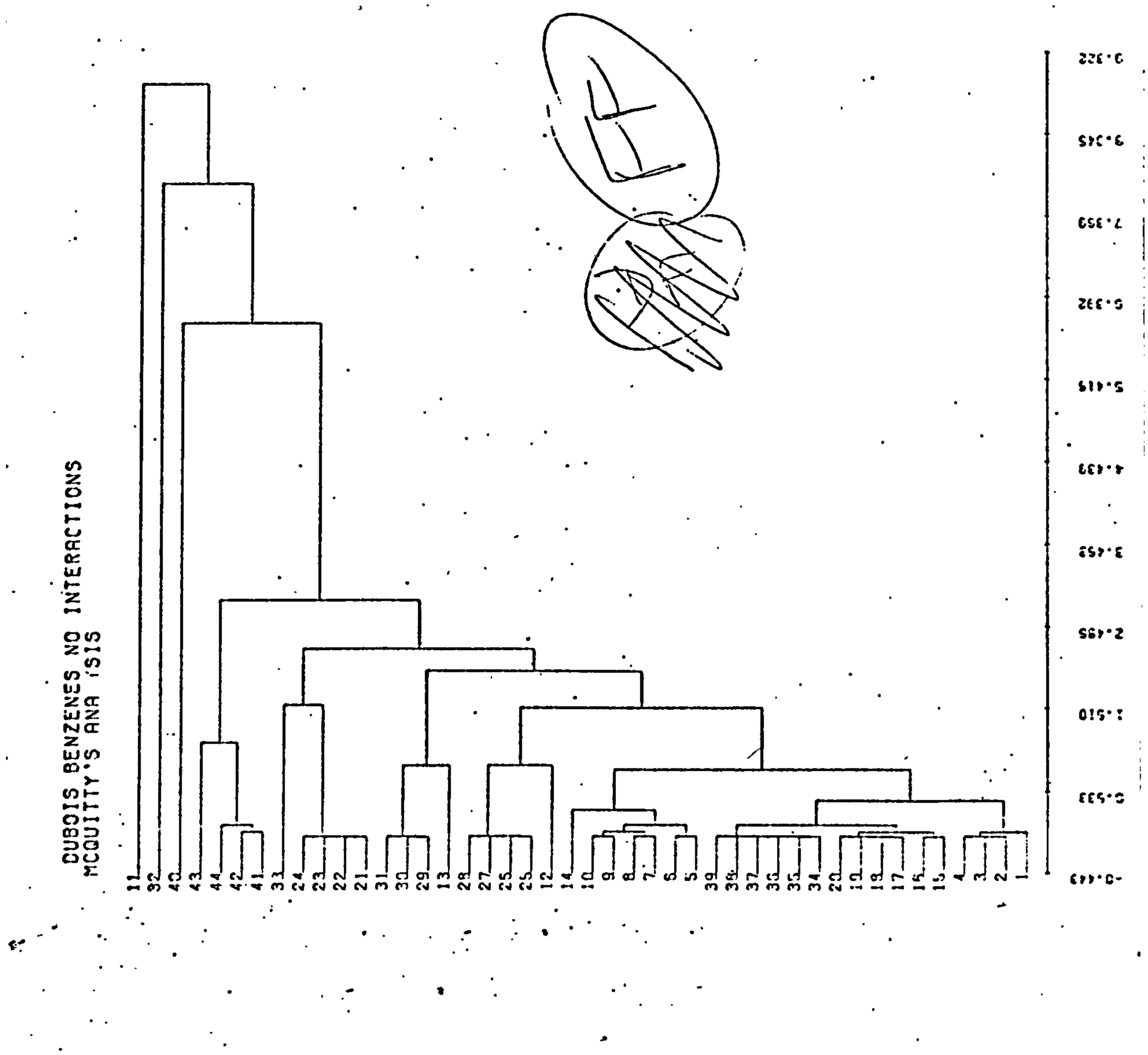

CUBOIS BENZENES NO INTERACTIONS
MCQUITTY'S ANA ISIS
11
32
40
43
44
42
41
33
24
23
22
21
31
30
29
13
28
27
26
25
12
14
10
9
8
7
6
5
39
38
37
36
35
34
20
19
18
17
16
15
4
3
2
1
0.322
3.245
7.359
5.332
5.415
4.439
3.452
2.495
1.510
0.000

Figure FF

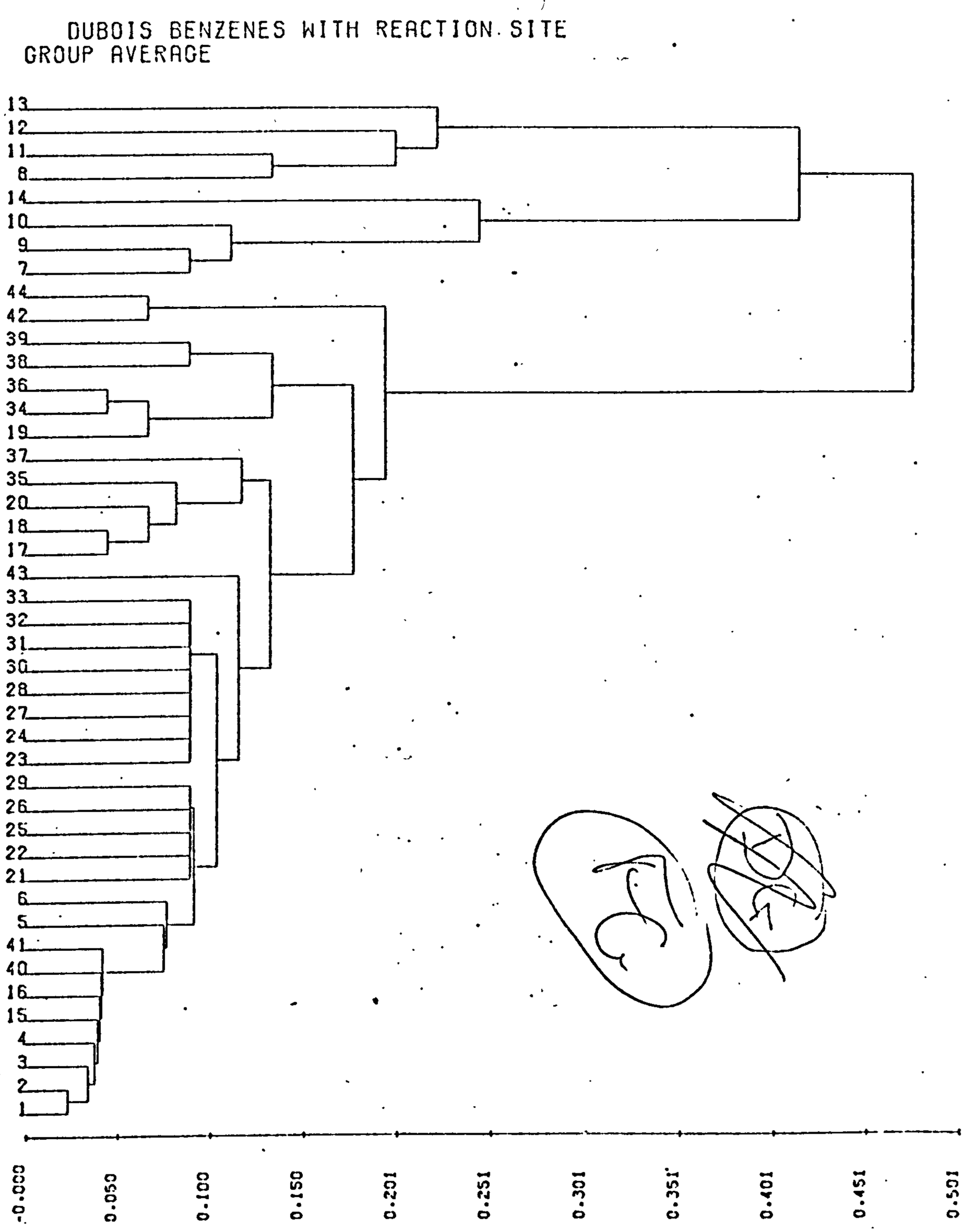
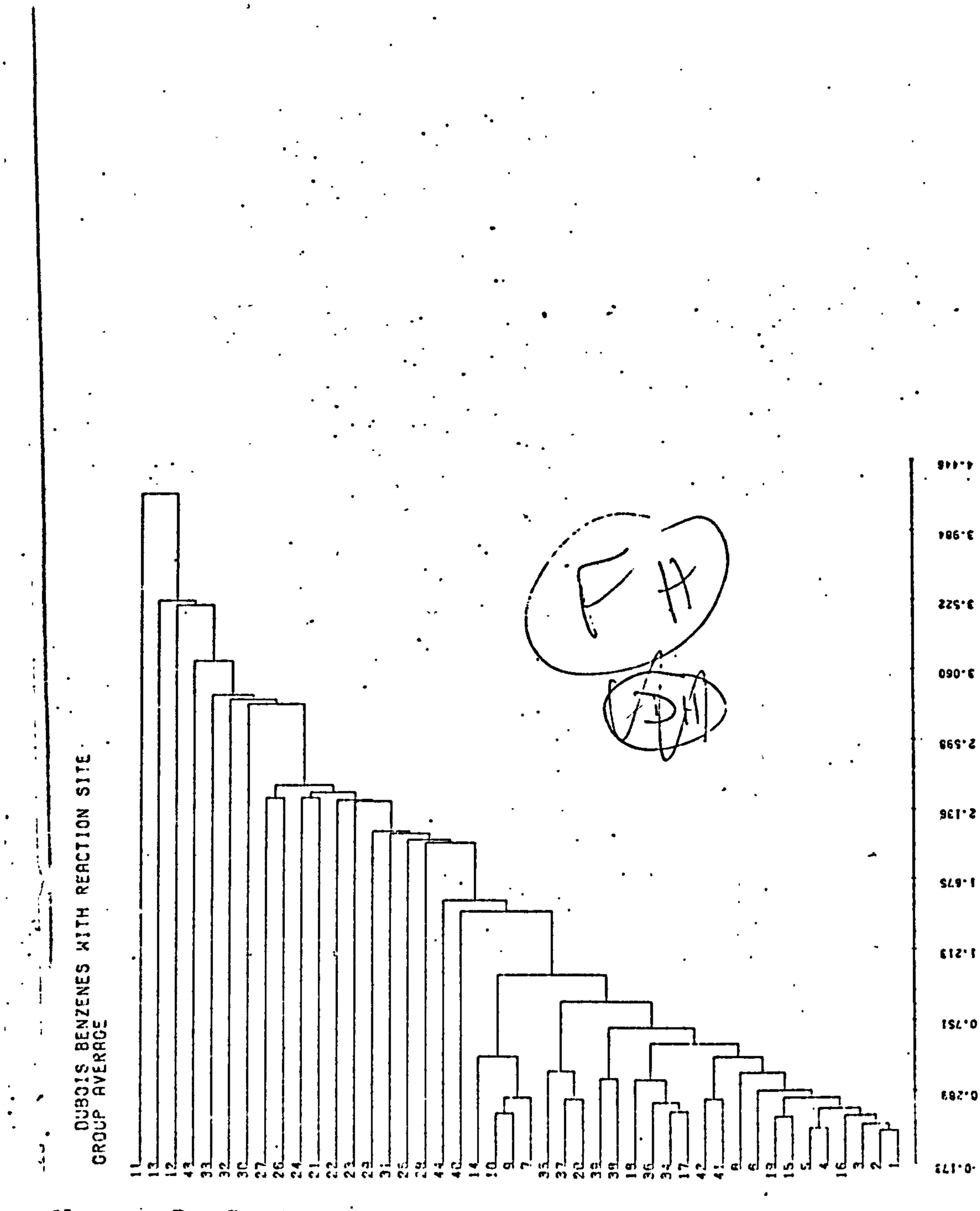

DUBOIS BENZENES WITH REACTION SITE
GROUP AVERAGE
13
12
11
8
14
10
9
7
44
42
39
38
36
34
19
37
35
20
18
17
43
33
32
31
30
28
27
24
23
29
26
25
22
21
6
5
41
40
16
15
4
3
2
1
-0.000
0.050
0.100
0.150
0.201
0.251
0.301
0.351
0.401
0.451
0.501
CALIFORNIA COMPUTER PRODUCTS, INC.
ANAHEIM, CALIFORNIA
CHART NO. 500

Figure FG



FH

FH

Figure FH

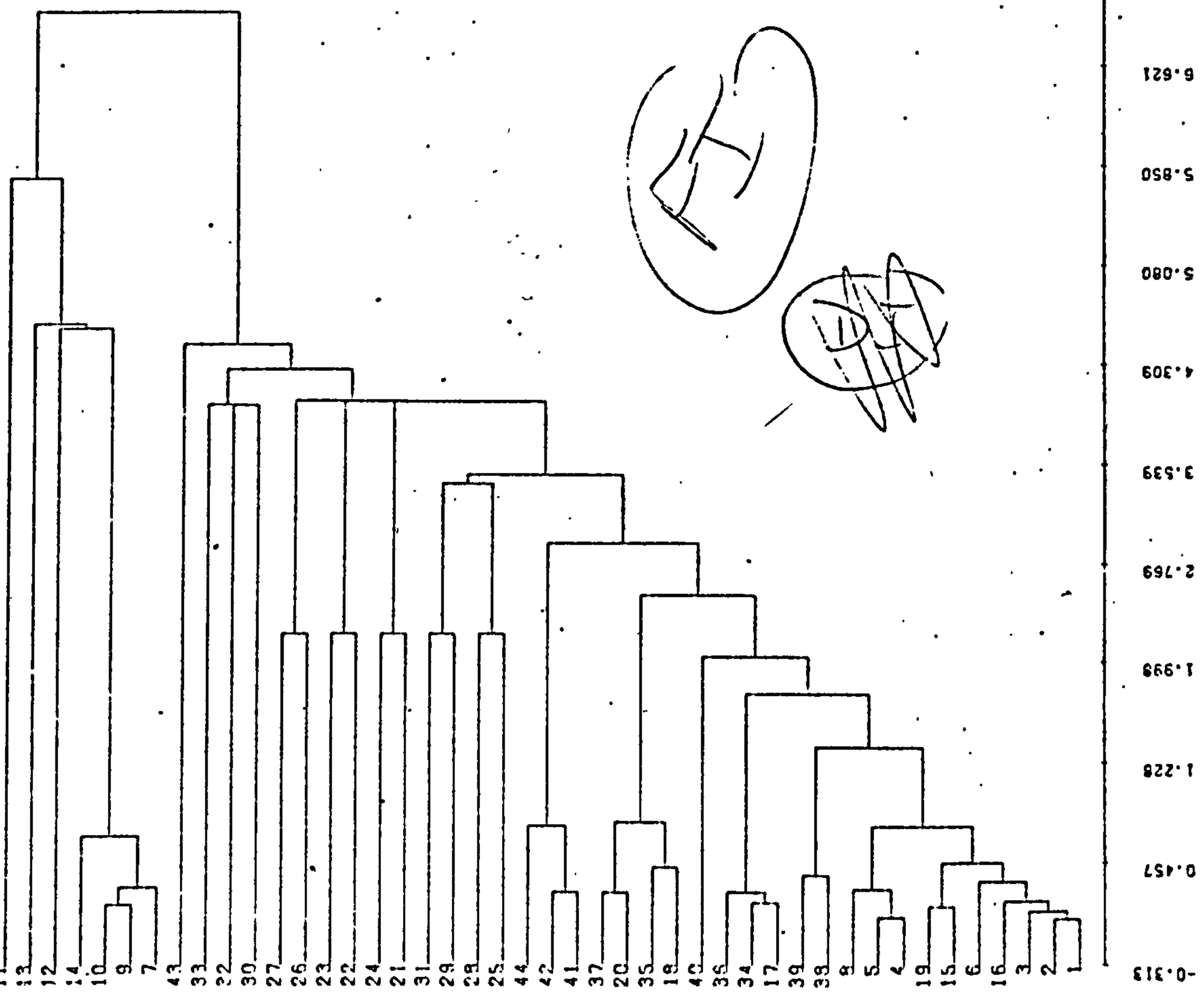

DUBOIS BENZENES WITH REACTION SITE
WARDS METHOD
11 13 12 14 10 9 7 43 33 32 30 27 26 23 22 24 21 31 29 28 25 44 42 41 37 20 35 18 40 36 34 17 39 38 8 5 4 19 15 6 16 3 2 1
7.391
6.621
5.850
5.080
4.309
3.539
2.769
1.998
1.228
0.457
-0.313

Figure FI

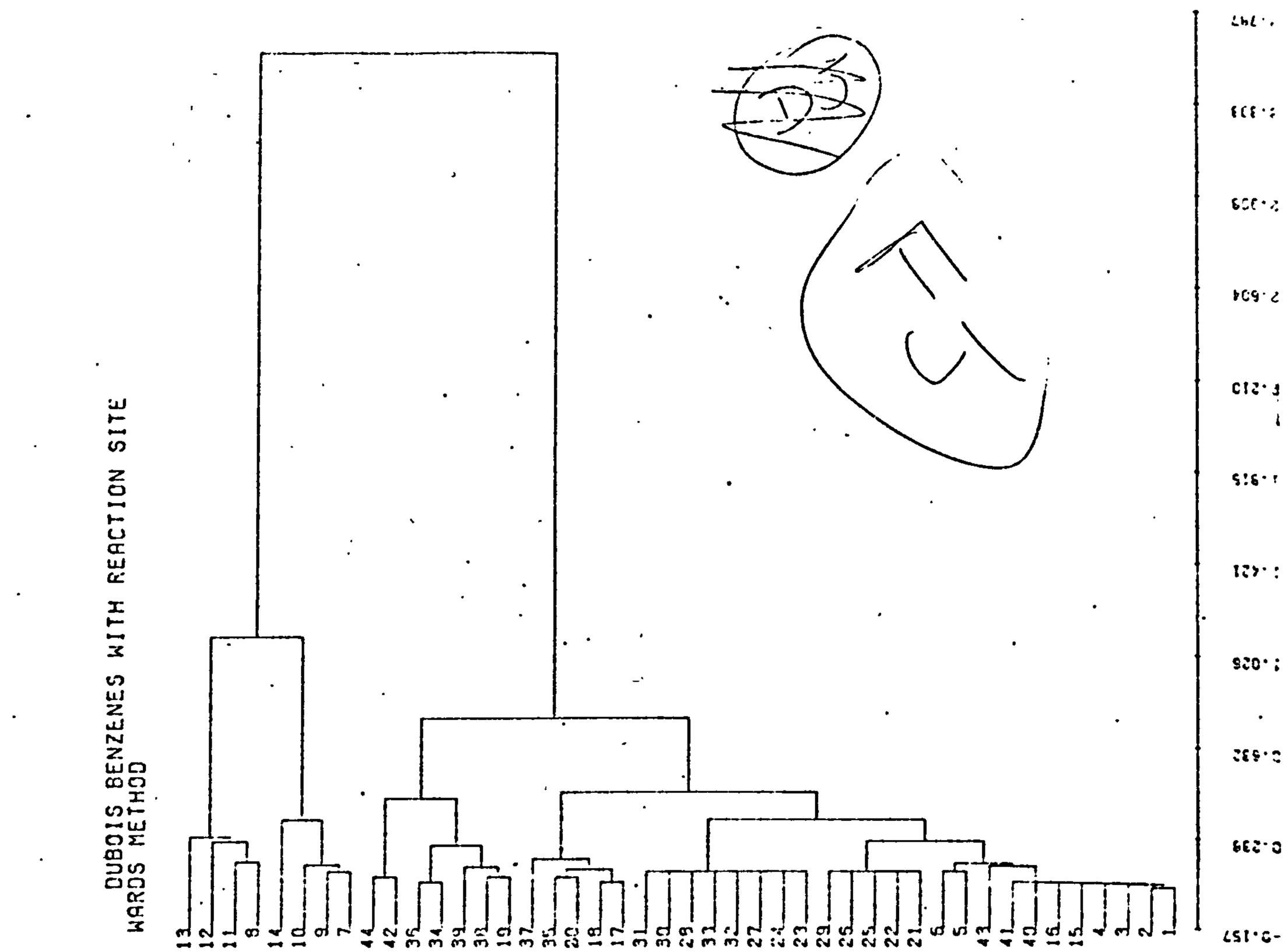


Figure F.1

CALCOMP OUTPUT FOR :LLIGWA.CLUSTX FROM FILE PRODUCED ON 7FEB77 AT 07.23.05

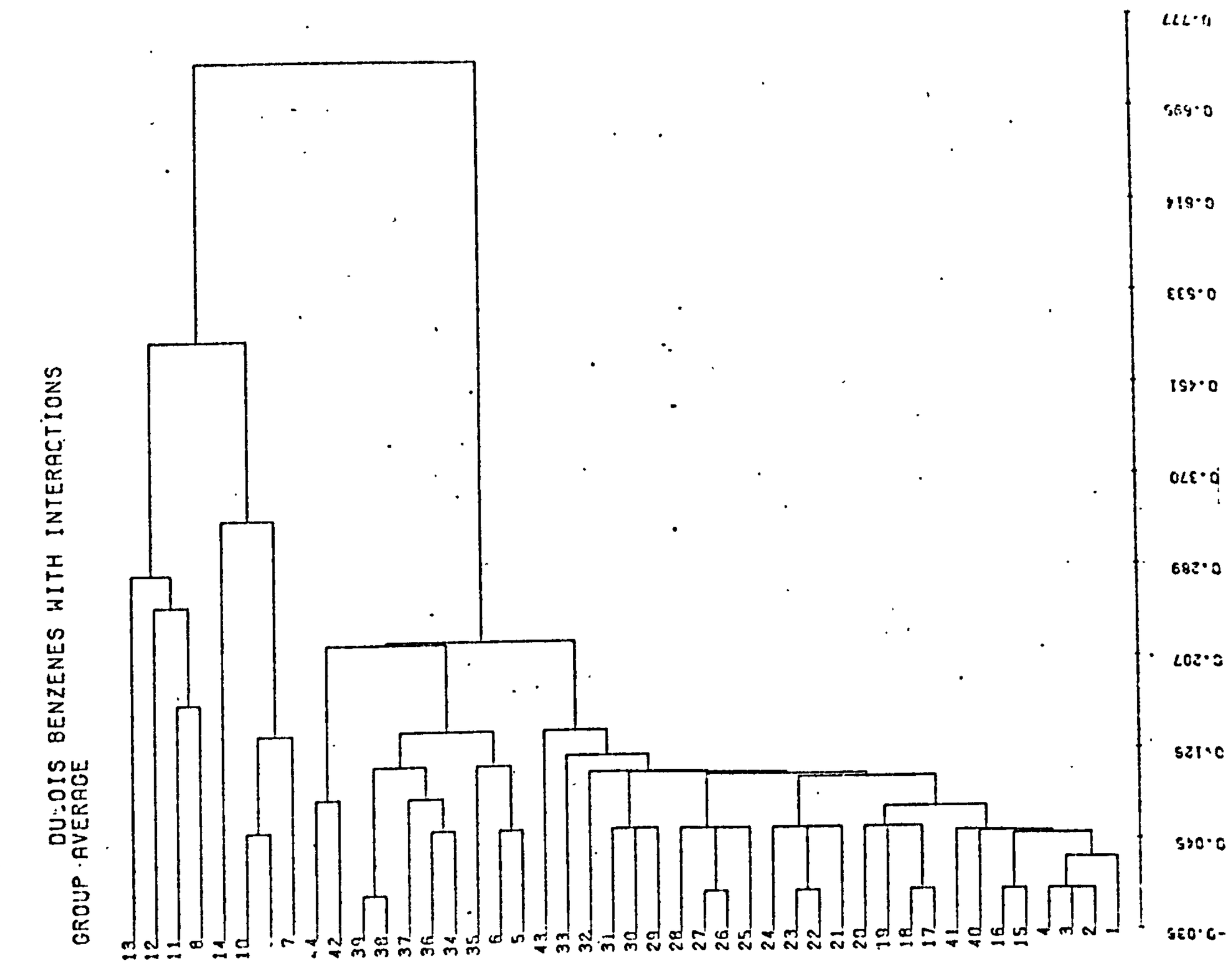


Figure FK

CALCOMP OUTPUT.FOR :LILGWA:CLUSTESTC 5FEB77 AT 09.59.05 FROM FILE PRODUCED ON

FK

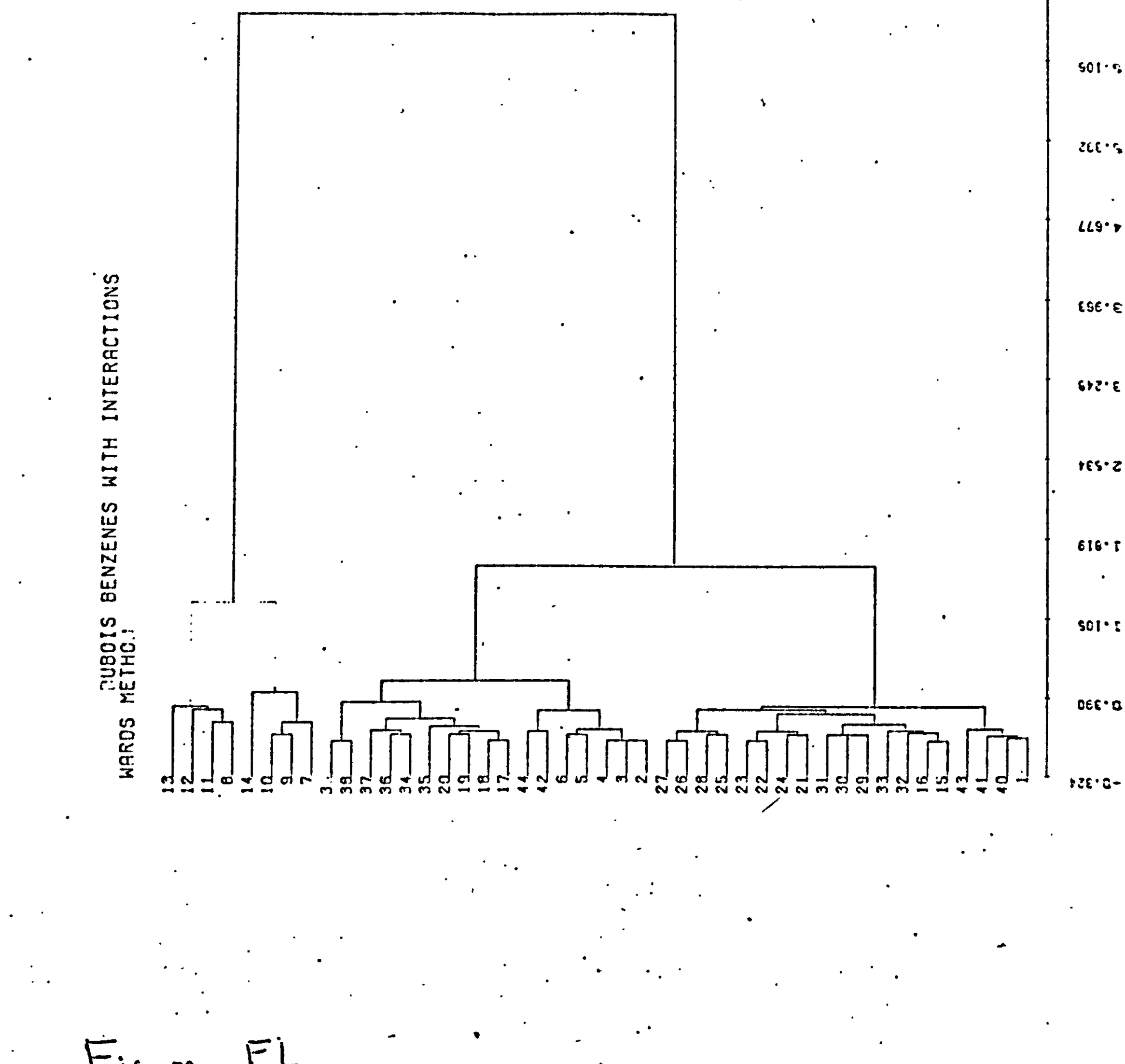

DUBOIS BENZENES WITH INTERACTIONS
WARDS METHOD
13 12 11 8 14 10 9 7 3 39 37 36 34 35 20 19 18 17 44 42 6 5 4 3 2 27 26 28 25 23 22 24 21 31 30 29 33 32 16 15 43 41 40 1
5.821
5.105
5.302
4.677
3.953
3.249
2.534
1.819
1.105
0.390
0.000
CHLCOMP OUTPUT FOR :LLIGWA.CLUSTESTC
FROM FILE PRODUCED ON 5FEB77 AT 11.20.31

Figure FL

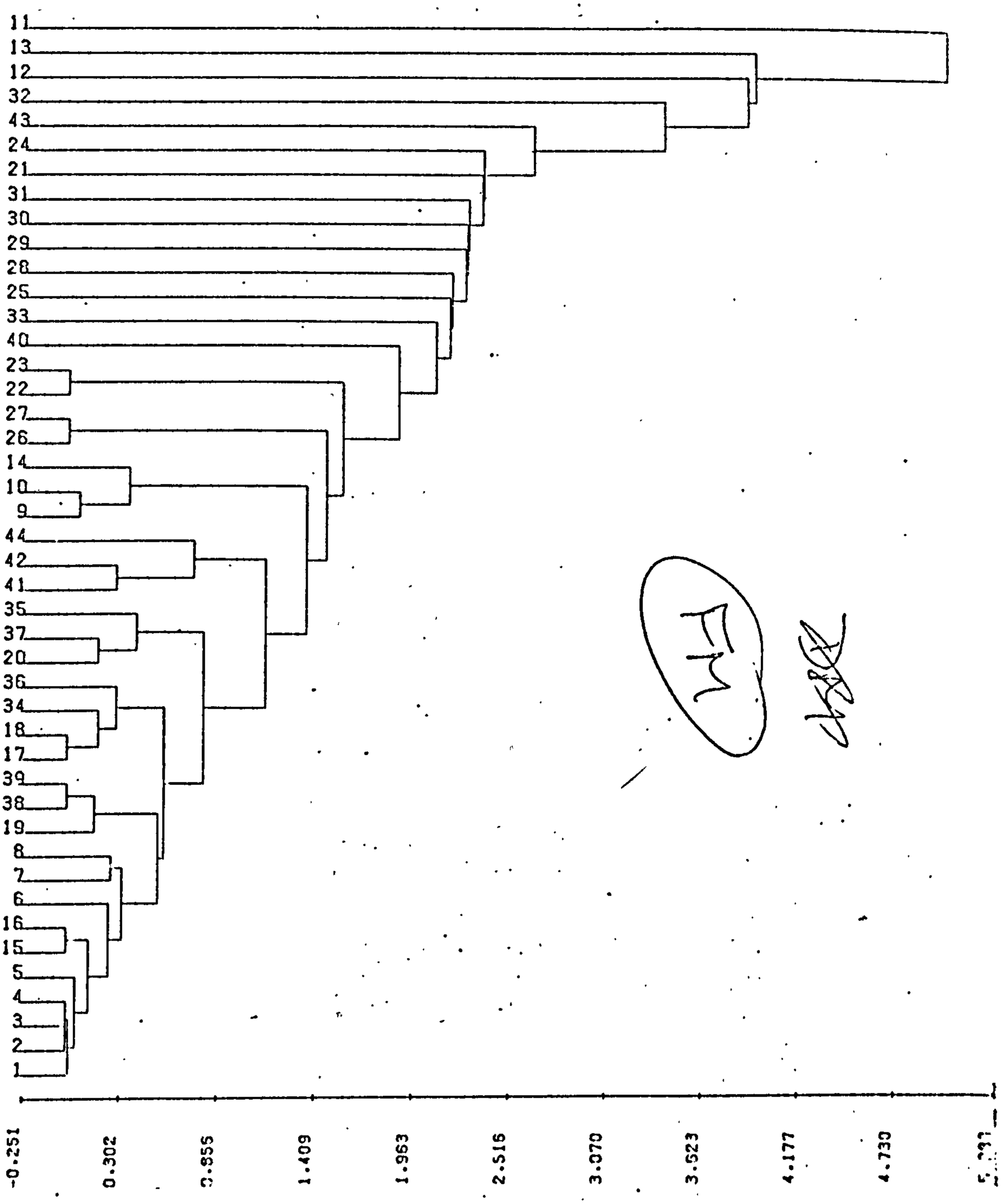


Figure FM

CALIFORNIA COMPUTER PRODUCTS, INC. ANAHEIM, CALIFORNIA CHART NO. 500

DUBOIS BENZENES WITH INTERACTIONS
WARDS METHOD

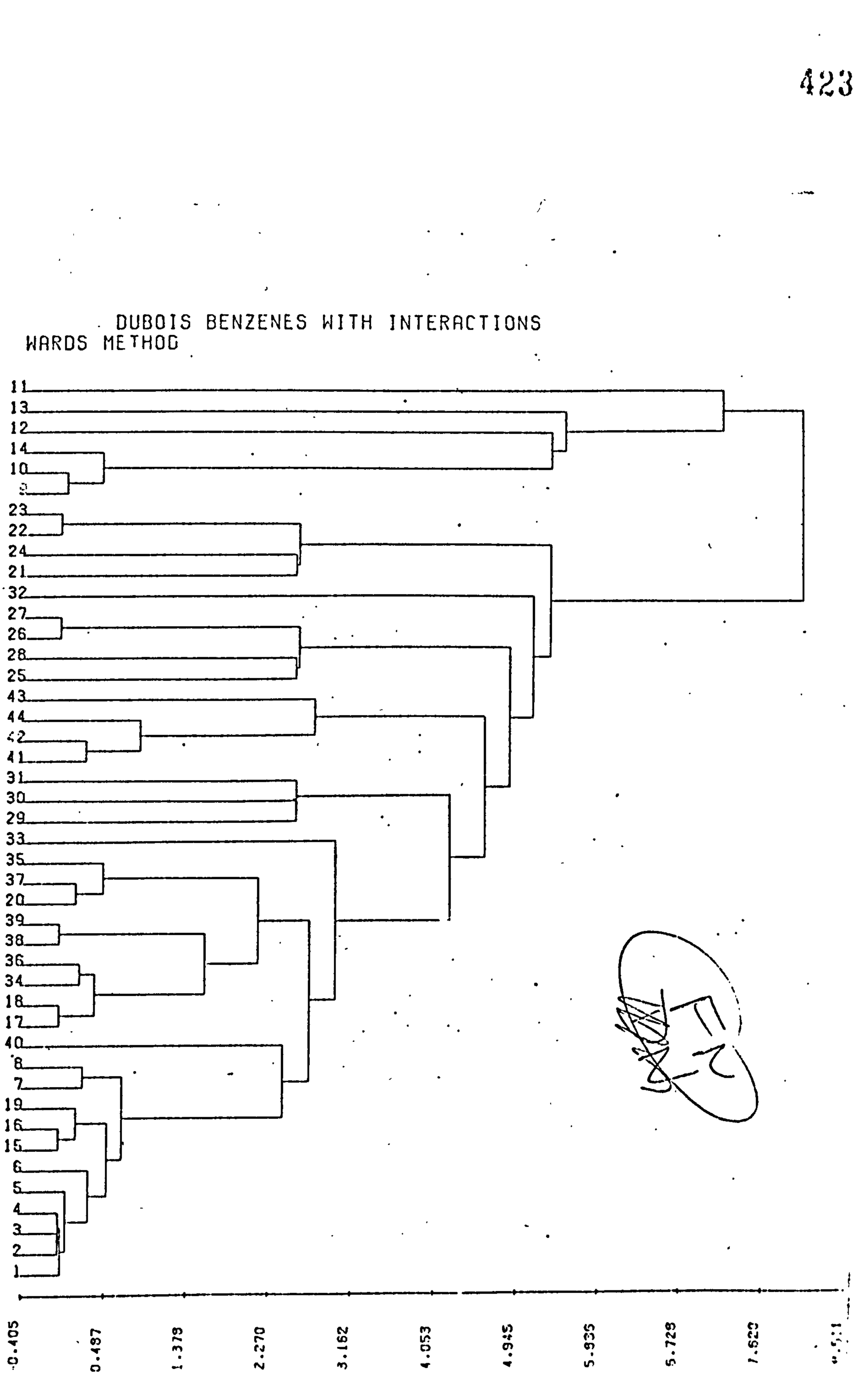
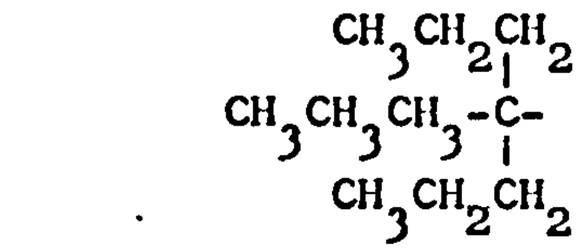
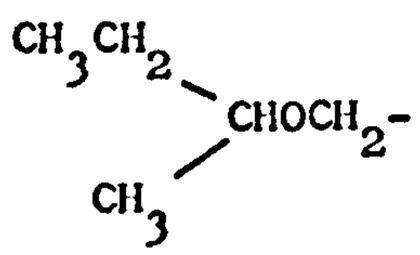
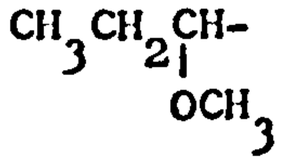
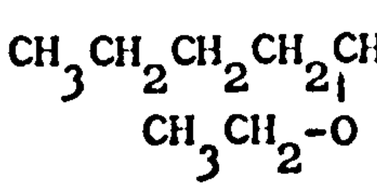
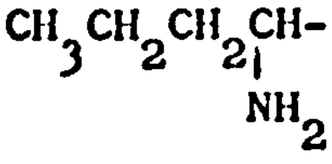
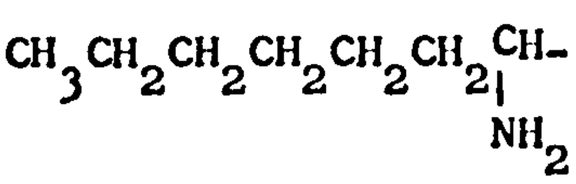
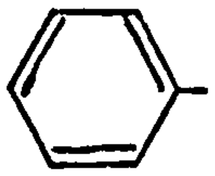
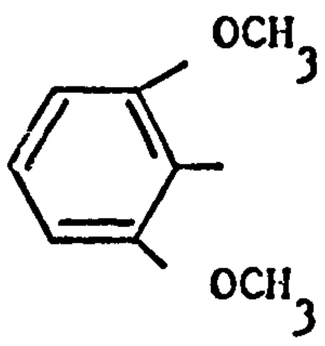
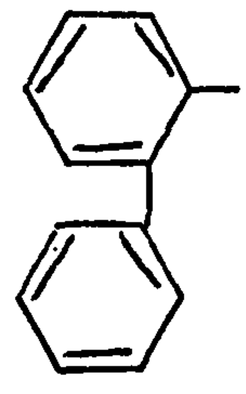
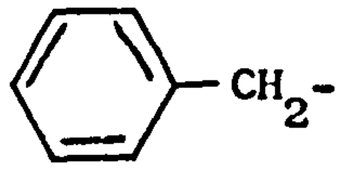
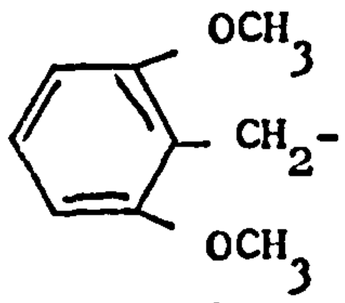
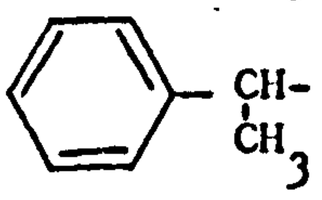
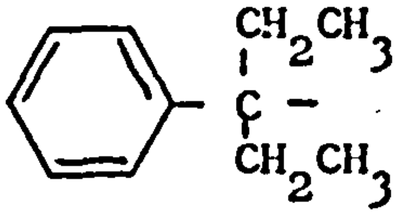
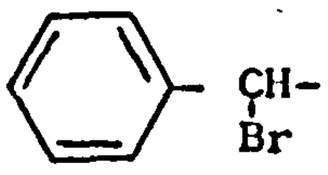
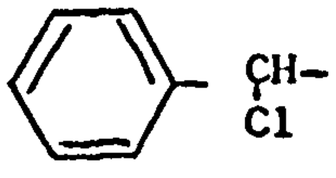
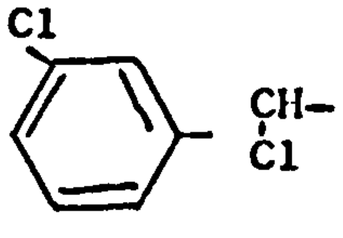
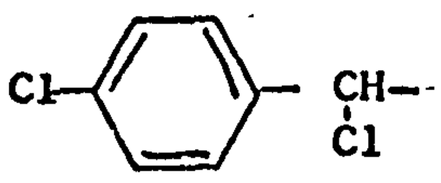
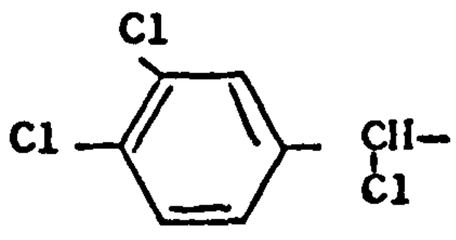
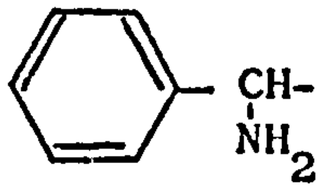
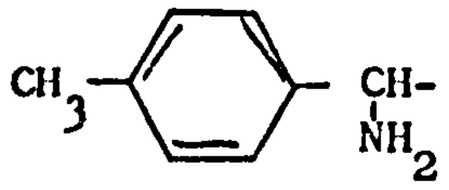
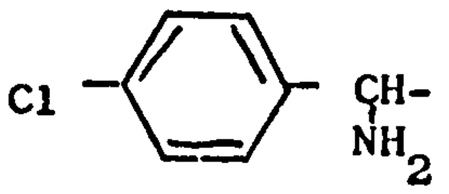
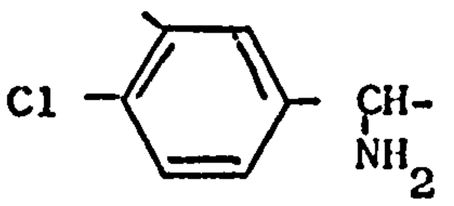
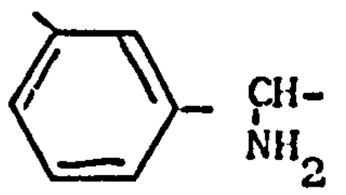
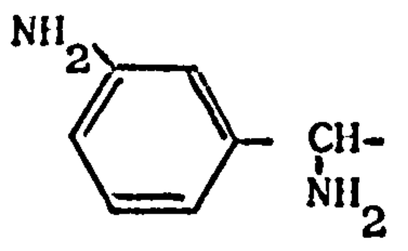
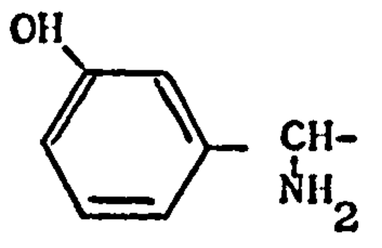
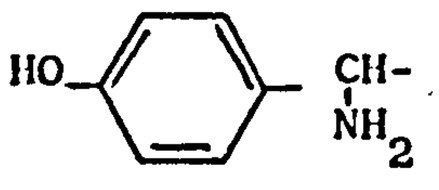
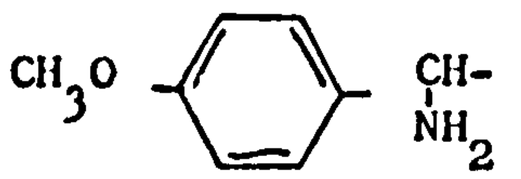
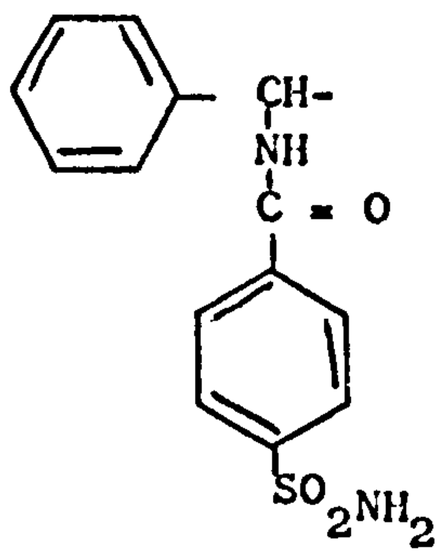


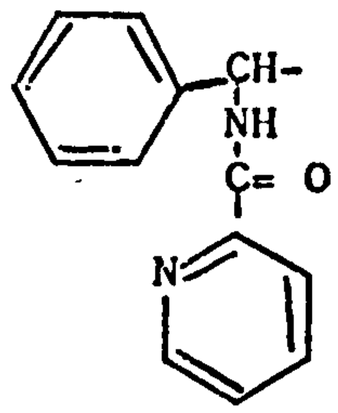
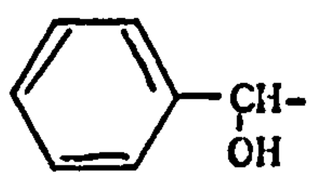
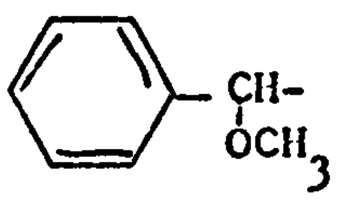
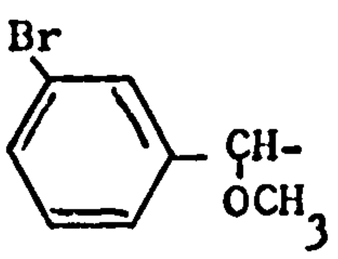
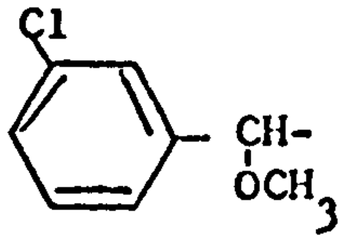
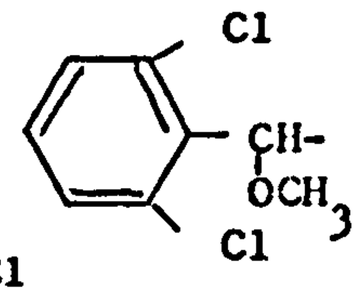
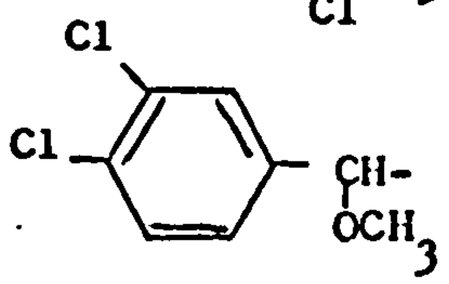
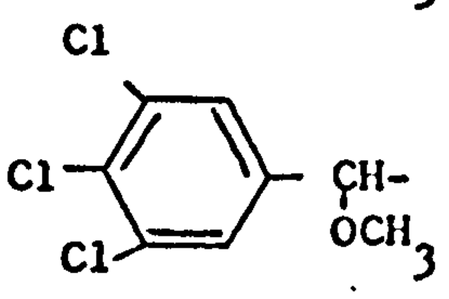
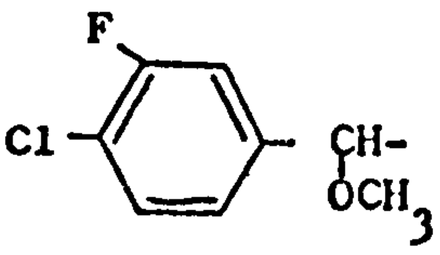
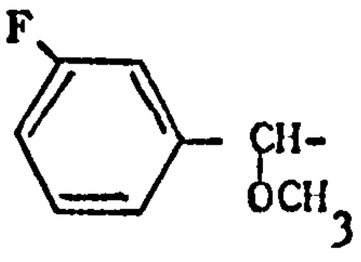
Figure 11

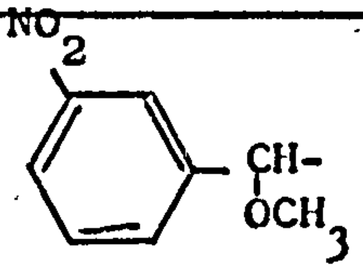
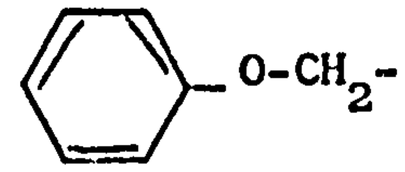
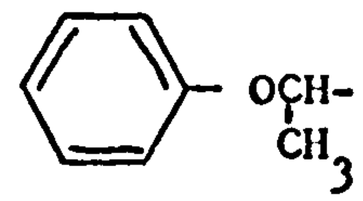
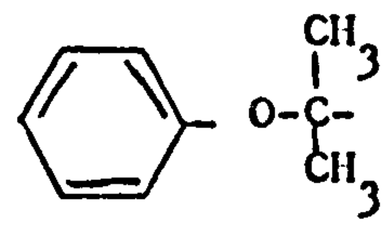
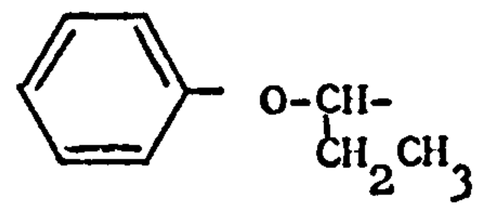
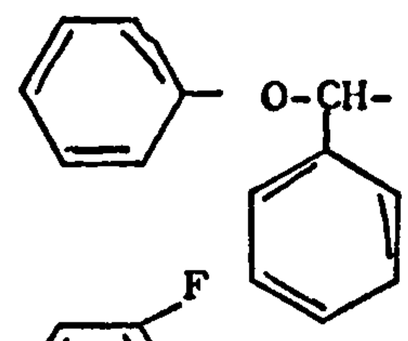
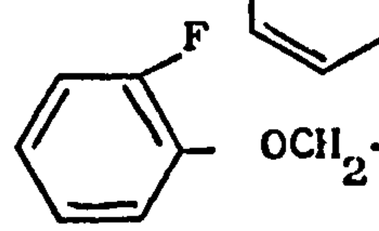
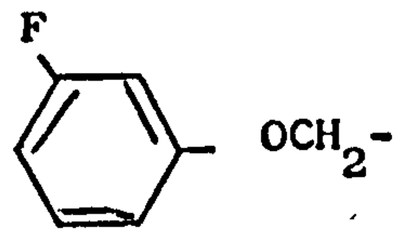
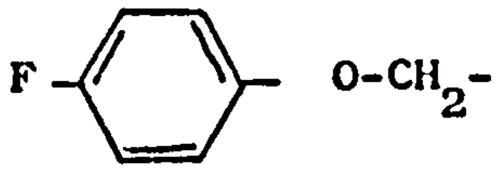
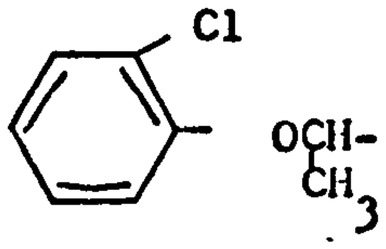
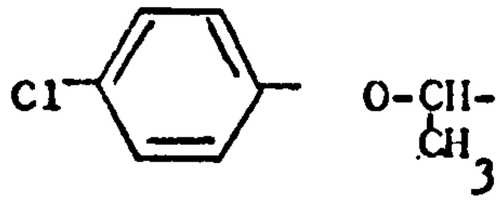
TABLES

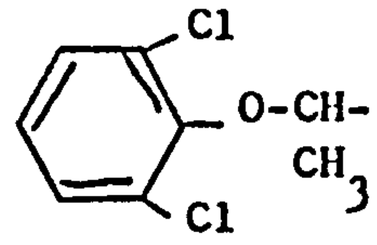
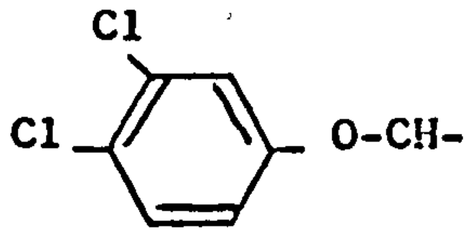
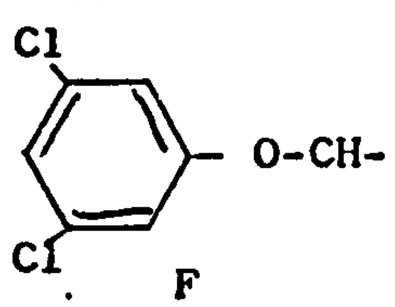
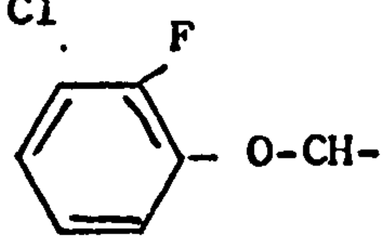
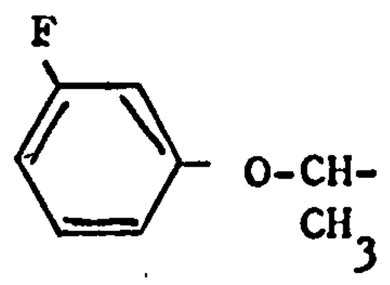
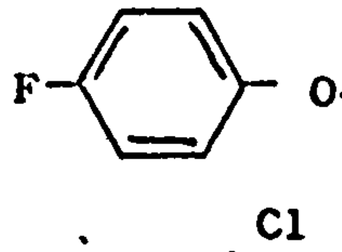
Structure Number	Side-chain Structure	log (b/f) observed[a]	log (b/f) estimated[b]
1	H-	-0.659	-0.659
2	$-CH_3$	-0.753	-0.654
3	$CH_3(CH_2)_5CH_2-$	1.085	0.936
4		1.144	1.414
5	CH_3OCH_2-	-1.110	-0.750
6	$CH_3CH_2OCH_2-$	0.154	-0.485
7	$CH_3CH_2CH_2CH_2OCH_2-$	0.154	-0.045
8		-0.052	-0.166
9		-0.602	-0.431
10		0.454	0.364
11	$CH_3CH_2OCH_2CH_2-$	-0.477	-0.220
12		-0.308	-0.636
13		0.292	0.159
14		0.826	0.179
15		-0.017	0.125

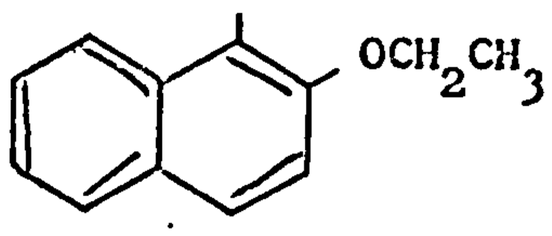
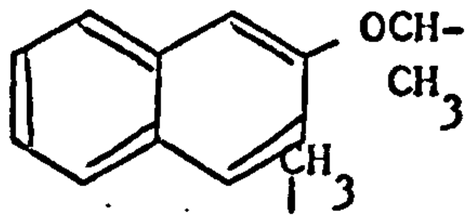
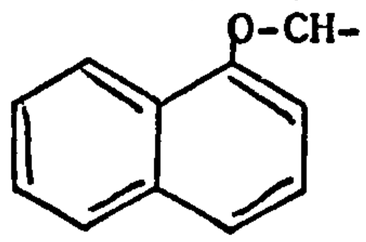
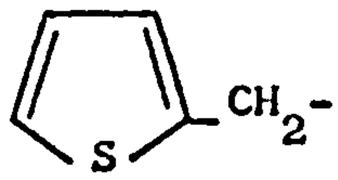
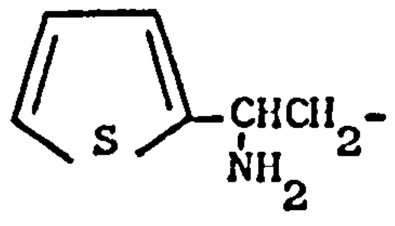
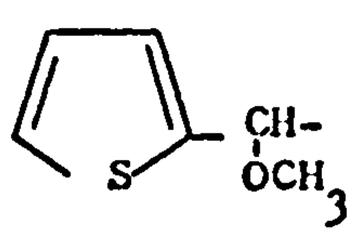
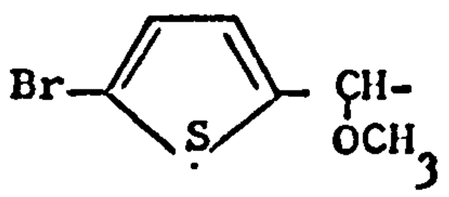
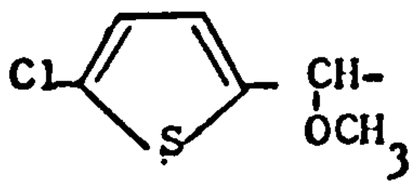
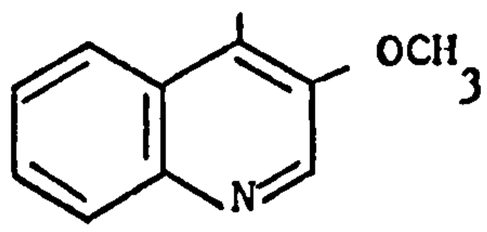
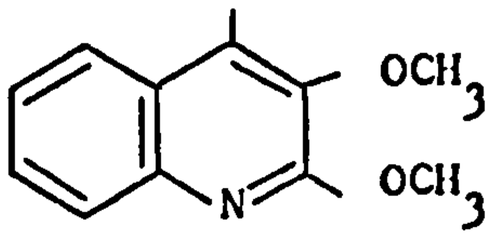
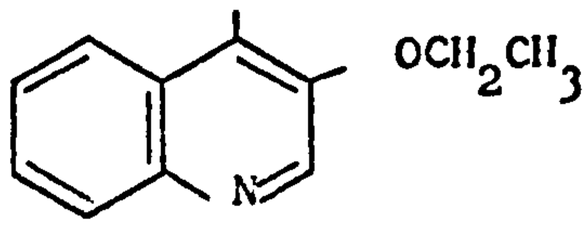
Structure Number	Side-chain Structure	log (b/f) observed[a]	log (b/f) estimated[b]
16		1.005	1.237
17		0.188	0.444
18		0.525	0.390
19		0.327	0.497
20		1.165	1.187
21		0.673	0.673
22		0.550	0.596
23		1.195	1.127
24		1.195	1.117
25		1.510	1.554

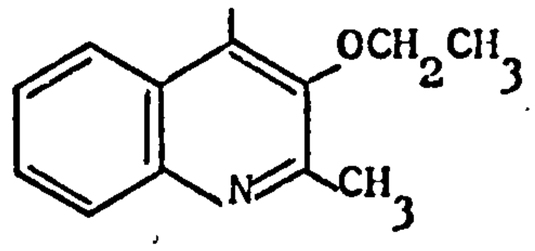
Structure Number	Side-chain Structure	log (b/f) observed[a]	log (b/f) estimated[b]
26		-0.659	-0.333
27		0.176	-0.212
28		0.087	0.243
29		0.664	0.625
30		-0.454	-0.271
31		-0.865	-0.865
32		-0.695	-0.657
33		-0.575	-0.613
34		-0.213	-0.069
35		-0.140	-0.140

Structure Number	Side-chain Structure	log (b/f) observed[a]	log (b/f) estimated[b]
36		0.231	0.231
37		0.056	0.056
38		0.213	0.136
39		0.865	0.877
40		0.689	0.667
41		0.720	0.706
42		1.061	1.094
43		1.440	1.440
44		0.689	0.626
45		0.269	0.199

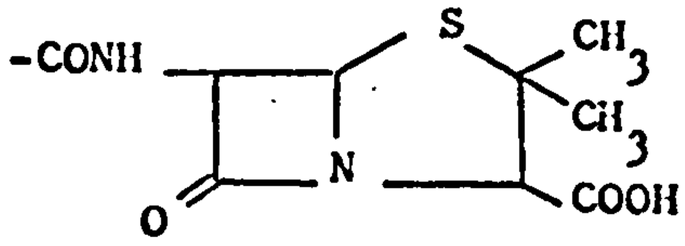
Structure Number	Side-chain Structure	log (b/f) observed[a]	log (b/f) estimated[b]
46		0.176	0.176
47		0.589	0.586
48		0.644	0.639
49		1.091	0.799
50		0.792	0.904
51		1.541	1.471
52		1.032	0.797
53		0.704	0.648
54		0.644	0.693
55		1.380	1.320
56		1.261	1.215

Structure Number	Side-chain Structure	log (b/f) observed[a]	log (b/f) estimated[b]
57		1.380	1.210
58		1.574	1.560
59		1.510	1.510
60		0.788	0.851
61		0.720	0.702
62		0.602	0.747
63		1.297	1.475
64		0.788	0.951
65		1.327	1.064
66		0.661	0.545
67		0.602	0.687

Structure Number	Side-chain Structure	log (b/f) observed[a]	log (b/f) estimated[b]
68		0.921	0.952
69		1.252	1.525
70		1.574	1.411
71		0.140	0.263
72		-0.327	-0.249
73		0.158	-0.044
74		0.940	0.928
75		0.707	0.719
76		0.122	0.110
77		0.207	0.204
78		0.362	0.374

Structure Number	Side-chain Structure	log (b/f) observed[a]	log (b/f) estimated[b]
79		0.466	0.469

Notes: penicillin nucleus is common



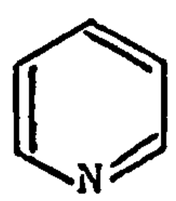
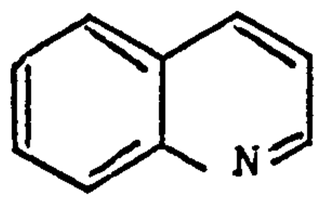
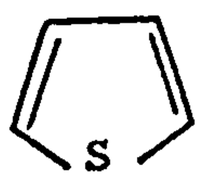
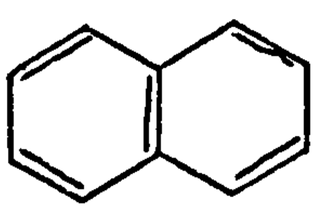
[a] from BIRD et al., 1967

[b] using structural feature set B

Penicillin serum binding

Observed and estimated activities

Table 1

Structural Feature	WLN string	Regression Coefficient	t statistic (60 degrees of freedom)
	T6NJ	excluded by regression program	
	T66 BNJ	1.09	5.84
	T5SJ	0.84	4.69
	L66J	1.81	10.67
	R	1.07	8.28
$-CH_3$	(1t)	-0.04	0.44
$-CH_2-$	(1c)	0.28	7.46
$-\overset{|}{C}H-$	Y	0.34	3.37
$-\overset{|}{\underset{|}{C}}-$	X	0.68	2.70
-CHO	VH	0.27	0.83
-CONH-	VM	-0.25	0.84
-OH	Q	-0.44	2.41
$-NH_2$	Z	-0.67	6.70
-O-	O	0.04	0.44
$-SO_2NH_2$	SWZ	-1.44	3.50
$-NO_2$	NW	-0.30	1.06

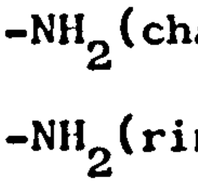
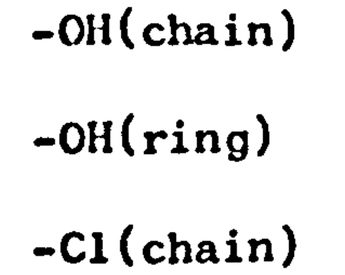
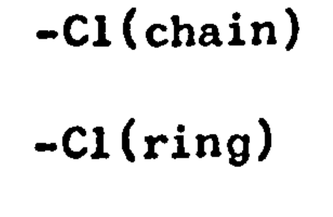
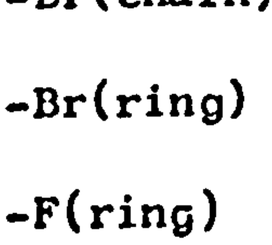
Structural Feature	WLN string	Regression Coefficient	t statistic (60 degrees of freedom)
-Cl	G	0.36	7.06
-Br	E	0.42	2.35
-F	F	0.11	1.01
regression constant		-0.59	1.30

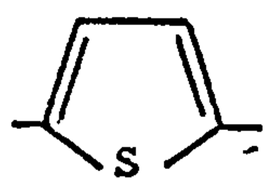
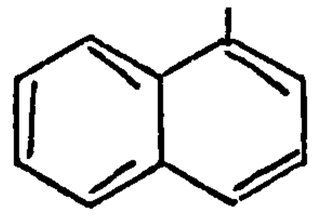
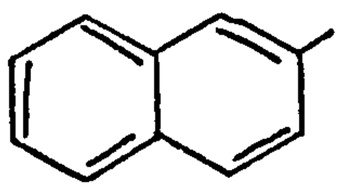
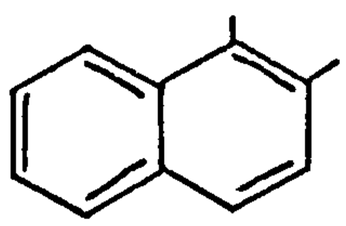
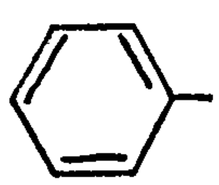
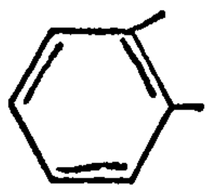
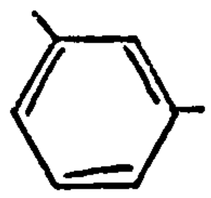
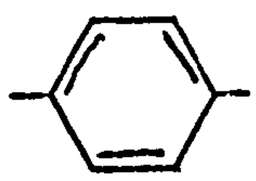
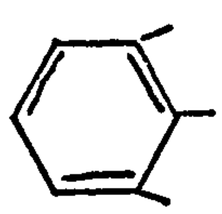
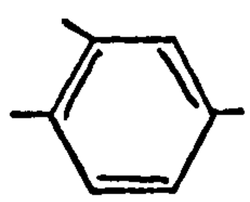
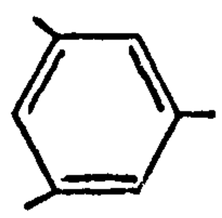
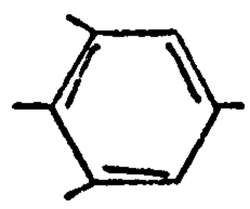
t = terminal, c = connective

Penicillin serum binding

Regression analysis results with structural feature set A

Table 2

Structural Feature	WLN string	Regression Coefficient	t statistic (45 degrees of freedom)
$-CH_3$	(1t)	-0.10	1.33
$-CH_2-$	(1c)	0.27	8.99
-CH-	Y	0.42	4.08
-C-	X	0.67	3.16
-CHO	VH	-0.10	0.43
$-SO_2NH_2$	SWZ	-1.33	3.96
$-NO_2$	NW	-0.13	0.51
-O-(chain)	O	-0.36	3.52
-O-(ring)	(1oc) O	0.14	1.67
$-NH_2$(chain)	Z	-0.93	7.81
$-NH_2$(ring)	(1oc) Z	-0.71	2.73
-OH(chain)	Q	-0.54	2.36
-OH(ring)	(1oc) Q	-0.50	2.54
-Cl(chain)	G	0.01	0.01
-Cl(ring)	(1oc) G	0.36	3.16
-Br(chain)	E	0.08	0.35
-Br(ring)	(1oc) E	0,57	2.59
-F(ring)	(1oc) F	-0.11	0.81
	T6NJ	excluded by regression program	
	T66 BNJ (D,E)	0.62	3.23
	T66 BNJ (C,D,E)	0.81	3.69
	T5 SJ	0.55	3.82

Structural Feature	WLN string	Regression Coefficient	t statistic 45 degrees of freedom)
	T5 SJ (B,E)	0.96	4.49
	L66J (B)	1.51	8.40
	L66J (C)	1.62	9.03
	L66J (B,C)	1.20	7.08
	R (A)	0.74	7.16
	R (A,B)	1.06	6.71
	R (A,C)	0.91	5.78
	R (A,D)	0.95	6.32
	R (A,B,E)	0.59	2.71
	R (A,C,D)	0.98	4.07
	R (A,C,E)	0.89	2.76
	R (A,C,D,E)	0.97	2.50

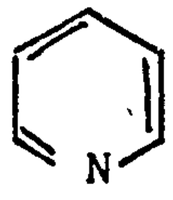
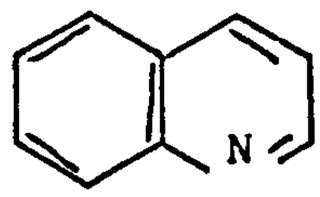
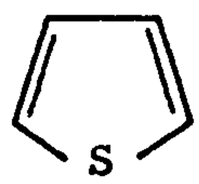
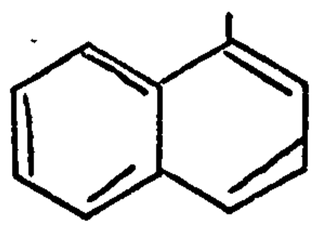
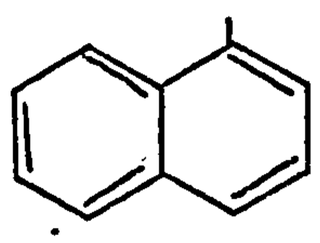
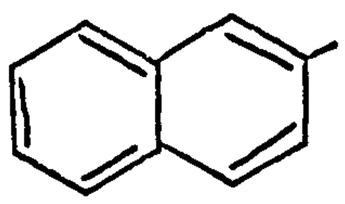
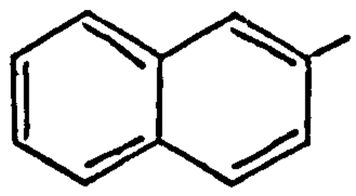
Structural Feature	WLN string	Regression Coefficient	t statistic 45 degrees of freedom)
regression constant	excluded by regression program		

Notes: c = connective, t = terminal, (loc) = WLN ring locant

Penicillin serum binding

Regression analysis results with structural feature set B

Table 3

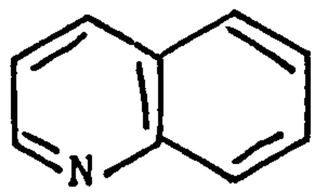
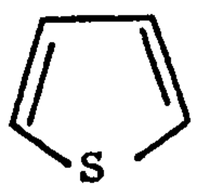
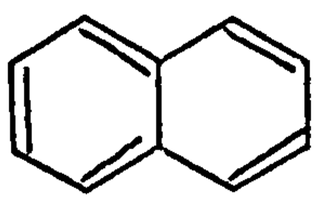
Feature Number	Structural Feature	
1		
2		
3		
4		(directly joined to penicillin nucleus)
5		(joined via chain to penicillin nucleus)
6		(directly joined to penicillin nucleus)
7		(joined via chain to penicillin nucleus)
8	CH_3-	
9	$-CH_2-$	
10	$-\overset{\vert}{C}H-$	
11	$-\overset{\vert}{C}-$	
12	-CHO	
13	-CONH-	
14	-OH	
15	$-NH_2$	
16	-O-	

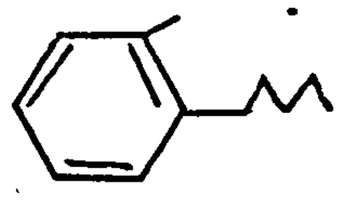
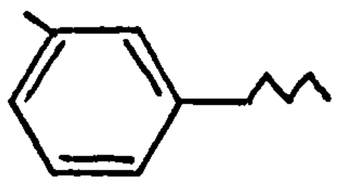
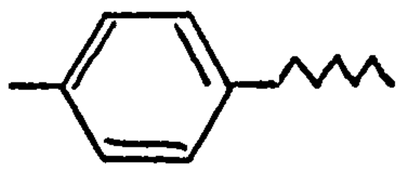
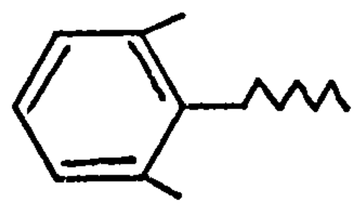
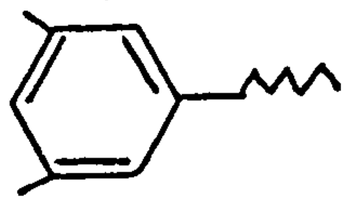
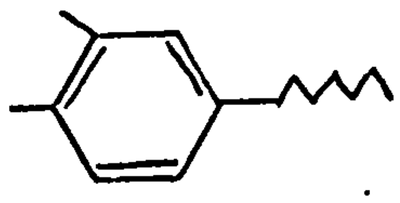
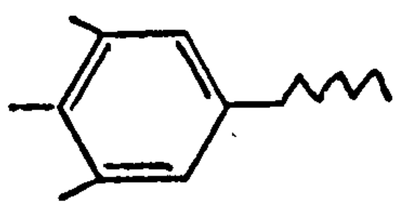
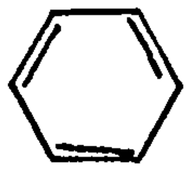
Feature Number	Structural Feature
17	$-SO_2NH_2$
18	$-NO_2$
19	-Cl
20	-Br
21	-F
22	

Penicillin serum binding

Structural Features in Set C

Table 4

Feature Number	Structural Feature
1	
2	
3	
4	
5	$-CH_3-$
6	$-CH_2-$
7	$-\overset{|}{C}H-$
8	$-\overset{|}{\underset{|}{C}}-$
9	-CHO
10	-CONH-
11	-OH
12	$-NH_2$
13	-O-
14	$-SO_2NH_2$
15	$-NO_2$
16	-Cl (on alkyl chain)
17	-Cl (benzene - <u>ortho</u>)
18	-Cl (benzene - <u>meta</u>)
19	-Cl (benzene - <u>para</u>)
20	-Cl (thiophene)
21	-Br (on alkyl chain)

Feature Number	Structural Feature	
22	-Br	(benzene - meta)
23	-Br	(thiophene)
24	-F	(benzene - ortho)
25	-F	(benzene - meta)
26	-F	(benzene - para)
27		halogen substituted
28		
29		
30		
31		
32		
33		
34		(without halogen substitution)

—M indicates chain to penicillin nucleus

- positions for halogens on benzene rings are relative to chain to penicillin nucleus

Penicillin serum binding

Structural Features in Set D

Table 5

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual Error	F Value
A	19	18 + constant	60	0.931	0.28	21.68
B	35	34	45	0.981	0.21	35.84
C	22	21 + constant	57	0.940	0.27	20.60
D	34	30 + constant	48	0.935	0.30	11.12
B (10% level)	35	25	54	0.976	0.22	43.39

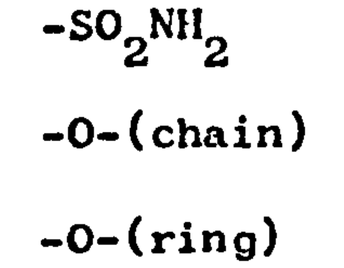
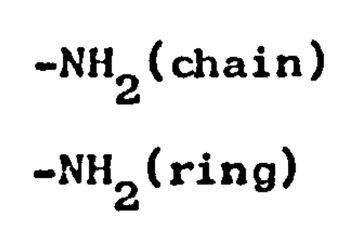
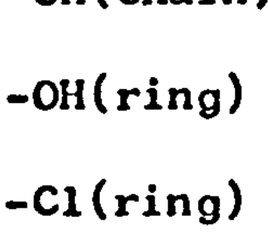
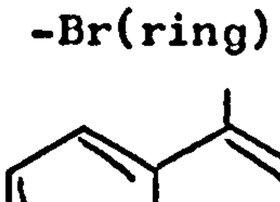
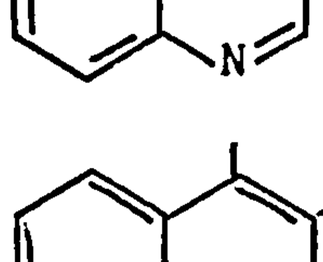
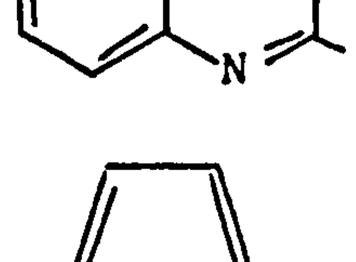
Number of structures = 79

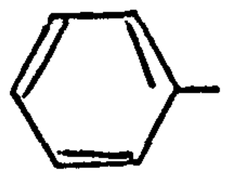
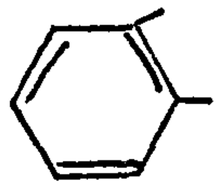
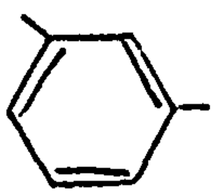
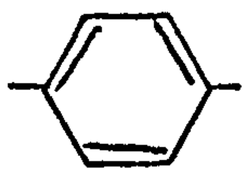
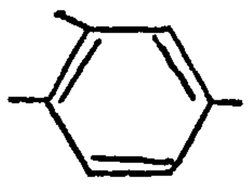
Range of log (b/f) values = 2.684

Penicillin serum binding

Overall regression results

Table 6

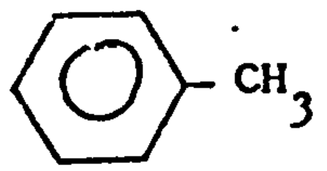
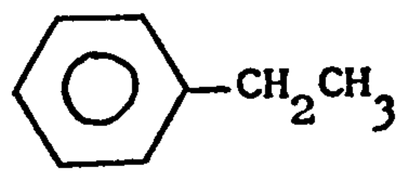
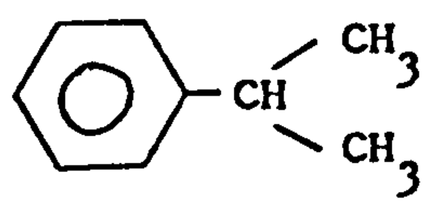
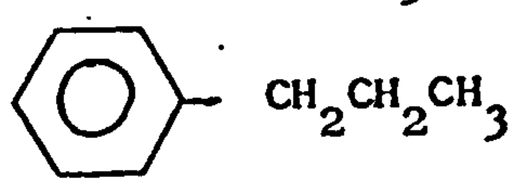
Structural Feature	Regression Coefficient	t statistic (54 degrees of freedom)
$-CH_2-$	0.23	9.15
$-\overset{\vert}{\underset{\vert}{C}}H-$	0.29	3.88
$-\overset{\vert}{\underset{\vert}{C}}-$	0.44	3.03
-CONH-	-0.48	5.43
$-SO_2NH_2$	-0.90	3.59
-O-(chain)	-0.40	4.92
-O-(ring)	0.21	3.72
$-NH_2$(chain)	-0.84	8.49
$-NH_2$(ring)	-0.54	2.25
-OH(chain)	-0.48	2.12
-OH(ring)	-0.31	1.80
-Cl(ring)	0.63	12.22
-Br(ring)	0.79	4.18
	0.39	2.30
	0.49	2.85
	0.53	3.74
	0.70	3.66
	1.41	8.18
	0.99	6.77

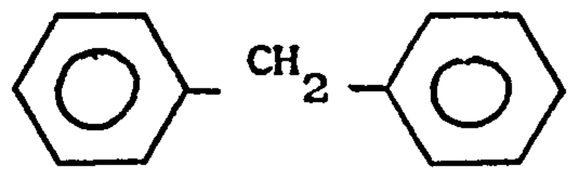
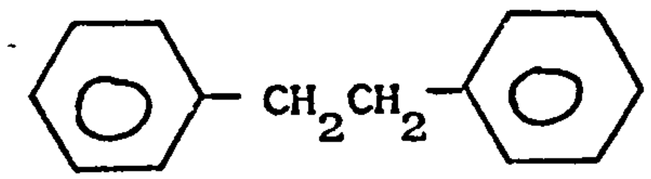
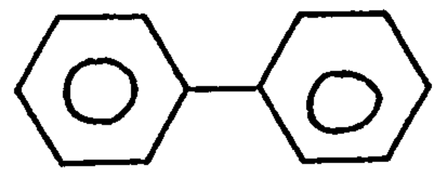
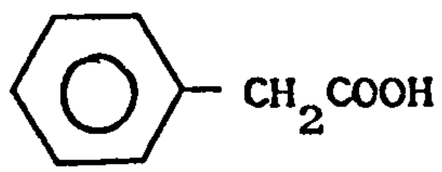
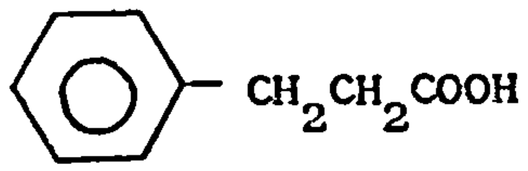
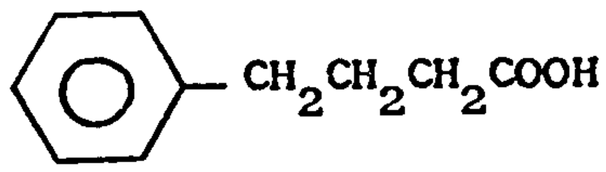
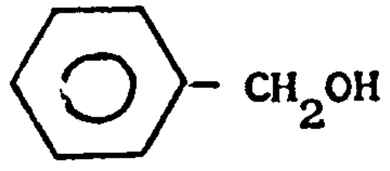
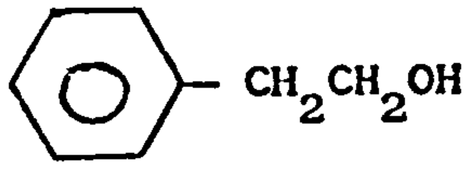
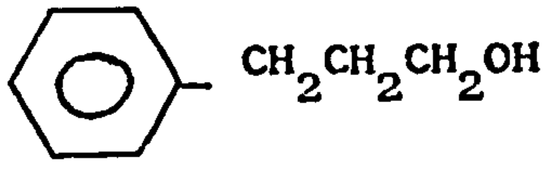
Structural Feature	Regression Coefficient	t statistic (54 degrees of freedom)
	0.72	8.25
	0.83	6.83
	0.70	6.88
	0.71	6.63
	0.43	3:49

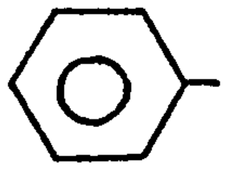
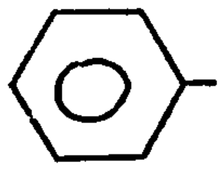
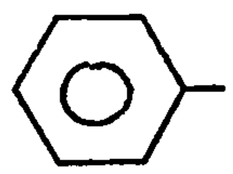
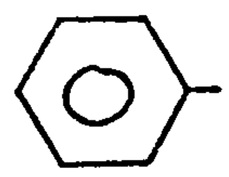
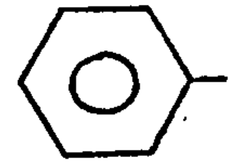
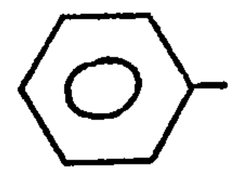
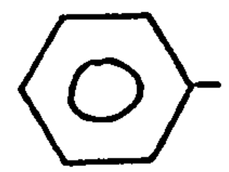
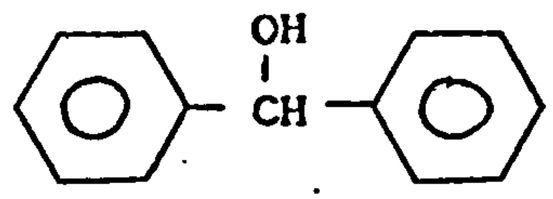
Penicillin serum binding

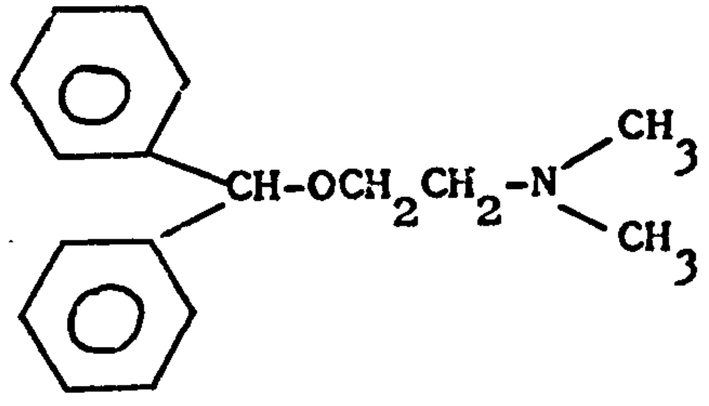
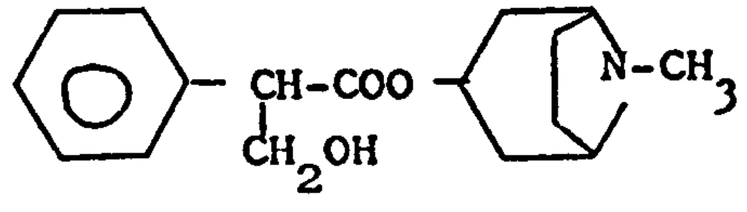
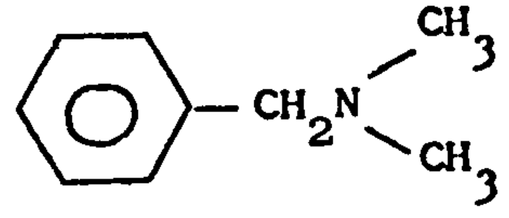
Structural features of set B included at 10%

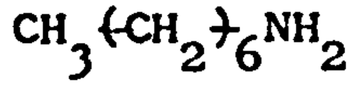
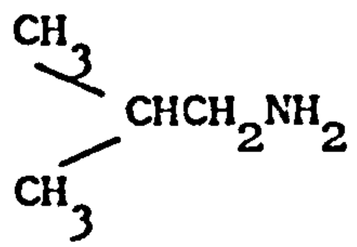
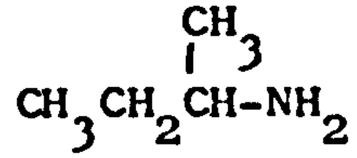
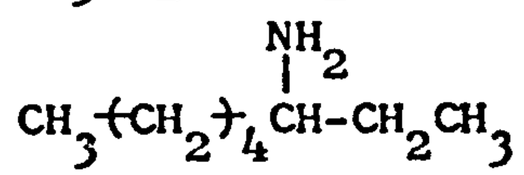
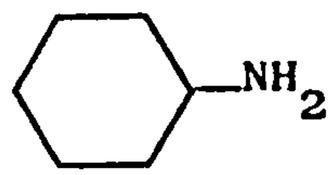
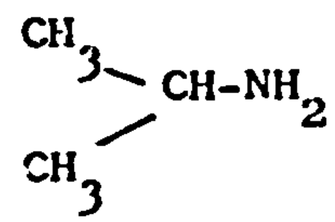
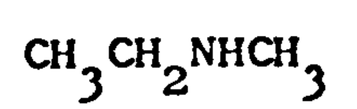
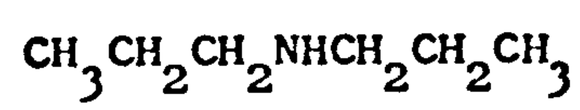
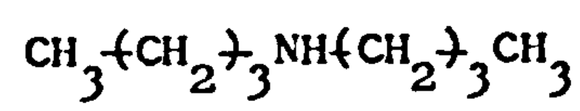
Table 7

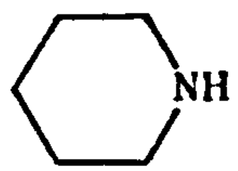
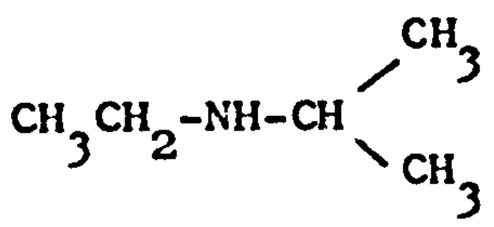
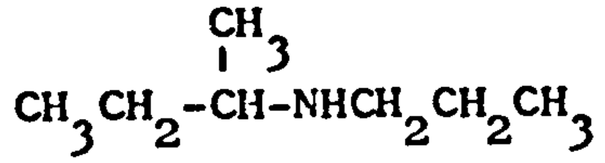
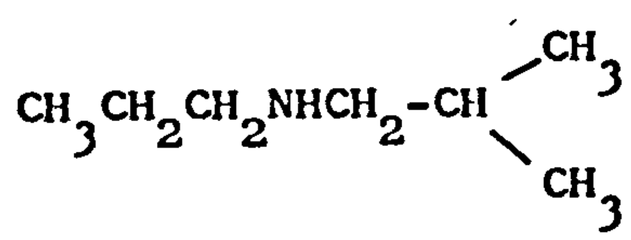
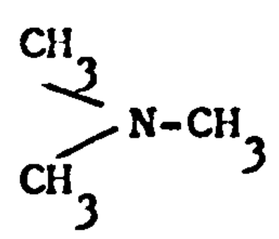
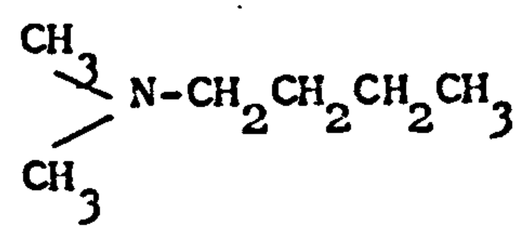
Structure Number	Structure	log P observed[a]	log P estimated[b]
1	CH_3COOH	-0.24	-0.18
2	$CH_3-COO-CH_3$	0.18	0.14
3	CH_3-CH_2-COOH	0.29	0.32
4	$CH_3(CH_2)_2COOH$	0.79	0.82
5	$CH_3CH_2COOCH_2CH_3$	1.21	1.14
6	$CH_3COOCH_2CH_3$	0.68	0.64
7	$CH_3(CH_2)_4COOH$	1.90	1.82
8	$CH_3(CH_2)_8COOH$	4.09	3.82
9	$HOOC(CH_2)_7COOH$	1.57	1.94
10		2.74	2.45
11		3.15	2.95
12		3.66	3.36
13		3.63	3.45

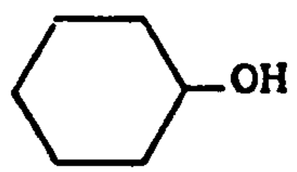
Structure Number	Structure	log P observed[a]	log P estimated[b]
14		4.14	4.19
15		4.81	4.69
16		4.04	3.69
17		1.41	1.56
18		1.84	2.06
19		2.42	2.56
20		1.10	1.13
21		1.36	1.63
22		1.88	2.13

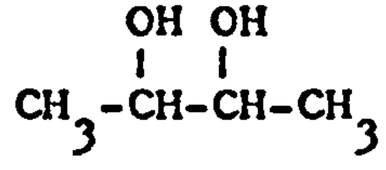
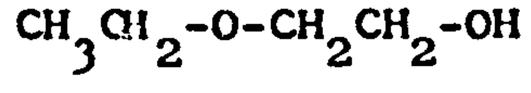
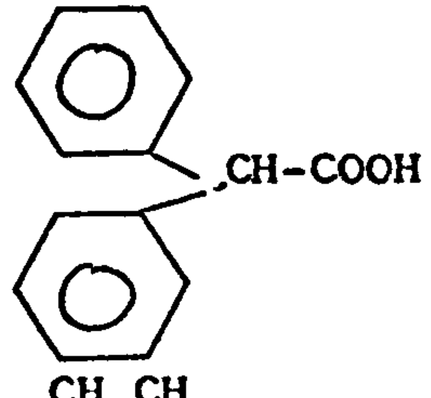
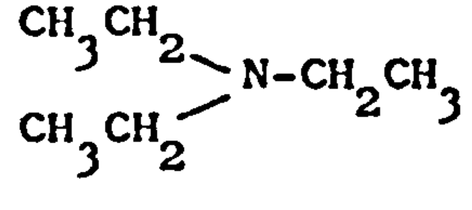
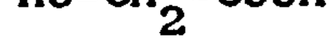
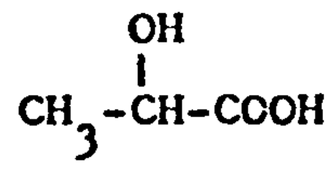
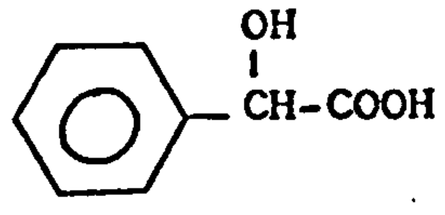
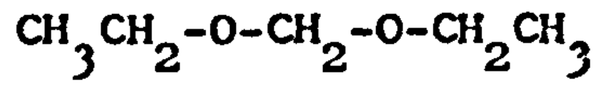
Structure Number	Structure	log P observed[a]	log P estimated[b]
23		1.09	1.16
24		1.41	1.66
25		1.83	1.88
26		1.96	1.88
27		2.32	2.38
28		2.77	2.88
29		2.70	2.81
30		2.66	2.78

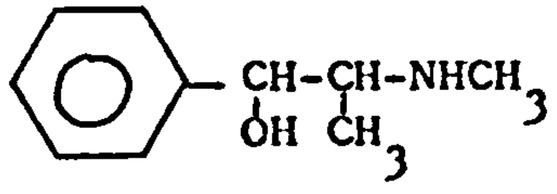
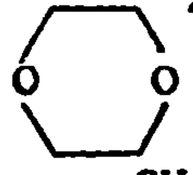
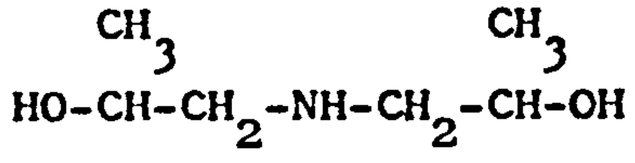
Structure Number	Structure	log P observed[a]	log P estimated[b]
31		3.34	3.39
32		1.81	1.81
33		1.95	1.88
34	CH_3NH_2	-0.57	-0.58
35	$CH_3CH_2NH_2$	-0.13	-0.08
36	$CH_3CH_2CH_2NH_2$	0.48	0.42
37	$CH_3CH_2CH_2CH_2NH_2$	0.83	0.92
38	$CH_3(CH_2)_4NH_2$	1.49	1.42
39	$CH_3(CH_2)_5NH_2$	2.02	1.92

Structure Number	Structure	log P observed[a]	log P estimated[b]
40		2.57	2.42
41		0.81	0.83
42		0.74	0.83
43		2.82	2.83
44		1.49	1.37
45		0.26	0.33
46		0.15	0.18
47		1.70	1.68
48		2.75	2.68

Structure Number	Structure	log P observed[a]	log P estimated[b]
49	$CH_3CH_2NHCH_2CH_3$	0.53	2.68
50	$CH_3(CH_2)_2NH(CH_2)_3CH_3$	2.12	2.18
51	$CH_3(CH_2)_3NHCH_3$	1.33	1.18
52		0.83	0.83
53		0.93	1.09
54		1.91	2.09
55		2.07	2.09
56		0.27	0.14
57		1.70	1.64

Structure Number	Structure	log P observed[a]	log P estimated[b]
58		1.23	1.35
59	CH_3OH	-0.74	-0.61
60	CH_3CH_2OH	-0.32	-0.11
61	$CH_3CH_2CH_2OH$	0.34	0.39
62	$CH_3(CH_2)_3OH$	0.88	0.89
63	$CH_3(CH_2)OH$	1.40	1.39
64	$CH_3(CH_2)_5OH$	2.03	1.89
65	$CH_3(CH_2)_7OH$	3.15	2.89
66	$CH_3CH_2\text{-}O\text{-}CH_2CH_3$	0.78	1.07
67	$CH_3CH_2CH_2\text{-}O\text{-}CH_2CH_2CH_3$	2.03	2.07
68	$CH_3CH_2\text{-}O\text{-}CH_2CH_2CH_2CH_3$	2.03	2.07
69	$(CH_3)_2CH\text{-}CH_2OH$	0.74	0.80
70	$CH_3CH_2\text{-}CH(CH_3)\text{-}OH$	0.61	0.80
71	$(CH_3)_2CH\text{-}CH_2CH_2OH$	1.16	1.30

Structure Number	Structure	log P observed[a]	log P estimated[b]
72		-0.92	-0.61
73		-0.54	-0.25
74		3.05	3.21
75		1.44	1.64
76		-1.11	-1.50
77		-0.62	1.09
78		0.56	0.15
79		0.84	0.43
80		-0.59	-0.56

Structure Number	Structure	log P observed[a]	log P estimated[b]
81		0.93	0.92
82		-1.31	-1.41
83		-0.42	-0.42
84		-0.82	-1.14

Notes: [a] from NYS et al., 1973 (log P values averaged as described)

[b] using structural feature set C

Diverse Structures Partition Coefficients

Observed and Estimated log P values

Table 8

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual Error	F Value
A	11	11 + constant	72	0.991	0.19	358.73
B	17	15	69	0.996	0.19	571.55
C	15	14 + constant	69	0.991	0.19	270.12

Number of structures = 84

Range of log P values = 6.12

All analyses at the 99.99% level.

Diverse Structures Partition Coefficients

Overall Regression Results

Table 9

Structural Feature	Regression Coefficient	t statistic (72 degrees of freedom)
CH_3-	0.70	15.26
-CH_2-	0.50	38.49
-CH-	0.20	4.25
-COOH	-0.68	9.48
-COO-	-1.07	12.60
-O-	-1.13	18.05
-OH	-1.12	15.41
-NH_2	-1.09	12.12
-NH-	-1.52	18.43
-N-	-1.76	17.70
benzene ring	1.94	32.05
regression constant	-0.19	1.55

Diverse Structures Partition Coefficients

Regression Analysis Results with Structural Feature Set A

Table 10

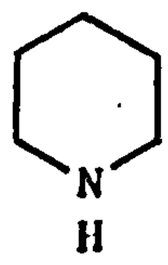
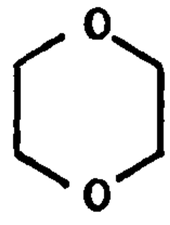
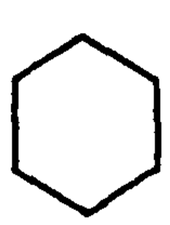
Structural Feature	Regression Coefficient	t statistic (69 degrees of freedom)	Perfectly correlated structural features
CH_3-	0.60	16.80	
-CH_2-	0.50	36.72	
-$\overset{|}{C}H$-	0.31	4.81	
-COOH	-0.78	15.52	
-COO-	-1.07	11.72	
-O-	-1.14	15.60	
-OH	-1.22	24.76	
-NH_2	-1.19	18.71	
-NH-	-1.53	17.33	
-$\overset{|}{N}$-	-1.67	14.00	
benzene ring	1.84	53.41	
CH_2	0.17	4.32	
$\overset{|}{C}H$	1.73	7.23	
CH	-0.78	6.03	*$\overset{|}{N}$*
NH	excluded by regression program		
O	-0.54	4.40	
regression constant	excluded by regression program		

Diverse Structures Partition Coefficients

Regression Analysis Results with Structural Feature Set B

Table 11

Note: * indicates ring bond

Structural Feature	Regression Coefficient	t statistic (69 degrees of freedom)
$-CH_3$	0.81	7.94
$-CH_2-$	0.50	36.72
$-\overset{\vert}{C}H-$	0.10	0.83
-COOH	-0.57	5.27
-COO-	-1.07	11.72
-O-	-1.14	15.60
-OH	-1.01	9.33
$-NH_2$	-0.98	8.49
-NH-	-1.53	17.33
$-\overset{\vert}{N}-$	-1.88	12.28
benzene ring	2.05	20.12
	0.84	3.92
	1.25	4.60
	excluded by regression program	
	2.77	16.06
regression constant	-0.42	2.19

Diverse Structures Partition Coefficients

Regression Analysis Results with Structural Feature Set C

Table 12

Structure Number	Substituents	Reaction Site Position	log k (observed)[a]	log k (estimated)[b]	log k (predicted)[b]
1	H	1	-5.569	-5.73	-6.03
2	1-Me	2	-2.745	-2.55	-2.40
3	1-Me	3	-4.328	-4.41	-4.53
4	1-Me	4	-1.959	-1.84	-1.75
5	$1,3\text{-Me}_2$	4	0.903	1.07	1.14
6	$1,4\text{-Me}_2$	2	-1.469	-1.45	-1.43
7	$1,2,3,\text{-Me}_3$	4	2.182	2.05	1.93
8	$1,3,5\text{-Me}_3$	2	3.954	3.72	3.34
9	$1,2,3,4\text{-Me}_4$	5	2.663	2.72	2.76
10	$1,2,3,5\text{-Me}_5$	4	4.439	4.48	4.49
11	$1,3,5\text{-Me}_3\text{-2-Et}$	4	4.288	4.29	-
12	$1,3,5\text{-Me}_3\text{-2-Cl}$	4	0.447	0.45	-
13	$1,3,5\text{-Me}_3\text{-2-Br}$	4	0.531	0.53	-
14	$1,2,3,4,5\text{-Me}_5$	6	4.954	4.98	5.04
15	1-MeO	4	3.980	4.13	4.31

Structure Number	Substituents	Reaction Site Position	log k (observed)[a]	log k (estimated)[b]	log k (predicted)[b]
16	1-MeO	2	2.300	2.27	2.22
17	1-MeO-2-Me	4	4.685	4.62	4.57
18	1-MeO-2-Me	6	2.837	2.75	2.65
19	1-MeO-3-Me	4	5.757	5.77	5.77
20	1-MeO-4-Me	2	2.922	2.84	2.76
21	1-MeO-2-F	4	1.586	1.59	-
22	1-MeO-3-F	4	3.157	3.16	-
23	1-MeO-3-F	6	2.799	2.80	-
24	1-MeO-4-F	2	-0.173	-0.17	-
25	1-MeO-2-Cl	4	1.399	1.40	-
26	1-MeO-3-Cl	4	2.766	2.77	-
27	1-MeO-3-Cl	6	1.570	1.57	-
28	1-MeO-4-Cl	2	-0.447	-0.45	-
29	1-MeO-2-Br	4	1.635	1.64	-
30	1-MeO-3-Br	6	1.162	1.16	-

Structure Number	Substituents	Reaction Site Position	log k (observed)[a]	log k (estimated)[b]	log k (predicted)[b]
31	1-MeO-4-Br	2	-0.404	-0.40	-
32	1-MeO-3-I	6	1.129	1.13	-
33	1,4-$(MeO)_2$	2	2.315	2.32	-
34	1-MeO-2,3-Me_2	4	6.010	6.08	6.16
35	1-MeO-2,4-Me_2	6	2.816	3.06	3.49
36	1-MeO-2,5-Me_2	4	6.190	6.03	5.79
37	1-MeO-3,4-Me_2	6	5.174	5.01	4.76
38	1-MeO-3,5-Me_2	4	7.140	7.14	7.13
39	1-MeO-3,5-Me_2	6	5.854	5.98	5.85
40	1-OH	4	4.608	4.61	-
41	1-NMe_2	4	8.336	8.31	8.18
42	1-NMe_2-3-Me	4	8.992	9.04	9.07
43	1-NMe_2-3-Br	4	7.772	7.77	-
44	1-NMe_2-3,5Me_2	4	9.531	9.51	9.75

Notes: [a] from DUBOIS et al., 1972

[b] using structural feature set D, including all compounds in the regression

[c] using structural feature set D, and hold-one-out procedure

Benzene derivatives reaction kinetics

Observed, estimated, and predicted log k values

Table 13

Structural Feature set	Number of Structural Features	Number included in analysis	Degrees of Freedom	Multiple correlation coefficient	Residual Error	F Value
A	10	9 + constant	34	0.905	1.59	17.09
B	10	9 + constant	34	0.905	1.59	17.09
C	24	23 + constant	24	0.987	0.70	47.70
D	45	31 + constant	12	70.999	0.17	196.16
D (10%)	45	24	20	70.999	0.19	416.04

Number of structures = 44

Range of log k values = 15.10

Benzene derivatives reaction kinetics

Overall regression analysis results

Table 14

Structural Feature	Regression Coefficient	t statistic (34 degrees of freedom)
benzene ring	common to all structures	
Me	1.88	7.30
Et	2.56	1.55
F	0.40	0.43
Cl	-0.35	0.43
Br	-0.55	0.68
I	-0.31	0.19
OH	8.51	4.79
OMe	5.34	8.48
NMe_2	11.29	11.26
regression constant	-3.90	4.90

Benzene Derivatives Reaction Kinetics

Regression Results for Structural Feature Set B

(99.99% level)

Table 15

Structural Feature	Regression Coefficient	t statistic (24 degrees of freedom)
benzene ring	common to all structure	
Me-o-RS	1.93	9.54
Me-m-RS	0.95	4.41
Me-p-RS	3.42	9.16
Et-m-RS	1.56	2.04
F-o-RS	-0.37	0.49
F-m-RS	-1.79	3.23
F-p-RS	1.33	1.73
Cl-o-RS	-0.76	1.00
Cl-m-RS	-2.11	4.61
Cl-p-RS	0.10	0.13
Br-o-RS	0.75	0.90
Br-m-RS	-1.99	4.35
Br-p-RS	-0.31	0.40
I-p-RS	-0.34	0.45

Structural Feature	Regression Coefficient	t statistic (24 degrees of freedom)
OH-p-RS	9.17	11.61
OMe-o-RS	6.03	16.17
OMe-m-RS	0.84	1.10
OMe-p-RS	8.09	20.83
NMe_2-p-RS	11.60	21.65
regression constant	-4.56	12.60

Note: RS indicates reaction site

Benzene Derivatives Reaction Kinetics

Regression Results for Structural Feature Set C

(99.99% level)

Table 16

Structural Feature	Regression Coefficient	t statistic (12 degrees of freedom)	Perfectly Correlated Structural Features
benzene ring	common to all structures		
Me-o-RS	3.18	23.81	
Me-m-RS	1.32	8.08	
Me-p-RS	3.89	26.96	
Et-m-RS	0.57	2.56	Me-o-Et, Me-p-Et
F-o-RS	-1.51	5.26	
F-m-RS	-2.44	11.22	
F-p-RS	excluded by regression program		
Cl-o-RS	-0.67	2.34	
Cl-m-RS	-2.71	12.48	
Cl-p-RS	excluded by regression program		
Br-o-RS	-0.54	2.27	Br-m-BMe_2
Br-m-RS	-2.67	12.29	
Br-p-RS	-1.10	5.08	Br-m-OMe

Structural Feature	Regression Coefficient	t statistic (12 degrees of freedom)	Perfectly Correlated Structural Features
I-p-RS	-1.14	5.23	I-m-OMe
OH-p-RS	10.34	45.97	
OMe-p-RS	7.99	43.55	
OMe-m-RS	0.05	0.23	OMe-p-OMe
OMe-p-RS	9.86	53.47	
NMe_2-p-RS	14.04	63.03	
Me-o-Me	-0.17	1.90	
Me-m-Me	-0.26	3.13	
Me-p-Me	-0.22	1.75	
Me-o-OMe	-0.83	4.66	
Me-m-OMe	-1.54	13.60	
Me-p-OMe	-0.74	3.69	
Cl-o-Me	-0.28	1.78	Cl-p-Me
Br-o-Me	-0.26	1.65	Br-p-Me
Cl-o-OMe	-0.02	0.07	

Structural Feature	Regression Coefficient	t statistic (12 degrees of freedom)	Perfectly Correlated Structural Features
Cl-m-OMe	-0.69	3.20	
Cl-p-OMe	excluded by regression program		
Br-o-OMe	0.17	0.60	
Br-p-OMe	excluded by regression program		
F-o-OMe	-0.11	0.37	
F-m-OMe	0.53	2.46	
F-p-OMe	excluded by regression program		
Me-m-NMe_2	-2.45	15.17	
regression constant	-5.73	40.54	

Note: RS indicates reaction site

Benzene Derivatives Reaction Kinetics

Regression Results for Structural Feature Set D

(99.99%. level)

Table 17

Structural Feature	Regression Coefficient	t statistic (20 degrees of freedom)	Perfectly Correlated Structural Features
benzene ring	-5.56	42.59	
Me-o-RS	3.06	22.81	
Me-m-RS	1.01	12.37	
Me-p-RS	3.76	28.12	
Et-m-RS	0.85	3.90	Me-o-Et, Me-p-Et
F-o-RS	-1.07	4.89	
F-m-RS	-2.59	16.13	
F-p-RS	0.42	2.00	
Cl-o-RS	-1.46	6.68	
Cl-m-RS	-2.89	21.30	
Cl-p-RS	-0.80	3.79	
Br-o-RS	-0.53	2.09	Br-m-NMe_2
Br-m-RS	-2.76	20.40	
Br-p-RS	-1.21	5.71	Br-m-OMe
I-p-RS	-1.25	5.87	I-m-OMe

Structural Feature	Regression Coefficient	t statistic (20 degrees of freedom)	Perfectly Correlated Structural Features
OH-p-RS	10.17	44.34	
OMe-o-RS	7.93	49.43	
OMe-p-RS	9.78	60.83	
NMe_2-p-RS	13.87	65.57	
Me-m-Me	-0.30	3.34	
Me-o-OMe	-0.65	5.00	
Me-m-OMe	-1.50	12.81	
Me-p-OMe	-0.52	3.39	
Me-m-NMe_2	-2.32	13.92	

Note: RS indicates reaction site

Benzene Derivatives Reaction Kinetics

Structural Features of set D included in 10% level analysis

Table 18

Variables	Number of variables included	Degrees of freedom	Multiple Correlation Coefficient	Residual Error	F Value
Number of substituents (ns)	1 + constant	42	0.366	3.12	6.50
Number of types of substituent (nt)	1 + constant	42	0.364	3.12	6.41
ns and nt	2 + constant	41	0.455	3.02	5.35

Benzene derivatives reaction kinetics

Correlations with simple variables

Table 19

Structure Number	Structure	pk value (observed)	pk value (estimated)[b]
	phenyl propionic acids		
1	H	4.66	4.65
2	m-Me	4.65	4.68
3	p-Me	4.69	4.68
4	m-Cl	4.58	4.56
5	p-Cl	4.61	4.56
6	m-OMe	4.65	4.71
7	p-OMe	4.69	4.71
8	p-NO_2	4.47	4.44
	phenoxy acetic acids		
9	H	3.15	3.16
10	o-Me	3.23	3.19
11	m-Me	3.20	3.19
12	p-Me	3.22	3.19
13	o-Cl	3.05	3.07

Structure Number	Structure	pk value (observed)	pk value (estimated)[b]
14	m-Cl	3.07	3.07
15	p-Cl	3.10	3.07
16	o-OMe	3.23	3.22
17	m-OMe	3.14	3.22
18	p-OMe	3.21	3.22
19	o-NO_2	2.90	2.95
20	m-NO_2	2.95	2.95
21	p-NO_2	2.89	2.95
22	O-Br	3.19	3.11
23	p-Br	3.15	3.11
24	p-CN	2.93	2.91
	<u>phenylthio acetic acids</u>		
25	H	3.38	3.39
26	o-Me	3.38	3.43
27	m-Me	3.39	3.43
28	p-Me	3.45	3.43

Structure Number	Structure	pk value (observed)	pk value (estimated)[b]
29	o-Cl	3.23	3.31
30	m-Cl	3.30	3.31
31	p-Cl	3.33	3.31
32	o-OMe	3.59	3.46
33	m-OMe	3.39	3.46
34	p-OMe	3.54	3.46
35	o-NO_2	3.10	3.19
36	p-NO_2	3.09	3.19
37	p-Br	3.33	3.35
38	o-SMe	3.57	3.47
39	p-SMe	3.52	3.47
40	m-CF_3	3.30	3.30
41	p-CN	3.12	3.14
42	p-NH_2	3.10	3.10
43	o-COOH	3.01	3.16

Structure Number	Structure	pk value (observed)	pk value (estimated)[b]
44	m-COOH	3.28	3.16
45	p-COOH	3.19	3.16
66	p-CMe_3	3.56[a]	3.48
67	m-SMe	3.27[a]	3.47
68	p-F	3.38[a]	3.35
69	m-F	3.31[a]	3.35
70	m-Br	3.33[a]	3.35
71	m-NO_2	3.40[a]	3.19
	<u>phenylseleno acetic acids</u>		
46	H	3.75	3.72
47	o-Me	3.76	3.76
48	m-Me	3.78	3.76
49	p-Me	3.83	3.76
50	o-Cl	3.57	3.64
51	m-Cl	3.64	3.64
52	p-Cl	3.68	3.64

Structure Number	Structure	pk value (observed)	pk value (estimated)[b]
53	o-OMe	3.87	3.79
54	m-OMe	3.73	3.79
55	p-OMe	3.86	3.79
56	o-OEt	3.90	3.79
57	p-OEt	3.86	3.88
58	o-NO_2	3.42	3.52
59	m-NO_2	3.55	3.52
60	p-NO_2	3.43	3.52
61	o-Br	3.58	3.68
62	p-Br	3.70	3.68
63	o-SMe	3.80	3.80
64	p-SMe	3.83	3.80
65	p-COOH	3.49	3.49
	<u>phenylsulfinyl acetic acids</u>		
72	H	2.53[a]	2.53

Structure Number	Structure	pk value (observed)	pk value (estimated)[b]
73	p-Me	2.54[a]	2.57
74	m-Me	2.56[a]	2.57
75	p-CMe_3	2.57[a]	2.62
76	p-OMe	2.55[a]	2.60
77	p-F	2.45[a]	2.48
78	p-Cl	2.46[a]	2.44
79	p-Br	2.46[a]	2.48
80	m-F	2.48[a]	2.48
81	m-Cl	2.46[a]	2.44
82	m-Br	2.47[a]	2.48
83	p-NO_2	2.39[a]	2.32
84	m-NO_2	2.42[a]	2.32
	<u>phenylsulphonyl acetic acids</u>		
85	H	2.31[a]	2.33
86	p-Me	2.34[a]	2.37
87	m-Me	2.37[a]	2.37

Structure Number	Structure	pk value (observed)	pk value (estimated)[b]
88	p-CMe_3	2.40[a]	2.43
89	p-OMe	2.40[a]	2.40
90	m-OMe	2.39[a]	2.40
91	p-F	2.38[a]	2.29
92	p-Cl	2.22[a]	2.25
93	p-Br	2.32[a]	2.29
94	m-F	2.23[a]	2.29
95	m-Cl	2.25[a]	2.25
96	m-Br	2.32[a]	2.29
97	p-NO_2	2.13[a]	2.13
98	m-NO_2	2.16[a]	2.13

Notes: [a]from PASTO et al., 1965, corrected as described in text
- remaining observed values are from PETTIT et al., 1968

[b]from the regression with set B structural features

Benzene Acid pk Values

Observed and estimated property values

Table 20

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual Error	F Value
A	16	15 + constant	82	0.972	0.17	93.54
B	20	18 + constant	79	0.996	0.07	545.32
C	too many variables for analysis to be performed					
D	44	29 + constant	68	0.408	0.71	0.47
E (99.99%)	64	49 + constant	48	0.997	0.07	162.53
E (10.00%)	64	17 + constant	80	0.996	0.06	584.71

Number of structures = 98

Range of pk values = 2.56

Benzene acid pk values

Overall regression results

Table 21

Structural Features	Regression Coefficient	t statistic (79 degrees of freedom)
benzene ring	common to all structures	
CH_2CH_2COOH	excluded by regression program	
OCH_2COOH	-1.49	52.39
SCH_2COOH	-1.25	45.47
$SeCH_2COOH$	-0.92	32.88
$SOCH_2COOH$	-2.12	70.43
SO_2CH_2COOH	-2.31	78.32
CH_3	0.04	1.19
CMe_3	0.09	1.97
F	-0.04	1.23
Cl	-0.09	2.73
Br	-0.04	1.30
CF_3	-0.09	1.30
NH_2	-0.29	4.11

Structural Features	Regression Coefficient	t statistic (79 degrees of freedom)
COOH	-0.23	5.39
CN	-0.25	4.64
NO_2	-0.21	6.50
OMe	0.07	2.11
OEt	0.16	2.87
SMe	0.07	1.81
regression constant	4.65	138.8

Benzene acid pk values

Regression results for set B structural features

Table 22

Structural Feature	Regression Coefficient	t statistic (80 degrees of freedom)
CH_2CH_2COOH	1.49	54.32
SCH_2COOH	0.23	10.52
$SeCH_2COOH$	0.57	26.24
$SOCH_2COOH$	-0.65	25.42
SO_2CH_2COOH	-0.82	35.09
F	-0.08	2.92
Cl	-0.13	6.68
Br	-0.08	3.76
CF_3	-0.13	1.96
NH_2	-0.33	5.06
NO_2	-0.26	12.63
CN	-0.29	6.26
COOH	-0.27	7.90
OEt	0.11	2.32
OMe-with-SCH_2COOH	0.08	2.01

Benzene acids pk values

Structural features from set E included in regression analysis and the 10% level

Table 23

Structural Feature	Regression Coefficient	t statistic (80 degrees of freedom)
CMe_3-with-SCH_2COOH	0.13	2.06
NO_2-with-$SOCH_2COOH$	0.11	2.15
regression constant	3.20	181.50

Structure Number	Structure	Heat of formation (kcal/mole)
1.	$CH_2=CH_2$	12.50
2.	$CH_3CH=CH_2$	4.88
3.	$CH_3CH_2CH=CH_2$	--0.03
4.	$CH_3CH=CHCH_3$ cis	-1.67
5.	$CH_3CH=CHCH_3$ trans	-2.67
6.	$CH_2=CH(CH_3)_2$	-4.04
7.	$CH_2=CHCH_2CH_2CH_3$	-5.00
8.	$CH_3CH_2CH=CHCH_3$ cis	-6.71
9.	$CH_3CH_2CH=CHCH_3$ trans	-7.59
10.	$CH_2=C(CH_3)CH_2CH_3$	-8.68
11.	$CH_2=CHCH(CH_3)_2$	-6.92
12.	$CH_3CH=C(CH_3)_2$	-10.17
13.	$CH_2=CH(CH_2)_3CH_3$	-9.96
14.	$CH_3CH=CHCH_2CH_2CH_3$ cis	-12.51
15.	$CH_3CH=CHCH_2CH_2CH_3$ trans	-12.88
16.	$CH_3CH_2CH=CHCH_2CH_3$ cis	-11.38
17.	$CH_3CH_2CH=CHCH_2CH_3$ trans	-13.01
18.	$CH_2=C(CH_3)CH_2CH_2CH_3$	-14.19
19.	$CH_2=CHCH(CH_3)CH_2CH_3$	-11.82
20.	$CH_2=CHCH_2CH(CH_3)_2$	-12.24
21.	$CH_3CH_2CH=C(CH_3)_2$	-15.98
22.	$CH_3CH=C(CH_3)CH_2CH_3$ cis	-14.86
23.	$CH_3CH=C(CH_3)CH_2CH_3$ trans	-15.08
24.	$CH_3CH=CHCH(CH_3)_2$ cis	-13.73
25.	$CH_3CH=CHCH(CH_3)_2$ trans	-14.69
26.	$CH_2=C(CH_3)_2$	-13.36

Structure Number	Structure	Heat of formation (kcal/mole)
27.	$CH_2=C(CH_3)CH(CH_3)_2$	-15.85
28.	$CH_2=CHC(CH_3)_3$	-14.70
29.	$(CH_3)_2C=C(CH_3)_2$	-16.68
30.	$CH_2=CH(CH_2)_4CH_3$	-14.89
31.	$CH_3CH=CH(CH_2)_3CH_3$ trans	-17.60
32.	$CH_3CH=CH(CH_2)_3CH_3$ cis	-16.90
33.	$CH_3CH_2CH=CHCH_2CH_2CH_3$ trans	-17.60
34.	$CH_3CH_2CH=CHCH_2CH_2CH_3$ cis	-16.90
35.	$CH_3CH_2CH=C(CH_3)CH_2CH_3$ cis	-18.60
36.	$CH_3CH_2CH=C(CH_3)CH_2CH_3$ trans	-19.22
37.	$CH_2=C(CH_3)CH_2CH(CH_3)_2$	-20.27
38.	$CH_2=CHCH_2C(CH_3)_3$	-19.20
39.	$(CH_3)_2C=CHCH(CH_3)_2$	-21.44
40.	$CH_3CH=CHC(CH_3)_3$ cis	-17.60
41.	$CH_3CH=CHC(CH_3)_3$ trans	-21.46
42.	$CH_2=C(CH_2CH_3)CH(CH_3)_2$	-19.25
43.	$CH_2=C(CH_3)C(CH_3)_3$	-20.67
44.	$CH_2=CH(CH_2)_5CH_3$	-19.82
45.	$CH_3CH_2CH=CHC(CH_3)_3$ cis	-21.77
46.	$CH_3CH_2CH=CHC(CH_3)_3$ trans	-26.16
47.	$CH_2=CH(CH_3)CH(CH_2CH_3)_2$	-24.40
48.	$CH_2=CH(CH_3)CH_2C(CH_3)_3$	-26.68
49.	$(CH_3)_2C=CHC(CH_3)_3$	-25.50
50.	$CH_2=C=CH_2$	45.90
51.	$CH_3CH=C=CH_2$	38.77

Structure Number	Structure	Heat of formation (kcal/mole)
52.	$CH_3CH_2CH=C=CH_2$	33.61
53.	$CH_3CH=C=CHCH_3$	31.79
54.	$CH_2=CHCH=CH_2$	26.33
55.	$CH_3CH=CHCH=CH_2$ cis	19.77
56.	$CH_3CH=CHCH=CH_2$ trans	18.77
57.	$CH_2=CHCH_2CH=CH_2$	25.41
58.	$CH_2=C(CH_3)CH=CH_2$	18.09
59.	$CH\equiv CH$	54.19
60.	$CH_3C\equiv CH$	44.32
61.	$CH_3CH_2C\equiv CH$	39.48
62.	$CH_3C\equiv CCH_3$	34.97
63.	$CH_3CH_2CH_2C\equiv CH$	34.50
64.	$CH_3CH_2C\equiv CCH_3$	30.80
65.	$CH\equiv CCH(CH_3)_2$	32.60
66.	$CH\equiv CCH_2CH_2C\equiv CH$	99.44
67.	$CH_3CH=CHC\equiv CH$ cis	60.60
68.	$CH_3CH=CHC\equiv CH$ trans	60.92
69.	$CH_3(CH_2)_5CH=CHC\equiv CH$ cis	36.25
70.	$CH_3(CH_2)_5CH=CHC\equiv CH$ trans	36.95

Table 24

Unsaturated Aliphatics Heats of Formation

Structures and Property Values

Structural feature set	Number of structural features	Number of included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
A	16	15 + constant	54	0.999	0.99	1797.30
B	17	16 + constant	53	>0.999	0.92	3310.01
C	17	16 + constant	53	>0.999	0.94	3310.01
D	18	17 + constant	52	>0.999	0.88	3056.53

Number of structures = 70

Range of property values = 126.12

Table 25

Unsaturated Aliphatics Heats of Formation

Overall Regression Results

Structural Feature	Regression Coefficient	t statistic (51 degrees of freedom)
CH_3-	0.64	0.95
-CH_2-	-4.74	50.86
-CH-	-11.99	17.82
-C-	-18.34	13.63
CH_2=CH_2 [a]	33.95	16.61
CH_2=CH-	24.83	22.83
-CH=CH-	16.89	23.50
CH_2=C-	15.93	24.27
-CH=C-	8.52	13.60
-C=C-	excluded by regression program	
CH_2=C=CH_2 [a]	67.37	32.97
CH_2=C=CH-	59.37	43.56
-CH=C=CH-	51.96	45.82
CH≡CH [a]	75.64	37.01
CH≡C-	65.22	61.07
-C≡C-	55.42	59.04
cis interaction	1.10	3.05
conjugation	-1.65	2.55
regression constant	-21.45	11.62

Note [a] unique feature representing single structure

Table 26

Unsaturated Aliphatics Heats of Formation

Regression results for structural feature set D

Structure Number	Structure	Heat of vaporization (kcal./g.f.w.)
1.	$CH_3(CH_2)_2CH_3$	5.02
2.	$CH_3CH(CH_3)CH_3$	4.61
3.	$CH_3(CH_2)_3CH_3$	6.39
4.	$CH_3CH_2CH(CH_3)_2$	6.03
5.	$CH_3(CH_2)_4CH_3$	7.54
6.	$CH_3(CH_2)_2CH(CH_3)_2$	7.14
7.	$CH_3CH_2CH(CH_3)CH_3CH_3$	7.24
8.	$CH_3CH(CH_3)CH(CH_3)_2$	6.96
9.	$CH_3CH_2C(CH_3)_3$	6.62
10.	$CH_3(CH_2)_5CH_3$	8.74
11.	$CH_3(CH_2)_3CH(CH_3)_2$	8.33
12.	$CH_3(CH_2)_2(CH_3)CH_2CH_3$	8.39
13.	$CH_3CH_2CH(CH_2CH_3)_2$	8.42
14.	$CH_3(CH_2)_2C(CH_3)_3$	7.75
15.	$CH_3CH_2CH(CH_3)CH(CH_3)_2$	8.18
16.	$CH_3CH(CH_3)CH_2CH(CH_3)_2$	7.86
17.	$CH_3CH_2C(CH_3)_2CH_2CH_3$	7.89
18.	$CH_3C(CH_3)_2CH(CH_3)_2$	7.66
19.	$CH_3(CH_2)_6CH_3$	9.92
20.	$CH_3(CH_2)_4CH(CH_3)_2$	9.48
21.	$CH_3(CH_2)_3CH(CH_3)CH_2CH_3$	9.52
22.	$CH_3CH_2CH_2CH(CH_3)CH_2CH_2CH_3$	9.48
23.	$CH_3CH_2CH_2CH(CH_2CH_3)_2$	9.48
24.	$CH_3(CH_2)_3C(CH_3)_3$	8.91
25.	$CH_3CH_2CH_2CH(CH_3)CH(CH_3)_2$	9.27
26.	$CH_3CH_2CH(CH_3)CH_2CH(CH_3)_2$	9.03

Structure Number	Structure	Heat of vaporization (kcal./g.f.w.)
27.	$CH_3CH(CH_3)CH_2CH_2CH(CH_3)_2$	9.05
28.	$CH_3CH_2CH_2C(CH_3)_2CH_2CH_3$	8.97
29.	$CH_3CH_2CH(CH_3)CH(CH_3)CH_2CH_3$	9.32
30.	$CH_3CH_2CH(CH_2CH_3)CH(CH_3)_2$	9.21
31.	$CH_3CH_2C(CH_2CH_3)_2CH_3$	9.08
32.	$CH_3CH_2CH(CH_3)C(CH_3)_3$	8.82
33.	$CH_3C(CH_3)_2CH_2(CH_3)_2$	8.40
34.	$CH_3CH_2C(CH_3)_2CH(CH_3)_2$	8.90
35.	$CH_3CH(CH_3)CH(CH_3)CH(CH_3)_2$	9.01
36.	$CH_3(CH_2)_7CH_3$	11.10
37.	$CH_3(CH_2)_8CH_3$	12.28
38.	CH_3OH	8.94
39.	CH_3CH_2OH	10.18
40.	CH_3CH_2OH	11.34
41.	$CH_3CH(OH)CH_3$	10.90
42.	$CH_3(CH_2)_3OH$	12.50
43.	$CH_3CH(CH_3)CH_2OH$	12.15
44.	$CH_3CH_2CH(OH)CH_3$	11.89
45.	$CH_3C(CH_3)_2OH$	11.14
46.	$CH_3(CH_2)_4OH$	13.61
47.	$CH_2CH_2CH_2CH(OH)CH_3$	12.56
48.	$CH_3(CH_2)_5OH$	15.00
49.	$CH_3(CH_2)_6OH$	16.20
50.	CH_3-CO-CH_3	7.37
51.	CH_3CH_2-CO-CH_3	8.34
52.	$CH_3CH_2CH_2$-CO-CH_3	9.14
53.	$CH_3CH(CH_3)$-CO-CH_3	8.82
54.	$CH_3CH_2CH_2$-CO-CH_2CH_3	10.01

Structure Number	Structure	Heat of vaporization (kcal./g.f.w.)
55.	CH_3CH_2-CO-$CH(CH_3)_2$	9.51
56.	CH_3CH_2-CO-$C(CH_3)_3$	10.12
57.	$CH_3CH(CH_3)$-CO-$CH(CH_3)_2$	9.93
58.	$CH_3C(CH_3)_2$-CO-$CH(CH_3)_2$	10.35
59.	CH_3-$(CH_2)_3$ CO $(CH_2)_3CH_3$	12.59
60.	$CH_3C(CH_3)_2$-CO-$C(CH_3)_3$	10.84
61.	$CH_3CH_2CH = CH_2$	4.92
62.	$CH_3CH = CH\ CH_3$ cis	5.40
63.	$CH_3CH = CH\ CH_3$ trans	5.16
64.	$CH_2 = C(CH_3)_2$	4.92
65.	$CH_3CH_2CH_2CH = CH_2$	6.09
66.	$CH_3CH_2CH = CH\ CH_3$ cis	6.41
67.	$CH_3CH_2CH = CH\ CH_3$ trans	6.38
68.	$CH_2 = C(CH_3)CH_2CH_3$	6.18
69.	$CH_2 = CH\ CH(CH_3)_2$	5.70
70.	$CH_3CH = C(CH_3)_2$	6.47
71.	$CH_3(CH_2)_3CH = CH_2$	7.34
72.	$CH_3CH_2CH_2CH = CH\ CH_3$ cis	7.54
73.	$CH_3CH_2CH_2CH = CH\ CH_3$ trans	7.56
74.	$CH_3CH_2CH = CHCH_2CH_3$ cis	7.49
75.	$CH_3CH_2CH = CHCH_2CH_3$ trans	7.56
76.	$CH_2 = C(CH_3)CH_2CH_2CH_3$	7.31
77.	$CH_2 = CHCH(CH_3)CH_2CH_3$	6.85
78.	$CH_2 = CHCH_2CH(CH_3)_2$	6.88
79.	$CH_3CH_2CH = CH(CH_3)_2$	7.57
80.	$CH_3CH = C(CH_3)CH_2CH_3$ cis	7.69
81.	$CH_3CH = C(CH_3)CH_2CH_3$ trans	7.51

Structure Number	Structure	Heat of vaporization (kcal./g.f.w.)
82.	$CH_3CH = CH\ CH(CH_3)_2$ cis	7.06
83.	$CH_3CH = CH\ CH(CH_3)_2$ trans	7.18
84.	$CH_2 = C(CH_3)CH(CH_3)_2$	6.99
85.	$CH_2 = CH\ C(CH_3)_3$	6.38
86.	$CH_3C(CH_3) = C(CH_3)_2$	7.80
87.	$CH_3(CH_2)_4CH = CH_2$	8.52
88.	$CH_3CH_2CH = C(CH_3)CH_2CH_3$ cis	8.73
89.	$CH_3CH_2CH = C(CH_3)CH_2CH_3$ trans	8.58
90.	$CH_2 = C(CH_3)CH_2CH(CH_3)_2$	7.93
91.	$CH_2 = CHCH_2C(CH_3)_3$	7.47
92.	$CH_3CH(CH_3)CH = C(CH_3)_2$	8.22
93.	$CH_3CH = CH\ C(CH_3)_3$ cis	7.81
94.	$CH_3CH = CH\ C(CH_3)_3$ trans	7.87
95.	$CH_3CH_2CH = CH\ C(CH_3)_3$ cis	8.88
96.	$CH_3CH_2CH = CH\ C(CH_3)_3$ trans	8.91
	benzene derivatives	
97.	H	8.09
98.	CH_3	9.08
99.	CH_2CH_3	10.10
100.	1,2 - $(CH_3)_2$	10.38
101.	1,3 - $(CH_3)_2$	10.20
102.	1,4 - $(CH_3)_2$	10.13
103.	$CH_2CH_2CH_3$	11.05
104.	$CH(CH_3)_2$	10.79
105.	1- CH_2CH_3,2-CH_3	11.40

Structure Number	Structure	Heat of vaporization (kcal./g.f.w.)
106.	1- CH_2CH_3,3-CH_3	11.21
107.	1- CH_2CH_3,4-CH_3	11.14
108.	1,2,3- $(CH_3)_3$	11.73
109.	1,2,4- $(CH_3)_3$	11.46
110.	1,3,5- $(CH_3)_3$	11.35
111.	1- OH,3-CH_3	14.75
112.	1- OH,2-CH_2CH_3	15.20
113.	1- OH,3-CH_2CH_3	16.30
114.	1- OH,2,4-$(CH_3)_2$	15.74
115.	F_6	8.53
116.	F_5	8.65
117.	1,2- F_2	8.65
118.	1,3- F_2	8.29
119.	1,4- F_2	8.51
120.	1,2- Cl_2	11.4
121.	1,3- Cl_2	11.1
122.	F	8.27
123.	Cl	9.63
124.	1- CH_3,F_5	9.78
125.	1- CH_3,4-F	9.42
126.	1- Cl,2-CH_2CH_3	11.3
127.	1- Cl,4-CH_2CH_3	11.5
	pyridine derivatives	
128.	H	9.61
129.	2- CH_3	10.15
130.	3- CH_3	10.62
131.	4- CH_3	10.83
132.	2,3- $(CH_3)_2$	11.70

Structure Number	Structure	Heat of vaporization (kcal./g.f.w.)
133.	2,4- $(CH_3)_2$	11.42
134.	2,5- $(CH_3)_2$	11.43
135.	2,6- $(CH_3)_2$	11.01
136.	3,4- $(CH_3)_2$	12.38
137.	3,5- $(CH_3)_2$	12.04

Table 27

Diverse Structures Heats of Vaporization

Structures and Property Values

Data Set	Number of Structures	Number of Structural Features	Number of variables included	Degrees of freedom	Multiple Correlation Coefficient	Residual Error	F Value
(1)	12	5	3 + const	8	0.995	0.22	264.67
(2)	11	5	3 + const	7	0.999	0.08	1164.91
(3)	23	6	4 + const	18	0.992	0.30	277.88
(4)	37	4	3 + const	33	0.997	0.12	1825.09
(5)	60	6	6	54	>0.999	0.27	4493.25

Table 28

Diverse Structures Heats of Vaporization

Overall Regression Results: Saturated Aliphatics

Structural feature set	Number of structural features	Number of variables included	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
A	9	8	28	>0.999	0.07	3497.37
B	10	9	27	>0.999	0.07	2997.75

Number of structures = 36

Range of property values = 3.99

Table 29

Diverse Structures Heats of Vaporization

Overall Regression Results for data subset (6)

Structural Feature	Regression Coefficient	t Statistic (28 degrees of freedom)
CH_3-	1.95	107.80
-CH_2-	1.15	71.26
-CH-	-0.09	2.17
-C-	-1.37	25.61
CH_2=CH-	1.90	39.98
-CH=CH-	1.36	28.38
CH_2=C-	1.10	20.38
-CH=C-	0.56	8.81
-C=C-	excluded by regression program	
regression constant	excluded by regression program	

Table 30

Diverse Structures Heats of Vaporization

Regression results for structural feature set A analysis of data subset (6)

Structural feature set	Number of structural features	Number of variables included	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
set C	8	7 + constant	23	0.996	0.23	408.25
set D	29	25 + constant	5	>0.999	0.03	249.85

Number of structures = 31

Range of property values = 8.21

Table 31

Diverse Structures Heats of Vaporization

Overall regression results for data subset (7)

Structural feature set	Number of structural features	Number of variables included	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
set E	2	1 + constant	8	0.881	0.43	27.74
set F	6	6 + constant	3	0.997	0.11	82.96

Number of structures = 10

Range of property values = 2.77

Table 32

Diverse Structures Heats of Vaporization

Overall regression results for data subset (8)

Structural features	Number of structural features	Number of variables included	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
set G	9	8 + constant	32	0.991	0.29	219.23
set H	18	17 + constant	23	0.998	0.15	337.22
set I	26	23 + constant	17	0.999	0.10	369.01
set J	25	23 + constant	17	0.999	0.15	369.01
set K	33	29 + constant	11	>0.999	0.07	379.03

Number of structures = 41

Range of property values = 8.21

Table 33

Diverse Structures Heats of Vaporization

Overall Regression Results for data subset (9)

Structural Feature	Regression Coefficient	t statistic (11 degrees of freedom	Perfectly correlated structural features
aromatic ring	common to all structures		
ring N	1.46	18.72	
CH_3-	0.99	17.59	
CH_3CH_2-	2.02	21.43	
$CH_3CH_2CH_2$-	2.96	31.54	
$-CH(CH_3)_2$	2.70	28.77	
-OH	6.20	61.56	
-F	0.22	4.97	
-Cl	1.54	16.42	
Me-ortho-ring N	-0.34	6.78	
Me-meta-ring N	0.20	4.06	
Me-para-ring N	0.26	3.99	
Me-ortho-Me	0.30	5.77	
Me-meta-Me	0.09	1.86	
Me-para-Me	0.03	0.47	
Me-ortho-Et	0.31	2.66	
Me-meta-Et	0.12	1.01	
Me-para-Et	0.05	0.41	
Me-ortho-OH	-0.62	4.80	Me-para-OH
Me-meta-OH	-0.53	4.09	
Et-ortho-OH	-1.10	10.92	
Et-meta-OH	excluded by regression program		
Me-ortho-F	0.03	0.45	Me-meta-F
Me-para-F	0.17	1.36	
Et-ortho-Cl	-0.34	2.49	

Structural Feature	Regression Coefficient	t statistic (11 degrees of freedom	Perfectly correlated structural features
Et-para-Cl	-0.14	1.04	
F-ortho-F	0.11	2.07	
F-meta-F	-0.25	4.50	
F-para-F	-0.03	0.34	
Cl-ortho-Cl	0.23	1.33	
Cl-meta-Cl	-0.07	0.43	
regression constant	8.09	132.10	

Table 34

Diverse Structures Heats of Vaporization

Regression Results for structural feature set K analysis of data subset (9)

Structural Feature	Regression Coefficient	t statistic (85 degrees of freedom)
$-CH_3$	1.61	32.08
$-CH_2-$	1.12	59.00
$-\overset{|}{C}H-$	0.22	2.38
$-\overset{|}{\underset{|}{C}}-$	-0.87	5.80
-OH	7.45	110.30
-C=O	3.41	40.08
$CH_2=CH-$	2.33	30.45
-CH=CH-	2.12	22.89
$CH_2=\overset{|}{C}-$	1.82	14.78
$-CH=\overset{|}{C}-$	1.62	10.49
$-\overset{|}{C}=\overset{|}{C}-$	1.37	4.57
regression constant	excluded by regression program	

Table 35

Diverse Structures Heats of Vaporization

Regression results for data subset (10)

(i) $CH_3CH_2CH_2CH_2CH_3$

(ii) $CH_3-CH(CH_3)-CH_2CH_3$

(iii) $CH_3-C(CH_3)_2-CH_3$

ΔH_v value	ΔH_v(i) - ΔH_v(ii)	ΔH_v(iii) - ΔH_v(i)
Observed	-0.42	-1.05
SSA subset (10)	-0.41	-1.02
Greenshield-Rossini	-0.19	-1.25
Laidler-Lovering	-0.32	-0.51

Table 36

Diverse Structures Heats of Vaporization

Comparisons of estimations (chain branching)

Structural feature set	Number of structural features	Number of variables included	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
set L	15	15 + constant	121	0.992	0.31	498.12
set M (99.99%)	39	37 + constant	99	0.997	0.21	443.94
set M (10.00%)	39	15 + constant	121	0.976	0.54	162.03

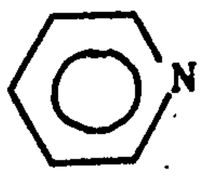
Number of structures = 137

Range of property values = 11.69

Table 37

Diverse Structures Heats of Vaporization

Overall regression results for data subset (11)

Structural Feature	Regression Coefficient	t statistic (99 degrees of freedom)	Perfectly correlated structural features
CH_3-	0.97	5.94	
-CH_2-	1.12	63.56	
-CH-	0.86	5.04	
-C-	0.40	1.21	
-OH	3.41	42.75	
F	0.26	2.01	
Cl	1.59	5.92	
CH_2=CH-	1.68	9.11	
-CH=CH-	2.11	24.45	
CH_2=C-	1.81	15.72	
-CH=C-	2.25	11.41	
-C=C-	2.63	6.58	
	6.74	32.21	
	8.27	24.62	
Me-ortho-ring N	-0.35	2.38	
Me-meta-ring N	0.20	1.33	
Me-para-ring N	0.26	1.35	
Me-ortho-Me	0.35	2.40	
Me-meta-Me	0.13	1.01	
Me-para-Me	0.09	0.49	
Me-ortho-Et	0.31	1.09	
Me-meta-Et	0.12	0.42	
Me-para-Et	0.05	0.18	
Me-meta-OH	-1.06	3.60	
Me-para-OH	-1.17	3.61	Me-ortho-OH
Et-ortho-OH	-1.73	5.85	

Structural Feature	Regression Coefficient	t statistic (99 degrees of freedom)	Perfectly correlated structural features
Et-meta-OH	-0.63	2.12	
F-ortho-F	0.11	0.65	
F-meta-F	-0.26	1.60	
F-para-F	-0.04	0.17	
Cl-ortho-Cl	0.18	0.35	
Cl-meta-Cl	-0.12	0.25	
Me-ortho-F	0.01	0.06	Me-meta-F
Me-para-F	0.16	0.65	
Et-ortho-Cl	-0.41	1.23	
Et-para-Cl	-0.22	0.64	

Table 38

Diverse Structures Heats of Vaporization

Regression Results for structural feature set M analysis of data subset (11) (99.99%)

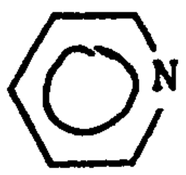
Structural Feature	Regression Coefficient	t statistic (121 degrees of freedom)
CH_3-	1.59	18.44
-CH_2-	1.01	24.77
-C-	-1.14	6.51
-OH	6.90	34.02
-C=O	3.03	15.58
F	1.14	5.75
Cl	2.45	12.04
CH_2=CH-	1.87	7.32
-CH=CH-	1.60	7.99
CH_2=C-	1.26	4.52
-CH=C-	1.04	4.51
	5.69	28.33
	8.39	24.91
Me-ortho-ring N	-0.72	2.79
F-meta-F	-0.86	4.08
regression constant	0.78	2.26

Table 39

Diverse Structures Heats of Vaporization

Regression Results for structural features of set M included in analysis of data subset (11) at the 10% level

Data subset	Structural Feature Set[a]	Predicted Value
(4)	-	8.67
(10)	-	8.82
(11)	set L	8.95
(11)	set M	8.82

structure: $CH_3-(CH_2)_5-CH_3$

observed value: 8.74

Note [a] where applicable

Table 40

Diverse Structures Heats of Vaporization

Property Prediction for n-heptane

Data subset	Structural Feature Set[a]	Predicted Value
(4)	-	7.84
(10)	-	7.81
(11)	set L	7.94
(11)	set M	7.81

structure: $CH_3CH_2CH_2C(CH_3)_3$

observed value: 7.75

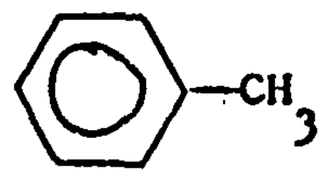
Note [a] where applicable

Table 41

Diverse Structures Heats of Vaporization

Property Prediction for 2,2 dimethyl pentane

Data subset	Structural Feature Set	Predicted Value
(7)	set C	9.28
(7)	set D	9.10
(9)	set G	9.28
(9)	set K	9.08
(11)	set L	9.04
(11)	set M	8.99

structure 

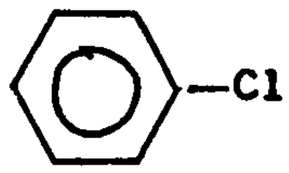
observed value: 9.08

Table 42

Diverse Structures Heats of Vaporization

Property Prediction for Toluene

Data subset	Structural Feature Set	Predicted Value
(7)	set C	9.67
(7)	set D	9.58
(9)	set G	9.67
(9)	set K	9.59
(11)	set L	9.58
(11)	set M	9.57

structure: 

observed value: 9.63

Table 43

Diverse Structures Heats of Vaporization

Property Predictions for Chlorobenzene

Data subset	Structural Feature Set[a]	Predicted Value
(2)	-	9.20
(3)	-	9.15
(5)	-	8.87
(10)	-	8.84
(11)	set L	8.92
(11)	set M	8.84

structure: $CH_3CH_2CH_2$-CO-CH_3

observed value: 9.14

Note [a] where applicable

Table 44

Diverse Structures Heats of Vaporization

Property Predictions for pentan-2-one

Data subset	Structural Feature Set[a]	Predicted Value
(1)	-	12.55
(3)	-	12.52
(5)	-	12.42
(10)	-	12.41
(11)	set L	12.17
(11)	set M	12.61

structure: $CH_3CH_2CH_2CH_2OH$

observed value: 12.50

Note [a] where applicable

Table 45

Diverse Structures Heats of Vaporization

Property Prediction of n-butanol

Data subset	Structural Feature Set	Predicted Value
(6)	set A	5.23
(6)	set B	5.24
(11)	set L	5.44
(11)	set M	5.34

structure: $CH_3CH=CHCH_3$

observed value: 5.40

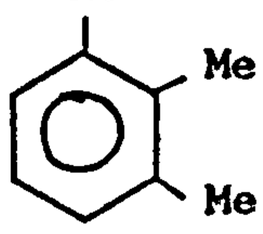
Table 46

Diverse Structures Heats of Vaporization

Property Prediction for cis-2-butene

Data subset	Structural Feature Set	Predicted Value
(9)	set G	11.35
(9)	set K	11.81
(11)	set L	10.92
(11)	set M	11.83

structure:



observed value: 11.73

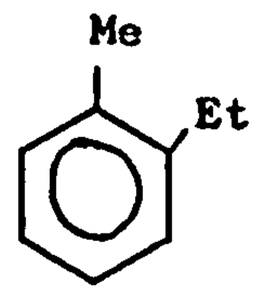
Table 47

Diverse Structures Heats of Vaporization

Property Prediction for 1,2,3 trimethyl benzene

Data subset	Structural Feature Set	Predicted Value
(9)	set G	11.14
(9)	set K	11.09[a]
		11.40[b]
(11)	set L	11.09
(11)	set M	11.09[a]
		11.44[b]

structure:



observed value: 11.40

Notes [a] ignoring Me-ortho-Et

[b] approximating Me-ortho-Et as equivalent to Me-ortho-Me

Table 48

Diverse Structures Heats of Vaporization

Property Prediction for 1-methyl,2-ethyl benzene

Source of value	n-heptane	2,2diMepentane	toluene	chlorobenzene	pentan-2-one	n-butanol
Observed	8.74	7.75	9.08	9.63	9.14	12.50
SSA (subset (11) (set M)	8.82	7.81	8.99	9.57	8.84	12.41
SSA (appropriate subset)	8.67	7.84	9.08	9.59	9.20	12.55
Laidler-Lovering	8.75	8.24	9.04	-	12.04	11.72
CH_2 increment	8.80	7.79	8.98	-	8.89	12.49
Wright's equation	8.65	7.83	9.36	10.18	9.67	-
Fightine's equation	8.76	9.97	9.15	10.10	9.40	12.55
Chen's equation	8.86	7.82	9.34	10.23	9.39	12.64

Table 49

Diverse Structures Heats of Vaporization

Comparison of simulated prediction methods

Structure number	Substituents	pKa	log solubility	log toxicity
2718	H	8.79	2.71	2.17
2727	5-CH_3	8.90	-	2.50
7489	5-C_2H_5	8.81	-	2.11
4098	5-tC_4H_9	9.12	-	2.16
3895	5-F	8.00	2.63	1.91
3990	4-Cl	7.57	2.88	1.94
2785	5-Cl	7.97	2.23	1.49
4307	4-Br	7.37	-	2.02
4173	5-Br	7.89	-	2.27
4762	5-I	7.71	-	2.49
4005	5-CN	7.16	2.73	1.32
3391	5-CF_3	7.52	1.98	1.55
4025	5-COOH	8.36	2.78	-
6727	5-$COOCH_3$	10.80	2.66	1.98
6717	5-$COOC_2H_5$	7.81	2.15	2.19
6735	5-$CONH_2$	7.78	3.18	2.14
6704	5-OH	9.00	3.34	2.31
4027	5-NH_2	4.54	-	2.52
4166	5-$NHCOCH_3$	8.30	1.88	1.99
4083	4-NO_2	6.84	2.93	1.66
2814	5-NO_2	6.68	2.32	1.20
10157	5-SO_2CH_3	6.90	3.45	2.42
4765	5-SO_2NH_2	7.12	-	1.90
4347	5-C_6H_5	8.36	1.95	2.42
3522	5,6-$ICH_3)_2$	9.18	1.23	2.53
6099	4,6-F_2	7.28	2.64	1.05

Structure number	Substituents	pKa	log solubility	log toxicity
5940	4-F;5-Cl	7.12	-	2.00
6082	4-F;6-Cl	6.76	1.97	1.49
6698	5-F;6-Cl	7.66	1.41	1.56
5876	4-Cl;6-F	7.11	2.66	0.92
5342	4-Br;6-F	7.10	2.85	1.30
3363	$4,5\text{-Cl}_2$	6.96	1.84	1.71
2983	$5,6\text{-Cl}_2$	7.40	1.32	1.49
3531	$4,6\text{-Cl}_2$	6.78	1.82	1.11
3048	$4,7\text{-Cl}_2$	6.24	1.56	1.01
5263	4-Cl;6-Br	6.69	2.18	1.33
5086	4-Br;6-Cl	6.76	1.90	1.38
7972	5-Br;6-Cl	7.25	1.56	1.48
6098	4-Cl;6-I	6.74	2.00	1.62
5500	4-I;6-Cl	6.82	2.11	1.84
4342	$4,6\text{-Br}_2$	6.77	2.00	1.24
9309	$5,6\text{-Br}_2$	7.26	1.43	-
4006	$4,7\text{-Br}_2$	5.99	1.48	1.72
4379	$4,5\text{-Br}_2$	6.73	-	-
4925	$4,6\text{-I}_2$	6.85	1.30	1.84
8452	4-OH,5-Cl	6.94	3.48	-
8613	5-Cl;6-OH	7.80	3.08	2.37
6920	$4\text{-OCH}_3\text{;5-Cl}$	7.63	2.38	1.10
5737	$4\text{-CH}_3\text{;5-Cl}$	8.22	1.90	1.97
4499	$5\text{-CH}_3\text{;6-Cl}$	8.26	1.92	2.37
5253	4-Br;5-CH_3	7.75	2.04	2.36
5792	$5\text{-CH}_3\text{;6-Br}$	8.22	1.48	2.44

Structure number	Substituents	pKa	log solubility	log toxicity
6801	4-Br;6-tC_4H_9	7.98	0.30	2.70
6700	4-Cl;6-CN	5.82	-	0.99
6917	4-CN;6-Cl	6.12	2.38	1.29
5855	4-CF_3;6-Cl	6.52	2.08	0.76
4892	4-NO_2;5-CH_3	7.10	2.43	1.50
5449	5-NO_2;6-CH_3	7.04	2.04	-
4317	4-NO_2;5-Cl	5.98	2.41	1.57
3534	4-NO_2;6-Cl	6.33	2.30	1.02
3946	4-Cl;6-NO_2	5.48	3.11	1.57
4064	5-Cl;6-NO_2	6.20	-	1.32
4570	4-Cl;7-NO_2	5.86	2.36	1.62
4782	4-NO_2;5-Br	5.90	-	1.64
4466	4-NO_2;6-Br	6.04	2.15	0.60
4487	4-Br;6-NO_2	5.45	-	1.67
4420	4,6-$(NO_2)_2$	4.96	-	1.62
3047	5,6-$(NO_2)_2$	4.96	2.48	1.74
3963	4-$NHCOCF_3$;6-SH	5.57	-	-
5585	4,7-$(OH)_2$	-	-	2.12
4066	5,6-$(OCH_3)_2$	8.95	-	2.23
5450	4,7-$(OCH_3)_2$	-	2.00	2.39
7925	4,5,6-F_3	6.69	2.93	-
7069	4,5,7-F_3	6.07	3.00	0.98
8589	4-Br;5-Cl;6-F	6.52	1.11	1.40
2786	4,5,6-Cl_3	6.18	1.63	0.72
2813	4,5,7-Cl_3	5.64	1.40	1.34
5381	4,5-Cl_2;6-Br	6.24	1.67	0.92

Structure number	Substituents	pka	log solubility	log toxicity
7960	4,6-Cl;5-Br	6.35	1.92	1.42
7021	4-Br;5,6-Cl_2	6.38	1.32	0.82
8246	4,6-Br_2;5-Cl	6.19	1.57	1.35
9294	4,5-Br_2;6-Cl	6.17	1.30	-
4690	4,5,7-Br_3	5.58	2.08	1.48
7974	4,5,6-Br_3	6.31	1.28	1.23
4348	4,6-Cl_2;5-CH_3	7.04	1.65	-
7355	4,7-Cl_2;5-CH_3	6.44	1.57	-
8614	5,6-Cl_2;4-OH	6.12	3.08	2.28
8489	4,6-Cl_2;5-OH	6.35	3.11	2.33
8380	4,6-Cl_2;5-OCH_3	6.66	2.04	1.45
8530	5,6-Cl_2;4-OCH_3	7.08	1.49	-
8528	4,5-Cl_2;7-OCH_3	6.29	1.08	-
10156	4-$OCONHC_2H_5$;5,6-Cl_2	9.71	2.04	2.53
10155	4-$OCON(CH_3)_2$;5,6-Cl_2	6.61	-	2.13
6073	4-CH_3;5-Cl;6-Br	7.43	1.54	1.98
4179	4-NO_2;5,6-CH_3	7.14	1.74	2.41
5428	4-NO_2;5-CH_3;6-Cl	6.21	1.50	2.05
5443	4-NO_2;5-CH_3;6-Br	6.16	1.54	2.41
4069	4-NO_2;5,6-Cl_2	5.25	-	2.48
4558	4,5-Cl_2;6-NO_2	5.24	-	2.26
4433	4,6-Cl_2;5-NO_2	4.86	2.20	2.08
4255	4,7-Cl_2;5-NO_2	4.42	2.67	-
4559	4,6(NO_2)$_2$;5-Cl	4.22	-	-
7621	4,5-Cl_2;6-CN	5.14	1.30	1.15
10815	4,5-Cl_2;6-SO_2CH_3	5.10	-	-

Structure number	Substituents	pKa	log solubility	log toxicity
11158	4,5-Cl_2;6-n$SO_2C_3H_7$	5.10	2.08	-
6787	4,5-Cl_2;7-OH	9.72	-	-
7372	4,5,6,7-F_4	5.37	2.71	1.11
7927	5-Cl;4,6,7-F_3	5.42	2.26	-
6100	4-F;5,6,7-Cl_3	5.36	1.23	0.44
6149	4,6-F_2;5,7-Cl_2	4.94	-	0.76
4178	4,6,7-Cl_3;5-F	5.38	1.48	0.89
6133	4,6-F_2;5,7-Br_2	5.31	-	1.28
5858	4,5,7-Br_3;6-F	5.24	1.48	1.34
7429	4,6,7-F_3;5-OCH_3	5.82	2.67	1.03
7695	4,6,7-F_3;5-CH_3	4.76	2.57	1.19
6132	4-Br;5,7-Cl;6-Br	5.50	1.75	1.25
5960	4-Cl;5,7-Br;6-F	5.33	1.41	1.20
2265	4,5,6,7-Cl_4	6.04	1.34	-0.19
3447	4-Br;5,6,7-Cl_3	5.27	1.49	1.05
5065	4,6,7-Cl_3;5-Br	5.04	1.34	1.11
5517	4,5-Cl_2;6,7-Br_2	5.40	1.65	0.92
5257	4,6-Cl_2;5,7-Br_2	5.31	0.95	-0.92
5690	4,7-Cl_2;5,6-Br_2	5.14	0.95	0.92
5924	4,7-Br_2;5,6-Cl_2	5.27	1.00	0.92
5824	4-Cl;5,6,7-Br_3	5.28	0.85	1.06
5739	4,5,7-Br_3;6-Cl	5.25	1.68	1.26
4792	4,5,6,7-Br_4	5.80	0.60	1.00
6061	4-CH_3;5,6,7-Cl_3	6.42	1.52	1.33
3234	4,6,7-Cl_3;5-CH_3	5.96	-	1.78
6331	4-CH_3;5,7-Cl_2;6-Br	6.46	1.04	1.72

Structure number	Substituents	pKa	log solubility	log toxicity
5738	4-CH_3;5-Cl;6,7-Br	6.54	0.90	1.77
9430	4-OCH_3;5,6,7-Cl_3	5.92	1.34	-
8529	4,6,7-Cl_3;5-OCH_3	5.57	1.58	2.01
6379	4,7-Cl_2;5-CH_3;6-OCH_3	6.32	1.51	-
3853	4,6,7-Cl_3;5-NO_2	4.14	2.18	1.70
4008	4,5,6-Cl_3;7-NO_2	4.32	1.26	1.82
3944	4,6-Cl_2;5,7-$(NO_2)_2$	3.40	2.43	2.53
3945	4,7-Cl_2;5,6-$(NO_2)_2$	2.96	2.48	-
4245	4,7-NO_2;5,6-Cl_2	3.20	-	-
5312	4,7-NO_2;5-CH_3;6-Cl	4.04	2.41	2.51
10158	4,6,7-Cl;5-NH_2	6.33	2.18	1.88
9729	4-NH_2;5,6,7-Cl_3	6.60	1.28	-
10354	4-$NHCOC_6H_5$;5,6,7-Cl_3	5.86	-	2.39
6123	4-$N(CH_3)_2$;5,6,7-Cl_3	6.25	0.69	1.82
11362	4-$N(C_2H_5)_2$;5,6,-Cl;7-Br	5.90	-	-
4378	4-SO_3H;5,6,7-Cl_3	7.04	-	2.77
4718	4-SO_2NH_2;5,6,7-Cl_3	5.92	-	2.35
10356	4-$SO_2N(CH_3)_2$;5,6,7-Cl_3	6.12	-	1.08
10355	4-$SO_2N(C_2H_5)_2$;5,6,7-Cl_3	6.12	1.11	1.42
10826	4-$SO_2N/(CH_2)_2OH/_2$;5,6,7-Cl_3	5.80	2.40	-
4695	4,5,6,7-$(CH_3)_4$	9.26	1.11	2.38
8135	4-OH;5,6,7-Cl_3	5.59	2.89	-
8131	4,6,7-Cl_3;5-OH	5.48	3.11	-
8051	4,7-$(OH)_2$;5,6-Cl_3	4.93	2.36	-

Structure number	Substituents	pKa	log solubility	log toxicity
8054	$4,7\text{-}(OCH_3)_2;5,6\text{-}Cl_2$	6.62	2.32	2.50
8087	$4,7\text{-}Cl_2;5,6\text{-}(OCH_3)_2$	6.00	1.89	2.50
9991	5-NHCOEt	-	-	1.54

Table 50

Structures and Property Values for Benzimidazole Derivatives

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
A	33	31 + constant	122	0.930	0.54	25.19
B	45	43 + constant	110	0.947	0.50	22.23
C	112	89 + constant	64	0.964	0.54	9.45
D	136	99 + constant	54	0.990	0.31	26.86

Number of structures = 154

Range of property values = 7.8 log units

Table 51

Overall regression analysis for pKa values of benzimidazole derivatives

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
A	25	24 + constant	98	0.863	0.38	11.92
B	36	35 + constant	87	0.890	0.36	9.47
D	107	83 + constant	39	0.961	0.33	5.67

Number of structures = 123

Range of property values = 3.18 log units

Table 52

Overall regression analysis results for solubilities of benzimidazole derivatives

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
A	29	28 + constant	100	0.784	0.41	5.67
B	40	39 + constant	89	0.812	0.41	4.42
D	118	89 + constant	39	0.953	0.32	4.34

Number of structures = 129

Range of property values = 3.0 log units

Table 53

Overall regression analysis results for toxicities of benzimidazole derivatives

Structural Feature	Regression Coefficient	t statistic (12 degrees of freedom)	Perfectly correlated structural features
ring system	common to all structures		
Me-o-RF	0.33	0.90	
Me-m-RF	0.36	1.25	
Et-m-RF	0.23	0.62	
CMez-m-RF	0.54	1.45	
Ph-m-RF	-0.22	0.59	
F-o-RF	-0.87	3.32	
F-m-RF	-0.39	1.59	
Cl-o-RF	-0.10	5.92	
Cl-m-RF	-0.64	3.87	
Br-o-RF	-1.22	6.66	
Br-m-RF	-0.82	4.40	
I-o-RF	-1.15	1.92	
I-m-RF	-0.87	2.33	
CF_3-o-RF	-1.42	4.19	Cl-m-CF_3
CF_3-m-RF	-1.06	2.84	
CN-o-RF	-1.48	2.47	
CN-m-RF	-1.42	3.80	
COOH-m-RF	-0.22	0.59	
OH-o-RF	-0.37	1.05	
OH-m-RF	0.95	2.91	
OMe-o-RF	-0.09	0.25	
OMe-m-RF	0.14	0.25	
NO_2-o-RF	-1.86	8.33	
NO_2-m-RF	-1.87	8.84	
NH_2-o-RF	-1.97	4.69	Cl-p-NH_2
NH_2-m-RF	-4.04	10.81	

Structural Feature	Regression Coefficient	t statistic (12 degrees of freedom)	Perfectly correlated structural features
NMe_2-o-RF	0.01	0.01	Cl-o-NMe_2, Cl-m-NMe_2, Cl-p-NMe_2
NHCOMe-m-RF	-0.28	0.75	
NHCOPh-o-RF	-0.39	1.13	Cl-o-NHCOPh, Cl-m-NHCOPh, Cl-p-NHCOPh
SH-m-RF	-3.01	8.06	$NHCOCF_3$-o-RF, SH-m-$NHCOCF_3$
$CONH_2$-m-RF	-0.80	2.14	
OCOMe-m-RF	2.22	5.94	
OCOEt-m-RF	-0.77	2.06	
OCONHEt-o-RF	2.37	6.86	Cl-o-OCONHEt, Cl-m-OCONHEt
$OCONMe_2$-o-RF	-0.73	2.12	Cl-o-$OCONMe_2$, Cl-m-$OCONMe_2$
SO_3H-o-RF	0.80	2.32	Cl-o-SO_3H, Cl-m-SO_3H, Cl-p-SO_3H
SO_2Me-m-RF	-1.68	4.50	
SO_2Pr-m-RF	-1.88	5.44	Cl-o-SO_2Pr, Cl-m-SO_2Pr
SO_2NH_2-o-RF	-0.33	0.95	Cl-o-SO_2NH_2, Cl-m-SO_2NH_2, Cl-p-SO_2NH_2
SO_2NH_2-m-RF	-1.46	3.91	
SO_2NMe_2-o-RF	-0.13	0.37	Cl-o-SO_2NMe_2, Cl-m-SO_2NMe_2, Cl-p-SO_2NMe_2
SO_2NEt_2-o-RF	-0.13	0.37	Cl-o-SO_2NEt_2, Cl-m-SO_2NEt_2
$SO_2N(C_2H_4OH)_2$-o-RF	-0.45	1.30	Cl-o-$SO_2N(C_2H_4OH)_2$, Cl-m-$SO_2N(C_2H_4OH)_2$, Cl-p-$SO_2N(C_2H_4OH)_2$

Structural Feature	Regression Coefficient	t statistic (12 degrees of freedom)	Perfectly correlated structural features
Me-o-Me	-0.15	0.28	
Me-m-Me	-0.13	0.21	Me-p-Me
Me-o-OMe	-0.11	0.18	
OMe-o-OMe	-0.02	0.02	
OMe-p-OMe	0.16	0.10	
Me-o-NO_2	-0.08	0.29	
Me-m-NO_2	-0.06	0.14	
NO_2-o-NO_2	0.19	0.57	
NO_2-m-NO_2	0.07	0.24	
NO_2-p-NO_2	-0.39	1.00	
OH-p-OH	-0.54	0.85	
F-o-F	-0.02	0.09	
F-m-F	-0.24	0.50	
F-p-F	-0.13	0.32	
F-o-Cl	0.01	0.01	
F-m-Cl	-0.03	0.15	
F-p-Cl	-0.18	0.49	
F-o-Br	-0.09	0.50	
F-m-Br	0.10	0.38	
F-p-Br	0.35	0.57	
Cl-o-Cl	0.04	0.30	
Cl-m-Cl	-0.14	0.86	
Cl-p-Cl	-0.12	0.51	
Cl-o-Br	0.11	0.85	
Cl-m-Br	-0.01	0.01	

Structural Feature	Regression Coefficient	t statistic (12 degrees of freedom)	Perfectly correlated structural features
Cl-p-Br	0.06	0.26	
Cl-m-I	0.03	0.06	
Br-o-Br	0.22	1.36	
Br-m-Br	0.18	0.86	
Br-p-Br	0.01	0.06	
I-m-I	0.29	0.33	
F-o-OMe	-0.19	0.56	F-m-OMe
F-o-CF_3	-0.12	0.44	F-m-CF_3
Cl-o-Me	-0.04	0.20	
Cl-m-Me	-0.20	0.64	
Cl-p-Me	0.08	0.20	
Cl-o-CN	-0.08	0.18	
Cl-m-CN	-0.34	0.73	
Cl-o-NO_2	0.03	0.21	
Cl-m-NO_2	-0.04	0.23	
Cl-p-NO_2	0.08	0.32	
Cl-o-NH_2	2.32	8.94	
Cl-m-NH_2	excluded by regression program		
Cl-o-OH	-0.91	4.18	
Cl-m-OH	0.34	1.07	
Cl-p-OH	1.52	3.71	
Cl-o-OMe	-0.09	0.30	
Cl-m-OMe	-0.21	0.73	
Cl-p-OMe	-0.16	0.37	
Cl-m-SO_2Me	-0.20	0.41	Cl-o-SO_2Me

Structural Feature	Regression Coefficient	t statistic (12 degrees of freedom)	Perfectly correlated structural features
Cl-p-SO_2NEt_2	excluded by regression program		
Br-o-Me	0.03	0.10	
Br-m-Me	-0.09	0.23	
Br-p-Me	0.12	0.25	
Br-m-CMe_3	0.08	0.17	
Br-o-NO_2	-0.01	0.01	
Br-m-NO_2	0.07	0.24	
Br-p-SO_2NEt_2	-0.20	0.42	
regression constant	8.58	41.07	

Notes RF = ring fusion point

O = ortho, m = meta, p = para

Table 54

pKa data for benzimidazole derivatives

Regression resultsfor set D structural features

Structural Feature	Regression coefficient	t statistic (98 degrees of freedom)	Perfectly correlated structural features
ring system	common to all structures		
Me	-0.43	5.86	
CMe_3	-1.85	4.78	
Ph	-0.59	1.49	
F	0.03	0.51	
Cl	-0.30	6.86	
Br	-0.38	7.51	
I	-0.47	2.82	
CF_3	-0.26	1.10	
OH	0.79	6.12	
OMe	-0.05	0.43	
NH_2	0.08	0.30	
NMe_2	-0.95	2.44	
$NHCOCH_3$	-0.67	1.69	
COOH	0.24	0.62	
OCOMe	0.12	0.31	
OCOEt	-0.39	0.98	
$CONH_2$	0.64	1.63	
OCONHEt	0.10	0.25	
NO_2	0.17	1.97	
SO_2Me	0.91	2.31	
SO_2Pr	0.14	0.36	
SO_2NEt_2	-0.54	1.38	
$SO_2N(C_2H_4OH)_2$	0.75	1.95	

Structural Feature	Regression coefficient	t statistic (98 degrees of freedom)	Perfectly correlated structural features
CN	-0.10	0.44	
regression constant	2.54	21.82	

Table 55

Solubility data for benzimdazole derivatives

Regression results for set A structural features

Structural Feature	Regression coefficient	t statistic (39 degrees of freedom)	Perfectly correlated structural features
ring system	common to all structures		
Me-oRF	0.23	0.48	
Me-m-RF	-0.13	0.29	
Ph-m-RF	-0.79	1.92	
CMe_3-m-RF	-2.14	5.49	Br-m-CMe_3
F-o-RF	-0.04	0.10	
F-m-RF	-0.18	0.69	
Cl-o-RF	-0.13	0.62	
Cl-m-RF	-0.82	4.12	
Br-o-RF	-0.30	1.35	
Br-m-RF	-0.60	2.85	
I-o-RF	-0.32	0.98	
I-m-RF	-1.12	3.44	
CF_3-o-RF	0.17	0.46	Cl-m-CF_3
CF_3-m-RF	-0.76	1.84	
OH-o-RF	0.23	3.17	
OH-m-RF	0.70	1.84	
OMe-o-RF	-0.26	0.45	
OMe-m-RF	-1.15	1.15	
CN-o-RF	excluded by regression program		
CN-m-RF	-0.01	0.01	
NH_2-o-RF	-0.70	1.76	Cl-p-NH_2
NH_2-m-RF	excluded by regression program		
NMe_2-o-RF	-0.94	2.58	Cl-o-NMe_2, Cl-m-NMe_2, Cl-p-NMe_2

Structural Feature	Regression coefficient	t statistic (39 degrees of freedom)	Perfectly correlated structural features
NO_2-o-RF	0.18	0.67	
NO_2-m-RF	-0.01	0.04	
COOH-m-RF	0.05	0.11	
OCOMe-m-RF	-0.08	0.18	
OCOEt-m-RF	-0.59	1.43	
$CONH_2$-m-RF	0.45	1.09	
OCONHEt-o-RF	0.72	1.95	Cl-o-OCONHEt, Cl-m-OCONHEt
NHCOMe-m-RF	-0.87	2.11	
SO_2Me-m-RF	0.72	1.75	
SO_2Pr-m-RF	0.06	0.17	Cl-o-SO_2Pr, Cl-m-SO_2Pr
SO_2NEt_2-o-RF	-0.53	1.45	Cl-o-SO_2NEt_2, Cl-m-SO_2NEt_2, Cl-p-SO_2NEt_2
$SO_1N(C_2H_4OH)_2$-o-RF	0.76	2.09	Cl-o-$SO_2N(C_2H_4OH)_2$, Cl-m-$SO_2N(C_2H_4OH)_2$, Cl-p-$SO_2N(C_2H_4OH)_2$
Me-o-Me	-1.24	1.38	
Me-m-Me	0.96	1.16	Me-p-Me
Me-o-NO_2	-0.46	1.22	
Me-m-NO_2	0.79	1.32	
Me-o-OMe	0.47	0.65	
OMe-o-OMe	1.30	1.13	
OMe-m-OMe	0.49	0.72	
OH-p-OH	-2.59	3.87	
NO_2-o-NO_2	-0.27	0.66	
NO_2-m-NO_2	0.48	0.89	

Structural Feature	Regression coefficient	t statistic (39 degrees of freedom)	Perfectly correlated structural features
NO_2-p-NO_2	-0.41	0.64	
F-o-F	-0.03	0.11	
F-m-F	0.12	0.24	
F-p-F	0.37	0.82	
F-o-Cl	-0.07	0.42	
F-m-Cl	0.21	0.75	
F-p-Cl	-0.51	1.03	
F-o-Br	-0.14	0.70	
F-m-Br	0.31	1.00	
Cl-o-Cl	0.23	1.38	
Cl-m-Cl	0.22	1.11	
Cl-p-Cl	-0.61	2.34	
Cl-o-Br	0.07	0.51	
Cl-m-Br	0.26	1.48	
Cl-p-Br	-0.27	1.07	
Cl-m-I	0.51	1.72	
Br-o-Br	-0.13	0.72	
Br-m-Br	0.24	0.92	
Br-p-Br	-0.09	0.30	
I-m-I	excluded by regression program		
Cl-o-Me	-0.12	0.39	
Cl-m-Me	0.07	0.18	
Cl-p-Me	-0.42	1.00	
Br-o-Me	-0.33	0.78	
Br-m-Me	-0.08	0.18	

Structural Feature	Regression coefficient	t statistic (39 degrees of freedom)	Perfectly correlated structural features
Br-p-Me	-0.34	0.69	
Cl-o-OMe	0.64	1.32	
Cl-m-OMe	-0.12	0.34	
Cl-p-OMe	-0.56	1.16	
Cl-o-OH	0.28	1.07	
Cl-m-OH	0.31	0.88	
Cl-p-OH	-0.56	1.19	
Cl-o-NO_2	0.17	1.02	
Cl-m-NO_2	0.30	1.54	
Cl-p-NO_2	-0.73	2.20	
Br-m-NO_2	-0.01	0.03	
Cl-o-CN	-1.18	1.71	
Cl-m-CN	0.47	1.29	
Cl-o-NH_2	0.34	1.88	
Cl-m-NH_2	excluded by regression program		
F-o-OMe	0.44	0.80	F-m-OMe
F-o-CF_3	0.19	0.62	F-m-CF_3
regression constant	2.74	11.05	

Notes RF = ring fusion point

o = ortho, m = meta, p = para

Table 56

Solubility data for benzimidazole derivatives

Regression results for structural feature set D

Structural feature	Regression coefficient	t statistic (89 degrees of freedom)
Me-o-RF	0.16	1.08
Me-m-RF	0.46	4.02
Et-m-RF	0.46	1.08
CMe_3-m-RF	0.84	2.71
Ph-m-RF	0.77	1.81
F-o-RF	-0.33	2.66
F-m-RF	-0.07	0.47
Cl-o-RF	-0.21	3.00
Cl-m-RF	-0.08	1.12
Br-o-RF	-0.12	1.33
Br-m-RF	-0.10	1.00
I-o-RF	0.03	0.11
I-m-RF	0.39	1.42
CF_3-o-RF	-0.81	1.93
CF_3-m-RF	0.08	0.26
CN-o-RF	-0.28	0.67
CN-m-RF	-0.33	1.29
OCOMe-m-RF	0.33	0.78
OCOEt-m-RF	0.54	1.27
OCONHEt-o-RF	1.04	2.45
$OCONMe_2$-o-RF	0.64	1.51
OH-o-RF	0.35	1.81
OH-m-RF	0.81	3.22
OMe-o-RF	0.34	2.31
OMe-m-RF	0.43	3.14
NO_2-o-RF	0.09	0.76
NO_2-m-RF	0.22	1.65

Structural Feature	Regression coefficient	t statistic (89 degrees of freedom)
NH_2-m-RF	0.80	2.65
NMe_2-o-RF	0.55	1.30
$CONH_2$-m-RF	0.49	1.15
NHCOMe-m-RF	0.34	0.80
NHCOEt-m-RF	-0.11	0.26
NHCOPh-o-RF	1.12	2.66
SO_2NH_2-o-RF	1.08	2.57
SO_2NH_2-m-RF	0.25	0.59
SO_2NMe_2-o-RF	-0.20	0.47
SO_2NEt_2-o-RF	0.15	0.35
SO_3H-o-RF	1.50	3.57
SO_2Me-m-RF	0.77	1.81
regression constant	1.65	13.60

Table 57

Toxicity data for benzimidazole derivatives.

Regression results for set B structural features

Structural Feature	Regression coefficient	t statistic (39 degrees of freedom)	Perfectly correlated structural features
ring system	common to all structures		
Me-o-RF	-0.18	0.45	
Me-m-RF	0.07	0.22	
Et-m-RF	-0.14	0.35	
CMe_3-m-RF	-0.09	0.22	
Ph-m-RF	0.17	0.44	
F-o-RF	-0.43	1.34	
F-m-RF	-0.56	1.85	
Cl-o-RF	-0.35	1.81	
Cl-m-RF	-0.52	2.81	
Br-o-RF	-0.30	1.37	
Br-m-RF	-0.38	1.75	
I-o-RF	0.63	1.02	
I-m-RF	0.24	0.62	
CF_3-o-RF	-0.97	2.77	Cl-m-CF_3
CF_3-m-RF	-0.70	1.78	
NO_2-o-RF	-0.84	3.35	
NO_2-m-RF	-0.82	3.18	
OH-o-RF	-0.06	0.90	
OH-m-RF	0.07	0.18	
OMe-o-RF	-0.23	0.37	
OMe-m-RF	1.12	1.35	
CN-o-RF	-0.46	0.73	
CN-m-RF	-0.93	2.36	
$CONH_2$-m-RF	-0.11	0.27	
OCOMe-m-RF	-0.27	0.68	

Structural Feature	Regression coefficient	t statistic (39 degrees of freedom)	Perfectly correlated structural features
OCOEt-m-RF	-0.06	0.14	
OCONHEt-o-RF	1.16	3.25	Cl-o-OCONHEt, Cl-m-OCONHEt
$OCONMe_2$-o-RF	0.77	2.14	Cl-o-$OCONMe_2$, Cl-m-$OCONMe_2$
NH_2-m-RF	0.27	0.70	
NMe_2-o-RF	0.90	2.54	Cl-o-NMe_2, Cl-m-NMe_2, Cl-p-NMe_2
NHCOMe-m-RF	-0.26	0.65	
NHCOEt-m-RF	-0.71	1.80	
NHCOPh-o-RF	1.47	4.13	Cl-o-NHCOPh, Cl-m-NHCOPh, Cl-p-NHCOPh
SO_3H-o-RF	1.85	5.20	Cl-o-SO_3H, Cl-m-SO_3H, Cl-p-SO_3H
SO_2Me-m-RF	0.17	0.44	
SO_2NH_2-o-RF	1.43	4.02	Cl-o-SO_2NH_2, Cl-m-SO_2NH_2, Cl-p-SO_2NH_2
SO_2NH_2-m-RF	-0.35	0.88	
SO_2NMe_2-o-RF	0.16	0.46	Cl-o-SO_2NMe_2, Cl-m-SO_2NMe_2, Cl-p-SO_2NMe_2
SO_2NEt_2-o-RF	0.50	1.41	Cl-o-SO_2NEt_2, Cl-m-SO_2NEt_2, Cl-p-SO_2NEt_2
Me-o-Me	0.14	0.22	
Me-m-Me	-0.03	0.04	Me-p-Me
Me-o-NO_2	0.20	0.64	
Me-m-NO_2	0.52	0.96	
NO_2-o-NO_2	1.13	2.22	
NO_2-m-NO_2	0.78	2.15	

Structural Feature	Regression coefficient	t statistic (39 degrees of freedom)	Perfectly correlated structural features
NO_2-p-NO_2	-0.07	0.11	
OMe-o-OMe	-2.22	1.40	
OMe-o-OMe	0.57	0.47	
OH-o-OH	excluded by regression program		
F-o-F	0.44	1.49	
F-m-F	-0.18	0.37	
F-p-F	-0.10	0.14	
F-o-Cl	0.34	2.13	
F-m-Cl	-0.10	0.36	
F-p-Cl	-0.26	0.64	
F-o-Br	0.30	1.64	
F-m-Br	-0.08	0.26	
F-p-Br	0.29	0.46	
Cl-o-Cl	0.15	0.96	
Cl-m-Cl	-0.25	1.38	
Cl-p-Cl	-0.18	0.68	
Cl-o-Br	0.28	2.11	
Cl-m-Br	-0.23	1.29	
Cl-p-Br	0.10	0.21	
Cl-m-I	-0.52	1.07	
Br-o-Br	0.20	1.08	
Br-m-Br	-0.31	1.07	
Br-p-Br	0.18	0.63	
I-m-I	-1.28	1.44	
Cl-o-Me	0.42	1.36	

Structural Feature	Regression coefficient	t statistic (39 degrees of freedom)	Perfectly correlated structural features
Cl-m-Me	0.10	0.21	
Cl-p-Me	0.07	0.16	
Br-o-Me	0.58	1.72	
Br-m-Me	0.10	0.25	
Br-p-Me	0.12	0.24	
Br-m-CMe_3	0.84	1.67	
F-o-CF_3	0.45	1.52	F-m-CF_3
Cl-o-OH	0.57	2.08	
Cl-m-OH	0.41	0.90	
Cl-o-OMe	-0.40	0.78	
Cl-m-OMe	0.93	1.89	
F-o-OMe	-0.53	1.13	F-m-OMe
Cl-o-CN	0.53	1.13	
Cl-m-CN	0.02	0.03	
Cl-o-NO_2	0.72	4.91	
Cl-m-NO_2	0.59	2.88	
Cl-p-NO_2	0.37	1.30	
Br-o-NO_2	0.61	1.52	
Br-m-NO_2	0.21	0.76	
Cl-o-NH_2	0.42	1.55	Cl-m-NH_2
regression constant	2.25	9.84	

Table 58

Toxicity data for benzimidazole derivatives

Regression results for set D structural features

Structural feature set	Number of structural features	Number included in in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
A	10	9 + constant	42	0.912	1.15	23.07
B	21	21 + constant	30	0.992	0.41	88.22
C	41	38 + constant	13	0.999	0.15	170.80
D	43	40 + constant	11	0.999	0.12	137.29
E	31	29 + constant	22	0.962	1.06	9.42

Number of structures = 52

Range of pKa values = 9.85

Table 59

pKa values for heterocyclic structures

Regression Analysis Results for subset of pyridine structures

Structural Feature	Regression Coefficient	t statistic (13 degrees of freedom)	Perfectly correlated structural features
Me-ortho-RING N	0.71	6.21	
Me-meta-RING N	0.47	4.14	
Me-para-RING N	0.75	5.50	
NH_2-ortho-RING N	1.67	9.83	
NH_2-meta-RING N	0.69	3.40	
NH_2-para-RING N	3.90	24.82	
NHMe-meta-RING N	0.95	2.81	
NHMe-para-RING N	4.37	23.35	
NMe_2-para-RING N	4.33	23.18	
Ome-ortho-RING N	-2.01	9.97	
Ome-meta-RING N	-0.41	2.05	
Ome-para-RING N	1.33	6.57	
SMe-ortho-RING N	-1.67	8.29	
SMe-meta-RING N	-0.84	4.18	
SMe-para-RING N	0.68	3.35	
NO_2-meta-RING N	-4.52	22.40	
NO_2-para-RING N	-3.68	18.26	
Cl-ortho-RING N	-4.13	29.25	Cl-p-NH_2
Cl-para-RING N	-1.41	7.00	
Br-meta-RING N	-2.38	11.80	
Br-para-RING N	-1.47	7.30	
Me-ortho-Me	0.08	0.65	
Me-Meta-Me	0.01	0.11	
Me-para-Me	0.02	0.18	
Me-ortho-NH_2	-0.30	2.24	

Structural Feature	Regression Coefficient	t statistic (13 degrees of freedom)	Perfectly correlated structural features
Me-meta-NH_2	-0.31	2.09	
Me-ortho-NHMe	-0.32	2.21	
Me-meta-NHMe	-0.78	5.22	
Me-ortho-NMe_2	-1.27	8.68	
NH_2-ortho-NH_2	-0.73	3.10	
NH_2-cara-NH_2	-1.10	3.98	
NH_2-ortho-NHMe	-0.78	2.71	
NO_2-ortho-NH_2	0.16	0.64	
NO_2-ortho-NHMe	0.05	0.19	
Cl-ortho-NH_2	0.29	0.89	Cl-meta-Cl
Cl-meta-NH_2	excluded by regression program		
Br-ortho-NH_2	0.23	0.85	
Br-ortho-NHMe	0.19	0.67	
Br-ortho-NMe_2	-0.72	2.53	
regression constant	5.29	40.38	

Table 60

pKa values for heterocyclic structures

Pyridine subset

Structural feature set C

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
F	10	10	50	0.983+	0.98	143.32+
G	21	21 + constant	38	0.934	0.96	12.37
H	56	47 + constant	12	0.974	1.08	4.72
I (99.99% level)	66	53 + constant	6	0.999	0.08	113.12
I (10% level)	66	30 + constant	29	0.993	0.36	68.32
J	53	47 + constant	12	0.984	0.84	7.79

Number of structures = 60

Range of pKa values = 8.73

Table 61
pKa values for heterocyclic structures
Regression analysis results for subset of pyrimidine structures

+correlation coefficient and hence F - value are relatively high due to the lack of a regression constant

Structural Feature	Regression Coefficient	t statistic (29 degrees of freedom)	Perfectly correlated features
2-Me	0.76	1.79	Me-meta*-Me
2-NH_2	2.23	9.98	
4-NH_2	4.09	21.90	
2-NHMe	2.53	6.72	
4-NHMe	2.14	11.66	
2-NMe_2	2.53	8.84	
4-NMe_2	4.56	17.14	
2-OMe	-0.91	3.22	
2-SMe	-0.90	3.65	
4-SMe	0.73	2.95	
2-Cl	-1.08	2.69	
4-Cl	-1.42	4.45	
5-Cl	-1.76	5.65	
5-Br	-1.83	8.38	
5-NO_2	-3.95	18.73	
NH_2-meta*-NH_2	-0.81	3.66	
NH_2-meta*-NMe_2	-0.67	1.83	
NH_2-meta-NH_2	-3.96	12.62	
NH_2-meta-NHMe	-1.68	5.55	
NHMe-meta-NMe_2	-2.08	4.45	
NH_2-meta*-OMe	1.03	3.15	
NMe_2-meta*-OMe	1.17	3.33	
NH_2-meta-OMe	-2.33	7.23	
NMe_2-meta-OMe	-2.08	4.61	
NH_2-meta-SMe	-2,60	7.07	
NMe_2-meta-SMe	-2.49	5.10	

Structural Feature	Regression Coefficient	t statistic (29 degrees of freedom)	Perfectly correlated feature
NH_2-meta-Cl	-2.42	6.50	
NMe_2-meta-Cl	-2.48	4.62	
Br-ortho-OMe	1.41	3.18	
Me-meta*-NH_2	0.37	2.05	
regression constant	1.77	10.96	

notes meta* denotes 4 - 6 meta interaction

4 denotes either 4 or 6 position

Table 62

pKa values for heterocyclic structures

Pyrimidine subset

Structural Feature Set I

Analysis at 10% level

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
K	7	7 + constant	49	0.817	1.35	14.05
L	19	18 + constant	38	0.878	1.28	7.10
M	33	26 + constant	30	0.934	1.08	7.89
N	41	34 + constant	22	0.955	1.04	6.71

Number of structures = 57

Range of pKa values = 9.16

Table 63

pKa values for heterocyclic structures

Regression analysis results for subset of diverse structures

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
O	11	11 + constant	157	0.884	1.22	21.04
P	30	29 + constant	139	0.891	1.26	18.46
Q	44	38 + constant	130	0.905	1.22	15.48
R (99.99% level)	87	78 + constant	90	0.949	1.09	10.45
R (10% level)	87	26 + constant	142	0.930	1.01	34.96

Number of structures = 169

Range of pKa values = 11.35

Table 64

pKa values for heterocyclic structures

Regression analysis results for total set of structures

Structural feature	Regression coefficient	t statistic (157 degrees of freedom)
RING N	-2.58	13.94
Me	0.72	5.09
NH_2	1.75	11.89
NHMe	1.98	7.38
NMe_2	2.85	8.14
OMe	-0.42	1.47
SMe	-0.58	1.81
Cl	-2.75	8.03
Br	-1.4	3.32
NO_2	-3.44	8.07
FUSED RING	-0.16	0.60
REGRESSION CONSTANT	7.85	20.40

Table 65

pKa values for heterocyclic structures

Total Set

Structural Feature Set Q

Structural Feature	Regression Coefficient	t statistic (142 degrees of freedom)	Perfectly correlated structural features
RING N-ortho-RING N	-1.41	5.44	
RING N-meta-RING N	-3.24	12.35	
RING N-para-RING N	-3.43	10.70	
NH_2-ortho-RING N	1.49	8.63	
NH_2-para-RING N	2.69	11.48	
NHMe-para-RING N	3.21	9.12	
NMe_2-ortho-RING N	1.66	5.63	
NMe_2-para-RING N	2.84	7.76	
Me-ortho-RING N	0.54	3.57	
Me-meta-RING N	0.59	3.57	
Cl-ortho-RING N	-2.95	6.25	
Cl-para-RING N	-1.01	1.99	
Br-meta-RING N	-1.28	5.29	
OMe-ortho-RING N	-0.60	2.76	
Sme-ortho-RING N	-0.74	3.18	
NO_2-meta-RING N	-2.46	10.22	
NO_2-para-RING N	-3.52	3.44	
NH_2-meta-NH_2	-2.41	6.51	
NH_2-meta-NHMe	-2.41	3.86	
NH_2-meta-NMe_2	-1.59	1.87	
NHMe-meta-NHme	-1.16	1.90	NHMe-ortho-NHMe
NHMe-meta-NMe_2	-3.22	2.85	
Cl-meta-NHMe	3.39	4.45	
Br-ortho-OMe	2.61	2.30	

Structural Feature	Regression Coefficient	t statistic (142 degrees of freedom)	Perfectly correlated structural features
RING N-meta-RING-FUSION	-0.31	1.94	
NH_2-ortho-RING-FUSION	1.67	3.87	
regression constant	5.13	31.00	

Table 66

pKa values for heterocyclic structures

Total Set Structural Feature Set R

Analysis at 10% level

Structure number	Structure (substituents)	Boiling Point (^{o}C, 760 mmHg)
	cyclohexane derivatives	
1.	H	81
2.	1-Me	101
3.	$1,1\text{-Me}_2$	120
4.	$c\text{-}1,2\text{-Me}_2$	130
5.	$t\text{-}1,2\text{-Me}_2$	123
6.	$c\text{-}1,3\text{-Me}_2$	120
7.	$t\text{-}1,3\text{-Me}_2$	124
8.	$c\text{-}1,4\text{-Me}_2$	124
9.	$t\text{-}1,4\text{-Me}_2$	119
10.	$1,1,3\text{-Me}_3$	136
11.	$1,1,4\text{-Me}_3$	135
12.	$r\text{-}1\text{-}c\text{-}3,5\text{-Me}_3$	139
13.	$r\text{-}1\text{-}c\text{-}3\text{-}t\text{-}5\text{-Me}_3$	142
14.	$r\text{-}1\text{-}c\text{-}2,3\text{-Me}_3$	151
15.	$r\text{-}1\text{-}c\text{-}2\text{-}t\text{-}3\text{-Me}_3$	151
16.	$r\text{-}1\text{-}t\text{-}2\text{-}c\text{-}3\text{-Me}_3$	146
17.	$r\text{-}1\text{-}c\text{-}2,4\text{-Me}_3$	146
18.	$r\text{-}1\text{-}c\text{-}2\text{-}t\text{-}4\text{-Me}_3$	146
19.	$r\text{-}1\text{-}t\text{-}2\text{-}c\text{-}4\text{-Me}_3$	145
20.	$r\text{-}1\text{-}t\text{-}2m4\text{-Me}_3$	142
21.	$1,1,3,3\text{-Me}_4$	155
22.	$1,1,4,4\text{-Me}_4$	153
23.	$c\text{-}1,1,3,5\text{-Me}_4$	152
24.	$t\text{-}1,1,3,5\text{-Me}_4$	156
25.	$r\text{-}1\text{-}c\text{-}2,3,5\text{-Me}_4$	169
26.	$r\text{-}1\text{-}t\text{-}2,4\text{-}c\text{-}5\text{-Me}_4$	161

Structure number	Structure (substituents)	Boiling Point (^{o}C, 760 mmHg)
27.	r-1-t-2-c-4,5-Me_4	165
28.	r-1-c-2-t-4,5-Me_4	170
29.	r-1-c-2,4,5-Me_4	170
	1,3 dioxan derivatives	
30.	H	105
31.	2-Me	110
32.	4-Me	114
33.	5-Me	118
34.	5,5-Me_2	127
35.	c-4,6-Me_2	126
36.	t-4,6-Me_2	137
37.	c-2,4-Me_2	119
38.	c-2,5	121
39.	t-2,5	127
40.	2,2-Me_2	125
41.	4,4-Me_2	133
42.	4,4,6-Me_3	143
43.	r-2-c-4,6-Me_3	129
44.	r-2-c-4-t-6-Me_3	138
45.	r-4-c-5-t-6-Me_3	157
46.	r-4-c-5,6-Me_3	148
47.	r-4-t-5-c-6-Me_3	148
48.	2,5,5-Me_3	132
49.	2,2,4-Me_3	132
50.	c-2,4,4,6-Me_4	139
51.	r-2-c-4-t-5,6-Me_4	159
52.	c-2,4,5,5-Me_4	147

Structure number	Structure (substituents)	Boiling Point (°C, 760 mmHg)
53.	c-2,2,4,6-Me_4	138
54.	2,2,5,5-Me_4	145
55.	r-2-c-4,5,6-Me_4	149
56.	r-2-c-4,6-t-5-Me_4	149
57.	t-4,5,5,6	173
58.	c-4,5,5,6	159
59.	c-2,4,4,5	156
60.	2,2,5	137

Notes

r = relative to

c = cis

t = trans

Table 67

Alicyclic Structures Boiling Points

Structures and Property Values

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
A	2	1 + constant	27	0.964	5.61	354.87
B	3	2 + constant	26	0.980	4.31	315.28
C	4	3 + constant	25	0.991	2.90	456.72
D	5	4 + constant	24	0.991	2.96	328.84
E	6	5 + constant	23	0.993	2.72	325.13
F	3	2 + constant	26	0.964	5.71	170.87
G	4	3 + constant	25	0.991	2.92	456.72
H	6	5 + constant	23	0.998	1.54	1146.55
I	9	8 + constant	20	0.998	1.61	623.13
J	9	8 + constant	20	0.998	1.50	623.13
K	12	11 + constant	17	0.999	1.03	771.57

Number of structures = 29

Range of boiling point values = 89 units

Table 68

Overall regression results for cyclohexane boiling points

Structural feature	Regression coefficient	t statistic (17 degrees of freedom)
ring system	common to all structures	
eq Me	19.19	28.77
ax Me	25.53	22.26
eq Me-gem-ax Me	-5.97	8.30
ax Me-ortho-eq Me	3.05	4.99
eq Me-ortho-eq Me	3.04	5.34
eq Me-meta-eq Me	-0.79	1.69
ax Me-meta-eq Me	-2.35	4.25
ax Me-meta-ax Me	-1.32	1.46
ax Me-para-eq Me	-2.48	4.06
eq Me-para-eq Me	-1.97	3.52
ax Me-para-ax Me	-0.73	0.67
regression constant	80.91	87.31

Notes

eq = equatorial

ax = axial

gem = geminal

Table 69

Boiling Points of Cyclohexanes

Regression results for structural feature set K

Structural feature set	Number of structural features	Number included in analysis	Degrees of freedom	Multiple correlation coefficient	Residual error	F value
L	3	1 + constant	29	0.859	8.28	81.64
M	4	2 + constant	28	0.872	8.07	44.43
N	5	3 + constant	27	0.902	7.24	39.28
O	8	6 + constant	24	0.964	4.73	52.57
P	5	3 + constant	27	0.935	5.96	62.56
Q	9	7 + constant	23	0.961	4.99	39.68
R	10	8 + constant	22	0.963	5.01	35.11
S	8	6 + constant	24	0.964	4.76	52.57
T	9	7 + constant	23	0.971	4.32	54.20
U	11	8 + constant	22	0.972	4.35	47.05
V	12	10 + constant	20	0.994	2.18	165.17
W	17	15 + constant	15	0.998	1.36	249.25
X	20	18 + constant	12	0.985	4.35	21.72

Number of structures = 31

Range of boiling point values = 68 units

Table 70

Overall regression analysis results for dioxan boiling points

Structural feature	Regression coefficient	t statistic (15 degrees of freedom
ring system	common to all structures	
O-meta-O	common to all structures	
2-eq Me	5.57	4.65
2-ax Me	21.92	13.49
4-eq Me	12.28	12.63
4-ax Me	24.39	16.74
5-eq Me	14.86	12.00
5-ax Me	15.15	9.68
eq Me-gem-ax Me	-6.92	7.30
eq Me-ortho-eq Me	2.13	2.92
eq Me-ortho-ax Me	2.91	4.14
ax Me-ortho-ax Me	5.42	4.03
eq Me-meta-eq Me	-1.54	2.27
eq Me-meta-ax Me	-3.27	4.32
eq Me-para-eq Me	1.32	1.07
eq Me-para-ax Me	-2.37	2.06
ax Me-para-ax Me	1.00	0.49
regression constant	103.77	90.52

Table 71

1,3 Dioxan Boiling Points

Regression Results for Structural Feature Set W

Structural feature	Regression coefficient	t statistic (40 degrees of freedom)
ring system	common to all structures	
ring O-meta-ring O	15.43	7.89
eq Me	18.66	12.14
ax Me	21.58	9.04
eq Me-gem-ax Me	-5.33	3.53
eq Me-ortho-eq Me	2.37	2.00
eq Me-ortho-ax Me	4.10	3.44
ax Me-ortho-ax Me	-8.46	2.73
eq Me-meta-eq Me	-1.21	1.19
eq Me-meta-ax Me	-1.34	1.15
ax Me-meta-ax Me	1.34	0.52
eq Me-para-eq Me	-0.90	0.64
eq Me-para-ax Me	-1.57	1.11
ax Me-para-ax Me	0.78	0.30
eq Me-ortho-ring O	-5.48	7.11
ax Me-ortho-ring O	-1.46	1.33
eq Me-meta-ring O	-0.82	1.13
ax Me-meta-ring O	-4.10	4.49
eq Me-para-ring O	-1.78	1.44
ax Me-para-ring O	4.78	2.76
regression constant	85.61	38.92

Table 72

Alicyclic Structure Boiling Points Regression Results