Chemical Complexity and Molecular Topology: The DARC Concepts and Applications

Jacques-Emile Dubois

Abstract

In the 1950s we found that all the sites of a well-defined carbon environment of the ketone function contribute to the expression of its properties. This justified a new description of ketones with a topological code. The ketone molecular structure was handled as a simple graph corresponding to an ordered rooted tree. We generalized this treatment to achieve a new, original nomenclature called DARC (Documentation and Automated Research of Correlations), valid for the whole of organic chemistry. This original coding breaks down the support graph of a structure into a series of local ordered rooted trees of limited size. It is simpler to handle segments, or ordered trees, with a computer than to treat the whole graph globally.

We show an idea stemming from fundamental research and the formalizing of physical observations that led to the birth of the DARC system. Its application and standardization rules took into account the diversity of operations that topology enabled us to predict, for managing both on-line files and for new applications dealing with the creation of new categories of chemical compounds. Topological coding enabled new correlations for physical, chemical, and biological properties.

The DARC epic of constituting large structural databases on-line (1970, 1985, and 1990) is evoked in the context of that era of computerizing chemical data. The usefulness of the FREL (Fragment of an Environment that is Limited) concept is confirmed in classical structural files as well as in files involving patents for structural chemistry.

The advantage of expressing organic structures topologically is illustrated by physical-chemical examples; information theory allows handling, with precision and speed, of complex data spread over vast areas of study. Topological concepts of local organization make it possible to account quantitatively for the idea of structural environments in various molecular spaces.

The place and the impact of topology on chemistry cannot be analyzed without some preliminary reminders of the present state of the art of structural organic chemistry. One needs to recall briefly some of its basic concepts and conventions in order to comprehend fully their interactions with or influence on topology. Topology is responsible for qualitative changes in structural organic chemistry while also providing quantitative advantages. In facilitating computerization, topology is instrumental in advancing many aspects of structural chemistry. To begin with, topology can change the representation of structures for more efficient handling in numerous complex situations. It modifies qualitatively some chemical concepts and helps to handle differences from standard representational notions.

Structural Chemistry Background

Two main concepts or principles constitute the backbone of structural chemistry. The first considers that molecules or structures result from joined groups of atoms. Some of them are more active and are usually considered as functions; others build the skeleton or framework. The other important concept is that of homogeneous families of structures organized either in ascending order, for homologues, or with a common global atomic identity, for isomers.

Over the years the paradigm of structural chemistry has been strengthened by rules and conventions so that it is easy to draw the structural diagram of a molecule in a unique standardized representation.

To facilitate communication and the elaboration of reference files of the numerous structures—those extracted from nature and those created by chemists—in such a way that they are clearly identified and classified, chemists use rules of constitution and graphic representation of structures to discuss the structures by systematic nomenclature and structural chemistry. These have developed in symbiotic ways, and the syntax of
systematic nomenclature bears a close relation to the ways in which structural chemistry is shaped and taught to future chemists. In particular, the virtual fragmentation of structures, or substructure partitioning, to elaborate the names of structures is often close to the chemical reactions and syntheses of those structures. In laboratories the numbering of atoms within a structure is a common part of daily work, and abstract services or compendia have adopted logical presentations similar to those used in the pedagogy of structural chemistry.

Real advantages are derived from the advances in systematic nomenclature. The advances come from the converging efforts of the International Union of Pure and Applied Chemistry (IUPAC), the Chemical Abstracts Service (CAS), Beilstein’s Handbook of Organic Chemistry, and other institutions that constantly enlarge the field of action of systematic nomenclature as science progresses. As a living science, though, chemistry continues to use many common or trivial names for each structure, mainly because of the difficulties encountered with the oral expression of systematic names, essentially derived to give a unique and complete description of structures. In certain frontier fields, researchers may have to work with some specific conventions not yet accepted by the systematic nomenclature. Thus, as in a natural language our chemical language recognizes local expressions but relies on a central organized and evolving language that retains the quality of systematic codes.

With computing, new horizons appeared in chemistry, both in its information-documentation aspects and in designing molecular syntheses. One could hope to handle efficiently the explosion of new compounds being synthesized and elucidated as well as the exponentially increasing amounts of analytical data produced by new physical methods. One could also foresee how to bring some involved ab initio calculations in line with structural chemistry’s basic simplifications.

The information revolution opened up a variety of avenues for structural chemistry; some of them were followed. Statistical surveys of large files helped identify substructures not considered chemically active fragments but nevertheless useful as screen or meta data in chemical design strategies. I propose here that the codes and computer-retrieval tools of structural data are instrumental in designing information management systems but should not themselves be called global information systems.

### Part I—Structural Information Resources before DARC

In this paper I will emphasize the academic origins of the DARC (Documentation et Automatisation des Recherches de Correlations [Documentation and Automated Research of Correlations]) Topological System.

### The Situation before DARC

In the 1960s data handling was restricted to the information institutions: CAS, patent offices, and industrial sectors. Scientists were users of primary and secondary information, but information research was not in the academic curriculum, certainly not part of the science faculties.

I became interested in aspects of information through my research in spectroscopy, which I pursued in the 1950s. Later I began new academic teaching areas in chemical and bioinformatics, but subsequently, we had to develop our DARC system, partly through external associations and with nonacademic partners.

### Spectroscopy: Evidence of an Extended Carbonyl Functional Group

In the 1950s at the University of Saarbrücken I began a research program to evaluate quantitatively the alkyl influence on the environment $E_X$ of the carbonyl group, leading to red shifts of the $n \rightarrow \pi^*$ transition in the ultraviolet range. For this project we had to consider synthesizing ketones belonging to the aliphatic ketone families with crowded and hypercrowded environments of the carbonyls in the $\alpha$ and $\beta$ positions, which meant changing the groups $R$ and $R_1$ in the general formula $RCOR_1$.

Our first deduction, based on experience, was that one could neglect as a first approximation those long-range influences that go farther than the carbons $\alpha$ and $\beta$—that is, the first and second carbons of the environment, later named A and B. Moreover, rather than evaluate alkyl group actions, we could identify the influence of the carbon sites in the A and B positions (Figure 1) and the practical absence of influence beyond these atoms.

The environment $E$ of the $-\text{C(CO)}-\text{C}$ root of the graph or focus of a family of aliphatic ketones consists of two segments or modules $E_1$ and $E_2$ contributions. One can identify the contributions of all atoms of the first layer of atoms of the $R$ and $R_1$ alkyl groups. Even the action of atoms B can be evaluated. The $\Delta \nu$ contributions are calculated by comparison with the acetone chosen as reference for the alkylacetone family. The
site contribution is achieved through a correlation over a large population of ketones. This validation process associates a $\delta\nu$ shift parameter with all A and B carbon positions.

These findings resulted in precise rules (the Dubois-Maroni rules) that were confirmed and improved during the project's continuation in my laboratory at the University of Paris. Our rules probably provided the first systematic labeling of the $E_B$ carbons, although much later similar contributions on the topological side became common practice with $C_{13}$ nuclear magnetic resonance spectroscopy, and each carbon has a $\delta_c$ shift value (chemical shift).

This research developed over a number of years, slowed by the difficulties of synthesizing crowded ketones. It confirmed our initial Dubois-Maroni prediction rules. They could be interpreted in favor of the concept of a carbon extended carbonyl function or as the validation of the environment ($E_B$) effect being active over a limited portion of the carbon framework. Limited but similar site interventions were provided by the Woodward-Fieser rules, and we decided to investigate them more fully with our topological concepts. Our results will be discussed later in this paper.

For our research programs we found systematic nomenclature cumbersome and of no help in our laboratory work. We also found it difficult to evaluate our various syntheses of ketones. Moreover, we had little information about crowding effects on the synthesis of encumbered species. Even our ketone database management, dealing with many similar compounds, was complex to operate.

All these constraints and results constantly stimulated us to look for better structural representation, both for direct communication as well as for computerized operations. Although we had success with the Dubois-Maroni rules, we asked whether we should be satisfied with those quantitative equations or try to seek a general theory to justify and explain them. We looked for an original quantitative handling of alkyl environments and opened a new paradigm in this field when, for interpretation, we turned to topological theory and ordered graph representation of our ketones. We found later that our original approach was similar to the directed graph notion (digraph) in which all edges are associated with pairs of ordered vertices.

**A Topological Representation of Alkyl Environments: DARC Concepts**

One Sunday morning in March 1964 I found a way to express our $A_B_i$ environmental sites by locating them in a special matrix (Figure 2) corresponding to our quantitative $A_i$ and $B_i$ validations by means of the Dubois-Maroni rules (Figure 1).

The special matrix adopted, called “E” for environment, has some aspects of the $A$ matrix (adjacency) and of the D matrix (distance) of classical graphs, as it implicitly introduces the concept of distances $A$ and $B$ for the first and second (concentric) layers of atoms. It is different from those $A$ and $D$ classical matrices, because we separated the three $B_i$ carbons as substituents of any $A_i$ site ($A_1$, $A_2$, $A_3$, or even $A_4$), easily designated by their connectivities ($p, s, t, q$, or primary, secondary, tertiary, and quaternary). Although formally at distance 2 from the focus, they each have their own environment $E(B_i)$, which justifies discretizing $B_1$, $B_2$, and $B_3$.

We proposed to identify an environment module called “Environment that is Limited, Concentric and
Ordered” (ELCOB). ELCOB is limited in B and described as $E_a = E_B$. The depth of $E_a$ can be varied but for canonical reasons, and later for statistical evaluation in chemistry and documentation, it is often standardized as $E_B$. This is a useful depth limit (also including $V_i$ row vectors starting at the $E_{A_1}$ level). The $E_B$ provides a convenient space to use local connectivity for structural validation. It groups neighbor atoms and next-neighbor atoms from a focus atom or functional group. Most numbering decisions are made at this level (Figure 2).

Whereas coding systems have usually been derived by compression, we have here a special property of the matrix (in which the $V_i$ vector rows are always in decreasing order to underline the priority of higher substituted vectors). Higher row vectors are senior to lower ones in the $A_iB_{ij}$ matrix. All these concepts could be conceived and expanded thanks to our physical data on ketones justifying their presentation in a specific $A_iB_{ij}$ matrix or $E_B$ matrix of spectral increments. An ELCO is similar to a limited rooted ordered tree.

**Linear Structural Descriptor: The DEL**

A matrix description is linearized for E site input by taking the summation values of the matrix columns. We chose this summation value to produce a linear descriptor without loss of information. This descriptor enables one to regenerate the matrix or produce a graphical display of a structure. It is called “Descriptor of an Environment that is Limited in Bi,” or $DEL_{Bi}$. It is the descriptor of the existence of an ELCOB. The size of the environment $E_B$ is chosen as a good optimization, based on physical data and computerization results.

The ELCOB concept can be used as an organizer of the representation of a ketone where the focus is $-C(CO)C-$ and its concentric environment is expressed as $E_B$, for example, EtCOEt = 1000(Fo)1000 = 1(Fo)1. In this case our representation is powerful and sufficiently clear to define 630 ketones limited in $B$, ranging from acetone to tetratertiobutyl acetone, $C-CO-C$ to $(C_{27}H_{54}O) C12(CCOCC)C12 = 3333(CCOCC)3333$. This DEL representation is simple and useful; one usually just retains the numeral part in a series of ketones.

**Generalization: Segmentation of a Structure with Several $E_B$**

Of course, one immediately wonders about the representation of higher ketones with a long alkyl chain, like the methylketone $CH_3COCH_{12}H_{25}$. What of sites beyond the active environment? How does one make a general language from a punctual coding? This can be achieved by including in our model the virtual segmentation of the environment in successive joined $E_B$ (Figure 3).
In fact we adopted a propagation principle of $E_B$ whereby a structural environment is gradually described by repetition of the $E_B$ segment concept. We can thus name any type of structure. It is easier to derive a numbering progressively, based on our local connectivity concept, than to do so with methodologies using “extended connectivity” that mobilize the whole structure in other naming languages. It is well-known that site-sequencing processes are more easily determined within an EB than over an entire structure.

**Structural Components: Structure, Substructure, and Hyperstructure**

A conceptual structural model based on topology, graph, and matrix emerged from our spectroscopy work on ketones. It gave us a coherent view of the interplay of three basic structural components: the substructure, the structure, and the embedding of these notions in a larger graph or filiation map of structure, named hyperstructure. The interplay facilities result mainly from our complex decision to mix structural edge distances within a limited ordered concentric graph descriptor. The synergy of those ideas found its source in our classical chemical heritage, and we compare them in Figure 4 with those of the trilogy “compound-function-family” and their evolution in time. These family notions influenced the development of systematic nomenclature as well as being useful for applications in SPER (structural and property energy relationship) and QSAR (quantitative structure-activity relationship). The progress here results from linking site validation to a strong coherent synergetic association structure-substructure-hyperstructure (Figure 5) presented topologically.

**DARC Unity and Objectives**

The environment concept with its physical origin led us to the larger interpretative paradigm combining different notions of ordered chromatic graphs and generative ordering functions, notions we found to be useful in several fields. We realized that they came from a central logical model that we decided to shape into a system.

As the original objectives concerned spectroscopic correlations (Dubois-Maroni rules) and although the generalization of the language was to lead to other important developments, we named the system DARC, or “Documentation et Automatisation des Recherches de Correlations” (Documentation and Automated Research of Correlations). To cover the diversity of concepts discovered in our research, three theories were conceived and presented together, as they are indeed interwoven.

Their separate development was very different, requiring sets of data often nonexistent or difficult to find in the literature.

**The DARC System: Three Basic Theories**

The DARC system encompassed three main theories closely linking structural representation and its use in different fields of chemistry. Thus the name of each theory combines one topological concept with specific application areas (Figure 6).

1. **Theory of Generation-Description.** This deals with the development and application of the language of chemistry knowledge: structuring meta-concepts. SS, S, and HS are managed with a unique topological grammar.

   **Figure 4.** Chemistry knowledge: structuring meta-concepts. SS, S, and HS are managed with a unique topological grammar.

   1—Partial ordered spaces
   ![functions](families)
   ![Compounds](Com)
   2—Partial ordered spaces
   ![Fragments](Populations, sets)
   ![functions, groups](Com)
   3—Total ordered spaces
   ![SS](S)(HS)
   ![substructures](structures)(hyperstructures)

   **Figure 5.** The DARC system: central role of the SS-S-HS (substructure-structure-hyperstructure) axis.
DARC-DUPEL (Description that is Uniline by Propagation of an Environment that is Limited in B). Generating the representation of structures and substructures and chemical database management are the important applications of this theory.

2. Theory of Population-Correlation. This involves the organization of spaces of states, such as the representation of structural populations and of the physical property spaces of those populations. Such organizations imply hybrid and complex representations, including classical and topological descriptors to use for QSPR (quantitative structure-property relationship) and QSAR strategies.

3. Theory of Topo-Information Correlations. DARC is a topological theory of environment effects in physical chemistry. The E matrix is used to derive topological validation of active sites within a structure-property space. Environment descriptors in theories 2 and 3 used the ELCOX idea, but because of the information community’s interest in data files and databases in which the concept of structural fragment is important, the ELCOX is often named FRELX, for “Fragment of an Environment Limited in X” (X is usually either A or B).

Part II—From DARC Concepts to Structural On-Line Services

Generation-Description Theory: The DARC-DUPEL Generator

The first theory and its common language DARC-DUPEL are better known than the others in the information profession. We will therefore elaborate on it more fully and only briefly present the two others that primarily concern scientific and academic applications.

To code a graphical formula G(X) in DARC language, one uses its topomodel described as: T(G) = focus (FO) + Environment (E). Both E and FO are coded similarly, but in documentation FO = 1 because it refers to one site only. The focus is equivalent to the root of a digraph.

Thus the coding of G is essentially the coding of E. This is done by a combined or generating process segmenting E in successive E₆ starting from the focus up to the terminal sites of G. The process obeys priority rules. It starts by creating a virtual graph or spanning tree (EXS), expressing the existence of a spanning S graph isomorph with G but with simple edges between the nodes and no ring closures. This underlying graph of all further information is called Existence Graph G(EX).

This basic topological organization of T(G) = T(EX) in segments is based on ordering rules of local connectivity. To order the sites, ordering decisions are based on the absolute priority given to the highest connectivity of all neighboring sites.

Vectors in matrices are in descending order of complexity (Figure 7). All chromatic information is added within the ELCOs of (EXS). This information includes the nature of bond-links (LI) of atoms (NA), geometry, chirality, conformations, electronegativities, and the ring closures.

The final representation can be stored in the DARC connectivity tables or as DELs, in the computer DEL = DEX/DLI/DNA/---. Since the algorithm generating the DARC code can also deliver the DEL, it is called DUPEL.

Overlapping Substructures: FREL

Whereas ELCO₆ were used as joint segments, it is interesting to create a larger number of different E₆ over the structure by choosing all sites as local foci. One then generates overlapping segments, which we prefer to name FRELs. The operation of overlapping fragments, here FRELs, was not used in past documentation strategies, although topological overlapping can lead to interesting observations. The FRELs on the sites are fascinating formal fragments for the query system of any database.

In the DARC system we consider that the union of all the overlapping FRELs of a G(X) constitutes another DARC name of the structure.

In this DUPEL algorithm we observe that the system complexity associates a well-defined, segmented topomodel with a syntactic process handling the final ordering. This recalls the Tosio Kitagawa information

<table>
<thead>
<tr>
<th>DB-KB</th>
<th>AI/Dgn</th>
<th>CONCEPTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>++++</td>
<td>+</td>
<td>S, HS, SS</td>
</tr>
<tr>
<td>++</td>
<td>++</td>
<td>P-S/HS; P-I</td>
</tr>
<tr>
<td>+</td>
<td>+++</td>
<td>P-S/HS/I</td>
</tr>
</tbody>
</table>
theory, whereby an information system must be influenced both by logical and linguistic processes.

**Canonicalization Procedure**

*Extended and Local Connectivity.* In the 1960s there was an urgent need to represent structures making use of the computer facilities that were developing so rapidly. The goals were clear, but the problem of transforming a graphic structure into a machine representation had no obvious solutions. Today it seems evident to state that the option was either to compress the information in linear codes or to accept the complexity of a tabular description (matrix approach). Competition and stimulation during those years were too important to be summarized in a few lines. Readers interested in this period will find two excellent books in my reference list of suggested reading (Lynch et al., 1971; Davis & Rush, 1974).

In early theoretical work topology played an important role, for example, in quantum chemistry, but ordering was usually associated with a specific project. Things became more complex with the intent to produce unique structure identifiers. This meant finding ways to canonicalize a structure matrix and consequently to obtain a reference ordering of all vertices. To adopt a complete description with matrices seemed cumbersome, even though the ever-increasing power of computers would clearly be of great help.

Two topology-inspired contributions helped solve these problems. With the idea of a connection table as a tabular-matrix representation, one could propose a description tool grouping all the interconnections (bonds) between the joint atoms of a structure. Moreover, other nonhydrogen atoms, complex bonds and geometry, chirality, conformations, electronegativities, and cyclic closure information are added to this basic connection table.
graph support. Then methods had to be found to designate, by convention, a unique canonical matrix among the many matrices resulting from all possible numberings of a structure.

Creation of or access to a unique matrix was proposed by two independent methods based on different ways of exploiting the vertex connectivities of a structure \( S \) on the level of supporting open graph \( G \).

In the early 1960s these two methods were found simultaneously. L. H. Morgan at CAS used an “extended connectivity” program to determine the unique focus of the underlying graph \( G \). With it came a sequential and unique ordering of all atoms of \( S \).

With the DARC system we developed a canonical search algorithm based on “local connectivity” using the DARC segmentation of the entire underlying graph \( G(\text{EX}) \) in successive ELCO\( _n \).

From its inception the “local connectivity” \( E_n \) concept was oriented to deal with chromatic graphs. The “existence graph” \( (G(\text{EX})) \) led to a sequential ordering of all vertices.

In both strategies any chromatic information is localized on the topological coordinates of the defining spanning tree graph. In both methods one ends with the equivalent rooted tree description, including a choice of a specific root.

**DARC-DUPEL and Chemical Finalization and Optimization**

Chemistry is complex and deals with many functional classes described by conventions. A formal language such as DARC needs to be adapted to integrate those conventions. To deal with the complexity of these numerous aspects, one needs to access large working files and to interact with their information specialists. Since France engaged very little in the chemical abstracting field and the French chemical industries did not participate in any consortium involved in structural research for abstracting or data filing, we developed our own documentation activities. Thus I created a small private association called the Association for Research and Documentation in Chemical Informatics (ARDIC) and hired professional computer people to develop and implement DARC structural tools and programs.

Simultaneously, with the DEL structural coding of a whole structure, the ARDIC group proposed a similar descriptor, FREL for substructures, handled as finite and ordered subgraphs or fragments. These DEL and FREL codes led to the creation of powerful metatoorganizing database tools and retrieval strategies. ARDIC became the normal interface for different exchange projects that we had launched before its existence. We reinforced our cooperation with the Basel Group of Pharmaceutical Industries, which began in the mid-1960s, and with CAS, begun early in 1970. The Basel Group lent us some connection table files, and we were able to start with ten thousand, then with five hundred thousand connection tables. We then began to create structural interconversion programs and to investigate various topological searches. Through these exchanges and lively cooperation we learned a good deal about functional types, conventions, and graphic structure norms for systematic nomenclature, in particular from the exceptional experience of CAS. We carried out a stimulating task of interconversion of their chemical and biological activities file (more than five hundred thousand connection tables) at the time of its computerization from the CAS program to DARC-DUPEL. We put a chemical and biological activities–DARC version on-line around 1975 to test it with industry users (Figure 8).

Structural information is input through a graphic terminal leading through the DUPEL program to the DEL and FREL files of all structures. The database consists of those files, generic keys, and a special organization of various topochromatic spaces of the total information (structural and literature). Query strategies are diversified, but the two main approaches are based either on FREL screening or on a topological search of an organized file tree. Answers include a final graphic edition of formulas.

During those years of rapid change extensive work in the IUPAC Interdivisional Committee on Machine
Documentation (that it was my privilege to chair from 1968 to 1976) created an international community of interest that benefited all its member experts. Bilateral exchanges were also fruitful. We remember with pleasure those years of warm and constructive discussion.

**ACS-CAS-CNIC Convention**

For the DARC the final stage of this epic came in 1978 when the Centre National d’Information en Chimie (CNIC; National Center for Chemical Information), a newly created French chemical organization, signed a cooperative agreement with CAS. We then received and transcoded the entire CAS registry and produced EURECAS (the European DARC version of the CAS file) as an on-line service monitored for two months by the development team of ARDIC. We then transferred the EURECAS and chemical and biological activities files and other by-products to the CNIC for QUESTEL, which became the developer vendor agent. In our EURECAS database the management system was based on an artificial intelligence tool called DARC-TSS (Topological Screen System). Partial extensions were managed both for specific populations (e.g., nuclear magnetic resonance) and for proper DARC-Markush generic structures.¹

Our adventure with large structural files for documentation that took place on the borders between academia and industry was over. It had been an exciting and instructive experience. Our vision of topology, free from user constraints, had given rise to constructive developments in the information field. Further DARC theories were to broaden our topological vision of chemistry and add new potentialities. The production and nature of topological chemical knowledge helped us to later develop elucidation and synthesis methodologies in parallel. We maintained an active interest in data and data-set handling, as we enjoyed testing the potential of the DARC topological tools. Indeed these further activities were the source of original documentation and communication tools.

**Part III—The DARC System: Population-Correlation Theory**

Property-structure correlations usually imply comparison of a certain number of similar structures. There is normally a relation between the structural set and the expected QSPR equations. For instance, practice stipulates the minimum number of specific substituents of the reference structure required to establish a Hammet correlation and thus define the minimum organized population of structural space data required for a good correlation. Compared with the enormous amount of work carried out in this field thus far, little has been done to describe adequately the populations of species to build for a correlation search.

**Space of States Representation**

Descriptions of spaces of states are important for managing these populations, particularly for the use of group-matching functions to detect QSPR relations. In scientific applications there is room for work on operational descriptions of spaces deploying population sets. These sets should include general and local descriptions, internal ordering rules, and operators to create filiations.

Associating structural spaces and property spaces is commonplace in physical chemistry. Within this population-correlation theory we use various population generators dealing with enumeration of species and their population structuring. In these structural spaces the concepts of filiation and regularity are important to detect errors and guide the choice of new tests to verify correlation and similarity trends.

**Frontiers of a Population Hyperstructure**

The frontiers of certain population spaces can be defined through the union of topological coordinates of individual species of the population. Graph mapping leads to the trace or imprint that corresponds to the frontiers of a population space. An example is given in Figure 9, which shows the imprint of a chlorocompound population. Such imprints or traces can be used to define a population of data. If the database is organized with internal subsystems of such types as a hyperstructure, it is easy to use a “trace” concept to develop smart queries.

**Generation of Filiation Graphs as Hyperstructures**

To generate all isomorphically distinct structures (trees) with vertices limited to four edges to create a specific filiation graph, one can use a progressive building procedure, that is, start from one vertex and make it grow by adding an extra vertex and so on. This procedure creates successive subgraphs with 1, 2, 3, and 4 neighbor nodes. They are parts of a filiation graph. In this filiation a compound of rank 3 is considered as an

¹ A Markush structure is a generic structure describing specific and nonidentified structures. One such structure may cover hundreds of thousands of identified candidates within a single patent coverage. It corresponds to a complex hyperstructure.
“anteriolog” of a vertex from the next rank, or rank 4. Anterior functions are similar to those father-son relationships used in certain presentations (Figure 10). Whereas in molecular graphs this relationship links sites within a structure, in DARC hyperstructures two independent structures are real neighbors with a father-child relationship.

Since the creation of ordered spaces is crucial to the DARC system for the strategic search of information based not only on enumeration but also on hyperstructure ranking, we needed an inference motor to perform the lengthy and recursive hyperstructure construction. We developed a DARC-GEANT program (GENERator by ANTeriology) that can be parametrized for different targets. It can be combined with physical constraint rules to build hybrid hyperstructure-properties (chemical or physical).

**Subsystems in Hyperstructure Filiation Graphs**

Carrying out local investigation of the hyperstructure graphs or maps is important. The entire hyperstructure can be virtually subdivided into local subtrees, and even line filiations are of interest. Various strategies can be used to examine the information contained in the hyperstructure and generate, if possible, new knowledge. One can use algorithms investigating the variations of properties with next, next nearest neighbor concepts or otherwise alter the entire internal organization to bring closer together, as a cluster, compounds with similar properties either in two-dimensional or three-dimensional hyperstructure.

**Topological Enumeration and Pruning Functions**

The use of generators leads to many products, particularly with isomers. These graphs correspond to candidate structures by recursive construction, but they are only mathematical visions as long as their vertices are but immaterial dots. To reach the real geometrical structures, one must introduce the atom volumes and the bond sizes. Other factors can be considered, such as interactions, attraction, or repulsion, between the vertices. These factors occur more intensively when the structures deviate from the norm of simple molecules. To obtain a more realistic generation, one uses “pruning
functions” to eliminate the unlikely graphs. Such functions are usually conservative and reject structure proposals that do not conform to existing substructures in the state-of-the-art chemical literature. These functions impose constraints that need to be defined. For crowded species we considered the difficulties involved in building structures when the local bulkiness of atom groups becomes too great. Such considerations are frequent with functional deviations caused by steric effects. But do we have valid parameters for expressing the bulkiness factor and its hindrance ability?

Real Structures and Steric Hindrance

Some neighboring groups introduce physical hindrances to a molecule’s behavior. Robert W. Taft proposed to measure such effects with a structural index called the Ė_storic or Ė_index. A scale of Ė for different groups was experimentally calculated through disparate measurements of reactivities, slowed down by the bulkiness of those neighboring groups. The scope of these parameters was too limited; we expanded it, normalized it on a key reaction, and derived Taft-DuBois factors Ė, which are extremely valuable to introduce steric hindrance factors in many organic reaction mechanisms. The Ė values correlate well with the ELCO Ė. In the following case study of aliphatic ketones we show how such Ė values are used for a complex description of alkyl behavior (Figure 11).

Case Study: Presence and Comparison of Aliphatic Ketones in a Triangular Space

In Figure 11 we present a diagram that shows the hyperstructure of aliphatic ketones where both alkyl groups R and R₁ are limited to the first ELCO_b (the longest alkyl chain being seven atoms); the first ketone is acetone, and the last is tetratertiobutyl acetone (C₂₇H₅₄O).

In this population of 630 ketones we have simple ones whose behavior is normal. However, since the bulkiness of their alkyl groups increases, some of their chemical properties vanish. Moreover, their syntheses become more difficult.

The triangular diagram shown here is interesting as a global information presentation of a large number of syntheses carried out in our laboratory. The ketones are located on the intersected squares of the x-y coordinates. On these coordinates we locate the R alkyl group (x) and the R₁ group (y), ranked by their atom numbers. One often finds alkyl isomers whose ranking is difficult. We learned by experience that their Taft-DuBois steric values (Ė values) can be used to rank them. To simplify this presentation, we ignore interactions between R and R₁.

If one specifies the methods used to synthesize ketones, one can either draw diagrams for each type of synthesis or, as here, group several methods in a triangular diagram. Limits of reaction syntheses are easily ascertained. Without going into detail, one sees rapidly that ketones become more difficult to create when their alkyl group becomes too bulky. The tetratertiobutyl is the most difficult to synthesize among the bisecondary ketones. It is thus far a unique synthesis. The rule that “difficulties increase from south to north” must be viewed with caution, bearing in mind the classification of ketones.

DARC Hyperstructure: Data and Database Quality Control

The quality of data can be assessed through the database organization and the existence of internal correlations either to spot “outliers” or to clarify uncertain data. One may have to restructure the data file organization and its screens to generate new knowledge beyond the general reductionist vision of “normalized” structures.

In 1990, after an extensive statistical search on large
data sets, Mark Johnson considered “automated associations in some QSAR databases” and even “posited to a group of automatable methods” potentially useful for routine predictions of properties. This was the first objective of the DARC system as, in 1965, we named it initially Documentation et Automatisation des Recherches de Correlation (Documentation and Automated Research of Correlations).

The first name of our system aroused a great deal of criticism when it was presented. Apparently it came too soon. We changed the name to Description, Acquisition, Retrieval, and Conception, or DARC, emphasizing the documentation or information aspect.

Today the emergence of knowledge through data mining of large databases can lead to original quality control through the discovery of knowledge keys and to improvements with the emergence of new classes in a random file. It can thus lead to more elaborate queries. A primitive or naive question based on the assumption of a “normal” situation may reveal a deeper complexity. Progress in the representation of data is essential to improve the search for and quality of correlation models.

Part IV—The DARC System: Topology-Information Theory

With this third theory our principal intention was to identify quantitatively the environment concept with topological variables. In fact this theory is one of “environment effect assessment,” despite the fact that it is often used as a QSPR method for property design since it is cast in the form of a similarity correlation.

We wanted to display the contribution of the limited $\text{E}_B$ as property vectors embedded in the
corresponding ordered E_b(M) matrix. Each site of the E_b(M) is thus associated with a property increment DP_i, which is a component of the property vector.

There is similarity between the quantum mechanics generation of local information on a structure and our intention to locate information on the E_b sites. The quantum mechanics calculation enriches the Hückel matrix for the final presentation of a structure with chromatic local information, for example, electron density mapping.

Our topo-information equations are also intended to enrich the skeletal structure with local complementary information concerning the E_b sites. The active environment is thus expressed by this contribution DP_i, or goal matrix. Whereas in quantum mechanics calculations the creation of information is theoretically produced, in the DARC approach the information is collected from the property of a population of structures.

**Topo-Information Equations**

*Focalization Model and PELCO Strategy.* The property (Pi) information is collected over an organized population of relevant structures, adding one by one local DI_p values in the A_B, sites of the E matrix E_b. In this model the collected DI_p are obtained by differences between values of successively ordered positions within E_b. These values correspond to the DARC step-by-step production of the E_b validation. Each step creates a reference for the next one. A site action (n+1) can be considered as a perturbation of ELCO^b, and such methods are therefore called PELCO.

The information acquisition process is controlled by the progressive layer matrix E_b(M). The information vector I(E) or matrix is obtained by the Dirac product of two vectors: one, T(E), characterizes the environment vector; and one, I(m), corresponds to the information collected over population P(m) of m entities. The n sites of T(E) correspond to the trace of all the sites of the ELCO description.

Several observations must accompany the presentation of two types of equations presented in Figure 12.

First, the quality of a topo-information correlation depends not only on the size of the population P(m) but also on the regular coverage of its hyperstructure. Second, contrary to linear free energy relationship (LFER) equations, the gradual validations of the E(sites) are carried out within a progressive environment of successive references. Third, approximations are available if the experimental data are deficient.

### Additivity Models

The simplest model is based on the additivity of the successive values of the A_B_l sites, but for certain properties and for a certain bulkiness E_b(M), one needs a model with nonlinear interactions between the A_B sites. There is a large variety of situations with nonlinear models of E_b interactions inside the topomodel. However, we are limiting this presentation to models with no interaction of the focus-environment type, avoiding neighboring group effects with functional focus. In our original spectroscopic work on ketones we almost had an additivity model, but as the number of ketones increased, we had to choose whether to handle the set with several subsystem equations or use a general PELCO with interactions centered on the bulkiest ketones. We compared these different linear free energy relations and our PELCO (perturbation of an ELCO) strategies. The LFER sums up a large amount of data in a complex network of linear rr equations, each with a specific reference; on the contrary the PELCO groups a very diverse population with a single PELCO-HS map.

#### Topo-Information Relations

**Environmental Property Estimates**

\[ I = I_o + I(E) \]

1. \[ I(E) = < T(E) | I(m) > \]
   - T(E) = Environment vector (matrix)
   - I(m) = Property vector for a population (P) of m entities
   - I(E) = Information vector for E
   - Vi or Ddi = defines one subgraph of the total E = ΣEi based on the focus Fo

**Site Interactions for Congested Environments**

2. \[ I(E) = < T(E) | M_{INT} | I(m) > \]

*Figure 12.* Classes of equations with site additivity contributions and site interactions.
Nonadditivity Models. A simple situation dealing with internal stress in the matrix can be handled as a generalization of equation 1 (in Figure 12) with the introduction of an interaction matrix $M_{\text{INT}}$.

Topo-Relations Case Studies

In the literature, authors are often satisfied with simple equations or even rules, primarily because the data are too few or not representative of the diversity to be found in this field. The following case study is presented to support these statements.

A Focalized Property Case: Ultraviolet Transition of $\alpha$-Enones. Empirical rules (referred to as the Woodward-Fieser rules) enable the prediction of the ultraviolet bands of conjugated systems; either dienes or $\alpha$-$\beta$ conjugated ketones show additive properties of the alkyl substituent effects on the $\pi \to \pi^*$ transition. These rules are useful for the structure determination of unknown compounds. They clearly specify the importance of the location of a substituent on the $\alpha$ or $\beta$ sites, but they do not specify its nature (most often a methyl). We decided to investigate this $\alpha$-ene population more closely for two reasons. We were interested in the topological aspect of one rule and its complementarity with our saturated ketone rules.

We studied a large collection of these enones with substituents both on the carbonyl side and on the double bond. All those $\alpha$-enones have two absorption bands corresponding to the $n \to \pi^*$ and $\pi \to \pi^*$ transitions.

The $n \to \pi^*$ of those structures had not been investigated before our work, and no systematic study had been carried out on DD$_2$ alkyl effects on the carbonyl side of the $\pi \to \pi^*$ transitions. We restrict this presentation to a specific population of $\beta$ substitutions on the $\pi \to \pi^*$ and compare our data with the Woodward-Fieser rule (Figure 13).

The global PELCO diagram shows cases with strict additive contributions for $\alpha$-enones with one $\beta_1$ substitution, called here $A_1$. The results are calculated in frequencies of $\nu \text{ cm}^{-1}$ and show an important deviation throughout this population.

This investigation of the K band sensitivity confirmed the Woodward-Fieser rule. It locates the rule in a broader analysis where one notes the importance of the $A_1$ contribution, but the Woodward-Fieser increment is insufficient in this position for higher alkyl substituents. Moreover, the substitution on the DD$_2$ side of the carbonyl can obscure the rule, since the sum of $E_{AB_1}$ has almost the same importance as the $A_1$ increment in DD$_2$. For example, $\Delta \nu = 2480 \text{ cm}^{-1}$ in DD$_2$ is similar to $A_1$ ($\Delta \nu = 2501$). As a result, for certain hindered ketones, if one hesitates, for an $S$ attribution, between two possible structures, $C_1$ or $C_2$, the Woodward-Fieser rule does not allow a clear-cut decision. This ambiguity was solved thanks to the study of the DD$_2$ environment of the $n \to \pi^*$ transitions, investigated with the Dubois-Maroni rules, presented here for saturated ketones. A similar PELCO

---

**Figure 13.** Alkyl groups are embedded on the DD$_1$ axis (carbonyl side) and on the DD$_2$ axis (ethylene side).
diagram shows that the n→π* is sensitive mainly to its DD, ELCO and also to the A₁ position located in DD₂.

These studies on two bands in ultraviolet illustrate the importance of “active environments” that correspond to different parts of the carbon framework. They also show that many rules are too simple but are nonetheless useful when applied to the usual simple cases. The true study of complexity must cover a broad and realistic set of cases. This example also confirms our theory that it requires a clear perception of the data-information-knowledge synergies, particularly if one wants to account for real problems in both their simple and complex situations.

The “focalized properties” group, where the functional group is in the focus and the population variations are essentially in the environment, as in this case, constitutes the largest number of situations of chemometric applications. Other classes exist, dealing with polyfocalization methods.

**Polyvision: Polyfocalization Models and LOGIC Method**

The situation is somewhat different when there is no obvious focus or when the environment of a focus is blurred by a very long aliphatic chain (E₈). The property to be investigated can be expressed by the polyvision of the framework. One can assume that the property is diffused throughout the framework and that we can estimate it by integrating local overlapping visions, an ELCO₈ or FREL₈. Each vertex is considered as a focus participating in the global property with its limited environment FREL₈. A molecule can be described with its local FREL-Del located on each vertex (Figure 14).

A topo-information equation called LOGIC (LOcal Information for Global Information Construction) helped us to derive FREL overlapping increments for the molecular volume property of the 70 prime alkyl groups. The training set included 300 compounds (C₆ to C₁₁), and we calculated the 335 dodecane isomers. Results are excellent (r = 0.9952); an example is shown in Figure 14. This treatment is only apparently additive, but it includes many interactions among the V FREL values.

**Population Ordering and Topological Approaches**

It is not easy to explain theoretically the ranking observed for a population of alkanes, which can be very large, particularly with their isomer classes. All experimental classifications show that some overlapping and “inversions” of structures detected by their properties are characteristic features that are difficult to interpret.

Topology can help deal with this complexity, but so far there is no definitive theoretical interpretation of the “branching effects.” More work and probably conventions need to be developed along those lines, and we suspect that the concept of DARC hyperstructure could provide interesting answers.

Other topological treatments have been attempted in this area. They are based on the exploitation of “topological indices,” which are expressions of the invariant parts of the matrix graph description. They are combined in order to reach a compressed topological representation as a “single number” for a structure. It is a tempting development as long as the crushing of information to produce indices does not entail too much information loss in the process. Ranking with single numbers seems easy. In fact, it is useful for analyzing a population ordering, but these methods have not succeeded in interpreting microscopic inversions.
Conclusion

Topology is a natural constituent of structural and theoretical chemistry. In explicit developments it has been instrumental in the computerization of chemistry and the birth of numerous information products.

Over the past fifty years topology has had positive effects on the representation and management of different subsets of chemistry: structures, substructures, large collections, and files of structural data. In particular, it has been used to simplify the computer interfaces used in chemical computer-aided design and correlation work. Our work has convinced us of the necessity of handling synchronously the topological expressions of σ (a single atom considered as a substructure, e.g., Cl, whereas C—CH₃; sσ is a limited form of substructure), substructure, structure, and hyperstructure; their interplay is needed to deal with chemistry in all its complexity. The three components of data, information, and knowledge are general and are valid for all disciplines. They are the communicating poles of the trilogy shown in Figure 15. Each of these poles has specific tools and products. They all fertilize and modify their neighbors, even as they themselves undergo modifications.

Each improvement of any of these poles can benefit the others. New data flows to databases structured by fragmentation and can produce better information on which knowledge can be based. Thus each pole, besides its own logistics, depends on topology to maintain an easy and harmonious cyclic data-information-knowledge flow, which results in a continuous and dynamic evolution, enhanced by faster, clearer communication. This is likely to remain the best hope for progress in chemistry and science in general.

Acknowledgments

I do not want to conclude without expressing my sincerest congratulations and my warmest gratitude to Mary Ellen Bowden—congratulations for her having organized this highly successful conference and gratitude for her hard work, patience, competence, and invaluable help in the editing of this conference paper. I also thank the Chemical Heritage Foundation for the invitation kindly extended to me, which I was more than happy to accept.

Suggested Reading

I recommend the following books for a good introduction to this field:

Structural Coding History: En Route to Computers


DARC System: Concepts and Theories


DARC Generation: Description Theory and Databases

Appendix: Supplementary Information on Topology, Graph, and Tree Concepts

The DARC system began with physical property searching, and the descriptive vocabulary developed for it was cast in language that chemists could assimilate. It opened the gates of computerized activities on chemical information. In fact DARC theories and tools can be better understood nowadays in terms of graph theory, as the original concept of chemical segmented environment corresponds to specific assignments of local rooted subtrees. Used worldwide for patent searching, the DARC system is useful wherever a large number of graphs are involved.

With this supplement, the reader can grasp in a different light the presentation of the DARC system and also better understand how it can be used in many fields, both in and outside of chemistry, where navigational networks and their libraries are seen as fundamental tools of an information system. In simple terms topology deals with invariant properties not sensitive to the deformation of objects or mathematical entities. The topological expression of a chemical formula is not tied to any particular version of that formula but strictly retains all the structured environments of all its atoms.

Topology expresses neighborhood relationships, their unions and intersections. A special development of topology theories called "analysis situs" studies the relative positions of points in a set without any metric consideration. The graph theory has provided excellent approaches in many fields, but they were usually limited to specific graphic problems until new approaches ensued from powerful recursive methods of computer programming. Some basic aspects used in this paper will be clarified hereafter.

Tree Concept: Rooted Ordered Trees and Graphs

One of the fundamental concepts in graph theory is the tree concept T. An open graph can be defined as comprising a distinctive element R1, called the root of T. The elements of T are disjoint trees—T1, T2, Tm—called subtrees. The roots of these T1, subtrees are successors of R.

Universal Address for Vertices of a Rooted Tree

The usual terminology is borrowed from genealogy. In a graph the nodes are parents, children, and siblings. In most applications properties are used to order a graph. In a genealogic presentation the dates of birth and of marriages are often used. This implies that in a "general
“tree” with three descendants the subtrees can be presented from left to right without consideration of the relative importance of the \( T_1 \) (1 site), \( T_2 \) (2 sites), and \( T_3 \) (3 sites) subtrees (order 1-2-3).

In a molecular formula, the same tree could be \( C_{T_1}, T_2, T_3 \), where \( T_1 = \text{methyl}, T_2 = \text{isopropyl}, \) and \( T_3 = \text{tertiobutyl alkyl group} \) (order 3-2-1). This ordering presented here is different from the usual DARC ordering.

**Traversing Operation and Graph Ordering (Search Tools)**

To describe a general tree, one carries out a traversing operation, going in turn from one node of a \( T_i \) to another node, before jumping to the next \( T_i \). Different traversing algorithms can be used; they are based on ordering sequences (preorder, order, or postorder).

To develop a search program of structures in chemistry, we proposed DARC rules to define the relative positions of \( T_i \) subtrees for a simple graph considered as a “general rooted tree.” Thus we define its root and order the subtrees \( T_i \) with regard to one another. With this process all the nodes are gradually numbered. It becomes easy to retrieve such an ordered graph from a database; one needs only a traversing algorithm. In the DARC system a basic algorithm takes advantage of the concentric description of the molecular graph: it is a multilayer traversing step. The final DARC name consists of the root number and the numeric descriptor of the graph environment.

**Concept of the Underlying Graph of a Chemical Formula**

A chemical formula can be viewed as a colored graph. Owing to the complexity of chemical formulas, one has to simplify their graph representation. One considers the underlying simple graph as a structure. Each non-hydrogen atom is a node; each bond is a simple edge. All other information (e.g., atom nature and multiple bonds) is considered as colors of this graph. They are subsequently taken into account and indexed to the vertices and edges of the underlying graph that is the whole spanning tree of the structure. They are accessories to the topological grasp of the existence tree (graph), or spanning tree.

To characterize each formula with such a “rooted tree underlying the structure,” one needs to open all the cycles encountered in the chemical formula. So in the labeling and ordering process is a “cut mechanism” that breaks some edges of the cyclic subgraph.

**Topological Indices of the Rooted Trees**

The traversing function used in the DARC system and the corresponding matrices lead to numerical descriptors of all structures. These site descriptors, based on topological environment concepts, are useful for building structural databases and for carrying out topological treatments valuable in assessing chemical and physical properties sensitive to structural environmental factors.

**Topological Structural Databases: Creation of Tree Information**

In the DARC topological system the chemical entities are represented by “rooted ordered trees” (ROTs); the nodes are ordered in this representation. The FREL and the DEL correspond to general spanning graphs handled as “general rooted ordered trees” (GROTs) made of subtrees corresponding to one root (FREL) or to a segmentation of the GROT (DEL) in sub-ROTs. The implementation of those trees and subtrees depends on a building process (editor or generator), using a set of abstract operators. The information products, resulting from such treatment of those trees, are the FREL, the DEL, the adjacency matrix, the adjacency list, or certain types of arrays associated with a rooted ordered tree. They are tools available for database management.

DARC used the tree-structured concepts for the first time in chemical documentation to search data files (e.g., EURECAS) with several tree-structured strategies (e.g., DUPEL). It is important to stress the coherence between input coding and query coding, the latter being converted from chemical structural questions to the DARC code in T forms.

**Internal Organization of Topo-Databases**

The organization of the ELCO-FREL “rooted ordered and limited trees” in a file generates a good deal of available information concerning the first and second spheres (A and B) around the focus (root) of the T graphs. Statistical analysis of such topological structural files helps to shape them, as it is advisable to derive special search keys based on occurrences of such local information. These keys are useful to increase search efficiency and speed of structural file searching.

For input, a formula is transformed into its DARC graph, which is kept as a numerical descriptor in the computer. Editing programs are required to edit the formula after retrieval for a final graphical reply to the user.
Since the first DARC proposals most of the developments using topological concepts retained the concentric or spherical arrangements of atoms in shells and implemented various T-structured data strategies. Supplementary information is sometimes added to the nodes. Little advantage has thus far been taken of the hyperstructure (especially hyperstructures of the substructures) potential structuring of a specific database.