Approaching the Logic of Synthesis Design

JAMES B. HENDRICKSON

Department of Chemistry, Brandeis University, Waltham, Massachusetts 02254

Received December 16, 1985 (Revised Manuscript Received June 9, 1986)

Synthesis is a branch of organic chemistry nearly as old as the science itself, and yet it is only in the last 2 decades that we have seen any discussion of the logic of synthesis design. Surprisingly, there had been no real effort prior to 1967 to formulate any logical stepwise protocol for synthesizing a compound, such as existed for carrying out an analysis, for example. Many syntheses at all levels of complexity had been accomplished, but no discussion of the thought process that led to a given choice of route was to be found. Synthetic chemists have been skeptical of such generalizations and have stayed close to the laboratory, aware that textbook generalities are imperfect. In the great renaissance of synthetic chemistry initiated by Woodward after the second world war much of the activity has been devoted to development of new reactions and methods, closely tied to laboratory practicality. Total syntheses have often been derived by exploiting such new methods as key reactions adapted to constructing some natural product family for which they were especially suited.

Interest in examining and formulating a logic for synthesis design began in several places at once in the late 1960s, owing perhaps to an awareness of the emerging capability of computers and the growth of the idea that synthesis design might be mechanized for computer application. In 1967, Corey published a ground-breaking article,[1] followed by another[2] in 1971, which explored concepts of synthesis design logic in general terms, and offered a clear description of transforms for deducing all last reactions to a target.

In 1969 Ireland[3] first specifically stated a basic axiom that has become the central mode of most computer systems for synthesis design: "If there is any key to success in planning a synthesis, it is to work the problem backwards." This was interpreted to mean that one should generate from the target structure all possible last reactions and their substrates and then repeat this process in turn on each of those, creating backwards from the target a "synthesis tree" of choices. Corey had described the logical mechanics for this process with his transforms.

This procedure of generating stepwise backwards all intermediates became the primary strategic operation in the computer programs that then developed:[9] LHASA from Corey and Wipke (1969)[4] and further developed by Corey's group,[5] and, labeled SECS, by Wipke's group;[6] an unlabeled program from Bersohn (1972);[7] and SYNCHEM from Gelernter (1973).[8] The key to the former two is an interactive mode in which the chemist–operator makes choices from the synthesis tree as soon as precursors are generated, while the latter two programs are executive, or noninteractive, in the sense that the choices are made by the program and the chemist is presented with the "best routes" when it is finished.

The synthesis tree generated by stepwise application of transforms is shown as commonly presented on the left side of Figure 1, the target (T) at the top, all first-level precursors as points aligned just below, their precursors at the second level below that, etc., with available starting materials circled back within the tree. The central problem of such a "blindly" generated tree lies in the vast number of intermediates, growing combinatorially with each level. If each molecule has an average of n reasonable precursors, then there must be n^k sequences, or synthetic routes, created at level k. If each has 40 precursors, there will be $40^5 > 100$ million routes of five steps, and most syntheses are over twice as long.

Thus, it is relatively easy to generate intermediates and routes mechanically, but the major task then becomes that of selection, of "pruning the tree" to a manageable number of results, based on applying some heuristic criteria. The tactic in the interactive programs is mainly one of leaving the selection to the chemist–operator, but at the first levels of the tree he cannot see which retrosynthetic starts lead to early discovery of starting materials and so of short pathways. In practice these programs are commonly used for only one or two steps ("What reactions will make compound A?"). The tactic in the executive programs is to predict yields or merit ratings of the various reactions in the database library. This tactic suffers from the obvious imprecision of such predicted ratings when comparing thousands of reactions.

A more subtle problem with these programs arises from the fact that they are functionality-directed. In many syntheses, dummy functional groups are used to direct target construction and then removed, leaving no

James B. Hendrickson graduated from Caltech in 1950 and did his Ph.D. thesis at Harvard with R. B. Woodward. He was then a posdoctoral fellow first in London with D. H. R. Barton and then at Harvard with Woodward. He joined the chemistry faculty at UCLA in 1957 and moved to Brandeis in 1963, where he is now professor of Chemistry. He has received Sloan, Guggenheim, and Fulbright Fellowships, and is interested in organic synthesis and the logic of synthesis design.

[1] Corey, E. J. *Pure Appl. Chem.* **1967**, *14*, 19.

[2] Corey, E. J. *Quart. Rev.* **1971**, *25*, 455.

[3] Ireland, R. E. *Organic Synthesis*; Prentice-Hall: Englewood Cliffs, NJ, 1969.

[4] Corey, E. J.; Wipke, W. T. *Science* **1969**, *166*, 178.

[5] Corey, E. J.; Long, A. K.; Rubinstein, S. D. *Science* **1985**, *228*, 408, and earlier references cited therein.

[6] Wipke, W. T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. *ACS Symp. Ser.* **1977**, No. 61.

[7] Bersohn, M. *Bull. Chem. Soc. Jpn* **1972**, *45*, 1897. Bersohn, M.; Esack, M.; Luchini, J. *Comput. Chem.* **1978**, *2*, 105. Bersohn, M.; MacKay, K. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 137.

[8] Gelernter, H.; Sridharan, N. S.; Hart, H. J.; Yen, S. C.; Fowler, F. W.; Shue, H. J. *Top. Curr. Chem.* **1973**, *41*, 113. Gelernter, H. J.; Sanders, A. F.; Larsen, D. L.; Agarwal, K. K.; Bovie, R. H.; Spritzer, G. A.; Searlemen, J. E. *Science* **1977**, *197*, 1041. Agarwal, K. K.; Larsen, D. L.; Gelernter, H. J. *Comput. Chem.* **1978**, *2*, 75.

[9] A very different, mathematically based approach to synthesis plans was developed in Munich by Ugi and Gasteiger.[10]

[10] Ugi, I.; Gillespie, P. *Angew. Chem., Int. Ed. Engl.* **1971**, *914*, 915. Dugundji, J.; Ugi, I. *Top. Curr. Chem.* **1973**, *39*, 19. Blair, J.; Gasteiger, J.; Gillespie, P. D.; Ugi, I. *Tetrahedron* **1974**, *30*, 1845. Gasteiger, J.; Jocum, C. *Top. Curr. Chem.* **1978**, *74*, 93.

0001-4842/86/0119-0274$01.50/0

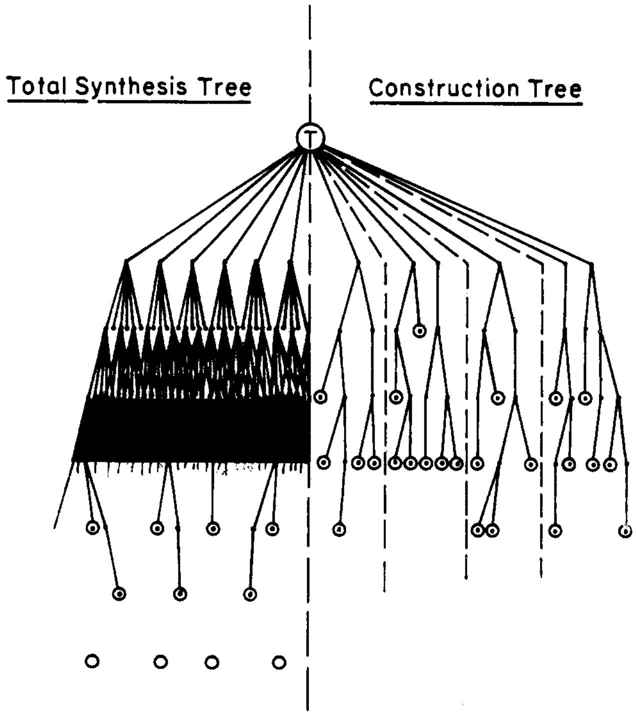
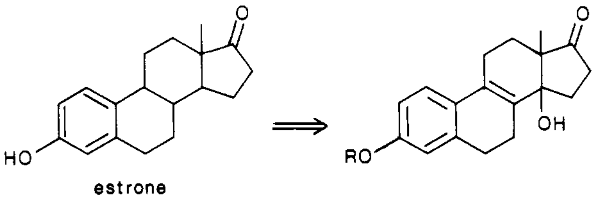


Figure 1. The synthesis tree.

trace in the target functionality. It is not clear, for example, what functionality-directed protocol would deduce this last step, used in a good synthesis of estrone.



It was the intellectual challenge of exploring and systematizing this logic of synthesis design that first attracted our attention in that early period.[11] This account reviews first the steps we took in trying to approach the logic in a new way and then the computer program (SYNGEN) which grew from it to articulate the approach and assess its results. It was clear that the combinatorial problem presented by the synthesis tree was enormous and that there exists a huge number of synthetic routes to any target of reasonable complexity. The question is, Which ones are the best? Accordingly, we must decide first on stringent criteria for selection of the best routes from the tree, and to seek these we must undertake a tree search which is assured of assessing all possibilities. The experience of tree searches in areas outside chemistry implies that the tree must first be simplified and subdivided to allow a full search in a reasonable time.

It also seemed clear that the mechanical generation of reactions stepwise backwards from the target was not the process used by synthetic chemists in their design of syntheses. Most syntheses appear to derive from a key reaction especially suited to construct the target skeleton. Such a central concept then determines the necessary starting materials, on the one hand, and the further forward reactions to the target on the other. Thus, only in a second phase of the reasoning are particular functional groups and reactions considered, both up to the key reaction and proceeding from it.

Skeletal Dissection

We modeled our logic on this idea of seeking key reactions. In simplest terms, synthesis is a skeletal concept consisting essentially of building a large molecule from a number of small starting materials. Hence, the key reactions are the construction reactions,[12] which serve to assemble the target skeleton from the skeletons of starting materials. The simplest gross description of any synthesis is simply the set of skeletal bonds which are constructed. This is called a *bondset*, or an *ordered bondset* if the order of their constructions is also indicated, as in the two estrone skeleton cases of Figure 2. In this view, the first step in synthesis design is to examine only the skeleton of the target, dissecting it with bondsets to find the most efficient mode of assembly of the pieces. Any defined bondset also shows directly the starting pieces needed, i.e., starting material skeletons, as shown in Figure 2.

Initial consideration of the skeleton only affords a major simplification of the total synthesis tree (left, Figure 1) to a *construction tree* (right, Figure 1), in which the points are skeletons only and the lines are constructions of particular skeletal bonds. Thus, many compounds of the same skeleton are coalesced in each point, and many different construction reactions forming the same skeletal bond are included in each line. The forked lines are reactions which link two skeletal pieces intermolecularly (*affixations*);[13] and the single, vertical lines are *cyclizations* (intramolecular).

While the original synthesis tree is an unordered and undiscriminating collection of all possible reactions, the simplified construction tree contains only the key constructions which assemble the target skeleton from starting skeletons. Furthermore, each full synthetic sequence is a separate *construction plan*,[13] separated by dotted lines in Figure 1. Each such plan is an independent smaller subtree that may be examined separately for its detailed chemistry. Hence, a full tree search now becomes manageable by subdividing it into a series of smaller tasks taken sequentially. Each construction plan corresponds to a particular ordered bondset, and vice versa. The ordered bondset for the Velluz[16] synthesis of estrone is shown in Figure 2 (top left) with its construction plan (I) shown below; either can be created from the other. Three other construction plans (II–IV) of different orders for the same bondset are also shown.

We can derive these construction plans systematically by dissecting the target skeleton all possible ways into smaller pieces, i.e., by designating bondsets. There is an enormous number of ways to do this;[13,14] probably far more than are generally appreciated. If we cut λ skeleton bonds out of a total of b bonds, there are $\binom{b}{\lambda}$ possible bondsets and each has $\lambda!$ possible orders. We did a survey of many published total syntheses which

[11] Hendrickson, J. B. *J. Am. Chem. Soc.* **1971**, *93*, 6847, 6854; *J. Chem. Ed.* **1978**, *55*, 216.

[12] A structure is composed of its skeleton and its functional groups, the skeleton understood as the framework of C–C σ-bonds. Reactions which create skeletal bonds are *constructions*; those which alter functional groups without affecting the skeleton are *refunctionalizations*.

[13] Hendrickson, J. B. *J. Am. Chem. Soc.* **1977**, *99*, 5439.

[14] Hendrickson, J. B. *J. Am. Chem. Soc.* **1975**, *97*, 5763.

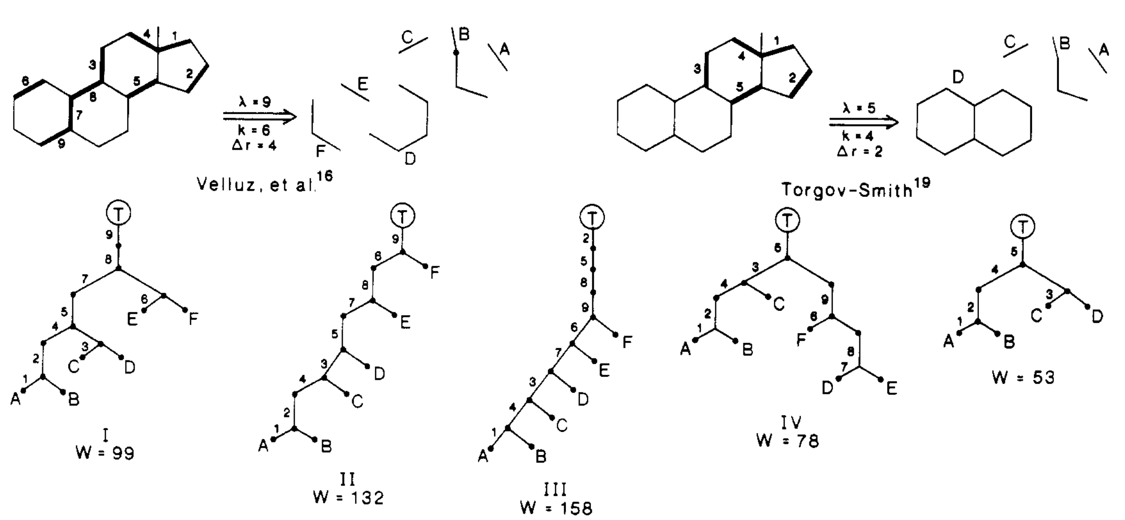


Figure 2. Bondsets, construction plans and weights for estrone (boldface = bonds constructed in numbered order).

shows that they commonly construct $^1/_3$–$^1/_4$ of all skeletal bonds in the target, with an average starting skeleton size (excluding aromatics) of only about three carbons incorporated into the target.

If we examine the 18-carbon skeleton of estrone in that light, an "average" synthesis should utilize six pieces of average C_3 size and so construct nine skeletal bonds ($\lambda = 9$).[15] There would be $\binom{21}{9}$ = 293 930 possible bondsets with 9! = 362 880 different orders to assemble each one. The bondset representing the actual synthesis of Velluz,[16] shown in Figure 2, is just one of the 107 billion possible. This ordered bondset corresponds to construction plan I shown below it.

It is an easy matter to dissect any skeleton all ways into bondsets and generate construction plans for all orders, but as so many are possible, it is critical to ascertain which are the best. It is possible to assess the efficiency of each plan by a simple calculation of the total weight of the several starting materials required, assuming a standard yield for each construction.[13,17] These relative weights are shown for the syntheses in Figure 2. When various weights are compared in this way, it is clear that the most efficient syntheses are fully convergent ones. Velluz himself first elucidated the idea of the convergent plan,[18] and his plan (I, Figure 2) is clearly more efficient than the two linear plans, II and III (which also show that early cyclization (II) is more efficient than late (III)). However, his plan is not *fully* convergent; the fully convergent plan is IV in Figure 2 and represents the most efficient assembly of the particular pieces (A–F) employed by his bondset. The calculation of relative weights of starting materials, in this way, constitutes a powerful tool for comparing the efficiency of different syntheses (convergent or linear), either at the level of skeletal assembly or when all detailed functional groups and their refunctionalizations are included.[13]

Fully convergent bondsets of any target skeleton are readily created by dissection of the target into two pieces, then cutting each of these in two again, etc. When all the pieces, so obtained, also correspond to skeletons of available starting materials, this creates an ordered bondset for a potential synthesis using those starting materials. The larger the found starting material skeletons, the fewer dissections necessary, hence the fewer constructions in the synthesis itself, which is then more efficient. Such an efficient, fully convergent synthesis of estrone is that of Torgov and Smith,[19] also shown in Figure 2 with the lowest calculated weight of all ($W = 47$). Unlike the stepwise backwards approach, this skeletal dissection procedure strikes down many levels into the synthesis tree to focus the search early on the closest available real starting materials.

At the outset, we saw that the huge size of the synthesis tree obliges us to apply very stringent criteria. The simplification and subdivision of the tree into the essential constructions of the skeleton, indeed, allows a dramatic reduction of the possible options if we restrict them to convergent bondsets of lowest weight from available starting skeletons. In the second phase of the analysis, we must examine the varieties of possible functional groups and reactions necessary to carry out the construction sequences of the bonds demanded by the few optimal bondsets. Here each single construction plan for skeleton assembly swells again to many possible routes dependent on the choice of actual construction reactions and the necessary functionality for their successful initiation. Here again, in the second phase, we must seek a powerful criterion to reduce the many possible synthesis routes to a select few.

The obvious hope that the reaction yields should afford such a criterion is dashed by the imprecision of prediction of these yields when used to compare so many choices. However, the simpler criterion—that the number of steps at least be minimal—is still a powerful selector.

If the fewest steps is the goal and the only obligatory reactions are constructions, then, obviously, the optimal syntheses are those composed only of sequential con-

[15] A skeleton with n atoms and r rings has $b = n + r - 1$ bonds. If there are k starting material pieces, or components, for a synthesis, then $\lambda = k + \Delta r - 1$, where λ is the number of bonds in the bondset.

[16] Velluz, L., et al. *C. R. Hebd. Seances Acad. Sci.* **1960**, *250*, 1084, 1510; **1963**, *257*, 3086.

[17] Since at the skeletal stage of analysis we have no functional groups, the number of carbons (n_i) for each starting skeleton, i, is used in place of molecular weight. The standard yield is taken as $y = 1/x$ and the total starting material weight is $W = \sum_i n_i x^{1_i}$, where 1_i is the tree level of skeleton i, i.e., number of construction steps it passes through to the target. Such weights are relative numbers serving to compare different construction plans, the lowest weights being the most efficient. Weights in Figure 2 were calculated for 75% yield, i.e., $x = 1.33$.

[18] Velluz, L.; Valls, J.; Mathieu, J. *Angew. Chem. Int. Ed. Engl.* **1967**, *6*, 788.

[19] Ananchenko, S. N.; Torgov, I. V. *Tetrahedron Lett.* **1963**, 1553. Smith, H. et al. *Experientia* **1963**, *19*, 394; *J. Chem. Soc.* **1963**, 5072.

struction reactions, with no refunctionalization needed to repair functional groups between constructions. This "ideal synthesis" is a very rare occurrence, but constitutes a sharply defined goal in selecting routes and also puts considerable demands on the choice of starting materials with correct functionality. We proposed, then, to seek those routes with fewest steps, minimizing the use of refunctionalization reactions.

Functionality and Reactions

The second phase of the analysis calls for a search among the very large variety of detailed functional groups and reactions. In large tree searches, when the units in the search space are too many, they must first be coalesced or abstracted into fewer "super units" or families of units. There is an analogy with maps here. As the mapped space enlarges, details are coalesced or omitted: town outlines become dots, houses are omitted, and information is lost. But when a particular desired place is located, a more detailed map of just that area can restore the detail. In the synthesis tree, the units in the search space are molecular structures and reactions, and these can be generalized into fewer descriptive units by coalescing trivial distinctions. Information so lost can be restored later when a small optimal group of units has been selected, but meanwhile, the tree has fewer units and is more readily searched exhaustively. This less detailed mapping is an important constraint for making a full search rapidly. Other existing programs employ full mapping, describing every atom in normal detail, so that many more fine distinctions must be separately examined, and a reaction database of thousands of known reactions must be prepared. This is ultimately too detailed a mapping to cope comprehensively with the size of the tree in a reasonable time.

We elected a simple numerical description of functionality,[11] giving each skeletal carbon a number to express its functional type; so any structure becomes a compact list of simple numbers, ordered by the numbering of the skeletal carbons. Thus, computer manipulation of many structures is very fast, and trivial distinctions are generalized (all leaving groups are one number, all acylating groups another, etc.). A reaction is expressed by its net structural change, which becomes simply the arithmetic change in such a functionality list from substrate to product, or vice versa. A "reaction" is then just a number list, across the involved carbons, which generates the substrate functionality list when added to the product list (or vice versa). This numerical generation of reaction products from substrates, or the reverse, has several advantages: it is a very fast process for the computer; all possible conversions become simply a set of all possible mathematical combinations so that none will be missed; no library or database of existing literature reactions need be laboriously created, updated, and searched; presently unknown conversions are generated as possible new chemistry; and, finally, mechanistic tests of reaction viability can also be made by quick numerical checks based on the nature of the functionality around the site of reaction.

The system developed to describe structures[11,20] defines four generalized kinds of attachment on any carbon atom: H for hydrogen (or other electropositive elements), R for σ-bond (skeletal bond) to another carbon; Π for a π-bond to carbon, and Z for a bond (π- or σ-) to electronegative heteroatom. For any carbon, then, the number of attachments of each kind is, respectively, h, σ, π and z, such that $h + \sigma + \pi + z = 4$. The functionality is then $\pi + z$ and, since the skeleton is given, σ is known and h derives by subtraction. The result is that the functionality on any carbon is expressed by two digits, z and π. The functionality list for any structure is then a $z\pi$-list of its carbons ordered by their skeletal numbering. For example, crotonic acid (ester, nitrile, etc.), linearly numbered (IUPAC rules), becomes 30.01.01.00 and its 2,4-dichloro derivative is 30.11.01.10. The fundamental nature of the description is substantiated by the observation that the oxidation state (x) at any carbon is given by $x = z - h$, and so the oxidation state change in any reaction is quickly calculated by $\sum \Delta x$ over all changing carbons.

Reactions are characterized in this system very clearly and simply. A *unit reaction* is defined as a unit exchange of attachments on one carbon, and can be expressed as two letters, the first being the kind of attachment bond which is made and the second being that which is broken. Thus, the reduction of alkyl halide to alkane is an HZ unit reaction, as is reduction of ketone to alcohol. The oxidation state change is found from $\Delta h = +1$, $\Delta z = -1$ and so $\Delta x = \Delta z - \Delta h = -2$. Some reactions must involve more than one carbon at the same time, as in alkene reduction, HΠ·HΠ ($\sum \Delta x = -1 + (-1) = -2$), and of course in all constructions, such as alkyllithium addition to ketone, which is RH·RZ, with $\sum \Delta x = 0$. There are 16 possible unit exchanges which may be written from combinations of the four kinds of attachments on one carbon.

This system makes possible a very clear and simple basis for characterizing and cataloguing all possible organic reactions in terms of their *net structural change*, i.e., exchange of attachment types at the several involved carbons. Such a system for organizing reactions[21] is analogous to the Beilstein system for organizing structures in that all possible reactions, presently known or unknown, have a defined place in the catalog. This can certainly be a very useful basis for defining, creating, and searching a compendium of organic reactions.

We can define any construction reaction as a combination of two half-reactions, one on each side of the constructed bond, one half nucleophilic (oxidative), the other half electrophilic (reductive). The carbons on each side of the bond will undergo RH, RZ, or RΠ exchanges. Those with RΠ will require the adjacent carbon also to exchange Π. In general, a strand of up to six carbons spanning the constructed bond (up to three on each side) contains all the carbons whose functionality changes in a construction. Thus, all possible construction reactions can be defined. There are three oxidative and three reductive half-reactions on three carbons or less, combining to nine basic construction unit reactions:

	$\sum \Delta x = +1$	$\sum \Delta x = -1$
simple	RH	RZ
π-addition	RΠ·ZΠ	RΠ·HΠ
allylic	RΠ·ΠΠ·ΠH	RΠ·ΠΠ·ΠZ

[20] Hendrickson, J. B. In *Mathematical and Computational Concepts in Chemistry*; Trinajstic, N., Ed.; Ellis Horwood: Chichester, W. Sussex, U.K. 1985; Chapter 11.

[21] Hendrickson, J. B. *J. Chem. Inf. Comput. Sci.* **1979**, *3*, 129.

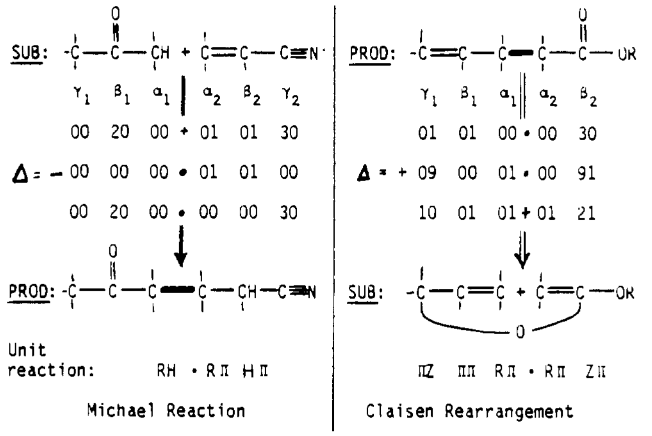


Figure 3. Examples of reaction generators.

Further, some real constructions involve another unit reaction, a spontaneous refunctionalization,[12] as in the elimination following aldol or Wittig constructions or the prior reduction of a halide to create a Grignard nucleophile. There are six of these composite constructions added to the nine simple ones in current use. Each of these 15 half-reactions is characterized by a *zπ-list generator* which, on addition to the $z\pi$-list of a product strand through the constructed bond, will create the two substrate $z\pi$-lists. Subtraction of the generator analogously creates product from substrates. This is illustrated for two constructions in Figure 3: one in the forward and one in the retrosynthetic direction.

This numerical description of functionality is very versatile and closely parallels real chemistry. It not only allows almost instant derivation of requisite substrates for particular reactions and allows all possible reactions to be systematically formulated, but there is also surprisingly little information loss from normal reaction description; part-structure representations of reactions are easily reconstituted from these number lists.

Furthermore, the digital expression has another useful aspect. Quick numerical checks of the values of h, σ, π and z on the carbons near the construction bond serve as rough tests of the mechanistic viability of a reaction. For each half-reaction we can test for necessary activating functionality, for correct regioselectivity, for interfering side reactions, etc. Here we find that coalescing all heteroatom types into the number z is too severe a condensation for mechanistic discrimination. Therefore, we further define, as a subset of z, the *function* of the attached heteroatom, as leaving group, electron-withdrawing group, or electron-donating group. This allows a more precise view of their effects on specific construction mechanisms.

Finally, we can also include skeletal nitrogens, essentially cyclic ones, by treating such nitrogens as "special carbons" with special tests applied to the viability of construction strands containing them. A special feature of the digital description allows all these restriction tests to be made at once in one quick bit-comparison test in the computer. A bit list of the functionality $z\pi$ and z-functions for all involved atoms for a given product is first assembled and then compared with a similarly assembled check-list for each half-reaction to see first, if required activation is present, and second, if objectionable functions on any atom are present to reject the reaction.

The SYNGEN Program

With this approach to a logical development in hand, we undertook to write a computer program embodying its principles. In summary, this amounts to an initial survey of a target skeleton to find the key construction sites, i.e., the most efficient, fully convergent plans for assembling the skeleton from the largest starting material skeletons in the catalog. Following this, actual construction reactions are generated and tested, sequentially through the bonds of each bondset back from target to starting materials. Routes which do not utilize catalog starting materials (functionality, as well as, skeleton), or which involve mechanistically nonviable construction reactions, are rejected. In effect, an optimal set of routes, consisting of only "ideal" (shortest) syntheses, is generated, and these are expressed with functionality still abstracted in numerical terms, still requiring further refinement, in detail, by the chemist.

The program developed, called SYNGEN (SYNthesis GENeration),[22,23] has been written in FORTRAN (about 6000 lines) for a minicomputer, the DEC 11/23. It requires about a megabyte of active memory and generally analyzes a target structure in under 10 min. The program involves neither operator interaction nor a database library of reactions, but it does include a catalog of available starting materials, currently including about 5000. It is presently being rewritten for greater efficiency, speed, and exportability on a micro-VAX computer.

The operation of SYNGEN may be illustrated with an example, shown in Figure 4. The economical Torgov–Smith synthesis of estrone[19] is an "ideal" synthesis of constructions only, up to a final intermediate with the target skeleton, which is then refunctionalized to estrone. That intermediate was labeled "testrone" and used as one model for program development, since the program must find known syntheses if it is to be realistic. The structure of testrone is entered by a simple drawing program, directly simulating a chemist's drawing, onto a Tektronix CRT screen. The drawing is done quickly and crudely and is then normalized by the computer to output a "clean" drawing with z-values entered on the carbons bearing heteroatoms, e.g., shown as (A) in Figure 4.

SYNGEN then proceeds to skeletal dissection, cutting the skeleton all ways into two pieces such that all contain at least three carbons. This is the first level, shown down the left side in Figure 4 with one such cut and ordered bonds B. The two pieces are now compared with skeletons in the starting material catalog: here C is not found, but D is found and so this partial bondset is marked for priority. Precursor C is now cut again all ways at second level, retaining only cuts giving found starting skeletons; one such set is E and F, and this line then creates one total ordered bondset, convergent in two levels of dissection.

For each such accepted bondset the functionalized target skeleton A is now retrosynthetically queried for viable construction reactions, shown down the right side of Figure 4. Among the two sequential constructions for bonds 1 and 2 (on left side) are several, like the one

[22] Hendrickson, J. B.; Braun-Keller, E.; Toczko, A. G. *Tetrahedron, Suppl.* **1983**, *37*, 359.

[23] Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. *J. Am. Chem. Soc.* **1985**, *107*, 5228.

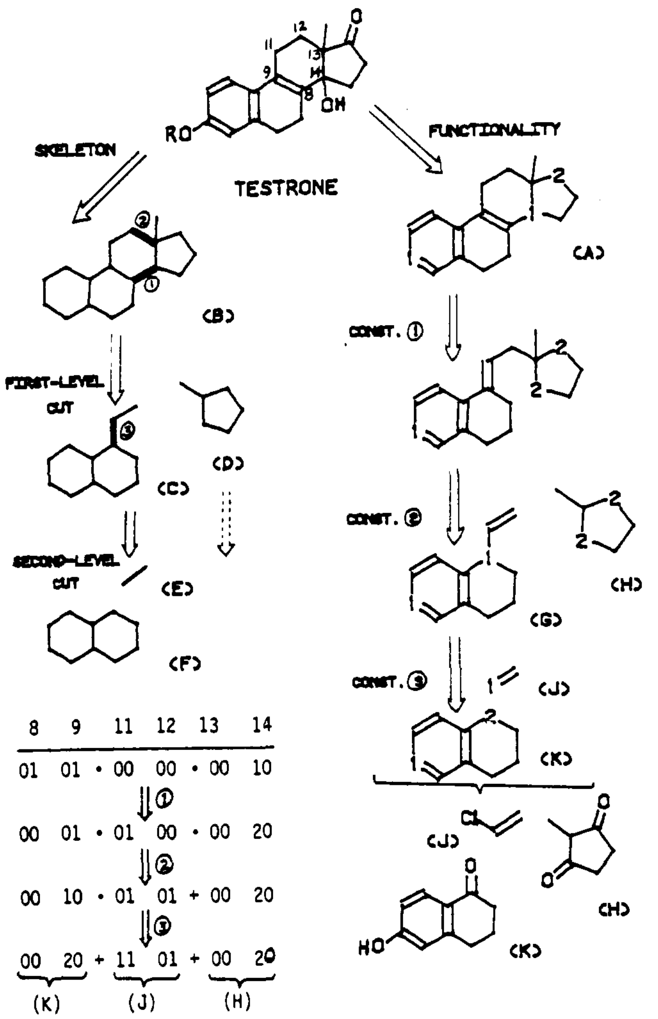


Figure 4. Analytical steps in generating a synthesis.

shown, which are flagged for priority since they are perceived as multiple constructions, i.e., a true annelation capable of occurring in one laboratory operation.[23] These generated precursors are G and H; the latter is now again searched in the catalog to determine if its functionality, as well as its skeleton, is found. Since the actual starting material H is found, the route retains its priority status. The dissection of G at bond 3 is now examined, and starting materials J and K are found for a particular construction reaction (vinyl organometallic addition to ketone). Hence, a full route has been found to effect the assembly of the bondset generated down the left side. The actual catalog starting materials H, J, and K are shown below in conventional notation. At the lower left in figure 4 are recorded the $z\pi$-lists generated in the sequence for A to starting materials H, J, and K, ordered over the changing carbons (#8, 9, 11, 12, 13, and 14).

All the results from the several successful bondsets are sorted and stored by SYNGEN for display from a second program (SYNOUT), in several ways, allowing a variety of different selections to be made and viewed from the optimal set of found routes. SYNOUT will sort routes by priorities, dubious constructions, groups of "chemically equivalent" routes, etc., and display bondsets, intermediates, starting materials, or reactions on command, plotting those of interest as shown in Figure 4 (the structures shown were taken directly from the plotter). For testrone, SYNGEN generated the known synthesis[19] of Figure 4 and also several other routes equally short, which appear to be chemically reasonable and without literature precedent. Although we have used estrone as a descriptive model, the results from many other targets are equally interesting.[23]

The procedure involves extensive comparison of generated compounds, both with each other and with the catalog set. For these comparisons we had to devise a new protocol which creates a unique canonical numbering of any skeleton for its unambiguous identification.[24] If two skeletons are each numbered by a rigorous canonical procedure, their comparison for identity is valid. However, for a skeleton of n atoms, there are $n!$ possible numberings. Since the skeleton is a graph, it may be fully characterized by its adjacency matrix, a symmetrical $n \times n$ connectivity matrix with 1 entries for a bond (ij) and 0 entries for no bond; there are $n!$ such matrices, one for each numbering of the n atoms. If the adjacency matrix entries are strung out into a linear binary string, we treat this string as a binary number. The particular matrix for which that is the *maximum* binary number is then unique. This maximum binary string, like the matrix, also fully characterizes the skeleton. The particular numbering resulting in a maximum binary string is called the maximal numbering and can be developed by a stepwise procedure growing the maximal matrix.[24] Any two skeletons with the same maximal binary string are then identical.

The catalog is kept as a listing of maximal binary strings for the skeletons of its compounds, ordered by skeletal size and listed in numerical order of the binary strings; such an ordered listing can be very rapidly searched. Since the program separates skeleton and functionality, each compound is characterized by its skeletal (connectivity) binary string followed by its $z\pi$-list of functionality ordered by the maximal numbering of the skeleton. Within symmetrical skeletons having more than one equivalent maximal numbering (with the same maximum binary string) the $z\pi$-lists further secure maximal numbering by being themselves maximized. This protocol not only allows rapid, unambiguous comparisons of structures for identity and rapid catalog search, but can produce a shorthand identification notation ("T/R-list"), compact for computer storage, which also perceives the rings (R) present and the overall spanning tree (T) of the skeleton which links them.[25]

SYNGEN generates only the shortest ("ideal") sequences of constructions; we were surprised at how many appear, even after sorting out minor variants ("chemical equivalents")[23] and flawed chemistry. Published syntheses, on the average, contain twice as many refunctionalizations as constructions and so are generally longer. Refunctionalizations are used to alter starting material functional groups before construction, to repair functionality between constructions, and to convert a fully constructed skeleton to the right functional groups, as in testrone to estrone above.

We considered that judicious introduction of some refunctionalizations would produce some useful new routes, somewhat longer, but still practical. To this end, another useful feature of the digital characterization

[24] Hendrickson, J. B.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171.

[25] Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. *Ibid.* **1984**, *24*, 195.

system allows the calculation of "*reaction distance*" between any two structures, i.e., the number of unit reactions (N) needed to convert one to the other.[26] This is a simple calculation summed over each changing carbon: $N = 1/2\sum_i (|\Delta h_i|) + (|\Delta z_i|)$. Here we can query any generated starting material to see if it can be made in one or two refunctionalizations from one in the catalog with the same skeleton. Routes from these materials may then be accepted with lower priority.

Refunctionalizations en route may be defined in another way with the "reaction distance" formula. This is done with a forward sequence generation instead of the normal retrosynthetic one (e.g., Figure 4). The optimal bondsets for skeletal dissection are created as before; then, for each bondset, we examine *all* the catalog starting materials of the right skeletons. For each of these we calculate the reaction distance of its carbons to their corresponding places in the target structure—a sum of construction and refunctionalization reaction distances. Those that have small enough distances are combined all ways into intermediates which are again tested for acceptable, converging reaction distance to target. In this way, when estrone itself is offered as target, the bondset of skeletons D, E, and F in Figure 4 affords the same routes via testrone and its final refunctionalization ($N = 2$ steps) to estrone.

We have several plans for improving and expanding SYNGEN output for the future. The simplest of these is not only to expand the starting material catalog, but also to grade it for categories of starting material cost, allowing the operator to select the cheapest synthetic routes. Another is to apply the Bertz concept of molecular complexity,[27] said to correlate with synthetic efficiency,[28] to the generated sequences. This will serve not only to test the concept, but also to use it as a pruning tool to give priority to least complex routes.

There are other skeletal dissection modes which are more efficient when multiple constructions[23] can be used, and these are being defined and incorporated. There are also particular cases in which C–C bond fragmentations or rearrangements are advantageous; these are not now employed, but are planned for the near future.[30]

The problem of stereochemistry has not been addressed, partly because of the perceived importance of other hitherto neglected aspects of synthesis, i.e., efficiency of steps and skeletal assembly. In the optimal set of all best routes generated, we presume that the chemist will recognize those amenable to stereocontrol, and methods for stereocontrol have so proliferated in recent years that they no longer constitute the synthesis design bottleneck they one represented. Even so, within the present system lie a number of devices which can be used to recognize stereochemistry and we shall incorporate these into route selection.

Finally, we plan to tie the SYNGEN output to a library database of reactions, not as in other programs to direct the selection of routes, but rather to afford specific literature validation for reactions already produced by the generation procedure in our program. The SYNLIB collection is especially well suited to this purpose since its algorithmic description of reactions can be smoothly spliced to that used in SYNGEN.[29] Indeed, the system of organizing reactions which has developed[21] from our digital description can be profitably applied to overlay a simple search catalog on the SYNLIB reaction collection, suitable than for finding subsets of reactions in less specifically defined reaction families.

Conclusion

What seems important in our approach is the application of several fundamental concepts of synthesis design specifically applied. The first of these is a clear focus on the optimal modes of skeletal assembly: a recognition of the enormous variety of possible assembly modes and a deliberate choice of the most efficient, i.e., the fully convergent plans. The skeleton of the target is specifically seen as the framework of carbon–carbon σ-bonds only. C–Z bonds which are simple appendages (e.g., alcohols and ketones) are naturally recognized as functional groups on the skeleton, but in this strict view the C–O–C and C–N–C linkages are also seen as functionality on the carbon skeleton since C–Z bonds are relatively much easier to make than C–C bonds. With the skeleton so defined, the assembly of the skeleton becomes strictly a matter of carbon–carbon bond construction reactions, and so these are clearly distinguished from refunctionalization reactions which do not alter the carbon skeleton. In our expanded approach, such heteroatom linkages, especially in rings, can also (and separately) be accepted as part of the target skeleton, but this is done clearly aware of the distinction; routes from either choice can then be separately generated and compared.

Another basic thrust of the approach is the deliberate contraction of the synthesis tree to make a full search more manageable. This is done in several ways. First, the focus on real starting materials in the synthesis tree allows the search to converge on them directly from the start, avoiding blind generation of routes stepwise backwards from the target. Second, the plan specifically seeks the shortest routes, hence the focus on the primary construction reactions, contracting the synthesis tree to a construction tree and separately examining the individual construction plans characterized by ordered bondsets, in particular, only the most efficient fully convergent bondsets or plans. Third, each of these plans is further simplified by abstracting the involved functional groups to simple numerical (digital) descriptions to coalesce groups into families by their synthetic function. Refinement of these groups to detailed chemistry is then left to the end, after an optimal set of synthetic routes has been selected.

Finally, there is a broader implication to the numerical description system introduced in that it affords a sharp, clear, mathematical basis for organizing all possible organic reactions into an ordered system suitable for creating a catalog of reactions analogous to the Beilstein system for cataloguing structures. The fundamental nature of our digital description system is emphasized by its accurate characterization both of the oxidation states of carbon compounds and of the reaction distances between them, i.e., the number of

[26] Hendrickson, J. B.; Braun-Keller, E. *J. Comput. Chem.* **1980**, *1*, 323.

[27] Bertz, S. *Chem. Commun.* **1981**, 818; *J. Am. Chem. Soc.* **1981**, *103*, 3599; **1982**, *104*, 5801.

[28] Bertz, S. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; 1983, p 206.

[29] Still, W. C., private communication.

[30] A manuscript has been submitted, systematizing the synthetic functions of fragmentations and rearrangements and their application to synthesis design.

unit reactions minimally needed to convert one to another.

The intent of the SYNGEN program is then to provide an optimal set of synthetic sequences to a target from real starting materials, and this set can then be used as a basis of comparison with routes otherwise created by the imaginations of practicing chemists. A specific analytical tool is provided, in the total weight of required starting materials, to make such a comparison. This is not to say that the sequential construction routes found here are necessarily the best; a truly elegant synthesis can still be shorter, but a basis for comparison is at least provided. In any case, our goal is not to replace "art in organic synthesis", but to provide standards of comparison against which true art will be more clearly seen.

The author gratefully acknowledges the computer expertise and enthusiasm of his coworkers: A. Glenn Toczko, David L. Grier, Elaine Braun-Keller, and Zmira Bernstein; and financial support provided by the National Science Foundation.

Anion States of Organometallic Molecules and Their Ligands

JUDITH C. GIORDAN,* JOHN H. MOORE, and JOHN A. TOSSELL

Chemistry Department, University of Maryland, College Park, Maryland 20742

Received August 12, 1985 (Revised Manuscript Received February 13, 1986)

Central to the understanding of the structure and reactivity of organometallic compounds is the need to characterize the properties of the ligand-to-metal bond. Heterogeneous catalysis at metal surfaces, homogeneous catalysis, the biological activity associated with metal centers (e.g., nitrogen fixation and the transport of O_2 in the blood), the ability of metal complexes to undergo substitution and isomerization reactions, and the ability to influence the regioselectivity of nucleophilic addition to unsaturated ligands are all governed by the reactivity and lability of specific ligand-to-metal bonds.

These bonds are intriguing in that they are basically coordinate covalent with the ligand supplying both of the required electrons. A theory describing complex formation between an electropositive metal and electron-donating ligands must account for the electron density distribution as well as the spatial orientation of the ligands. Pauling suggested back-bonding as a mechanism for delocalization of electron density on the metal.[1]

In back-bonding, electron density from the metal is transferred onto the ligand, thus reinforcing the bond between the two centers and reducing the magnitude of charge separation. When incorporating this idea into molecular orbital theory, one distinguishes two types of ligand orbitals, those of σ-symmetry and those of π-symmetry with respect to the metal–ligand bond axis. A metal–ligand σ complex forms as a result of overlap between ligand and metal atomic orbitals; this overlap is however modified by interaction of the metal–complex orbitals with filled or unfilled orbitals of π-symmetry on the ligand.[2] This is illustrated in Figure 1 for the case of a typical octahedral complex. The metal d orbitals are split in the field of the ligands to give the differentiated σ-complex orbitals t_{2g} and e_g.

For many important ligands there is also the possibility of a π interaction with the σ-complex. This would be the case for CO, which has π^* orbitals of appropriate symmetry, or for ligands such as PR_3 where orbitals on the ligating atom also have π symmetry with respect to the metal-ligand bond axis. This interaction results in a stabilization of the occupied bonding orbitals of the complex (t_{2g}) at no cost in energy since the t_{2g}^* orbitals which are concurrently destabilized are unoccupied. Thus, there is a redistribution of charge, or a back-bonding, away from the electropositive metal center onto unoccupied orbitals of the ligand with a concomitant increase in bond energy.

This back-bonding may give rise in some cases to only subtle energy differences in the molecule while in others, as in the case of some metal carbonyls, larger energy changes can occur resulting in complexes for which the first optical transition is so high in energy as to render the complex colorless. In addition, the extent of back-bonding between the metal and certain ligands governs ligand lability in competitive substitution reactions.

To appreciate back-bonding fully requires a knowledge of both the energy and orbital character of the participating unoccupied ligand orbital. Electron transmission spectroscopy (ETS) is an experimental technique which can aid in this.[3] This method measures gas-phase electron affinities corresponding to

Judith C. Giordan received her Ph.D. in organic chemistry from the University of Maryland in 1980. After a year as an Alexander von Humboldt Postdoctoral Fellow with Prof. Hans Bock at the University of Frankfurt, Germany, and short stints as a Visiting Professor of Chemistry at Dartmouth College and as a Gillette Corporation Fellow, she joined Polaroid Corporation where she is now a senior scientist engaged in applied photographic research and development of new photosystems.

John H. Moore received his Ph.D. in physical chemistry from The Johns Hopkins University in 1967. He spent 2 additional years at Johns Hopkins developing the technique of ion impact spectroscopy and then joined the faculty of the University of Maryland in 1969, where he is Professor of Chemistry. He has been a JILA Visiting Fellow and a Program Officer with the National Science Foundation. His research work includes a variety of electron spectroscopies as well as instrument development for space exploration.

John A. Tossell received a Ph.D. in physical chemistry from Harvard University in 1972 and did postdoctoral work in mineralogy with Professor R. G. Burns at MIT for 2 years. In 1973 he joined the faculty at the University of Maryland where he is now an Associate Professor. His current research focuses on quantum mechanical studies of inorganic compounds and minerals.

* Current address: Polaroid Corporation, Waltham, MA 02254.

[1] Pauling, L. *The Nature of the Chemical Bond*, 3rd ed.; Cornell University: New York, 1960.

[2] Huheey, J. E. *Inorganic Chemistry, Principles of Structure and Reactivity*, 3rd ed; Harper and Row: New York, 1978; pp 429–432.

[3] Jordan, K. D.; Burrow, P. D. *Acc. Chem. Res.* 1978, *11*, 341.

0001-4842/86/0119-0281$01.50/0 © 1986 American Chemical Society