

Classification of Synthetic Pathways in Organic Chemistry

JAKOB HERMANN WINTER

Hoechst AG, D-6230 Frankfurt am Main 80, West Germany

Received May 30, 1984

While different methods exist to generalize single reaction steps, there are no methods available to generalize entire real synthetic pathways, which are visually comprehensible and permit machine storage and retrieval. To solve this problem, rules are given for a stepwise approach to the total generalization of structures. Synthetic pathways may thus be generalized by the generalization of all the reactants involved. This reveals strategies of synthesis and leads to the classification of possible synthetic pathways.

INTRODUCTION AND DISCUSSION OF PROBLEM

Organic chemistry has grown to the point where there is an overwhelming number of known reactions and syntheses; hence, the problem of classifying these reactions has become much more important. To classify means to develop general categories and gather related single items under these class concepts. In order to recognize generalized reaction types, we usually specify, in the structural formula, that part of the compound that is directly involved in structural change during a reaction while generalizing the rest of the molecule.¹ Sometimes we also specify the parts of the molecule that do not themselves change but that have a significant influence on the reaction. Generalization is achieved, for example, by omission ($\text{-COOH} \rightarrow \text{-COCl}$), by the use of the nomenclature for class concepts such as alkyl, aryl, etc., and sometimes by the use of special schemes. Auxiliary substances and reaction conditions may be more or less exactly specified, generalized, or even omitted entirely. However, formalized reaction types of more theoretical importance are also obtained by generalization and/or schematization of the reacting structural parts, while the other structural parts are omitted. Such generalizations are sometimes reduced to a description of the making and breaking of specific bonds or types of bonds; in other cases, the reaction type is specified by a mechanistic explanation in terms of intermediates such as ions and radicals.² Here we are not primarily interested in classifying isolated reaction steps but rather bringing (a) complete series of successive reaction steps and (b) synthetic routes, from starting material to target compound, in their entirety into a classification.

One might, for example, describe a sequence of reaction steps by simply listing reaction types in succession. This method is entirely adequate where there is no need to provide a representation that is readily comprehensible on visual inspection as, for example, in storage for computer retrieval. However, this sequential listing of reaction types is not a generalized description of a synthetic route. In order to generalize synthetic pathways, one might omit intermediates. Thus, $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ would be generalized to $A \xrightarrow{2 \text{ steps}} C \xrightarrow{2 \text{ steps}} E$. However, the omitted intermediates are no longer available for retrieval when a generalization is stored in this form. Since no one can tell which intermediates have been omitted, a reliable search for such generalizations would be impossible.

The problem is now definable. It is necessary to generalize a reaction step sequence and especially a synthetic pathway in such a way that all starting materials, intermediates, and end products are included. This means the total generalization, not only of the complete route but also of each individual compound. This must also be done in a way that yields a comprehensible visual representation that clearly reveals the characteristic features of the pathway. Further, in order to

permit machine storage and retrieval, generalization rules should be applied that are compatible with the topological description of structures, which has proved successful in practice.

TOTAL GENERALIZATION OF STRUCTURAL FORMULAS

To solve this problem, one should start by examining the way a synthetic chemist looks at a structural formula. He looks at both the skeleton of the molecule and any functional group that may be present. The latter, however, may be to a greater or lesser extent part of the skeleton. An internal ether group, for example, is both a functional group and part of the skeleton. A methoxy group on a ring structure, however, is not normally considered to belong to the skeleton nor are attached hydroxy groups, sulfo groups, etc. The skeleton of an organic compound is usually considered to comprise the ring-forming atoms, chain-forming carbon atoms, and heteroatoms embedded in a carbon chain; however, marginal groups can be considered as functional groups or as part of the skeleton or both. Sometimes, rings only serve to protect functional groups and are thus themselves essentially functional groups. In the generalization procedure, it is necessary to take into account all of the foregoing possibilities.

Figure 1 presents a set of rules for a stepwise approach to the total generalization of structures. In the first step, all nonskeleton-forming atoms are disregarded. Skeleton-forming atoms are described in these different ways: (1) Skeleton-forming carbon atoms that are bound only to carbon or hydrogen atoms and that are linked to other atoms only by a single and/or aromatic bonds are described by a dot. In an aromatic ring system, the neighboring atoms may be heteroatoms without affecting the description. (2) Skeleton-forming heteroatoms are described by a square with a central dot, regardless of what other atoms they are bound to. (This also means that atoms that are linked to the heteroatom but are not part of the skeleton are omitted as mentioned above.) (3) Skeleton-forming carbon atoms that are linked to at least one heteroatom and/or to any other atom by at least one nonaromatic multiple bond are described by small circles with a central dot. In this way, all functional groups, including carbon atoms with nonaromatic multiple bonds to other carbon atoms, are accounted for. The bonds between the skeleton-forming atoms are covered by the following rules: All bonds between skeleton-forming atoms are represented by a single line. In the case of aromatic bonds, an enclosed ring is used to symbolize the aromatic character.

These rules reduce the structural formula to a description of the skeleton-forming atoms and a representation of the molecular skeleton as a simple sequence of σ -bonds. The only additional feature specified is the aromatic character. All functional groups and nonaromatic carbon-carbon multiple

a) Symbols for generalizations relevant to storage

(or as angle formed by two bond lines) means C saturated or aromatic, otherwise bound to C or H.

⊙ means functional group with central C atom and also C joined to another C by a double or triple bond and maybe in addition to C or H.


□ means skeleton-forming hetero-atom. May also be linked with non skeleton-forming hetero-atom.

⊕ means + and ⊖ means -, thus more general than either one.

A means □ and ⊕, thus more general than either one.

— means σ-bond.

 means aromatic bond

 means aromatic bond possibly present.

b) Higher level generalizations


γ means any open chain of C-atoms.

γ̇ means any open chain of C-atoms possibly present.

⌋ means any part of a ring of C-atoms, ⌋̇ if aromatic.

⌋̇ means any part of a ring of C-atoms possibly present, if aromatic.

Note: If □ or A is superimposed on one of these symbols.

eg  or , then the atoms are hetero-atoms or atoms

of any kind, respectively. The symbol ⊕ indicates

functional groups.

Figure 1. Rules for the total generalization of structural formulas.

bonds are generalized by the same symbol without special regard to oxidation states. This takes into account the fact that in general they are relatively easy to interconvert and usually do not present critical problems in synthesis. However, the double role of skeleton-forming, nonaromatically bonded heteroatoms as skeleton-forming atoms on the one hand and parts of functional groups on the other is still expressed. In the next section, we shall see that these rules are not only easy to apply but that they lead to a very clear, comprehensible, and, hence, readily classifiable representation of synthetic routes.

First, however, let us consider hierarchically higher levels of generalization. It seems reasonable that at the next higher level all hydrocarbon groups and all functional groups should be generalized in the same way. In Figure 1, a double circle is used for this purpose. At the next higher level, every group, including the skeleton-forming single atom, is described by the letter A. Still further generalizations are possible, as shown in Figure 1. Chiral centers and charges are indicated in the usual way. Other symbols may be added to indicate stereo-selective syntheses so that they may be recognized as such if needed.

GENERALIZED DESCRIPTION OF SYNTHETIC PATHWAYS

Figure 2 shows a synthetic route for grandisol, described in the customary specific manner (without the reaction conditions being specified). Figure 3 shows the same route in a first-level generalization according to the rules in Figure 1. In cases of mere functional-group transformations, the formulas of starting material and product often become identical at this stage of generalization, so that one formula may represent two or more succeeding structures. This shortens the whole description of the route, since only the skeleton-transforming reactions remain. The result is that the entire pathway is easier to grasp and retain in one's memory (Figure 4). Nevertheless, all formulas represent the compounds involved in a general way. Figure 5 gives further examples of generalized structural formulas.

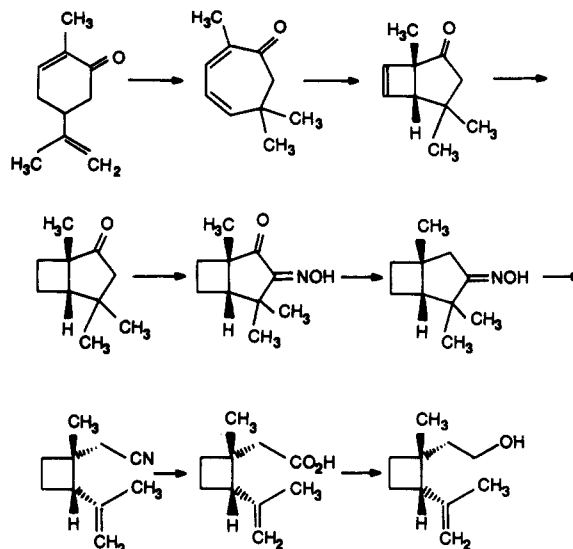


Figure 2. Normal description of a synthesis pathway, in this case with grandisol as the target molecule (without reaction conditions).

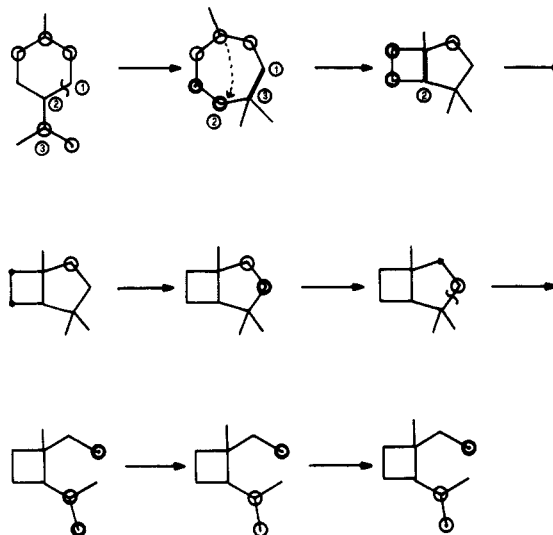


Figure 3. Description of the previously shown synthesis pathway (see Figure 2) in a first-level generalization. (The transformation may be indicated by a wavy line across the bond being broken and by making the new bonds and symbols heavier and darker. Identifying numbers and arrows may also give added clarity.)

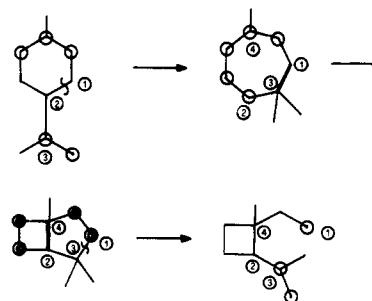


Figure 4. Further generalized description of the synthesis pathway in Figure 3, thereby revealing the strategy of skeletal change.

CLASSIFICATION PROCEDURES

In order to classify, one may group all individual cases into general categories, which means classifying merely by ordering. For scientific purposes, it is more valuable to develop a classification system and then fit individual cases into it. To create a classification system, it is necessary to know which features are characteristic enough to be ordered systematically.

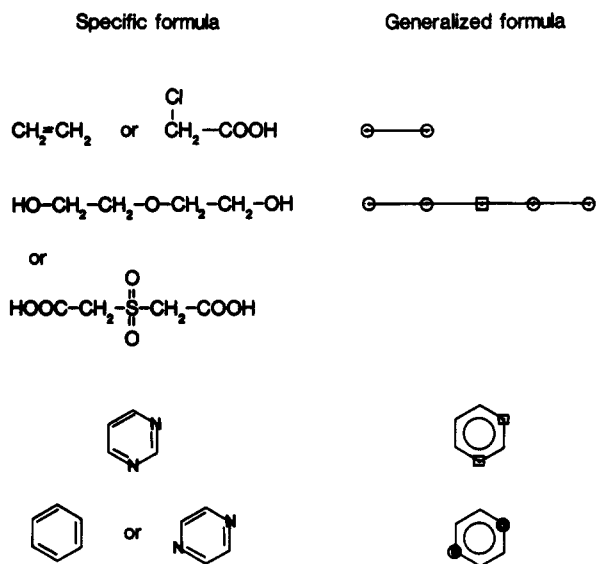


Figure 5. Further example of generalized structural formulas.

Consequently, one usually begins by ordering individual cases in order to discover which are the important and characteristic features. As shown above, for synthetic routes it is primarily the skeleton-transforming reaction that should be used to develop a classification system. Figures 2–4 show a typical synthetic route that begins with a ring expansion of a six- to a seven-membered ring, which then undergoes a 1,2-fusion to form a four- and a five-membered ring. The final step is the opening of the five-membered ring. These skeleton-transforming reactions would still be typical of this synthetic route even if the functional groups and their reactions were entirely different, if it were not the keto group (Figure 2) that activated ring enlargement in the first reaction but some other functional group and if fusion occurred by way of a condensation reaction. All these modifications in functional group reactions would leave this synthetic pathway classified in the same generalized form. The typical skeleton-transforming reactions taken together represent the strategy of the pathway. Changing their sequence or replacing one by another changes the strategy. Such changes may result in a relatively linear pathway becoming branched⁴ and vice versa.

It is worth noting in this connection that synthetic routes may be of a type in which compounds are continuously synthesized or degraded or, in effect, a mixture of both. An interesting strategy is that of large cycles (cf. citric acid cycle) in which starting materials are introduced at certain points in the cycle and products removed at others, while most of the intermediates are derived from one another repeatedly in a closed pathway.

Another important question is which length of synthesis pathways should form the elements of the classification system; i.e., should parts of a sequence or the entire sequence be entered in the system? The latter seems desirable, but it is also useful to base the classification on shorter sections. In the extreme case, the element is reduced to a single reaction step. This is then no longer a pathway, but it means that the rules of generalization in Figure 1 can also be used for the classification of individual reaction steps. Particularly for such applications, it may be desirable to introduce intermediate levels of generalization between the customary explicit description and the first level of generalization discussed above, for example, one in which the skeleton-forming heteroatoms are specified and/or functional groups are differentiated ac-

cording to their degree of oxidation. Where necessary, such questions must be decided ad hoc on the basis of the reaction in question or on the basis of synthetic routes that are typical of certain skeleton-forming heteroatoms and/or when changes in the degree of oxidation are important. (Of course, another way of generalizing a single reaction step is to describe specifically the reacting substructures and generalize the rest of the molecule according to the given rules. Still another way would be merely generalizing the reacting substructures and removing all other parts of the reactants.)

The specific description of skeleton-forming heteroatoms (regardless of the extent to which nonskeleton-forming atoms or groups of atoms are bound to them) in particular opens up the possibility of adapting the method to the requirements of inorganic chemistry, where carbon atoms are no longer of central importance. One can even create a special system, where in "silicon chemistry", for example, the dot symbolizes the silicon atom instead of carbon or in "phosphorus chemistry" it would stand for the phosphorus atom etc. and all other atoms, including carbon, would be looked upon as heteroatoms.

CONCLUSIONS

The number of known reaction steps recorded in the literature is growing at a dramatic rate, so also is the number of longer synthetic pathways, for example, of those relating to the complete synthesis of natural products and their analogues. In consequence, it is becoming imperative to develop a reasonable and practicable method of classification. The generalizations required for this purpose must be readily comprehensible and susceptible of machine storage and retrieval. The method represented here satisfies these conditions and renders it easier to recognize the strategies already used in organic chemistry and those that are still required. Additionally, it has been shown that it is possible to extend the approach in an analogous manner to silicon chemistry, phosphorus chemistry, etc., thus revealing its general applicability in chemistry.

ACKNOWLEDGMENT

I thank Dr. Robert Fugmann, Hoechst AG, and Dr. Douglas Mass, Beilstein-Institut, very heartily for their interest in this work and their valuable comments.

REFERENCES AND NOTES

- (1) See textbooks of organic chemistry. A procedure for documentation of reactions: Fugmann, R.; Bitterlich, W. "Reaktionendokumentation mit dem GREMAS-System." *Chem.-Zig.* **1972**, *96*, 323–330. Further discussions about classification and documentation of reactions are in the following: Winter, J. H. "Chemische Syntheseplanung in Forschung und Industrie"; Springer-Verlag: Berlin, 1982. Winter, J. H. "Maschinelle Klassifikation im Problembereich der chemischen Synthese." "Studien zur Klassifikation"; Indeks Verlag: Frankfurt/Main, 1983; Vol. 13, pp 61–75.
- (2) See, e.g., Blair, J.; Gasteiger, J.; Gillespie, C.; Gillespie, P. D.; Ugi, I. "Representation of the Constitutional and Stereochemical Features of Chemical Systems in the Computer-Assisted Design of Synthesis." *Tetrahedron* **1974**, *30*, 1845–1859. Satchell, D. P. N. "The Classification of Chemical Reactions." *Naturwissenschaften* **1977**, *64*, 113–121. Hendrickson, J. B. "A Systematic Characterization of Structures and Reactions for Use in Organic Synthesis." *J. Am. Chem. Soc.* **1971**, *93*, 6847–6854. Hendrickson, J. B. "A Systematic Organization of Synthetic Reactions." *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 129–136. March, J. "Advanced Organic Chemistry, Reactions, Mechanisms, and Structures," 2nd ed.; McGraw-Hill-Kogakusha: Tokyo, 1977.
- (3) Ayer, W. A.; Browne, L. M. "The Transformation of Carvone into Racemic Grandisol." *Can. J. Chem.* **1974**, *52*, 1352–1360; *Synform* **1983**, *1*, 9–10.
- (4) See converging synthesis pathways in Velluz, L.; Vals, J.; Mathieu, J. "Spatial Arrangement and Preparative Organic Synthesis." *Angew. Chem., Int. Ed. Engl.* **1967**, *6*, 778–789.