# Organic Synthesis − Art or Science?[†]

Christoph Rücker,*,[‡] Gerta Rücker,[§] and Steven H. Bertz*,[‖]

Department of Mathematics, Universität Bayreuth, D-95440 Bayreuth, Germany, Department of Rehabilitative
and Preventative Sports Medicine, Universität Freiburg, D-79106 Freiburg, Germany, and
Complexity Study Center, Mendham, New Jersey 07945

The LHASA rules for finding strategic bonds in polycyclic target structures are analyzed with respect to the following question: *Do the strategic bonds tend to give the greatest simplification upon disconnection, as measured by recently introduced indices of molecular complexity?* The answer is *yes*, at least for the more general rules. This result implies that the bonds most useful for retrosynthetic disconnection can now be identified by a simple calculation rather than by application of a body of rules. It is concluded that organic synthesis, as far as described by these rules, has a mathematical basis and consequently can be considered a science as well as an art.

## INTRODUCTION

An art is a human activity circumscribed by expressions such as "skill in doing anything as the result of knowledge and practice",[1] "a body or system of rules serving to facilitate the carrying out of certain principles",[1] "implying a personal, unanalyzable creative force",[2] and "beautiful or thought-provoking works produced through creative activity".[3] All this can certainly be said about organic synthesis, which therefore seems to be an art. But what about a science?

A science is "a department of systematized knowledge ... covering general truths or the operation of general laws"[4] based on "collection of data through observation and experiment, and the formulation and testing of hypotheses".[4] Is organic synthesis a science? The first thing a novice in the field learns is that every compound of significant complexity requires its own individual synthesis, and seemingly similar intermediates often behave very differently. The difficulty in reducing organic synthesis to a body of general truths or laws is illustrated by the sheer size of synthesis programs such as LHASA (Logic and Heuristics Applied to Synthetic Analysis).[5,6]

Looking for elements of science in organic synthesis, we take our cue from a sentence written by German philosopher Immanuel Kant in 1786: "*I state, however, that in any particular natural philosophy there is only as much real science as there is mathematics to be found in it.*"[7]

In 1975, E. J. Corey and co-workers published a seminal paper entitled "General Methods of Synthetic Analysis. Strategic Bond Disconnections for Bridged Polycyclic Structures",[5] wherein they defined the term "strategic bond" and gave a set of rules for finding such bonds in fused or bridged polycyclic ring systems. In their words, *strategic bonds* are "those ring bonds which lead to the most accessible and/or simplest precursor structures", when disconnected in the retrosynthetic direction. Likewise, in their 1989 book Corey and Cheng defined *strategic disconnections* as "those which lead most effectively to retrosynthetic simplification".[6] No definition of molecular simplicity or retrosynthetic simplification was given; indeed they noted, "Since synthetic accessibility is not subject to rigorous mathematical definition, the techniques for the recognition of the most strategic disconnections have been derived heuristically".[5,8]

The rules for assigning strategic bonds were built into the synthesis planning computer program LHASA, and they have been used successfully, either manually or as part of the program, for planning a plethora of syntheses, as exemplified in the aforementioned book.[6] Thus, we believe that their value is well established.

Today, thanks to progress in chemical graph theory,[9,10] we have mathematical measures of the complexity of molecular structures.[11,12] Two molecular complexity indices of proven value are the *total number of substructures* $N_T$[13] and the *total walk count twc*.[14] In recent work we used these measures to find the simplest and thus most useful precursor structures for synthetic targets[15] and to evaluate general principles of structure simplification.[16] Since structure simplification is the key concept of retrosynthetic analysis, we felt these indices to be potentially useful to probe the rules of retrosynthesis themselves.

Synthetic accessibility is an aspect of synthetic complexity, which is distinct from molecular complexity, as recently emphasized.[12,17] For a given compound molecular complexity depends only on structure, thus it does not change over time. In contrast, synthetic complexity is a function of the progress of synthetic chemistry, and it tends to decrease over time, although a quantitative measure is not at hand.

Of course, the synthetic chemist is primarily interested in synthetic complexity, so that it would be understandable if the LHASA rules were dominated by experience, i.e., by the state of the art of synthesis and the derived heuristics. However, we were curious whether, as implied by the L in LHASA, the rules or a subset thereof might reflect molecular complexity and thus correlate with complexity indices.
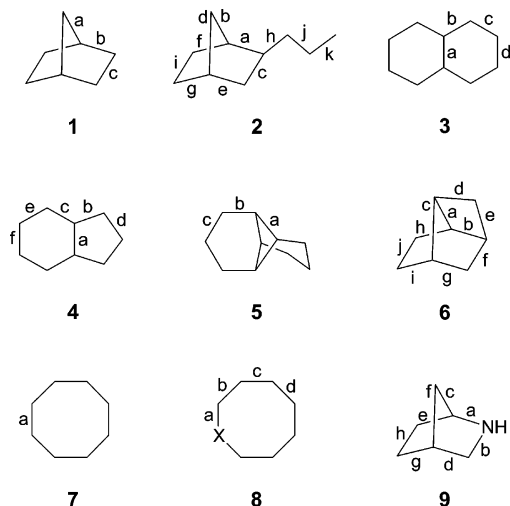
* Corresponding author phone: +49-921-553386; fax: +49-921-553385; e-mail: christoph.ruecker@uni-bayreuth.de (C.R.); sbertz@complexitystudycenter.org (S.H.B.).
† Dedicated to Dr. George W. A. Milne, a former long-term Editor-in-Chief of JCICS.
‡ Universität Bayreuth.
§ Universität Freiburg.
‖ Complexity Study Center.

**Figure 1.** Structures of compounds **1**−**9** used to analyze the LHASA rules.

Herein, we first analyze the rules for strategic bonds (hereafter referred to simply as "the rules")[5] by using simple examples to answer the following question: *Are the bonds selected by the rules those that produce the largest decrease in molecular complexity upon disconnection?* Then, we will consider for the purpose of validation some more realistic synthetic targets, taken from the classic paper by Corey et al.[5] Finally, we will attempt to answer the larger question: *How strongly is a limited but central part of organic synthesis, as reflected in the rules, based on discrete mathematics?* Or if we accept Kant's view, the question is *to what extent is organic synthesis a science in contrast to an art?*

## RESULTS AND DISCUSSION

Figure 1 shows simple compounds **1**−**9**, taken in part from ref 5, with their bonds labeled alphabetically. The letter used to symbolize a bond also represents (in conjunction with the number) the compound missing that bond and the disconnection resulting in that compound, depending on the context. Table 1 lists the *twc* and $N_T$ values of compounds **1**−**9** as well as those of all precursors obtained from them by cleavage of one bond. A rank is also given for each precursor, and the higher rank (lower number) corresponds to the higher absolute value of $\Delta C = C(\text{precursor}) - C(\text{target})$, where the complexity $C$ is *twc* or $N_T$. Since $C(\text{target})$ is the same for all the precursors of a given target molecule, the bond with the higher rank corresponds to the precursor having the lower value of $C(\text{precursor})$.

The rank according to *twc* determines the labeling of bonds in each structure in Figure 1, with bond *a* giving the greatest simplification upon disconnection, bond *b* the second greatest, and so on. (In a symmetrical structure such as **1**, only one of the equivalent bonds of each type is labeled.) This lettering scheme reveals at a glance the usefulness of each bond in retrosynthesis. Finally, for most disconnections in Table 1, a comment is made as to whether that bond is strategic by the rules, and if not which rule is violated.

The first two observations to be made about Table 1 are (i) all the disconnections are ordered according to their degree of simplification, and (ii) the two measures of molecular complexity result in similar—but not always identical—

**Table 1.** Molecular Complexity Data *twc* and $N_T$ for Compounds **1**−**9** and Synthetic Precursors Lacking One Bond

| | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|
| **1** | 1974 | | 131 | | |
| a | 1094 | 1 | 60 | 1 | strategic |
| b | 1180 | 2 | 61 | 2 | strategic |
| c | 1340 | 3 | 70 | 3 | rule 2A |

| | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|
| **2** | 40634 | | 401 | | |
| a | 16060 | 1 | 121 | 1 | strategic |
| b | 19764 | 2 | 174 | 5 | strategic |
| c | 20464 | 3 | 157 | 3 | rule 2A |
| d | 21222 | 4 | 183 | 7 | strategic |
| e | 21334 | 5 | 169 | 4 | strategic |
| f | 22192 | 6 | 175 | 6 | strategic |
| g | 23936 | 7 | 187 | 8 | strategic |
| h | 25712 | 8 | 137 | 2 | rule 1 |
| i | 26910 | 9 | 214 | 9 | rule 2A |
| j | 34190 | 10 | 223 | 10 | rule 1 |
| k | 38342 | 11 | 311 | 11 | rule 1 |

| | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|
| **3** | 29886 | | 357 | | |
| a | 10220 | 1 | 101 | 1 | rule 4 |
| b | 15908 | 2 | 135 | 2 | strategic |
| c | 21222 | 3 | 183 | 3 | rule 2A |
| d | 22676 | 4 | 199 | 4 | rule 2A |

| | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|
| **4** | 12520 | | 258 | | |
| a | 4590 | 1 | 82 | 1 | rule 4 |
| b | 6792 | 2 | 109 | 3 | strategic |
| c | 6902 | 3 | 100 | 2 | strategic |
| d | 8650 | 4 | 141 | 5 | rule 2A |
| e | 9050 | 5 | 133 | 4 | rule 2A |
| f | 9628 | 6 | 144 | 6 | rule 2A |

| | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|
| **5** | 69992 | | 997 | | |
| a | 29886 | 1 | 357 | 1 | strategic |
| b | 44676 | 2 | 399 | 2 | rule 3 |
| c | 55586 | 3 | 531 | 3 | rule 2A |

| | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|
| **6** | 9838 | | 468 | | |
| a | 4846 | 1 | 186 | 2/3 | strategic |
| b | 5142 | 2 | 186 | 2/3 | strategic |
| c | 5186 | 3 | 177 | 1 | strategic |
| d | 5924 | 4 | 220 | 8 | rule 3 |
| e | 6136 | 5 | 226 | 9 | rule 3 |
| f | 6244 | 6 | 213 | 7 | strategic |
| g | 6348 | 7 | 212 | 5/6 | strategic |
| h | 6504 | 8 | 206 | 4 | rule 3 |
| i | 6818 | 9 | 212 | 5/6 | rule 3 |
| j | 7546 | 10 | 250 | 10 | rule 2A |

| | twc | | $\Delta twc$ | | $N_T$ |
|---|---|---|---|---|---|
| **7** | 2032 | | | | 65 |
| a | 1254 | | 778 | | 36 |

| | twc (X≙2) | rank | $\Delta twc$ | twc (X≙1) | rank | $\Delta twc$ | $N_T$ |
|---|---|---|---|---|---|---|---|
| **8** | 9582 | | | 3573 | | | 65 |
| a | 3664 | 1 | 5918 | 1643 | 1 | 1930 | 36 |
| b | 7022 | 2 | 2560 | 2283 | 2 | 1290 | 36 |
| c | 8258 | 3 | 1324 | 2610 | 3 | 963 | 36 |
| d | 8636 | 4 | 946 | 2729 | 4 | 844 | 36 |

| | twc (N≙2) | rank | twc (N≙1) | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|
| **9** | 5210 | | 2839 | | 131 | | |
| a | 2188 | 1 | 1395 | 1 | 61 | 3-6 | strategic |
| b | 2670 | 2 | 1656 | 2 | 70 | 7/8 | strategic |
| c | 3676 | 3 | 1728 | 3 | 60 | 1/2 | strategic |
| d | 3698 | 4 | 1792 | 4 | 61 | 3-6 | strategic |
| e | 3810 | 5 | 1834 | 6 | 61 | 3-6 | strategic |
| f | 4002 | 6 | 1829 | 5 | 60 | 1/2 | strategic |
| g | 4208 | 7 | 1961 | 7 | 61 | 3-6 | strategic |
| h | 4396 | 8 | 2133 | 8 | 70 | 7/8 | rule 2A |

orderings of the precursor compounds. Such discrepancies are not unexpected, given that these measures count very different graph theoretical entities.

The rules are formulated as a set of six necessary conditions for a bond to be strategic, so that all rules must be met simultaneously.[5] Additionally, a sufficient condition (the C-heterobond procedure) is given for heteroatom-carbon bonds (vide infra).[18]

**Rule 1.** " ... a strategic bond must be in a four-, five-, six-, or seven-membered 'primary' ring. A primary ring is one which cannot be expressed as the envelope of two or more smaller rings bridged or fused to one another." For our purposes it is convenient to split this rule into two parts: ***Rule 1A.*** A strategic bond must be in a ring. ***Rule 1B.*** The ring must be a four- to seven-membered primary ring.

To check rule 1A for a possible mathematical basis, we consider a typical bridged hydrocarbon, bicyclo[2.2.1]heptane **1**, its 2-propyl derivative **2**, and the potential precursors derived via cleavage of each bond. Acyclic bonds *h*, *j*, and *k* in **2** occupy ranks 8, 10, and 11 (out of eleven) according to *twc* and 2, 10, and 11 according to $N_T$. Thus, the bond with the highest rank is a ring bond, and chain bonds occupy lower ranks. The same observations apply to compounds **15**, **16**, and **24**−**27** below and to 1-propyl- and 7-propylbicyclo-[2.2.1]heptane as well as to alkyl-substituted monocyclic alkanes (not shown).[19] Colloquially speaking, a ring bond is more "central" than a chain bond, and the presence of the latter has relatively little effect on the complexity of a molecule. This is the essence of rule 1A, derived here from observations on graph-theoretical measures of molecular complexity. We conclude that rule 1A is strongly mathematical.

As far as rule 1B is concerned, the difference between four- to seven-membered rings on one hand and three-membered or eight-membered and higher rings on the other is not predicted from complexity considerations. For example, calculating the molecular complexities (*twc* or $N_T$) of precursors obtained by one-bond cleavages of 1-propyl- or 2-propylbicyclo[1.1.0]butane (three-membered rings) and 1-propyl- or 2-propylbicyclo[1.1.1]pentane (four-membered rings) did not uncover any difference in behavior. Likewise, the orderings of bonds by *twc* or $N_T$ in propylcycloheptane (seven-membered ring) and propylcyclooctane (eight-membered ring) are completely analogous. We conclude that rule 1B has no mathematical basis; rather it follows exclusively from the empirical observation of the "relative ease of formation of common-sized rings".[5]

**Rule 2A.** "A strategic bond must be directly attached to another ring (exo to another ring)."

A typical example of the application of this rule is bicyclo-[4.4.0]decane **3**. The highest rank is attributed by both *twc* and $N_T$ to *a*, the central bond; however, this bond is not considered strategic by LHASA owing to rule 4 (vide infra). Both indices also agree on the second through last place rankings. The #2 bond for disconnection is *b*, which is directly attached (exo) to another ring. Cleavage of bonds *c* or *d* (not exo to another ring) results in two appendages, and these disconnections are ranked third and fourth (last), respectively. The results for bicyclo[4.3.0]nonane **4** are analogous. Both indices agree that fusion bond *a* gives the simplest precursor upon disconnection. They also agree that

bonds *b* and *c* (exo to another ring) are #2 and #3, while bonds *d*, *e*, and *f* (not directly attached to another ring) occupy the lowest ranks.

The same effect is seen in compounds **1**, **2**, **5**, **6**, and **9**−**27** (vide infra), e.g., bonds *c* in **1** and *i* in **2** are not involved in a ring junction and occupy the lowest rank as far as ring bonds are concerned. Such bonds are farther from the center of the molecule, and consequently they have less influence than the more central ones. From these observations it can be concluded that rule 2A is strongly based on graph theory, i.e., on mathematics.

**Rule 2B.** " ... strategic bonds may not be exo to rings of that [i.e., three-membered] size."

This exception is "due to the paucity of ring closure methods in which bonds are formed to preexisting three-membered rings".[5] It is obviously not based on graph theory but is a reflection of the state of the art at the time the rules were promulgated.

**Rule 3.** "... strategic bonds should be in the ring (or rings) which exhibits the greatest degree of bridging."

Several annotations are given; the two most important ones are (i) "the maximum bridging ring (or rings) is selected from the set of 'synthetically significant rings', which is defined as the set of all primary rings (see rule 1) plus all secondary rings less than eight membered", and (ii) "the maximum bridging rings of a molecule are defined as those rings which are bridged at the greatest number of sites. This definition is *not* equivalent to defining maximal bridging rings as those containing the greatest number of bridgehead sites." For comparison, the book gives a different version of this rule:[6] "a strategic bond must be endo to (within) a ring of maximum bridging. ... The ring of maximum bridging is usually that synthetically significant ring containing the greatest number of bridgehead atoms. ... Synthetically significant rings ... is the set of all primary rings plus all secondary rings less than 7-membered." We do not comment here on the subtle differences between these two versions of the rule, as they do not affect our conclusions. Rather, we investigate whether the examples reveal a graph-theoretical basis for the rule, i.e., whether the bonds classified as strategic by Corey et al. are the same as the bonds ranked highly by *twc* or $N_T$.

In tricyclo[4.4.0.0$^{2,7}$]decane **5** the ring exhibiting the greatest degree of bridging is the central four-membered ring. It is comprised of four copies of bond *a*, which is in fact ranked highest by both *twc* and $N_T$. In tricyclo[3.3.0.0$^{2,7}$]-octane **6** the ring of greatest bridging is the five-membered ring, comprised of bonds *a*, *b*, *c*, *f*, and *g*. These bonds are ranked 1, 2, 3, 6, and 7 (out of ten) by *twc* and 1, 2, 3, 7 and (tied for) 5/6 by $N_T$. In tricyclo[4.2.1.0$^{3,7}$]nonane **10** (Figure 2) the ring of greatest bridging is the six-membered secondary ring, comprised of two copies each of bonds *a*, *c*, and *e*. These bonds are ranked 1, 3, and 5 (out of seven) according to *twc* and 1, 2, and 5 according to $N_T$ (Table 2). Again, bond *a* is not considered strategic by LHASA, owing to rule 4 (vide infra), but this need not concern us when analyzing rule 3.

In our opinion, these results exhibit a reasonable agreement between rule 3 and the graph theoretical bond rankings, pointing to a mathematical basis for this rule. Generally speaking, rule 3 selects bonds in the most central ring, the atoms of which tend to have high connectivities. Bonds

Organic Synthesis – Art or Science?

*J. Chem. Inf. Comput. Sci., Vol. 44, No. 2, 2004* **381**
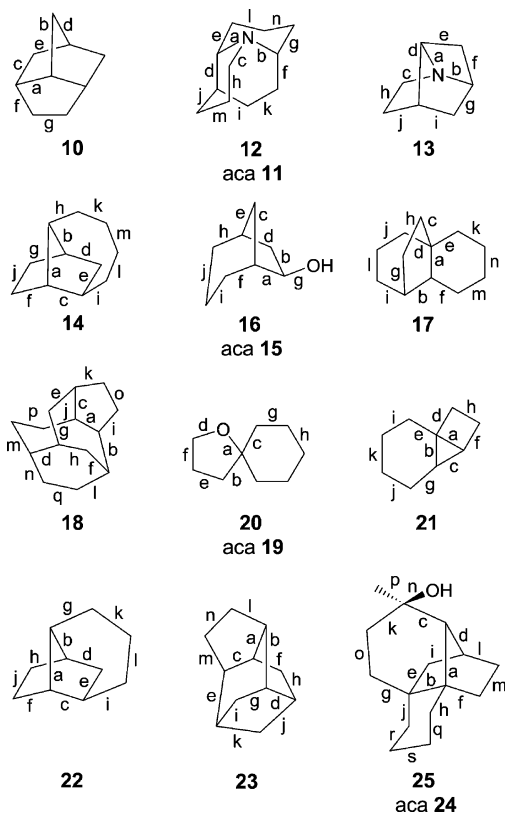
**Figure 2.** Structures of compounds **10**−**25** used to validate the results. The acronym aca stands for "all carbon analog".

between atoms of high connectivity contribute more to *twc* than bonds between atoms of low connectivity,[14a] e.g., those in appendages. As with rule 1, the exclusion of certain ring sizes, e.g., eight-membered and higher[5] (seven-membered and higher in the book version[6]) is not graph theoretical.

**Rule 4.** "... any bond common to a *pair* of bridged or fused primary rings whose envelope is eight-membered or larger cannot be considered strategic."

This rule is fundamentally nonmathematical. In fact, it is contramathematical, since the bonds excluded by it are typically the most central ones, which are the best ones to disconnect according to graph theoretical analysis (vide supra **3** and **4**). The only purpose of this rule is "to avoid the formation of rings having greater than seven members during antithetic bond cleavage",[5] i.e. rings that were synthetically inaccessible when the rules were promulgated.[20]

**Rule 5.** "Bonds within aromatic rings are not considered to have potential strategic character."

In the present field of application, bridged polycyclic targets, this rule is not relevant. It is based on the concept of an indivisible "building block", obviously a very practical and intuitive heuristic.

**Rule 6A.** "If a cyclic arc linking a pair of common atoms ... contains a chiral carbon atom, then none of the bonds in the cyclic arc may be considered strategic."

**Rule 6B.** "This part effectively modifies the operation of part A for the special case of bonds *directly attached* to a chiral center, when that is the *only* chiral center on the arc linking the two common atoms. Such bonds are not excluded from the set of possible strategic bonds."

This rule is concerned with chirality, which is certainly a mathematical concept; however, it is not treatable in terms of graph theoretical measures *twc* or $N_T$.

As a preliminary conclusion, we may state the following: *The more important rules, viz. 1, 2, and 3, as far as they describe general features, are all paralleled by molecular complexity and are therefore firmly based on graph theory, i.e., mathematics.* The exceptions, rule 4 and parts of rules 1−3, all deal with the inaccessibility of certain ring sizes, which is an aspect of synthetic complexity. They are less important today than in the past and may be predicted to be of even less concern in the future.[20]

**The C−Heterobond Procedure.** "To the set of strategic bonds determined by application of rules 1−6 above is added the collection of bonds in the cyclic network between carbon and O, N, and S which satisfy rules 2B, 4, 5, and 6."[5]

While rule 4 is maintained, the strict criteria of rules 1, 2A, and 3 are not applied here. Consequently, most cyclic C−X single bonds (X = O, N, S) are strategic as a result of "the rich and powerful chemistry for forming bonds between carbon and these heteroatoms". Surprisingly, although this rule is not graph-theoretical in origin, its consequences are amenable to modeling by graph theoretical index *twc*. A heteroatom is distinguished by one or more loops in the molecular graph, which means the corresponding diagonal element in the adjacency matrix is set at a positive integer, $a_{ii} > 0$. Consequently, the number of walks in a heteroatom-containing molecule is higher than in the corresponding hydrocarbon.

Comparison of cyclooctane **7** and hetero-analogue **8** shows that cleavage of bond *a*, incident to the heteroatom in the latter, results in a Δ*twc* that is higher on a percent basis than cleavage of one of the (equivalent) bonds in the former, whereas the Δ*twc* values of bonds *b*, *c*, and *d* are *lower* on a percent basis. The Δ*twc* values for bonds *a*, *b*, *c*, and *d* in **8** decrease monotonically; thus, a heteroatom most markedly impacts the molecular complexity contributions of bonds directly attached to it and less so the contributions of bonds farther away. This is analogous to the LHASA procedure, whereby the heteroatom renders incident bonds strategic, but does not affect bonds farther away.

There is some arbitrariness, or looking at it another way, some opportunity for fine-tuning in the particular choice of $a_{ii}$ for characterizing a given heteroatom. (In general we use $a_{ii} = 2$ for N, $a_{ii} = 1$ for O, and $a_{ii} = 0$ for C.) For structure **8** in Table 1, values of *twc* and Δ*twc* are given for one or two loops on the heteroatom. The *twc* values with two loops are much higher than the corresponding values with one. Interestingly, the decrease in the heteroatom effect with increasing distance is more rapid for the structure with the higher number of loops. Rankings in heterocycles can sometimes be interchanged by variation of the heteroatom parameter $a_{ii}$, depending on the balance between heteroatom and constitutional effects, e.g., ranks 5 and 6 in 2-aza-norbornane **9**.

While in norbornane **1** the highest rank is assigned by *twc* to bonds *a* in the short bridge, the heteroatom effect results in the two C−N bonds in its 2-aza derivative **9** being ranked above the corresponding bonds *c* and *f*. The high rank of bond *b* in **9** is analogous to it being made strategic by the LHASA C-heterobond procedure.

Figure 2 contains additional polycycles taken from ref 5, along with the corresponding all-carbon analogue (aca) of each heterocycle. (Note that **6** is the aca of **13**.) Table 2 gives their index data. As in Figure 1, the bonds in Figure 2 are

**Table 2.** Molecular Complexity Data *twc* and $N_T$ for Compounds **10−25** and Synthetic Precursors Lacking One Bond

| | *twc* | rank | $N_T$ | rank | comments | | *twc* | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **10** | 23902 | | 708 | | | d | 15128 | 4 | 323 | 6 | rule 3 |
| a | 11724 | 1 | 252 | 1 | rule 4 | e | 15432 | 5 | 312 | 5 | strategic |
| b | 13908 | 2 | 303 | 3 | rule 3 | f | 15894 | 6 | 310 | 4 | rule 3 |
| c | 14886 | 3 | 301 | 2 | strategic | g | 17974 | 7 | 371 | 7 | rule 2A |

| | *twc* | rank | $N_T$ | rank | comments | | *twc* | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **11** | 374444 | | 2167 | | | g,j | 246240 | 5 | 855 | 4 | rule 3 |
| a | 140748 | 1 | 615 | 2 | rule 4 | k | 266986 | 6 | 1030 | 6 | rule 2A |
| b,d | 165630 | 2 | 609 | 1 | rule 4 | h,l | 286152 | 7 | 1095 | 7 | rule 2A |
| c,e | 223262 | 3 | 798 | 3 | rule 3 | m,n | 291556 | 8 | 1114 | 8 | rule 2A |
| f,i | 235504 | 4 | 856 | 5 | strategic | | | | | | |

| | *twc* (N≅2) | rank | $N_T$ | rank | comments | | *twc* (N≅2) | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **12** | 3593126 | | 2167 | | | h | 2949500 | 8 | 1095 | 11/12 | rule 2A |
| a | 994012 | 1 | 615 | 3 | rule 4 | i | 3149112 | 9 | 856 | 8/9 | strategic |
| b | 1125982 | 2 | 609 | 1/2 | rule 4 | j | 3168704 | 10 | 855 | 6/7 | rule 3 |
| c | 1529068 | 3 | 798 | 4/5 | strategic | k | 3220880 | 11 | 1030 | 10 | rule 2A |
| d | 2519894 | 4 | 609 | 1/2 | rule 4 | l | 3262892 | 12 | 1095 | 11/12 | rule 2A |
| e | 2768430 | 5 | 798 | 4/5 | rule 3 | m | 3270780 | 13/14 | 1114 | 13/14 | rule 2A |
| f | 2787764 | 6 | 856 | 8/9 | strategic | n | 3270780 | 13/14 | 1114 | 13/14 | rule 2A |
| g | 2814810 | 7 | 855 | 6/7 | rule 3 | | | | | | |

| | *twc* (N≅2) | rank | $N_T$ | rank | comments | | *twc* (N≅2) | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **13** | 35366 | | 468 | | | f | 26840 | 6 | 226 | 9 | rule 3 |
| a | 14962 | 1 | 186 | 2/3 | strategic | g | 27674 | 7 | 213 | 7 | strategic |
| b | 15772 | 2 | 186 | 2/3 | strategic | h | 29564 | 8 | 250 | 10 | rule 2A |
| c | 19984 | 3 | 206 | 4 | strategic | i | 29632 | 9 | 212 | 5/6 | strategic |
| d | 25714 | 4 | 177 | 1 | strategic | j | 30202 | 10 | 212 | 5/6 | rule 3 |
| e | 26572 | 5 | 220 | 8 | rule 3 | | | | | | |

| | *twc* | rank | $N_T$ | rank | comments | | *twc* | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **14** | 160910 | | 1419 | | | g | 99122 | 7 | 629 | 9 | rule 3 |
| a | 64382 | 1 | 456 | 2 | rule 4 | h | 101676 | 8 | 493 | 4 | rule 3 |
| b | 70648 | 2 | 474 | 3 | strategic | i | 108568 | 9 | 517 | 5 | rule 3 |
| c | 71362 | 3 | 441 | 1 | rule 4 | j | 112796 | 10 | 748 | 12 | rule 2A |
| d | 89904 | 4 | 587 | 7 | strategic | k | 130672 | 11 | 694 | 10 | rule 2A |
| e | 91374 | 5 | 576 | 6 | strategic | l | 132844 | 12 | 706 | 11 | rule 2A |
| f | 93648 | 6 | 599 | 8 | rule 3 | m | 138336 | 13 | 765 | 13 | rule 2A |

| | *twc* | rank | $N_T$ | rank | comments | | *twc* | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **15** | 14578 | | 313 | | | f | 8584 | 6 | 127 | 2 | strategic |
| a | 6792 | 1 | 109 | 1 | strategic | h | 9050 | 7 | 133 | 5/6 | strategic |
| c | 7450 | 2 | 129 | 3 | strategic | i | 10670 | 8 | 163 | 8 | rule 2A |
| e | 7788 | 3 | 133 | 5/6 | strategic | j | 10798 | 9 | 165 | 9 | rule 2A |
| d | 8040 | 4 | 131 | 4 | rule 6 | g | 11258 | 10 | 187 | 10 | rule 1A |
| b | 8184 | 5 | 135 | 7 | rule 2A | | | | | | |

| | *twc* (O≅1) | rank | $N_T$ | rank | comments | | *twc* (O≅1) | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **16** | 17359 | | 313 | | | f | 10794 | 6 | 127 | 2 | strategic |
| a | 7796 | 1 | 109 | 1 | strategic | g | 11266 | 7 | 187 | 10 | rule 1A |
| b | 9511 | 2 | 135 | 7 | rule 2A | h | 11595 | 8 | 133 | 5/6 | strategic |
| c | 9562 | 3 | 129 | 3 | strategic | i | 13271 | 9 | 163 | 8 | rule 2A |
| d | 10124 | 4 | 131 | 4 | rule 6 | j | 13471 | 10 | 165 | 9 | rule 2A |
| e | 10165 | 5 | 133 | 5/6 | strategic | | | | | | |

| | *twc* | rank | $N_T$ | rank | comments | | *twc* | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **17** | 485020 | | 2298 | | | h | 324710 | 8 | 1143 | 10 | rule 2A |
| a | 141578 | 1 | 602 | 1 | rule 4 | i | 326982 | 9 | 930 | 7 | strategic |
| b | 213782 | 2 | 674 | 2 | strategic | j | 363478 | 10 | 1144 | 11 | rule 2A |
| c | 235504 | 3 | 856 | 6 | strategic | k | 377442 | 11 | 1122 | 9 | rule 2A |
| d | 252728 | 4 | 786 | 4 | strategic | l | 381236 | 12 | 1192 | 13 | rule 2A |
| e | 259268 | 5 | 740 | 3 | rule 3 | m | 389664 | 13 | 1154 | 12 | rule 2A |
| f | 295548 | 6 | 804 | 5 | rule 3 | n | 408966 | 14 | 1260 | 14 | rule 2A |
| g | 304472 | 7 | 984 | 8 | strategic | | | | | | |

| | *twc* | rank | $N_T$ | rank | comments | | *twc* | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **18** | 3772500 | | 12786 | | | i | 2395246 | 9 | 5146 | 13 | rule 3 |
| a | 1766838 | 1 | 3762 | 2 | rule 4 | j | 2411242 | 10 | 4631 | 7 | rule 3 |
| b | 1903434 | 2 | 3309 | 1 | strategic | k | 2516420 | 11 | 5394 | 14 | rule 3 |
| c | 1909592 | 3 | 4002 | 4 | rule 4 | l | 2579502 | 12 | 4940 | 11 | rule 3 |
| d | 2266358 | 4 | 3782 | 3 | rule 4 | m | 2700420 | 13 | 4917 | 10 | rule 3 |
| e | 2310856 | 5 | 4567 | 6 | rule 4 | n | 2733390 | 14 | 5090 | 12 | rule 3 |
| f | 2334892 | 6 | 4784 | 9 | rule 4 | o | 2786564 | 15 | 6493 | 17 | rule 2A |
| g | 2354706 | 7 | 4494 | 5 | rule 4 | p | 2854090 | 16 | 5802 | 15 | rule 2A |
| h | 2368372 | 8 | 4781 | 8 | rule 4 | q | 2927682 | 17 | 6083 | 16 | rule 2A |

**Table 2** (Continued)

| | twc | rank | $N_T$ | rank | comments | | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **19** | 36986 | | 377 | | | g | 25864 | 4 | 181 | 3 | rule 2A |
| a,b | 15908 | 1 | 135 | 2 | strategic | f | 27652 | 5 | 219 | 6 | rule 2A |
| c | 16060 | 2 | 121 | 1 | strategic | h | 29440 | 6 | 211 | 5 | rule 2A |
| d,e | 25108 | 3 | 198 | 4 | rule 2A | | | | | | |

| | twc (O≙1) | rank | $N_T$ | rank | comments | | twc (O≙1) | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **20** | 63285 | | 377 | | | e | 45723 | 5 | 198 | 5/6 | rule 2A |
| a | 17926 | 1 | 135 | 2/3 | strategic | f | 46459 | 6 | 219 | 8 | rule 2A |
| b | 31323 | 2 | 135 | 2/3 | strategic | g | 49143 | 7 | 181 | 4 | rule 2A |
| c | 32500 | 3 | 121 | 1 | strategic | h | 54552 | 8 | 211 | 7 | rule 2A |
| d | 35218 | 4 | 198 | 5/6 | strategic | | | | | | |

| | twc | rank | $N_T$ | rank | comments | | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **21** | 35552 | | 636 | | | f | 21888 | 6 | 303 | 7 | rule 2B |
| a | 12520 | 1 | 258 | 4 | strategic | g | 23700 | 7 | 240 | 3 | rule 2B |
| b | 12718 | 2 | 236 | 2 | strategic | h | 25478 | 8 | 357 | 11 | rule 2A |
| c | 15316 | 3 | 263 | 5 | rule 1B | i | 29114 | 9 | 315 | 8 | rule 2A |
| d | 19388 | 4 | 271 | 6 | rule 2B | j | 29984 | 10 | 329 | 9 | rule 2A |
| e | 20650 | 5 | 212 | 1 | rule 2B | k | 31232 | 11 | 354 | 10 | rule 2A |

| | twc | rank | $N_T$ | rank | comments | | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **22** | 63492 | | 1035 | | | g | 40634 | 7 | 401 | 4 | rule 3 |
| a | 27426 | 1 | 351 | 2 | rule 4 | h | 40746 | 8 | 461 | 9 | rule 3 |
| b | 29886 | 2 | 357 | 3 | strategic | i | 43074 | 9 | 419 | 5 | rule 3 |
| c | 30048 | 3 | 343 | 1 | strategic | j | 45890 | 10 | 547 | 12 | rule 2A |
| d | 37326 | 4 | 437 | 7 | strategic | k | 50388 | 11 | 537 | 10 | rule 2A |
| e | 37700 | 5 | 433 | 6 | strategic | l | 50940 | 12 | 543 | 11 | rule 2A |
| f | 38658 | 6 | 441 | 8 | rule 3 | | | | | | |

| | twc | rank | $N_T$ | rank | comments | | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **23** | 275300 | | 4145 | | | h | 179860 | 8 | 1797 | 11/12 | rule 3 |
| a | 143382 | 1 | 1485 | 4 | rule 4 | i | 182336 | 9 | 1797 | 11/12 | rule 3 |
| b | 144906 | 2 | 1385 | 2 | strategic | j | 184490 | 10 | 1799 | 13 | strategic |
| c | 146858 | 3 | 1425 | 3 | rule 4 | k | 185588 | 11 | 1788 | 10 | strategic |
| d | 153666 | 4 | 1511 | 5 | rule 4 | l | 190258 | 12 | 1764 | 8/9 | rule 3 |
| e | 158258 | 5 | 1365 | 1 | strategic | m | 193470 | 13 | 1764 | 8/9 | rule 3 |
| f | 172616 | 6 | 1737 | 6 | rule 3 | n | 215940 | 14 | 2163 | 14 | rule 2A |
| g | 175502 | 7 | 1754 | 7 | rule 3 | | | | | | |

| | twc | rank | $N_T$ | rank | comments | | twc | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **24** | 75479984 | | 46074 | | | j | 47926840 | 10 | 14814 | 6 | rule 3 |
| b | 20548036 | 1 | 11866 | 2 | rule 4 | l | 51551922 | 11 | 20302 | 12 | rule 3 |
| a | 22112186 | 2 | 12299 | 3 | rule 4 | k | 53742738 | 12 | 17570 | 10 | rule 2A |
| d | 35027134 | 3 | 14777 | 5 | strategic | m | 55227978 | 13 | 23721 | 16 | rule 2A |
| c | 38048242 | 4 | 9834 | 1 | rule 3 | o | 56973898 | 14 | 20913 | 13 | rule 2A |
| e | 38282854 | 5 | 17199 | 8 | strategic | q | 63157804 | 15 | 22425 | 14 | rule 2A |
| f | 40277044 | 6 | 17430 | 9 | rule 3 | p,n | 63508016 | 16 | 25188 | 18 | rule 1A |
| g | 44085212 | 7 | 15714 | 7 | rule 3 | r | 64125750 | 17 | 22589 | 15 | rule 2A |
| h | 44218504 | 8 | 14486 | 4 | rule 3 | s | 67324890 | 18 | 25126 | 17 | rule 2A |
| i | 44559918 | 9 | 19148 | 11 | strategic | | | | | | |

| | twc (O≙1) | rank | $N_T$ | rank | comments | | twc (O≙1) | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **25** | 84700228 | | 46074 | | | j | 55893681 | 10 | 14814 | 6 | rule 3 |
| a | 25846745 | 1 | 12299 | 3 | rule 4 | k | 57780399 | 11 | 17570 | 10 | rule 2A, 6 |
| b | 26196600 | 2 | 11866 | 2 | rule 4 | l | 59255569 | 12 | 20302 | 12 | rule 3 |
| c | 39108427 | 3 | 9834 | 1 | rule 3 | m | 63405621 | 13 | 23721 | 16 | rule 2A |
| d | 40212131 | 4 | 14777 | 5 | strategic | n | 63508031 | 14 | 25188 | 18/19 | rule 1A |
| e | 45517424 | 5 | 17199 | 8 | strategic | o | 63525510 | 15 | 20913 | 13 | rule 2A, 6 |
| f | 47446694 | 6 | 17430 | 9 | rule 3 | p | 69219976 | 16 | 25188 | 18/19 | rule 1A |
| g | 50882725 | 7 | 15714 | 7 | rule 3, 6 | q | 71872719 | 17 | 22425 | 14 | rule 2A |
| h | 51675982 | 8 | 14486 | 4 | rule 3 | r | 72949420 | 18 | 22589 | 15 | rule 2A |
| i | 51828989 | 9 | 19148 | 11 | strategic | s | 76279514 | 19 | 25126 | 17 | rule 2A |

labeled in order of decreasing rank according to *twc* except for the aca, which have the same labeling scheme as the corresponding heterocycles in order to facilitate comparison. Thus, the rank differences between a heterocycle and its aca reflect the heteroatom effect.

These data are used to assess the overall agreement between the *twc* ranking and the LHASA rules, as applied to more realistic synthetic targets. In the case of ideal agreement, all bonds identified as "strategic" in a particular target should appear above all others (see the "comments" column). If, on the other hand, strategic character and *twc* rank are independent, then the strategic bonds should be randomly distributed over the list ordered by *twc*.

Examination of the tables reveals that neither ideal case obtains. Nevertheless, in most of the structures, a tendency for the strategic bonds to appear in the upper part of the

**384** *J. Chem. Inf. Comput. Sci., Vol. 44, No. 2, 2004*

RÜCKER ET AL.
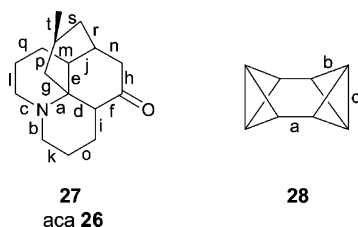


**Figure 3.** Structures of compounds **26**–**28** used as illustrations. The acronym aca stands for "all carbon analog".

table is observed. The single most important cause of disagreement is rule 4. Several other discrepancies are attributable to rule 3. From the point of view of *twc*, rule 3 is too restrictive. It usually limits the set of potential strategic bonds to those within one ring, thereby excluding other centrally located bonds, which might have as great or even greater impact on molecular complexity and thus simplification. Therefore, we also envision as potentially valuable bonds for retrosynthetic disconnection those that have a high *twc* rank—even when they are excluded by rule 3. Such bonds fulfill the express intent of the rules, viz. to most efficiently lead to simplification, better than some of the bonds included under rule 3.

In Figure 3 and Table 3 we show structures and data for Corey's most complex example, lycopodine **27**, along with aca **26** and octabisvalene **28**, a very small pentacyclic compound. The former two are included to cover the complete set of examples in ref 5, and the latter to exemplify the fact that for a set of rules one can always find cases where they are not useful.

Bond *h* in tetracycle **26** is a rare case of a bond with a relatively high *twc* rank (#4) that violates rule 2A, as it is not exo to any ring. The reason is the adjacency of this bond to a double bond, which enhances its impact on *twc*.

Pentacycle **28** contains three distinct kinds of C–C bonds, none of which qualifies as strategic by the LHASA rules. Bond *a*, ranked highest by both *twc* and $N_T$, is doubly forbidden as strategic by rule 2B. In second place according to both *twc* and $N_T$ is bond *b*, which is likewise eliminated by rule 2B, while bond *c* violates rule 1B. In the actual synthesis, bond *b* was formed last, which corresponds to disconnecting it first in retrosynthesis.[21]

**Statistical Considerations.** The 284 distinct bonds in compounds **1**–**28** form a data set that can be subjected to statistical treatment in order to either confirm or falsify our conclusions derived from qualitative observations.

First of all, to what extent do the *twc* and $N_T$ complexity scales agree? Since $N_T$ is not sensitive to heteroatoms, we excluded the heterocyles from this part of the study. Moreover, cyclooctane **7**, which only contains one type of bond, does not allow for a meaningful correlation. The $r^2$ values for linear correlation between *twc* and $N_T$ within the remaining 19 hydrocarbons vary between 0.520 for **21** and 0.988 for **3**, with all but the former above 0.700.

Since the values of both indices increase rapidly with increasing graph size (number of atoms), they have to be converted to logarithms in order to obtain a meaningful correlation of the two scales across the whole data set. For the 191 distinct precursors derived from the 20 hydrocarbon targets by one-bond disconnections, the $r^2$ value for linear correlation between log(*twc*) and log($N_T$) is 0.954. Figure 4 gives an impression of this correlation.

**Table 3.** Molecular Complexity Data *twc* and $N_T$ for Compounds **26**–**28** and Synthetic Precursors Lacking One Bond

| | *twc* | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|
| **26** | 703342396 | | 82186 | | |
| d | 218415126 | 1 | 18198 | 2 | rule 4 |
| f | 230290474 | 2 | 23518 | 6 | strategic |
| e | 305095824 | 3 | 20534 | 3 | rule 4 |
| h | 348351966 | 4 | 31978 | 14 | rule 2A |
| a | 369412664 | 5 | 17818 | 1 | rule 3 |
| j | 411846018 | 6 | 22626 | 4 | rule 4 |
| g | 415006480 | 7 | 23326 | 5 | rule 3 |
| n | 438820962 | 8 | 29470 | 10 | strategic |
| i | 462751624 | 9 | 29382 | 9 | rule 3 |
| m | 528588952 | 10 | 29330 | 8 | rule 3 |
| r | 536741516 | 11 | 28976 | 7 | rule 3 |
| p | 556875826 | 12 | 31358 | 13 | rule 2A |
| b | 560195102 | 13 | 30798 | 11 | rule 3 |
| c | 564359398 | 14 | 30938 | 12 | rule 3 |
| s | 587173100 | 15 | 33618 | 15 | rule 2A |
| o | 599895774 | 16 | 40808 | 16 | rule 2A |
| k | 619379914 | 17 | 41280 | 18 | rule 2A |
| q | 621540132 | 18 | 40916 | 17 | rule 2A |
| l | 628666850 | 19 | 41452 | 19 | rule 2A |
| t | 657475542 | 20 | 49588 | 20 | rule 1A |

| | *twc* (N≜2, O≜1) | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|
| **27** | 7435986449 | | 82186 | | |
| a | 1533579349 | 1 | 17818 | 1 | rule 4 |
| b | 3505820281 | 2 | 30798 | 11 | strategic |
| c | 3541427739 | 3 | 30938 | 12 | strategic |
| d | 3758416139 | 4 | 18198 | 2 | rule 4 |
| e | 4504308653 | 5 | 20534 | 3 | rule 4 |
| f | 4761065101 | 6 | 23518 | 6 | strategic |
| g | 5216904019 | 7 | 23326 | 5 | rule 3 |
| h | 5898299965 | 8 | 31978 | 14 | rule 2A |
| i | 6120518389 | 9 | 29382 | 9 | rule 3 |
| j | 6121365185 | 10 | 22626 | 4 | rule 4 |
| k | 6186470779 | 11 | 41280 | 18 | rule 2A |
| l | 6250618787 | 12 | 41452 | 19 | rule 2A |
| m | 6463042215 | 13 | 29330 | 8 | rule 3 |
| n | 6471575467 | 14 | 29470 | 10 | strategic |
| o | 6687762133 | 15 | 40808 | 16 | rule 2A |
| p | 6690427655 | 16 | 31358 | 13 | rule 2A |
| q | 6804392145 | 17 | 40916 | 17 | rule 2A |
| r | 6873702885 | 18 | 28976 | 7 | rule 3 |
| s | 7029643595 | 19 | 33618 | 15 | rule 2A |
| t | 7258576737 | 20 | 49588 | 20 | rule 1A |

| | *twc* | rank | $N_T$ | rank | comments |
|---|---|---|---|---|---|
| **28** | 26232 | | 1860 | | |
| a | 16914 | 1 | 595 | 1 | rule 2B |
| b | 17234 | 2 | 826 | 2 | rule 2B |
| c | 17442 | 3 | 917 | 3 | rule 1B |

More interesting is the question of dependence or independence of the LHASA criteria and the graph theoretical measures. In Figure 5 the mean rank numbers according to *twc* of the strategic and nonstrategic bonds (assigned by the rules) are plotted versus the target identification numbers. In all applicable cases the strategic bonds have higher rankings (lower rank numbers) than the nonstrategic ones. (*N.B.*, there are no bonds in **7**, **8**, and **28** that are strategic by rules 1–6.) The corresponding picture for $N_T$ (Figure 6) is similar except for compound **11**, which has only one kind of strategic bond (ranked #5) and a mean rank of 4.43 for its nonstrategic bonds.

For quantitatively comparing the binary LHASA results (strategic vs nonstrategic) with the continuous scales for *twc* and $N_T$, the latter were dichotomized by taking the median of *twc* or $N_T$ for each target structure as the border between "high impact" and "low impact" bonds. (As above, com-
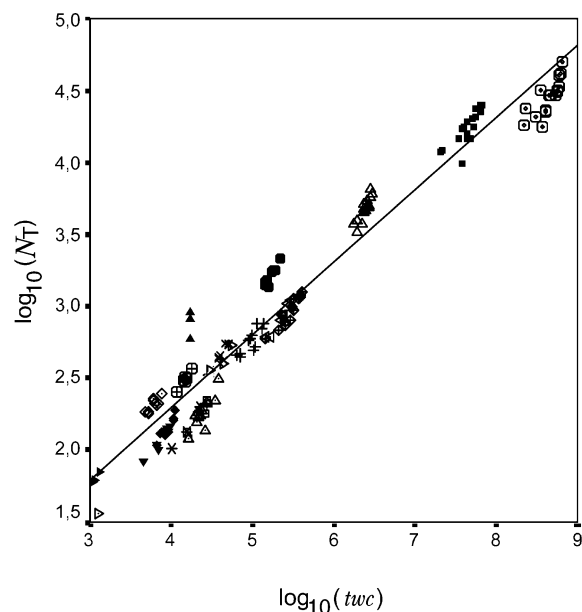
**Figure 4.** Plot of $\log_{10}(N_T)$ vs $\log_{10}(twc)$ for all hydrocarbon precursors (i.e., all distinct bonds in compounds **1**−**7**, **10**, **11**, **14**, **15**, **17**−**19**, **21**−**24**, **26**, and **28**). Bonds within a given target structure are represented by the same symbols.
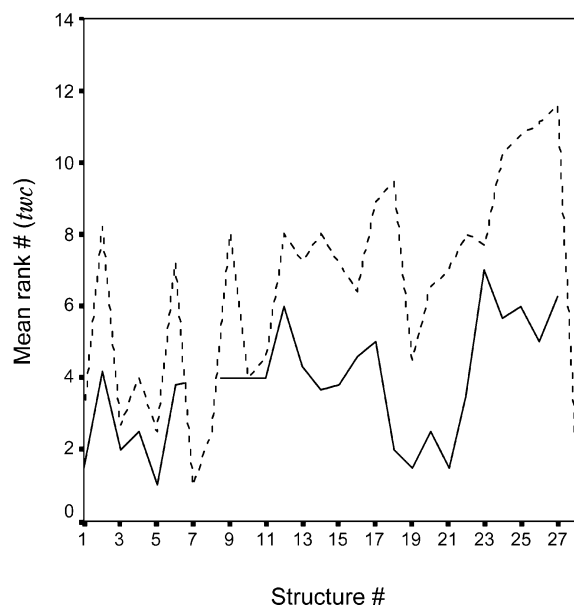


**Figure 5.** Mean rank numbers by *twc* of LHASA strategic bonds (solid line) and nonstrategic bonds (dashed line) vs compound ID number.

pounds **7**, **8**, and **28** were left out here.) Then for each compound a 2 × 2 contingency table was constructed, with the upper cells marked "high impact" and the lower cells "low impact", the left cells "strategic" and the right cells "nonstrategic". If the two categories were independent (null hypothesis), then one would expect the ratio between the left and right entries in the upper row to equal the corresponding ratio in the lower row. This would imply that if the left cell is more populated than the right cell in the upper row, then the same should be the case for the lower row.

The data evidently violate the null hypothesis expectation, in that for *twc*, the main diagonal is more populated for 24
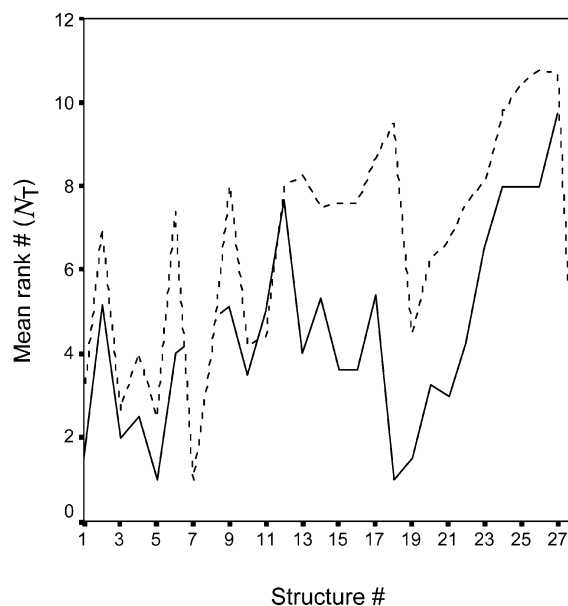


**Figure 6.** Mean rank numbers by $N_T$ of LHASA strategic bonds (solid line) and nonstrategic bonds (dashed line) vs compound ID number.

of 25 targets. The exception is one case of equal population (**23**). The appropriate statistical method for this situation is the Mantel-Haenszel test,[22] which utilizes the information contained in all the four-cell tables. This test allows us to conclude that *twc* and the LHASA criteria are decidedly *not* independent. Rather, the estimated probability for a high-impact bond to be strategic is 7.6 times that for a low-impact bond, with the 95% confidence interval ranging from 3.8 to 15.3.

The corresponding treatment of the $N_T$ data gives a somewhat weaker but still compelling case: there are 2 cases (out of 25) of equal and 5 cases of higher population on the secondary diagonal. The estimated probability for a high-impact bond to be strategic is 5.2 times that for a low-impact bond, with the 95% confidence interval ranging from 2.7 to 10.2.

### CONCLUSION

The LHASA rules for strategic bonds are a blend of those aimed at maximal simplification and those aimed at avoiding inaccessible ring sizes. The former were found to be paralleled by graph theoretical measures of molecular complexity, but not the latter. Bonds leading to maximal retrosynthetic simplification can be identified by simple calculations, and an ordering of such bonds is almost always achieved. In contrast, a yes-or-no decision is all that can be obtained from the LHASA rules.

From the beginning, a total correspondence between the LHASA rules and molecular complexity, measured here by the number of walks *twc* or the number of substructures $N_T$, was not to be expected. On the other hand, a complete lack of correspondence between molecular complexity and the LHASA rules would indicate total irrelevance of the former for organic synthesis. Perhaps surprising to some, our analysis reveals a high degree of correlation between graph theoretical molecular complexity indices and the rules for strategic bonds, which were originally found heuristically. This means that organic synthesis, at least as far as bridged

polycyclic compounds are concerned, is to a significant degree determined by the underlying mathematics, and following Kant,[7] to that extent can be considered a science. Other heuristics of organic synthesis are also based on graph theory, e.g., the principle of convergence, which is not detailed here, but is implicit in our previous results.[14a,17] Thus, our conclusions are potentially true for larger regions of organic synthesis, as well.

There is undoubtedly more of mathematics in the LHASA rules than highlighted here, e.g., chirality (rule 6) is fundamentally a mathematical concept. Furthermore, one might ask about a mathematical basis for the rules concerned with ring sizes (4 and parts of 1, 2, and 3), which have heretofore been considered "heuristic". After all, chemists have theories, ultimately based on physics, for the observation that some ring sizes are more favorable than others. Likewise, there certainly must be a physical basis underlying the "rich and powerful chemistry for forming bonds between carbon and ... heteroatoms". At present, such aspects seem to be beyond the reach of simple indices.

For more than a century, synthetic chemists have felt their profession to be both an art and a science—and a craft, as well. Our analysis adds some numerical evidence to this intuition, putting it on a more solid foundation. We conclude with another sentence written by Immanuel Kant in 1786: *"... so chemistry cannot become more than systematic art or experimental teachings, indeed never real science, since its principles ... are not amenable to the application of mathematics."*[23] In his time, chemistry gradually coalesced into existence from its predecessors, alchemy and the arts and crafts; however, neither Kant nor anyone else could foresee the extraordinary evolution of chemistry in the two centuries that followed.

## REFERENCES AND NOTES

(1) *The Oxford English Dictionary*, 2nd ed.; Simpson, J. A., Weiner, E. S. C., Eds.; Clarendon Press: Oxford, 1989.

(2) *Webster's Third New International Dictionary*; Gove, P. B., Ed.; Merriam: Springfield, MA, 1968.

(3) *Encarta World English Dictionary*; Soukhanov, A. H., Ed.; St. Martin's Press: New York, 1999.

(4) *Webster's New Collegiate Dictionary,* 150th Anniversary ed.; Woolf, H. B., Ed.; Merriam: Springfield, MA, 1981.

(5) Corey, E. J.; Howe, W. J.; Orf, H. W.; Pensak, D. A.; Petersson, G. General Methods of Synthetic Analysis. Strategic Bond Disconnections for Bridged Polycyclic Structures. *J. Am. Chem. Soc.* **1975**, *97*, 6116−6124.

(6) Corey, E. J.; Cheng, X.-M. *The Logic of Chemical Synthesis*; Wiley: New York, 1989; pp 37, 42.

(7) Kant, I. *Metaphysische Anfangsgründe der Naturwissenschaft*; Riga, 1786. In the German original: "Ich behaupte aber, daß in jeder besonderen Naturlehre nur so viel eigentliche Wissenschaft angetroffen werden könne, als darin Mathematik anzutreffen ist."

(8) Use of the expression "the most strategic disconnections" would seem to indicate that the authors thought there should be more and less strategic disconnections.

(9) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.

(10) Corey and many other chemists refer to this area as "chemical topology"; see also ref 15, footnote 7.

(11) Nikolić, S.; Trinajstić, N.; Tolić, I. M.; Rücker, G.; Rücker, C. On Molecular Complexity Indices. In *Complexity in Chemistry: Introduction and Fundamentals*; Bonchev, D., Rouvray, D. H., Eds.; Taylor & Francis: London, 2003; pp 29−89.

(12) Bertz, S. H. Complexity of Molecules and Their Synthesis. In *Complexity in Chemistry: Introduction and Fundamentals*; Bonchev, D., Rouvray, D. H., Eds.; Taylor & Francis: London, 2003; pp 91−156.

(13) Index $N_T$ counts all connected substructures from methane to the molecule itself, e.g., *n*-butane has four methane, three ethane, two propane, and one *n*-butane subgraph, so that its $N_T$ is 10. Note that substructures are named after the molecule in which any free valences have been replaced by hydrogen. A computer program that finds all connected substructures of a given structure is available from C.R. For more information, see: (a) Bertz, S. H.; Herndon, W. C. The Similarity of Graphs and Molecules. In *Artificial Intelligence Applications in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; American Chemical Society: Washington, DC, 1986; pp 169−175. (b) Bertz, S. H.; Sommer, T. J. Rigorous Mathematical Approaches to Strategic Bonds and Synthetic Analysis Based on Conceptually Simple New Complexity Indices. *Chem. Commun.* **1997**, 2409−2410. (c) Bertz, S. H. Complexity of Synthetic Reactions. The Use of Complexity Indices to Evaluate Reactions, Transforms and Disconnections. *New J. Chem.* **2003**, *27*, 860−869. (d) Bertz, S. H. Complexity of Synthetic Routes: Linear, Convergent and Reflexive Syntheses. *New J. Chem.* **2003**, *27*, 870−879.

(14) Index *twc* counts all walks of length 1 to $n - 1$ in a molecular graph, where *n* is the number of non-hydrogen atoms. A walk is an alternating sequence of atoms and bonds beginning and ending with an atom. For example, *n*-butane contains six walks of length 1 (1-2, 2-1, 2-3, 3-2, 3-4, 4-3), 10 walks of length 2 (1-2-3, 1-2-1, 2-3-4, 2-3-2, 2-1-2, 3-4-3, 3-2-1, 3-2-3, 4-3-4, 4-3-2), and 16 walks of length 3 (1-2-3-4, 1-2-3-2, 1-2-1-2, 2-3-4-3, 2-3-2-3, 2-3-2-1, 2-1-2-1, 2-1-2-3, 3-4-3-4, 3-4-3-2, 3-2-3-4, 3-2-3-2, 3-2-1-2, 4-3-4-3, 4-3-2-3, 4-3-2-1), so that its *twc* is 32. An efficient computer program that calculates walk counts without actually counting walks is available from C.R. For more information, see: (a) Rücker, G.; Rücker, C. Walk Counts, Labyrinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 99−106. (b) Rücker, G.; Rücker, C. On Finding Nonisomorphic Connected Subgraphs and Distinct Molecular Substructures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 314−320, 865. (c) Rücker, G.; Rücker, C. Substructure, Subgraph, and Walk Counts as Measures of the Complexity of Graphs and Molecules. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1457−1462. (d) Gutman, I.; Rücker, C.; Rücker, G. On Walks in Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 739−745.

(15) Bertz, S. H.; Rücker, C.; Rücker, G.; Sommer, T. J. Simplification in Synthesis. *Eur. J. Org. Chem.* **2003**, 4737−4740.

(16) Bertz, S. H.; Rücker, C. In Search of Simplification: The Use of Topological Complexity Indices to Guide Retrosynthetic Analysis. *Croat. Chem. Acta*, in press.

(17) (a) Bertz, S. H.; Wright, W. F. The Graph Theory Approach to Synthetic Analysis: Definition and Application of Molecular Complexity and Synthetic Complexity. *Graph Theory Notes of New York* **1998**, *35*, 32−48. (b) Bertz, S. H.; Sommer, T. J. Applications of Graph Theory to Synthesis Planning: Complexity, Reflexivity and Vulnerability. In *Organic Synthesis: Theory and Applications*; Hudlicky, T., Ed.; JAI Press: London, 1993; Vol. 2, pp 67−92.

(18) In this paper we use the version of the rules from ref 5; later they were reformulated in a manner less amenable to our analysis (ref 6).

(19) Bond *h* in **2** has an atypically high rank for a chain bond, since it is rather central in the molecule, close to the bicyclic skeleton and to a bridgehead. This effect is seen even more dramatically in 1-propyl-bicyclo[2.2.1]heptane, where the innermost acyclic bond is directly attached to a bridgehead. This bond is ranked #4 out of eight by *twc* and #1 by $N_T$, i.e., higher than any cyclic bond.

(20) Progress in organic synthesis can be amazingly rapid. E.g., a bond's location in a four-membered ring was not sufficient to pass the equivalent of rule 1 in the 1971 version of the rules, and "rings having greater than six members" were to be avoided at that time, see: Corey, E. J. Computer-assisted Analysis of Complex Synthetic Problems. *Q. Rev., Chem. Soc.* **1971**, *25*, 455−482. By 1975 four- and seven-membered rings were acceptable, see ref 5.

(21) (a) Rücker, C. Phenylsulfonylsubstituierte Octabisvalene, Synthesen und Reaktionen. *Chem. Ber.* **1987**, *120*, 1629−1644. (b) Rücker, C.; Trupp, B. Pentacyclo[5.1.0.0.$^{2,4}$0.$^{3,5}$0$^{6,8}$]octane (Octabisvalene). *J. Am. Chem. Soc.* **1988**, *110*, 4828−4829.

(22) Mantel, N.; Haenszel, W. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *J. Nation. Cancer Inst.* **1959**, *22*, 719−748.

(23) See ref 7. In the German original: "... so kann Chymie nichts mehr als systematische Kunst, oder Experimentallehre, niemals aber eigentliche Wissenschaft werden, weil die Principien derselben ... der Anwendung der Mathematik unfähig sind."