Simplification in Synthesis[‡]

Steven H. Bertz,*[a] Christoph Rücker,*[b] Gerta Rücker,[c] and Toby J. Sommer[d]

Dedicated to Prof. E. J. Corey on the occasion of his 75th birthday

Keywords: Synthesis design / Synthetic methods / Retrosynthesis

Mathematics Applied to Synthetic Analysis (MASA) is a useful addition to Logic and Heuristics Applied to Synthetic Analysis (LHASA), as it can be used to calculate the simplification afforded by alternative disconnections of a target molecule. One-bond and two-bond disconnections that are more efficient as far as simplification is concerned than those identified by the LHASA criteria have been found by using MASA.

The prevailing paradigm for synthesis planning is *retrosynthetic analysis*,[1−5] which involves starting with the target molecule and working backwards, breaking bonds until 'simple' starting materials are recognized.[6] This approach is especially useful for polycyclic systems, which are among the most complex. As cogently stated by Corey and Cheng, "The existence of alternative bond paths through a molecular skeleton as a consequence of the presence of cyclic subunits gives rise to a *topological complexity* which is proportional to the degree of internal connectivity."[3] (The emphasis on topological complexity is ours.) Unfortunately, this concept has not been rigorously defined,[7−10] and consequently it is not as useful as it might be. A meaningful method for quantifying topological complexity and thus the degree of simplification is of obvious theoretical and practical value, given the size of the synthetic enterprise.[11]

As implied by the above quote, topological complexity is the aspect of molecular complexity that results from connectivity.[7] There are other interesting aspects of molecular complexity,[12] but topology is the foundation.[13] Over the past several years, two independent approaches to the quantification of molecular complexity C have been developed. *All possible substructures* N_T counts all connected substructures,[14−17] isomorphic and nonisomorphic,[18] and the *total walk count twc* counts all walks of length 1 to $n - 1$, where n is the number of non-hydrogen atoms in a molecule.[19,20] While they are based on different mathematical formalisms, N_T and *twc* give similar results for ordering hydrocarbons (vide infra), which suggests that both tap a deeper underlying reality.[21,22] By using an axiomatic system,[23] it can be shown that isobutane ($N_T = 11$, $twc = 36$, see Exp. Sect.) is more complex than butane ($N_T = 10$, $twc = 32$), and both indices order these simple molecules correctly.[24] As will be appreciated upon comparison of these values with those of typical synthetic targets (vide infra), computer programs for the calculation of N_T and *twc* are indispensable.[17,19]

Tricyclo[7.3.1.0^{1,5}]tridecane (**1**) was the final example used to illustrate one-bond and two-bond ('bond-pair') disconnections of bridged and spiro systems according to the criteria of LHASA.[3] Our labeling scheme is the same except we have added letters *i-o*, and we also denote disconnections by the bonds that are broken. Table 1 summarizes the complexities N_T and *twc* of the precursors that result from all possible one-bond disconnections of target **1** along with their ranks in parentheses. Typically, N_T and *twc* values are highly correlated, and for precursors *a*−*o* in Table 1, the correlation coefficient is $r = 0.884$.

Rank is determined by the degree of simplification in the retrosynthetic direction, calculated as the change in complexity $\Delta C = C(\text{precursors}) - C(\text{target})$, where the target is the desired product of a synthetic reaction and the precursors are the starting materials. In this paper $C = N_T$ or *twc*. The best disconnection (#1) is the one that gives the greatest simplification, i.e., the negative number with the greatest absolute value for ΔC. There is agreement between N_T and *twc* as to the #1 and #2 disconnections: *b* and *a*,

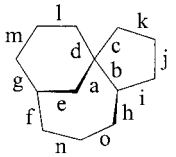
[‡] Applications of Discrete Mathematics to Chemistry, 21. Part 20: Ref.[16]

[a] Complexity Study Center,
Mendham, NJ 07945, USA
Fax: (internat.) + 1-973-628-3401
E-mail: sbertz@complexitystudycenter.org
[b] Department of Mathematics, University of Bayreuth,
95440 Bayreuth, Germany
Fax: (internat.) + 49-921-553385
E-mail: Christoph.Ruecker@uni-bayreuth.de
[c] Department of Sports Medicine, University of Freiburg,
79106 Freiburg, Germany
[d] Department of Biochemistry, Brandeis University,
Waltham, MA 02454, USA

Supporting information for this article is available on the WWW under http://www.eurjoc.org or from the author.

 DOI: 10.1002/ejoc.200300558

Table 1. One-bond disconnections of **1**



1

Bond	N_T	*twc*
− (**1**)	3087	1098050
a	863 (2)	441870 (2)
b	711 (1)	344554 (1)
c	1083 (6)	517860 (3)
d	979 (3)	523666 (4)
e	1039 (4)	653798 (5)
f	1187 (7)	815880 (11)
g	1243 (9)	805638 (10)
h	1079 (5)	689448 (7)
i	1239 (8)	671584 (6)
j	1645 (15)	818512 (12)
k	1593 (14)	782520 (8)
l	1453 (11)	805074 (9)
m	1541 (13)	873122 (14)
n	1485 (12)	902250 (15)
o	1449 (10)	870716 (13)

respectively, which reduce branching by breaking a bond to the quaternary center, as illustrated in Figure 1. The #3 disconnection according to N_T (#4 according to *twc*) is *d*, which also breaks a bond to the quaternary carbon. We

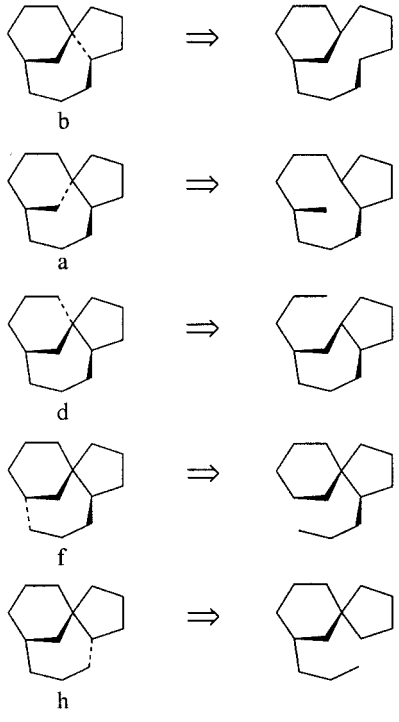


Figure 1. One-bond disconnections: the #1−3 topological strategic bonds (dashed lines) according to MASA (*b*, *a* and *d*, respectively), and the heuristic strategic bonds according to LHASA (*f* and *h*)

refer to the bonds rated highly by MASA (Mathematics Applied to Synthetic Analysis) as the *topological strategic bonds*.

The *heuristic strategic bonds* identified by LHASA are *f* and *h*,[3] each of which breaks a bond to one of the two tertiary centers. They are ranked lower (between #5 and #11) by N_T and *twc*, since they effect less reduction in the degree of branching. (The best way to reduce overall branching is to reduce the number of bonds to the most branched atom.) These disconnections preserve the spiro-[4.5]decane ring system (see Figure 1).

Topological considerations may be overridden by heuristic (e.g., mechanistic) ones, depending upon the state of the art. Disconnection *b* preserves a bicyclo[7.3.1]tridecane bridged ring system, which contains a 10-membered ring, and disconnection *a* preserves a bicyclo[7.3.0]dodecane fused ring system, which contains a nine-membered ring. They are consistent with LHASA criterion 1,[3] "A strategic bond must be ... within a primary (i.e. non-peripheral, or non-perimeter) ring of 4−7 members and *exo* to a primary ring larger than three-membered," but they violate criterion 2, "A disconnection is not strategic if it involves a bond common to two bridged [or fused[5]] primary rings and generates a new ring having more than 7 members." As progress is made in medium and large ring synthesis,[25−27] the LHASA prohibition will weaken, and the heuristics will come into alignment with the mathematics.[28]

Disconnection *d*, which we rank third overall, also reduces branching at the quaternary center, but does not comport with LHASA criterion 3:[3] "A strategic bond must be *endo* to (within) a ring of *maximum bridging*." The ring of maximum bridging in this case is the seven-membered one (*a*-*b*-*h*-*o*-*n*-*f*-*e*), which is a primary ring within the limits (four to seven members) set by criterion 1 (vide supra); however, bond *d* is *exo* to it rather than *endo*. While it is excluded from the set of strategic bonds by the LHASA criteria, bond *d* is in a six-membered primary ring, which is generally the most favorable size to form.[29] All things considered, disconnection *d* is an excellent compromise between topology and the state of the art.

Compound **1** itself is not in the CAS database; however, several derivatives are known. Figure 2 summarizes the key ring-forming steps from the literature syntheses of molecules that contain **1** as a substructure. Reaction (a) is taken from a synthesis of portulal,[30] and it forms bond *d*, which was one of the top three (out of fifteen) calculated by our method (vide supra). Reaction (b) is taken from the chemistry of laurenene,[31] and it forms bond *g*, which was not selected by LHASA or MASA. Disconnection *g* is ranked #9 by N_T and #10 by *twc*, which are in the range (#5−11) of the LHASA one-bond disconnections (vide supra). Reactions (c) and (d) are based on two-bond disconnections, as discussed below.

The strategic two-bond disconnections based on MASA and LHASA are illustrated in Figure 3. According to N_T, the best topological two-bond disconnections are *b*,*c* and *b*,*f* ($\Delta N_T = -2868$), which result in identical precursors. (They are placed third by $\Delta twc = -910658$.) According to

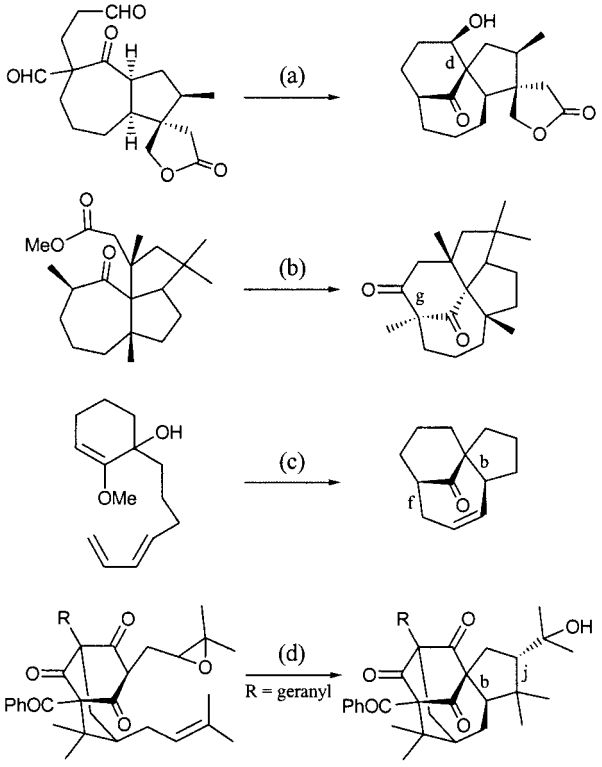


Figure 2. Syntheses of the tricyclo[7.3.1.0^{1,5}]tridecane skeleton from the literature: (a) portulal intermediate,[30] (b) laurenene transformation,[31] (c) synthetic reaction,[32] (d) biosynthetic reaction;[33] the bonds formed are labeled in the products

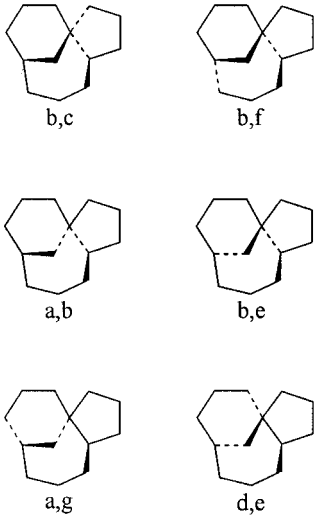


Figure 3. Two-bond disconnections: the best topological strategic bonds (dashed lines) according to N_T (*b*,*c*, and *b*,*f*) and *twc* (*a*,*b* and *b*,*e*), and the heuristic strategic bonds according to LHASA (*a*,*g* and *d*,*e*)

twc, the best topological two-bond disconnections are *a*,*b* and *b*,*e* ($\Delta twc = -954144$), which also result in identical precursors. (They are placed a close second by $\Delta N_T = -2862$.) At this point, it is useful to note that there are 105 possible two-bond disconnections of **1**, and that these four are in the top 5 % for both indices. (The *twc*-N_T correlation coefficient for all 105 precursors is $r = 0.892$.) According to LHASA, the heuristic two-bond disconnections are *a*,*g* ($\Delta N_T = -2732$, $\Delta twc = -758784$) and *d*,*e* ($\Delta N_T = -2820$, $\Delta twc = -819454$), which are rated significantly lower (cf. Supporting Information).

It appears that LHASA puts a premium on preserving the spiro system in the one-bond disconnections, which corresponds to forming it early in the synthesis plan; however, the LHASA two-bond disconnections eliminate the spiro system, which corresponds to forming it late in the plan. Thus, there is a fundamental contradiction in the heuristic approach. In contrast, the preferred topological two-bond disconnections (*a*,*b*; *b*,*c*; *b*,*e*; *b*,*f*) all break bond *b*, which is the best topological one-bond disconnection.

Disconnection *b*,*f* (#1 according to N_T) corresponds to a [4+3] cycloaddition, which has been reduced to practice by Harmata et al., as shown in Figure 2(c).[32] Sampsonione F has been proposed to arise via biosynthetic reaction (d), which forms bond set *b*,*j*,[33] not selected by LHASA or MASA. The LHASA strategic two-bond disconnections, *a*,*g* and *d*,*e*, correspond to intramolecular Diels–Alder reactions; however, no literature examples were found for these disconnections.

Figure 4 contains the structures used to illustrate strategic bond disconnections in the seminal paper that introduced the concept.[5] (The presentation may be different.) The heuristic (LHASA) strategic bonds are indicated by

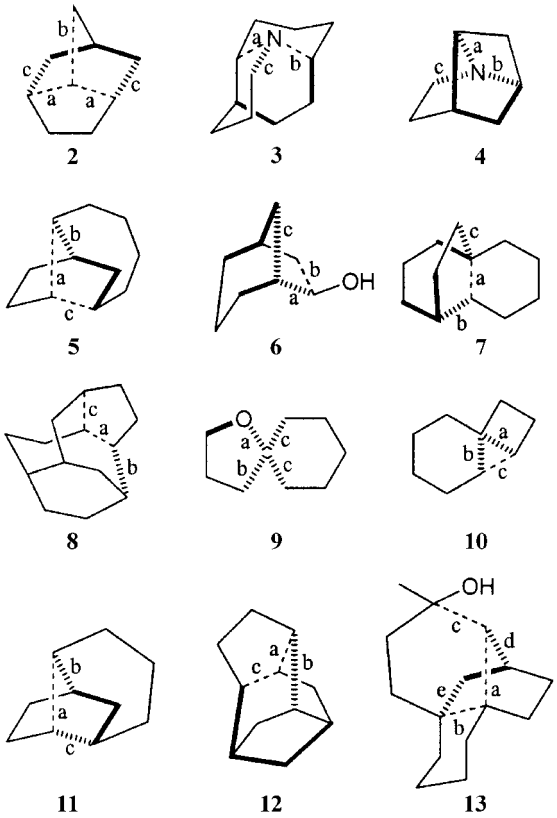


Figure 4. Topological (dashed or hashed) and heuristic (hashed or bold) strategic bonds for typical polycyclic examples (after ref.[5]); the order of simplification is indicated by *a*, *b*, *c*, etc.; the hashed (dashed and bold) bonds are those identified by both MASA and LHASA, e.g., *a*-*c* in **4**

bold lines, and the topological (MASA) strategic bonds, calculated by using *twc*, are denoted by dashed lines. The hashed bonds, indicating bold and dashed, are both heuristic and topological. The topological strategic bonds shown are those that are ranked #1−3 (*a*−*c*) except in the case of **13**, where #1−5 (*a*−*e*) are included, since this structure has more bonds. In every case at least one heuristic strategic bond is selected by the topological method. In a majority of the cases (7 out of 12), 50 % or more of the LHASA strategic bonds are identified by MASA. Especially noteworthy is **8**, which has only one heuristic strategic bond, and it is among those selected by our method. An additional advantage of MASA is that it allows strategic bond disconnections to be ranked, whereas LHASA does not, e.g., six of ten bonds in **4** are equally strategic by the heuristic rules.[5] Our method identifies bond *a* as most strategic, *b* as second and *c* as third. All three are C−N bonds, which are preferred mechanistically.

The use of general indices of molecular complexity appears to be a promising approach to the determination of topological strategic bonds. It is useful for ranking the LHASA (heuristic) strategic bonds as well as stimulating consideration of alternative disconnections. As the state of the art develops, the heuristics may be expected to converge with topology, and MASA can hasten that day, as it can highlight fruitful areas for research and development.

Experimental Section

General Remarks: Values of the indices N_T and *twc* were calculated by using the computer programs already published.[17,19] To illustrate the indices used in this paper, N_T and *twc* have been derived for butane and isobutane. Substructures are usually named for the stable molecules with the same skeletons. Based on the definition,[14] a structure is a substructure of itself. For butane the substructures are methane (4), ethane (3), propane (2) and butane (1), and for isobutane they are methane (4), ethane (3), propane (3) and isobutane (1). In the following lists the numbers refer to carbon atoms (IUPAC numbering), and the (directed) walks that are also (directed) paths are italicized. The walks in butane are *1-2*, *2-1*, *2-3*, *3-2*, *3-4*, *4-3*, 1-2-1, *1-2-3*, 2-1-2, 2-3-2, *2-3-4*, *3-2-1*, 3-2-3, 3-4-3, *4-3-2*, 4-3-4, 1-2-1-2, 1-2-3-2, *1-2-3-4*, 2-1-2-1, 2-1-2-3, 2-3-2-1, 2-3-2-3, 2-3-4-3, 3-2-1-2, 3-2-3-2, 3-2-3-4, 3-4-3-2, 3-4-3-4, *4-3-2-1*, 4-3-2-3, 4-3-4-3. The walks in isobutane are *1-2*, *2-1*, *2-3*, *3-2*, *2-4*, *4-2*, 1-2-1, *1-2-3*, *1-2-4*, 2-1-2, 2-3-2, 2-4-2, *3-2-1*, 3-2-3, *3-2-4*, *4-2-1*, *4-2-3*, 4-2-4, 1-2-1-2, 1-2-3-2, 1-2-4-2, 2-1-2-1, 2-1-2-3, 2-1-2-4, 2-3-2-1, 2-3-2-3, 2-3-2-4, 2-4-2-1, 2-4-2-3, 2-4-2-4, 3-2-1-2, 3-2-3-2, 3-2-4-2, 4-2-1-2, 4-2-3-2, 4-2-4-2. For undirected walks or paths, the number of directed ones is simply divided by 2.

[1] J.-H. Fuhrhop, G. Li, *Organic Synthesis: Concepts and Methods*, Wiley-VCH, Weinheim, **2003**, p. 425−459.
[2] M. B. Smith, *Organic Synthesis*, McGraw-Hill, New York, **1994**, p. 980−1098.
[3] E. J. Corey, X.-M. Cheng, *The Logic of Chemical Synthesis*, Wiley, New York, **1989**.
[4] S. Warren, *Organic Synthesis: The Disconnection Approach*, Wiley, Chichester, **1982**.
[5] E. J. Corey, W. J. Howe, H. W. Orf, D. A. Pensak, G. Petersson, *J. Am. Chem. Soc.* **1975**, *97*, 6116−6124.
[6] In some cases, bonds may be formed for strategic reasons.[30]
[7] Topology is defined broadly in chemistry. The term is commonly used, as by Corey and Cheng,[3] to describe the effects of connectivity; its more rigorous applications are in the areas of graph-theoretically non-planar compounds,[8] molecular knots,[9] chirality,[10] etc.
[8] C. Rücker, M. Meringer, *MATCH−Commun. Math. Comput. Chem.* **2002**, *45*, 153−172.
[9] *Molecular Catenanes, Rotaxanes and Knots: A Journey Through the World of Molecular Topology* (Eds.: J.-P. Sauvage, C. Dietrich-Buchecker), Wiley, New York, **1999**.
[10] E. Flapan, *When Topology Meets Chemistry: A Topological Look at Molecular Chirality*, Cambridge University Press, Cambridge, **2000**.
[11] D. Bradley, *New Sci.* **1997**, *156*, 40−43.
[12] S. H. Bertz in *Complexity in Chemistry: Introduction and Fundamentals* (Eds.: D. Bonchev, D. H. Rouvray), Taylor & Francis, London, **2003**, p. 91−156.
[13] S. Nikolić, N. Trinajstić, I. M. Tolić, G. Rücker, C. Rücker in *Complexity in Chemistry: Introduction and Fundamentals* (Eds.: D. Bonchev, D. H. Rouvray), Taylor & Francis, London, **2003**, p. 29−89.
[14] A substructure of molecule *M* is a structure that has all its atoms and bonds in *M*, see S. H. Bertz, T. J. Sommer, *Chem. Commun*. **1997**, 2409−2410. When lone pairs of electrons are included, N_T is called N_T(lpe).[12,16]
[15] S. H. Bertz, W. C. Herndon in *Artificial Intelligence Applications in Chemistry* (Eds.: T. H. Pierce, B. A. Hohne), American Chemical Society, Washington, DC, **1986**, p. 169−175.
[16] S. H. Bertz, *New J. Chem.* **2003**, *27*, 870−879 and references cited therein.
[17] G. Rücker, C. Rücker, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 314−320.
[18] Isomorphic structures have the same adjacency matrix for some labeling, see S. H. Bertz, T. J. Sommer, *Chem. Commun.* **2003**, 1000−1001.
[19] A walk is an alternating series of non-unique points (atoms) and lines (bonds) beginning and ending with points, see G. Rücker, C. Rücker, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1457−1462.
[20] I. Gutman, C. Rücker, G. Rücker, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 739−745.
[21] The 'underlying reality' is a poset (partially ordered set); see D. J. Klein, *MATCH−Commun. Math. Comput. Chem.* **2000**, *42*, 7−21.
[22] S. Nikolić, I. M. Tolić, N. Trinajstić, I. Baučić, *Croat. Chem. Acta* **2000**, *73*, 909−921. See ref. [27] therein.
[23] S. H. Bertz, *Discrete Appl. Math.* **1988**, *19*, 65−83.
[24] Butane and isobutane each contain 6 nontrivial, undirected paths (see Exp. Sect.), which illustrates why the number of paths is not a robust measure of complexity.
[25] A. J. McCarroll, J. C. Walton, *Angew. Chem. Int. Ed.* **2001**, *40*, 2224−2248.
[26] G. Mehta, V. Singh, *Chem. Rev.* **1999**, *99*, 881−930.
[27] G. A. Molander, *Acc. Chem. Res.* **1998**, *31*, 603−609.
[28] This expectation is based in part on changes in the LHASA rules between the first version and later ones,[3,5] see E. J. Corey, *Q. Rev., Chem. Soc*. **1971**, *25*, 455−482.
[29] C. D. Johnson, *Acc. Chem. Res.* **1993**, *26*, 476−482. See also ref.[1], p. 111−133 and ref.[2], p. 601−611.
[30] T. Tokoroyama, K. Matsuo, H. Kotsuki, R. Kanazawa, *Tetrahedron* **1980**, *36*, 3377−3390.
[31] R. E. Corbett, J. R. Guild, D. R. Lauren, R. T. Weavers, *Aust. J. Chem.* **1991**, *44*, 1139−1143.
[32] M. Harmata, G. Bohnert, L. Kürti, C. L. Barnes, *Tetrahedron Lett.* **2002**, *43*, 2347−2349.
[33] L.-H. Hu, K.-Y. Sim, *Tetrahedron* **2000**, *56*, 1379−1386.

Received September 4, 2003

