

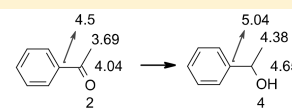
# Molecular Complexity and Retrosynthesis

John R. Proudfoot\*<sup>1</sup>

Boehringer Ingelheim Pharmaceuticals Inc, 900 Ridgebury Road, P.O. Box 368, Ridgefield, Connecticut 06877, United States

**S** Supporting Information

**ABSTRACT:** An atom-environment complexity measure,  $C_A$ , to assess local changes in complexity during synthetic transformations is described. The complexity measure is based on applying Shannon's equation to the number and diversity of paths up to two bonds in length emanating from an atom node. The method requires no explicit accounting for bond type, stereochemistry, ring membership, symmetry, or molecular size.  $C_A$  varies with expectation across a number of basic reaction examples and may identify the key disconnections to guide retrosynthesis.



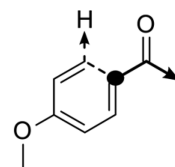
Comparative changes in local calculated complexity,  $C_A$ , for reactants and products during reaction transformations

$$C_A = - \sum p_i \log_2 p_i + \log_2 N$$

In the classic work *The Logic of Chemical Synthesis*, Corey and Cheng emphasized the value of considering molecular complexity in synthesis planning and pointed to molecular size, element and functional group content, cyclic connectivity, stereocenter content, chemical reactivity, and structural instability as key parameters.<sup>1</sup> While chemical reactivity and structural instability affect success in the physical manipulation of compounds, the remaining features explicitly relate to the graphical representation of the target structure. As such, they can be quantified, and graph theoretical and substructure feature approaches to assessing molecular complexity along a synthesis route have been developed.<sup>2–11</sup> Recently, Sarpong applied a network complexity analysis to identify key disconnections in the retrosynthesis of weisaconitine D using a maximally bridged ring feature as the key determinant of complexity.<sup>12</sup>

We recently disclosed an atom environment approach to measuring molecular complexity and applied it to a representation of approved drugs.<sup>13</sup> While that disclosure focused on comparisons of complexity across a broad set of drug molecules, it was based on a measure of the complexity,  $C_A$ , associated with each atom environment in a molecule. As such, the method shows regions in a structure that are of higher complexity and possibly of highest value for disconnection.

The method generates all of the paths up to length 2 (i.e., assemblies of atoms connected by one or two bonds, including those terminating in hydrogen atoms) emanating from each atom. It aggregates three features of each atom node, the atom type represented as atomic number, the total number of connections, and the number of non-hydrogen connections. These latter two are the X and D features in SMARTS. Each path is then represented in a form such as [#06&X3&D3] ~ [#06&X3&D2] ~ [#01&X1&D1], where ~ represents a bond connection of any type. Figure 1 provides a detailed example, representing the paths emanating from the atom indicated in the structure shown. The equation  $C_A$  (Figure 1), where  $p$  is the fractional occurrence of each path type and  $N$  represents the total number of paths emanating from the atom, provides the complexity of the individual atom environment. The



**Paths:**

**[#06&X3&D3]~[#06&X3&D2]~[#01&X1&D1]**  
**[#06&X3&D3]~[#06&X3&D2]~[#01&X1&D1]**  
**[#06&X3&D3]~[#06&X3&D2]~[#06&X3&D2]**  
**[#06&X3&D3]~[#06&X3&D2]~[#06&X3&D2]**  
**[#06&X3&D3]~[#06&X3&D3]~[#06&X4&D1]**  
**[#06&X3&D3]~[#06&X3&D3]~[#08&X1&D1]**

$$C_A = - \sum p_i \log_2 p_i + \log_2 N$$

$$\begin{aligned} C_A &= -[2/6(\log_2 2/6) + 2/6(\log_2 2/6) + 1/6(\log_2 1/6) + 1/6(\log_2 1/6)] + \log_2 6 \\ &= -(-0.53 - 0.53 - 0.43 - 0.43) + 2.58 \\ &= 4.5 \end{aligned}$$

**Figure 1.** Paths from the indicated atom with highlighted paths in boldface and sample calculation of  $C_A$ .

equation is essentially that applied by Bertz<sup>14</sup> to the complexity of whole molecules which, through addition of the term for the total number of paths emanating from the node atom, is an extension of Shannon's equation.<sup>15</sup> The method requires no explicit accounting for bond type, stereochemistry, ring membership, symmetry, or molecular size. A representative calculation for one atom node is given in Figure 1.<sup>16</sup>

Figure 2 illustrates the application  $C_A$  to some simple synthetic transformations. In reaction 1, a Friedel–Crafts acylation,  $C_A$  of the atom to which the acyl group becomes attached increases from 3.84 to 4.5 and the adjacent atom  $C_A$  also increases, from 4.24 to 4.64. The values for the other atom nodes in the parent structure do not change, a consequence of limiting the method to paths only up to length 2. In the

**Received:** March 27, 2017

**Published:** May 31, 2017

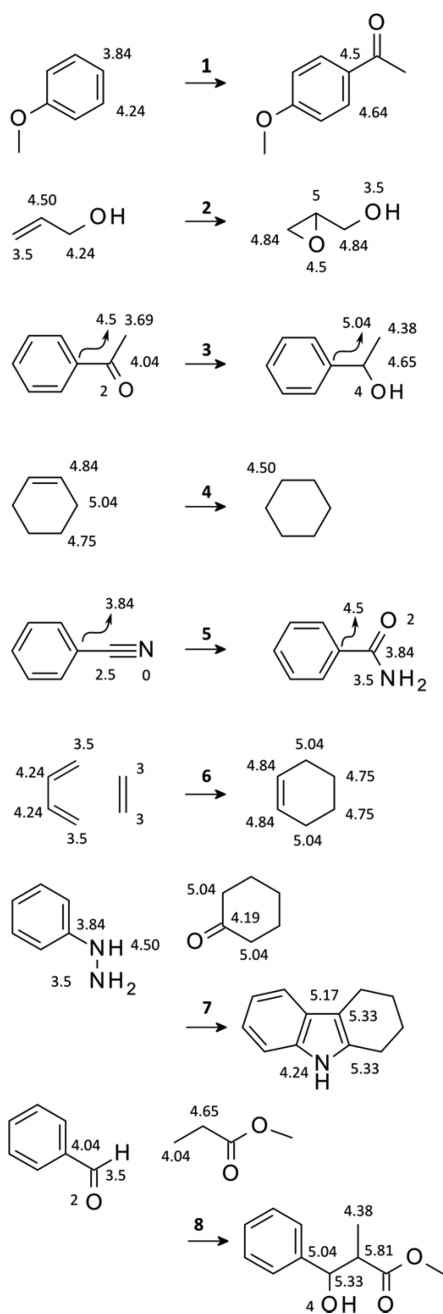


Figure 2.  $C_A$  applied to some example reactions.

epoxidation reaction 2,  $C_A$  for the olefin precursor atoms and the allylic atom increase. In the reduction reaction 3, all of the atom nodes around the reaction center increase in complexity, while for the reduction reaction 4, all are decreased. This latter result is a consequence of the decreased diversity of path types in the product molecule. Note that  $C_A$  of the nitrogen atom node in the reactant for reaction 4 is 0, since there is only one path emanating from this atom and the equation reduces to  $\log_2(1*(1/1)) + \log_2(1)$ , i.e., 0. Note also that the maximum attainable complexity for a tetravalent atom is  $12(1/12)\log_2(1/12) + \log_2 12$ , or 7.168. Cyclization reactions 6 and 7 show that  $C_A$  increases for all the nodes involved in ring formation. In the enolate–aldol reaction 8, the complexity of the chiral centers generated in the product is higher than the corresponding precursor atoms. Since these changes in  $C_A$  reflect expectation across a range of transformations in the forward reaction sense

we considered that it might provide guidance on retrosynthetic disconnections in more complex systems, where that guidance is independent of relationship to reactants or reactions.

As noted above, a network complexity analysis was applied to the retrosynthesis of weisaconitine D to identify complex sites as a guide for disconnection. That analysis focused on bridged ring systems as centers of complexity and identified the disconnection indicated by the dashed line on the molecular structure in Figure 3. A  $C_A$  map ranking the 10 most complex

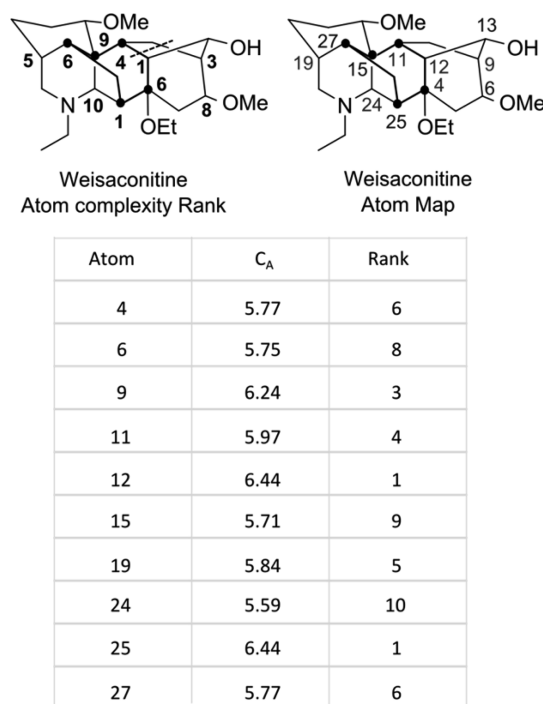


Figure 3. Top 10 most complex atoms of weisaconitine D.

atom nodes is also overlaid on the structure. Three of the top four complex atom environments in the molecule form part of the two bond disconnection selected as the key retrosynthesis step.

We also surveyed the syntheses of strychnine, once described as the most complex substance known for its molecular size,<sup>17</sup> in relation to the  $C_A$  map. The top 10 most complex atom environments in strychnine are indicated in Figure 4 and

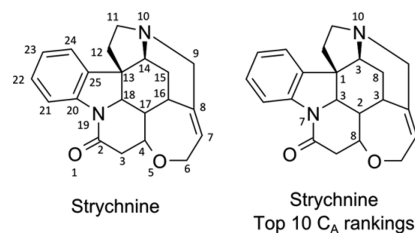
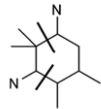
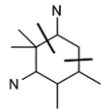
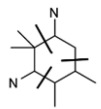
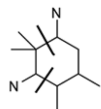
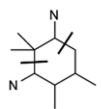
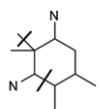
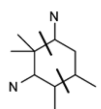
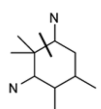


Figure 4.

localize to the central carbocyclic ring. Table 1 provides a summary of this ring construction across the various syntheses.<sup>18</sup> Three begin with this ring intact as a monocycle (no disconnection, column 2). Somewhat unexpectedly, two of these are significantly longer than the median 15 step length, essentially because of the number of subsequent manipulations required on this central ring. Many of the remaining syntheses

Table 1

Synthesis	Ring Disconnection(s)	Ring Step	Total Steps	
Vanderwal <sup>19</sup> Reissig <sup>20</sup> Bodwell <sup>21</sup> Padwa <sup>22</sup>	13-14 17-18		3 2 5 5	9 10 12 12
MacMillan <sup>23</sup>	13-14 15-16		4	12
Vollhardt <sup>24</sup>	13-14 17-18 15-16		6	13
Bosch <sup>25</sup>	none			14
Martin <sup>26</sup>	13-14 17-18		10	15
Rawal <sup>27</sup>	13-18 14-15		8	15
Andrade <sup>28</sup>	13-12 17-18		3	15
Kuehne <sup>29</sup>	13-14 17-16		6	19,21
Mori <sup>30</sup>	None			22
Overman <sup>31 32</sup> Fukuyama <sup>33</sup> Magnus <sup>34</sup> Woodward <sup>35 36</sup>	13-14		18 20 13 15	24 25 28 29
Shibasaki <sup>37</sup>	None			31

begin with the 13–18 connection as part of an indole system and then employ a concurrent 13–14 and 17–18 connection, involving the most complex atom environments of the final system, in the ring formation step. Most of the sequences of median length or shorter employ this approach. Three routes employ a disconnection from atom 15, the least complex atom environment in the ring; however, this is always concurrent with the 13–14 disconnection.

## CONCLUSION

An atom environment measure of molecular complexity generates a complexity value  $C_A$  for each atom environment in a molecule.  $C_A$  changes logically across a range of molecular transformations and may have application in retrosynthesis to identify regions of higher complexity that can logically be prioritized for disconnection.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.joc.7b00714.

Data for Figure 2 (TXT)

Complete  $C_A$  maps for all the molecules shown (PDF)

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: john.proudfoot@discoverybytes.com.

### ORCID

John R. Proudfoot: 0000-0002-5922-2228

### Notes

The author declares no competing financial interest.

## REFERENCES

- Corey, E. J.; Cheng, X.-M. *The Basis for Retrosynthetic Analysis. The Logic of Chemical Synthesis*; Wiley: New York, 1989; pp 2–16.
- Bertz, S. H. Convergence, molecular complexity, and synthetic analysis. *J. Am. Chem. Soc.* **1982**, *104*, 5801–5803.
- Barone, R.; Petitjean, M.; Baralotto, C.; Piras, P.; Chanon, M. Information theory description of synthetic strategies. A new similarity index. *Synthesis* **1998**, *1998*, 1559–1583.
- Whitlock, H. W. On the Structure of Total Synthesis of Complex Natural Products. *J. Org. Chem.* **1998**, *63*, 7982–7989.
- Barone, R.; Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Model.* **2001**, *41*, 269–272.
- Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.
- Li, J.; Eastgate, M. D. Current complexity: a tool for assessing the complexity of organic molecules. *Org. Biomol. Chem.* **2015**, *13*, 7164–7176.
- Hendrickson, J. B. Systematic synthesis design. 6. Yield analysis and convergency. *J. Am. Chem. Soc.* **1977**, *99*, 5439–5450.
- Bertz, S. H.; Sommer, T. J. Applications of Graph Theory to Synthesis Planning: Complexity. In *Reflexivity, and Vulnerability In Organic Synthesis – Theory and Applications*; Hudlicky, T., Ed.; JAI Press, Inc.: Greenwich, CT, 1993; Vol. 2, pp 67–92.
- Bertz, S. H. Complexity of synthetic routes: Linear, convergent and reflexive syntheses. *New J. Chem.* **2003**, *27*, 870–879.
- Wender, P. A.; Miller, B. L. *Towards the Ideal Synthesis: Connectivity Analysis and Multibondforming Processes In Organic Synthesis – Theory and Applications*; Hudlicky, T., Ed.; JAI Press, Inc.: Greenwich, CT, 1993; Vol. 2, pp 27–66.
- Marth, C. J.; Gallego, G. M.; Lee, J. C.; Lebold, T. P.; Kulyk, S.; Kou, K. G.; Qin, J.; Lilien, R.; Sarpong, R. Network-analysis-guided synthesis of weisaconitine D and liljestrandinine. *Nature* **2015**, *528*, 493–498.
- Proudfoot, J. A Path Based Approach to Assessing Molecular Complexity. *Bioorg. Med. Chem. Lett.* **2017**, *27*, 2014–2017.
- Bertz, S. H. The first general index of molecular complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599–3601.
- Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

(16) The method is implemented in Pipeline Pilot version 9.5. The protocol is deposited at the user exchange site <https://exchange.sciencecloud.com/exchange/>.

(17) Robinson, R. Molecular structure of strychnine, brucine, and vomicine. *Progress in Organic Chemistry*; Academic Press: New York, 1952; Vol. 1, pp 1–21.

(18) The site <http://www.synarchive.com/> (accessed 2/1/2017) provided a straightforward access to the data provided in the table.

(19) Martin, D. B. C.; Vanderwal, C. D. A synthesis of strychnine by a longest linear sequence of six steps. *Chem. Sci.* **2011**, *2*, 649–651.

(20) Beemelmans, C.; Reissig, H.-U. A Short Formal Total Synthesis of Strychnine with a Samarium Diodide Induced Cascade Reaction as the Key Step. *Angew. Chem., Int. Ed.* **2010**, *49*, 8021–8025.

(21) Bodwell, G. J.; Li, J. A Concise Formal Total Synthesis of (±)-Strychnine by Using a Transannular Inverse-Electron-Demand Diels ± Alder Reaction of a [3](1,3)Indolo[3](3,6)pyridazinophane. *Angew. Chem., Int. Ed.* **2002**, *41*, 3261–3262.

(22) Zhang, H.; Boonsombat, J.; Padwa, A. Total Synthesis of (±)-Strychnine via a [4 + 2]-Cycloaddition/Rearrangement Cascade. *Org. Lett.* **2007**, *9*, 279–282.

(23) Jones, S. B.; Simmons, B.; Mastracchio, A.; MacMillan, D. W. C. Collective synthesis of natural products by means of organocascade catalysis. *Nature* **2011**, *475*, 183–188.

(24) Eichberg, M. J.; Dorta, R. L.; Lamottke, K.; Vollhardt, K. P. C. The Formal Total Synthesis of (±)-Strychnine via a Cobalt-Mediated [2+ 2 + 2] Cycloaddition. *Org. Lett.* **2000**, *2*, 2479–2481.

(25) Solé, D.; Bonjoch, J.; García-Rubio, S.; Peidró, E.; Bosch, J. Enantioselective Total Synthesis of Wieland - Gumlich Aldehyde and (–)-Strychnine. *Chem. - Eur. J.* **2000**, *6*, 655–665.

(26) Ito, M.; Clark, C. W.; Mortimore, M.; Goh, J. B.; Martin, S. F. A Biomimetic Approach to the Strychnos Alkaloids. A Novel, Concise Synthesis of (±)-Akuammicine and a Route to (±)-Strychnine. *J. Am. Chem. Soc.* **2001**, *123*, 8003–8010.

(27) Rawal, V. H.; Iwasa, S. A Short, Stereocontrolled Synthesis of Strychnine. *J. Org. Chem.* **1994**, *59*, 2685–2686.

(28) Sirasani, G.; Paul, T.; Dougherty, W., Jr.; Kassel, S.; Andrade, R. B. Concise Total Syntheses of (±)-Strychnine and (±)-Akuammicine. *J. Org. Chem.* **2010**, *75*, 3529–3532.

(29) Kuehne, M. E.; Xu, F. Total Synthesis of Strychnan and Aspidospermatan Alkaloids. 3. The Total Synthesis of (±)-Strychnine. *J. Org. Chem.* **1993**, *58*, 7490–7497.

(30) Mori, M.; Nakanishi, M.; Kajishima, D.; Sato, Y. A Novel and General Synthetic Pathway to Strychnos Indole Alkaloids: Total Syntheses of (–)-Tubifoline, (–)-Dehydrotubifoline, and (–)-Strychnine Using Palladium-Catalyzed Asymmetric Allylic Substitution. *J. Am. Chem. Soc.* **2003**, *125*, 9801–9807.

(31) Knight, S. D.; Overman, L. E.; Pairedeau, G. Synthesis applications of cationic aza-Cope rearrangements. 26. Enantioselective total synthesis of (–)-strychnine. *J. Am. Chem. Soc.* **1993**, *115*, 9293–9294.

(32) Knight, S. D.; Overman, L. E.; Pairedeau, G. Asymmetric Total Syntheses of (–)- and (+)-Strychnine and the Wieland - Gumlich Aldehyde. *J. Am. Chem. Soc.* **1995**, *117*, 5776–5788.

(33) Kaburagi, Y.; Tokuyama, H.; Fukuyama, T. Total Synthesis of (–)-Strychnine. *J. Am. Chem. Soc.* **2004**, *126*, 10246–10247.

(34) Magnus, P.; Giles, M.; Bonnert, R.; Kim, C. S.; McQuire, L.; Merritt, A.; Vicker, N. Synthesis of Strychnine via the Wieland-Gumlich Aldehyde. *J. Am. Chem. Soc.* **1992**, *114*, 4403–4405.

(35) Woodward, R. B.; Cava, M. P.; Ollis, W. D.; Hunger, A.; Daeniker, H. U.; Schenker, K. The Total Synthesis of Strychnine. *J. Am. Chem. Soc.* **1954**, *76*, 4749–4751.

(36) Woodward, R. B.; Cava, M. P.; Ollis, W. D.; Hunger, A.; Daeniker, H. U.; Schenker, K. The Total Synthesis of Strychnine. *Tetrahedron* **1963**, *19*, 247–288.

(37) Ohshima, T.; Xu, Y.; Takita, R.; Shimizu, S.; Zhong, D.; Shibasaki, M. Enantioselective Total Synthesis of (–)-Strychnine Using the Catalytic Asymmetric Michael Reaction and Tandem Cyclization. *J. Am. Chem. Soc.* **2002**, *124*, 14546–14547.