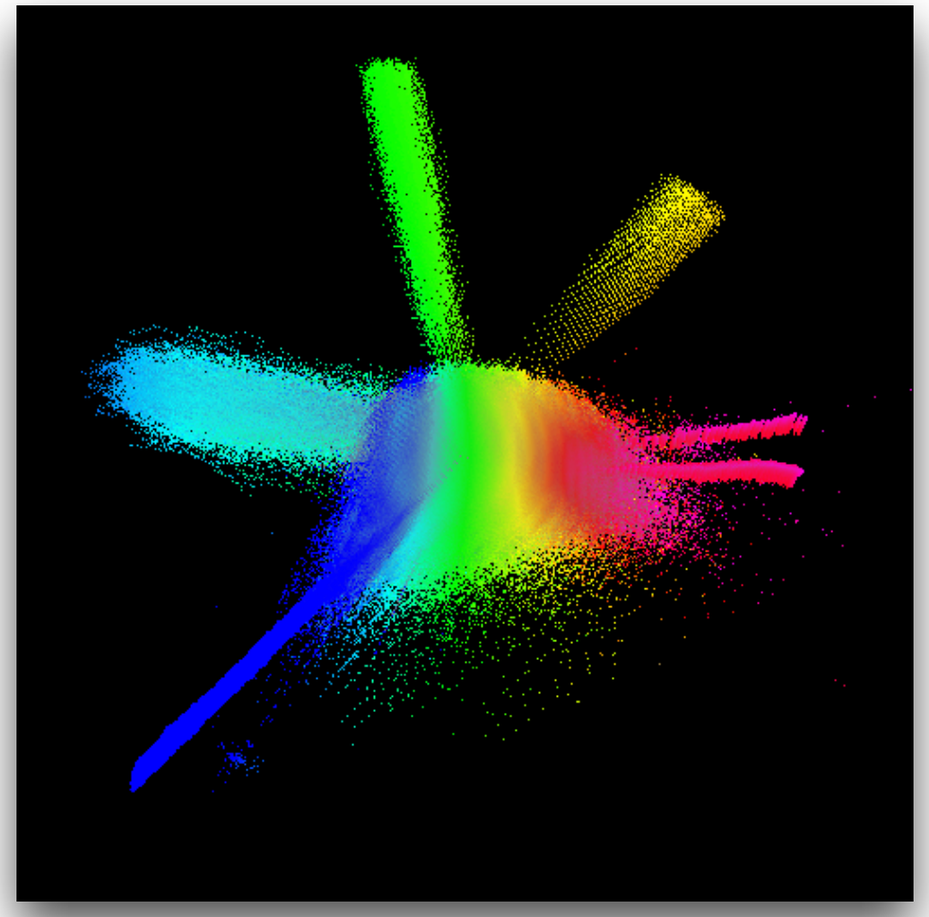
Quantifying Molecular Complexity



Johanna Masterson

Sorensen Group Meeting

Literature - August 21, 2020

ACS Chem. Neurosci. 2012, 3, 9, 649–657

nature

Article | Published: 16 October 2019

Concise asymmetric synthesis of (−)-bilobalide

Meghan A. Baker, Robert M. Demoret, Masaki Ohtawa ✉ & Ryan A. Shenvi ✉

Nature **575**, 643–646(2019) | Cite this article

nature

Article | Published: 16 October 2019

Concise asymmetric synthesis of (−)-bilobalide

Meghan A. Baker, Robert M. Demoret, Masaki Ohtawa ✉ & Ryan A. Shenvi ✉

Nature **575**, 643–646(2019) | Cite this article

Synthesis and Mechanistic Interrogation of *Ginkgo biloba* Chemical Space en route to (–)-Bilobalide

Robert M. Demoret,[1,‡] Meghan A. Baker,[1,‡] Masaki Ohtawa,[1,†] Shuming Chen,[2] Ching-Ching Lam,[2] Stefano Forli,[3] Kendall N. Houk,[2] Ryan A. Shenvi[1*]

[1]Department of Chemistry, Scripps Research, 10550 North Torrey Pines Road, La Jolla, California 92037, United States
[2]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA
[3]DISCoBio, Scripps Research, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

Nature, 2019, 575, 643-646
ChemRxiv, 2020, 10.26434/chemrxiv.12132939.v2

Molecular Complexity: Two General Types

Intrinsic Molecular Complexity

- A measure of the complexity of a molecule's structure
(i.e. heteroatoms, stereochemistry, etc.)
- Advancing methods and technologies do not alter intrinsic molecular complexity

J. Chem. Inf. Model. 2016, 56, in 462-470

Molecular Complexity: Two General Types

Intrinsic Molecular Complexity

- A measure of the complexity of a molecule's structure
(i.e. heteroatoms, stereochemistry, etc.)
- Advancing methods and technologies do not alter intrinsic molecular complexity

Extrinsic Molecular Complexity

- A measure of the complexity of the synthesis of a molecule
(i.e. how difficult certain motifs are to synthesize)
- Advancing methods and technologies do alter extrinsic molecular complexity

J. Chem. Inf. Model. 2016, 56, 462-470

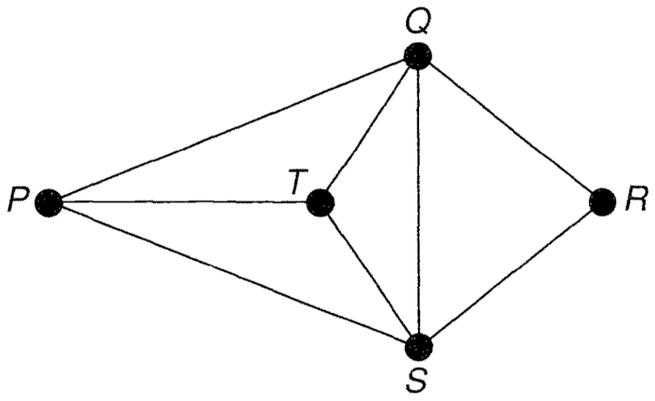
Graph Theory: Introduction

- The study of *graphs*
- *Graph*: a set of vertices and the edges connecting those vertices
- *Subgraph*: a graph within a larger graph

Robin J. Wilson, *Introduction to Graph Theory*, 1996, 4th ed.

Graph Theory: Introduction

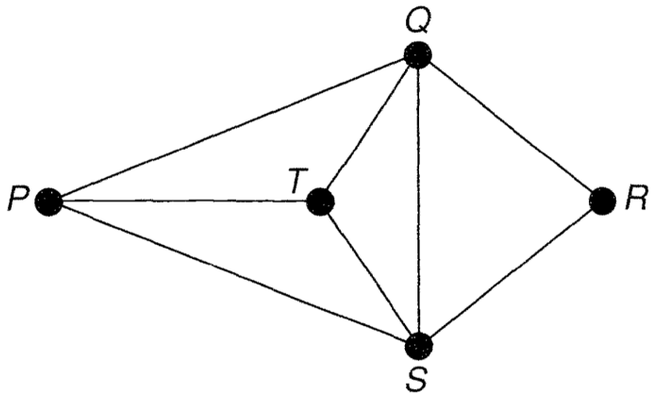
- The study of *graphs*
- *Graph*: a set of vertices and the edges connecting those vertices
- *Subgraph*: a graph within a larger graph

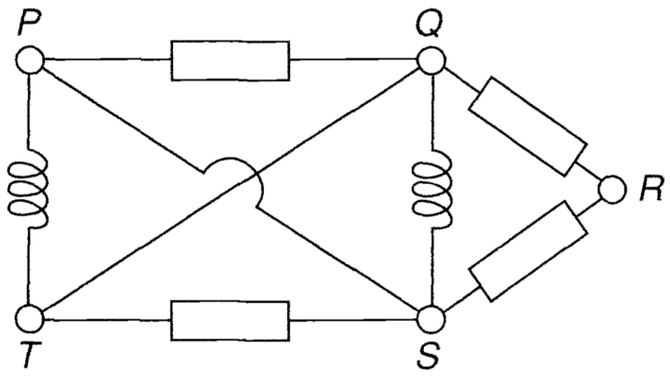


Robin J. Wilson, *Introduction to Graph Theory*, 1996, 4th ed.

Graph Theory: Introduction

- The study of *graphs*
- *Graph*: a set of vertices and the edges connecting those vertices
- *Subgraph*: a graph within a larger graph

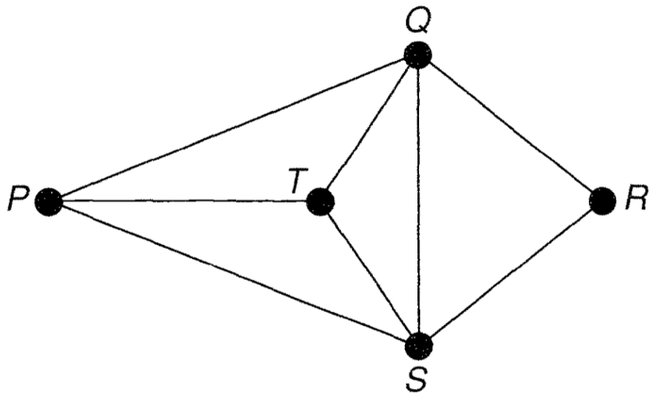


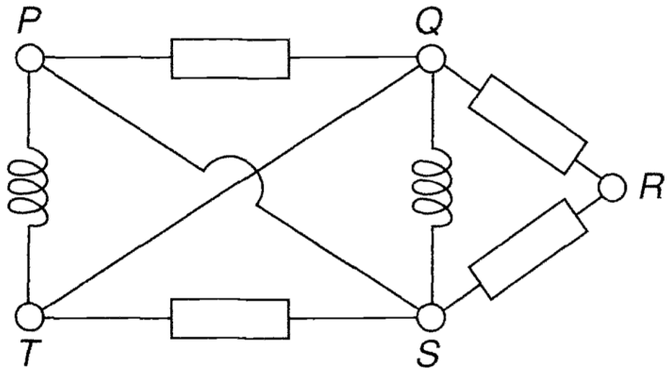


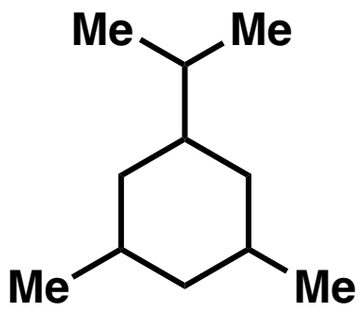
Robin J. Wilson, *Introduction to Graph Theory*, 1996, 4th ed.

Graph Theory: Introduction

- The study of *graphs*
- *Graph*: a set of vertices and the edges connecting those vertices
- *Subgraph*: a graph within a larger graph



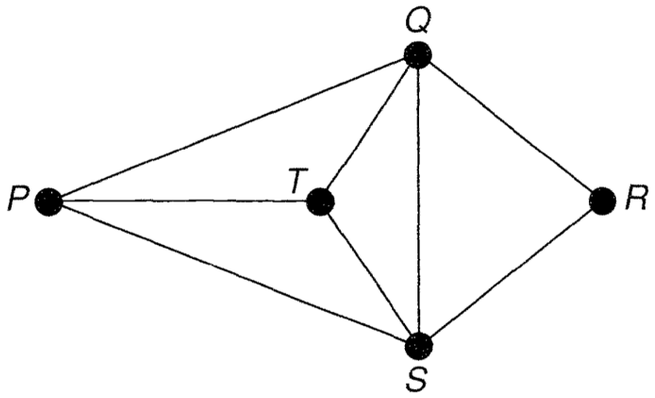


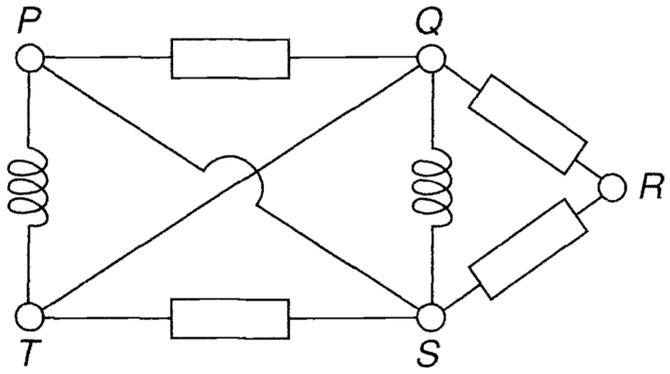


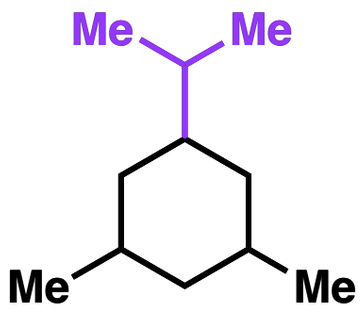
Robin J. Wilson, *Introduction to Graph Theory*, 1996, 4th ed.

Graph Theory: Introduction

- The study of *graphs*
- *Graph*: a set of vertices and the edges connecting those vertices
- *Subgraph*: a graph within a larger graph







Robin J. Wilson, *Introduction to Graph Theory*, 1996, 4th ed.

Information Theory: Introduction

- The study of the *storage and communication of information*
- Claude Shannon's "A Mathematical Theory of Communication" (1948)
  - Information is stored in *variables* which can assume different *values*
  - *Information entropy*: the amount of information in a variable, or how much storage is required to store a variable
  - *Shannon's entropy*: formula which tells the number of bits needed to store the information in a given variable

Understanding Shannon's Entropy Metric for Information, https://arxiv.org/pdf/1405.2061.pdf

Information Theory: Introduction

- The study of the *storage and communication of information*
- Claude Shannon's "A Mathematical Theory of Communication" (1948)
  - Information is stored in *variables* which can assume different *values*
  - *Information entropy*: the amount of information in a variable, or how much storage is required to store a variable
  - *Shannon's entropy*: formula which tells the number of bits needed to store the information in a given variable

$$H(X) = -\sum P_X(x)\log P_X(x)$$

Entropy

Variable

Probability of outcome x

Information encoded in outcome x

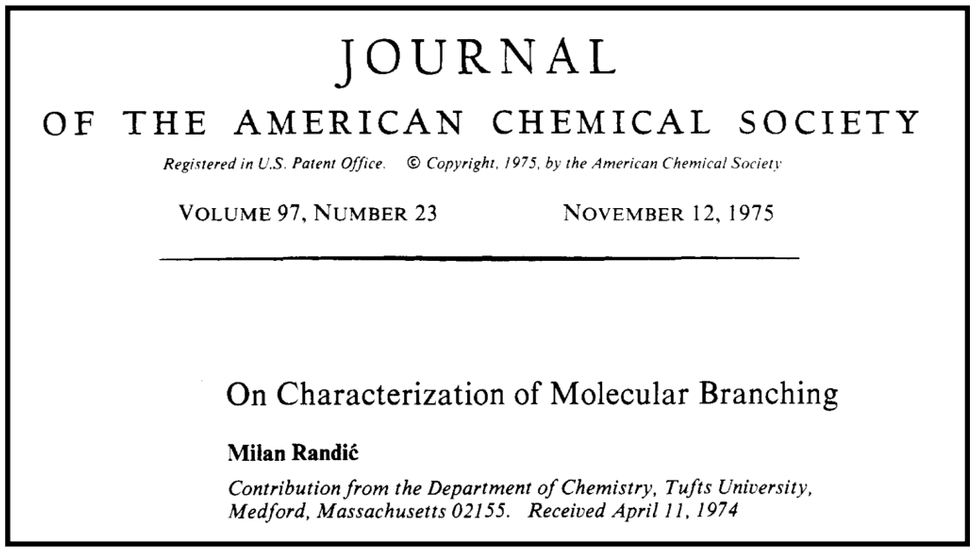
Understanding Shannon's Entropy Metric for Information, https://arxiv.org/pdf/1405.2061.pdf

Molecular Complexity: A History

Molecular Complexity: A History

J. Chem. Inf. Model. 2016, 56, 3, 462–470

Randić: On the Characterization of Molecular Branching



JOURNAL

OF THE AMERICAN CHEMICAL SOCIETY

VOLUME 97, NUMBER 23 NOVEMBER 12, 1975

On Characterization of Molecular Branching

Milan Randić

Contribution from the Department of Chemistry, Tufts University, Medford, Massachusetts 02155. *Received April 11, 1974*

- Aimed to address the ambiguity in degree of branching between hydrocarbons
- Developed an index for assessing degree of branching in the context of graph theory
  - Index was built by assigning each bond type in acyclic hydrocarbons a value - a molecule's branching index is given by the sum of the values of the bonds

J. Am. Chem. Soc. 1975, 97, 23, 6609–6615

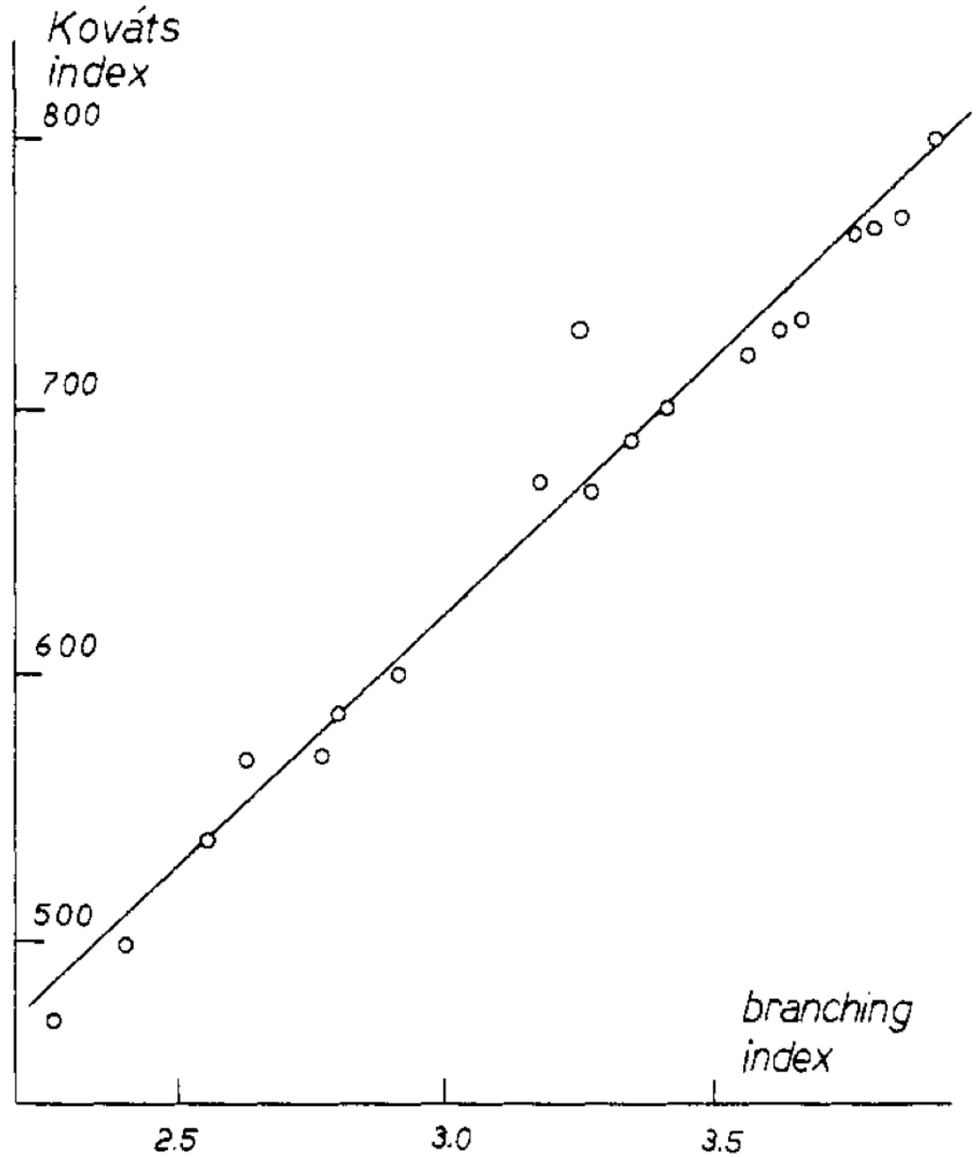
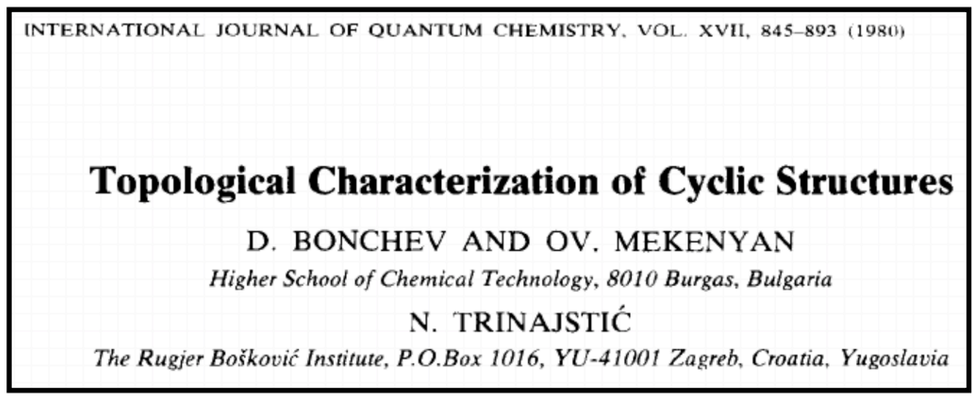


Figure 2. A correlation between the empirical branching index of Kováts and the topological branching index of this work.

Bonchev: Topological Characterization of Cyclic Structures



INTERNATIONAL JOURNAL OF QUANTUM CHEMISTRY, VOL. XVII, 845–893 (1980)

Topological Characterization of Cyclic Structures

D. BONCHEV AND OV. MEKENYAN
Higher School of Chemical Technology, 8010 Burgas, Bulgaria
N. TRINAJSTIĆ
The Rugjer Bošković Institute, P.O.Box 1016, YU-41001 Zagreb, Croatia, Yugoslavia

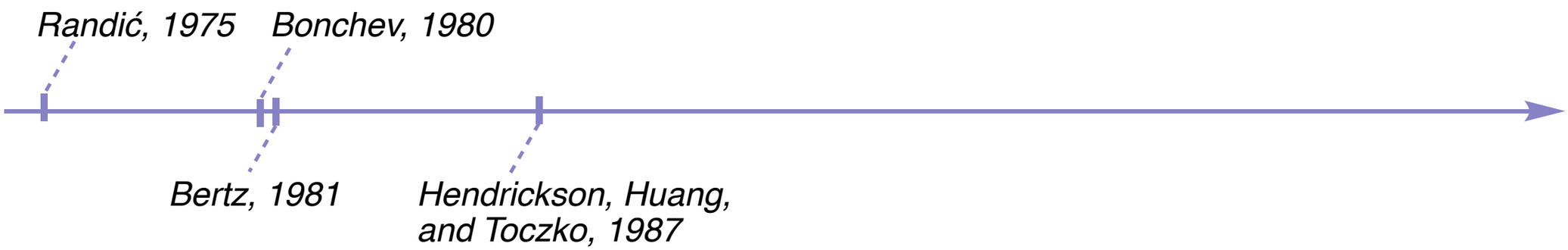
- Used the concept of topological distance in graphs to determine molecular cyclicity

Int. J. Quantum Chem. 1980, 17, 5,845-893

Molecular Complexity: A History

J. Chem. Inf. Model. 2016, 56, 3, 462–470

Molecular Complexity: A History



J. Chem. Inf. Model. 2016, 56, 3, 462–470

Bertz: The First General Index of Molecular Complexity

The First General Index of Molecular Complexity

Steven H. Bertz

Bell Laboratories,
Murray Hill, New Jersey 07974
Received July 18, 1980

Synthetic chemists have been defining a "complex molecule" in the way that many people define art: they know it when they see it. While the features which contribute to the complexity of a molecule have been discussed,[1–4] no unified index has been formulated which takes into account the size, symmetry, branching, rings, multiple bonds, and heteroatoms characteristic of a complex molecule. This communication shows how concepts from graph theory and information theory can be combined to create the first general index of molecular complexity.

$$C = C_\eta + C_E$$

$$C_\eta = 2\eta \lg \eta - \sum_i \eta_i \lg \eta_i$$

$$C_E = E \lg E - \sum_j E_j \lg E_j$$

- Built on the work of Rashevsky, who first applied information theory to graph theory in attempting to define the information content of organisms
- Based on Shannon entropy, where the information containing variables considered are *connections* and *heteroatoms*
  - The sum of these two entropies yields the total complexity
- Immediately recognized the use in assessing synthetic transformations

Bull. Math. Biophys. 1955, *17* (3), 229–235.
J. Am. Chem. Soc. 1981, 103, 12, 3599–3601
J. Chem. Inf. Comput. Sci. 1987, 27, 2, 63–67

Bertz: The First General Index of Molecular Complexity

The First General Index of Molecular Complexity

Steven H. Bertz

Bell Laboratories,
Murray Hill, New Jersey 07974
Received July 18, 1980

Synthetic chemists have been defining a "complex molecule" in the way that many people define art: they know it when they see it. While the features which contribute to the complexity of a molecule have been discussed,[1–4] no unified index has been formulated which takes into account the size, symmetry, branching, rings, multiple bonds, and heteroatoms characteristic of a complex molecule. This communication shows how concepts from graph theory and information theory can be combined to create the first general index of molecular complexity.

$$C = C_\eta + C_E$$

$$C_\eta = 2\eta \lg \eta - \sum_i \eta_i \lg \eta_i$$

$$C_E = E \lg E - \sum_j E_j \lg E_j$$

- Built on the work of Rashevsky, who first applied information theory to graph theory in attempting to define the information content of organisms
- Based on Shannon entropy, where the information containing variables considered are *connections* and *heteroatoms*
  - The sum of these two entropies yields the total complexity
- Immediately recognized the use in assessing synthetic transformations

Bull. Math. Biophys. 1955, *17* (3), 229–235.
J. Am. Chem. Soc. 1981, 103, 12, 3599–3601
J. Chem. Inf. Comput. Sci. 1987, 27, 2, 63–67

Bertz: The First General Index of Molecular Complexity

The First General Index of Molecular Complexity

Steven H. Bertz

Bell Laboratories,
Murray Hill, New Jersey 07974
Received July 18, 1980

Synthetic chemists have been defining a "complex molecule" in the way that many people define art: they know it when they see it. While the features which contribute to the complexity of a molecule have been discussed,[1–4] no unified index has been formulated which takes into account the size, symmetry, branching, rings, multiple bonds, and heteroatoms characteristic of a complex molecule. This communication shows how concepts from graph theory and information theory can be combined to create the first general index of molecular complexity.

$$C = C_\eta + C_E$$

$$C_\eta = 2\eta \lg \eta - \sum_i \eta_i \lg \eta_i$$

$$C_E = E \lg E - \sum_j E_j \lg E_j$$

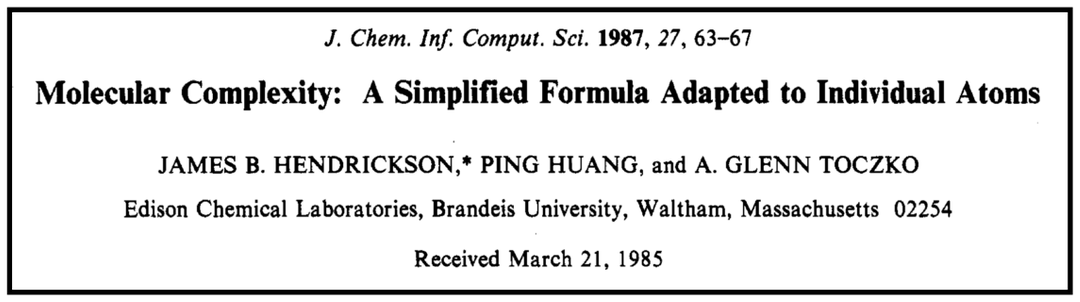
- Built on the work of Rashevsky, who first applied information theory to graph theory in attempting to define the information content of organisms
- Based on Shannon entropy, where the information containing variables considered are *connections* and *heteroatoms*
  - The sum of these two entropies yields the total complexity
- Immediately recognized the use in assessing synthetic transformations

Bull. Math. Biophys. 1955, *17* (3), 229–235.
J. Am. Chem. Soc. 1981, 103, 12, 3599–3601
J. Chem. Inf. Comput. Sci. 1987, 27, 2, 63–67

Now that it is possible to calculate a number for any molecule which measures its complexity, it is possible to calculate the change in complexity, ΔC_T (hereafter symbolized as Δ), upon going from reactant to product in the course of a chemical reaction. The increases in complexity for the Diels–Alder reaction between butadiene and *p*-benzoquinone and the Weiss reaction[13] of glyoxal with dimethyl 3-ketoglutarate are calculated in Figure 1. This example shows that it is still possible to invent powerful synthetic reactions[14] and the calculation of complexity can aid in recognizing them. For a *functional group interchange*[1] such as the conversion of cyclohexene to cyclohexanol ($\Delta = 1.1$), the change in complexity is small. The process of calculating Δ can be repeated for all the steps in a synthetic sequence, thus providing a means to gauge progress toward a complex target molecule.

J. Am. Chem. Soc. 1981, 103, 12, 3599–3601

Hendrickson, Huang, Toczko: A Simplified Formula Adapted to Individual Atoms



J. Chem. Inf. Comput. Sci. **1987**, *27*, 63–67

Molecular Complexity: A Simplified Formula Adapted to Individual Atoms

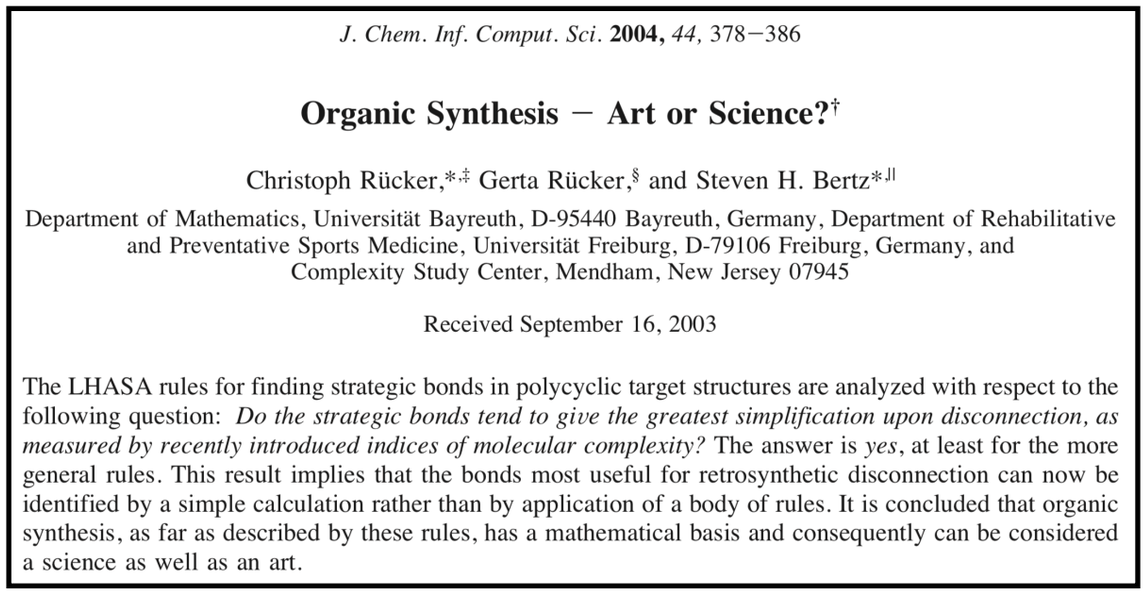
JAMES B. HENDRICKSON,* PING HUANG, and A. GLENN TOCZKO

Edison Chemical Laboratories, Brandeis University, Waltham, Massachusetts 02254

Received March 21, 1985

- Interested in the applications of Bertz's index to computer assisted synthesis planning
- Simplified the Bertz index by calculating based on properties of individual atoms (i.e. how many protons attached to each atom)
- Wrote a program which reproduced the Bertz index for a given molecule based on the simplified formulas

Bertz: Art or Science?



J. Chem. Inf. Comput. Sci. **2004,** *44,* 378−386

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

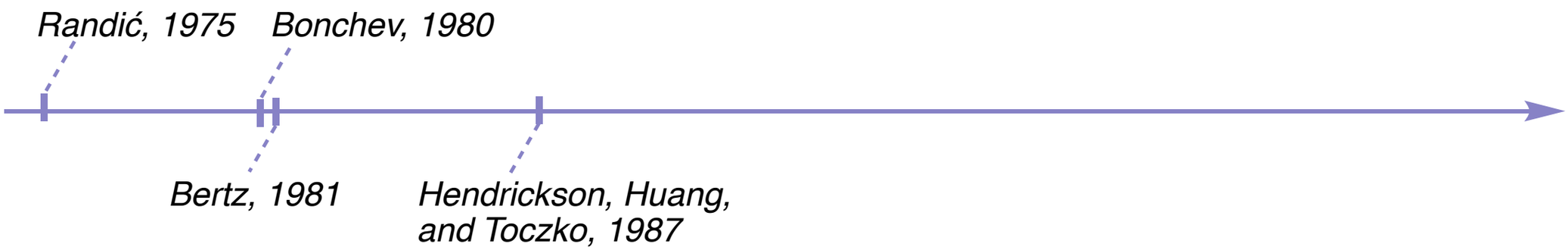
Received September 16, 2003

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.

- Analyzed the LHASA (Logic and Heuristics Applied to Synthetic Analysis) rules within the context of molecular complexity indicators (twc and N_T)
- Used a set of polycyclic compounds to determine most strategic disconnections and whether these adhered to the LHASA rules
- Found most of the rules to be validated by the mathematical principles of complexity, and therefore determined synthesis was both an art and a science

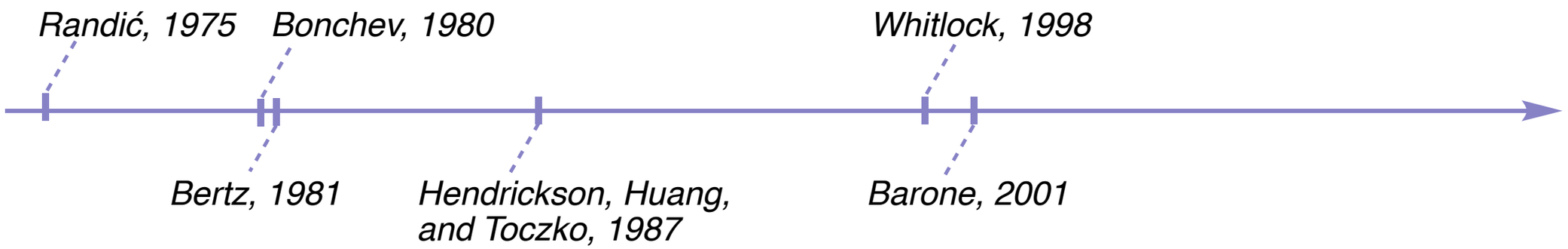
J. Chem. Inf. Comput. Sci. 2004, 44, 2, 378–386

Molecular Complexity: A History



J. Chem. Inf. Model. 2016, 56, 3, 462–470

Molecular Complexity: A History



References

Whitlock: On the Structure of Total Synthesis of Complex Natural Products

J. Org. Chem. **1998,** *63,* 7982−7989

On the Structure of Total Synthesis of Complex Natural Products

H. W. Whitlock

Department of Chemistry, University of Wisconsin, Madison, Wisconsin 53706

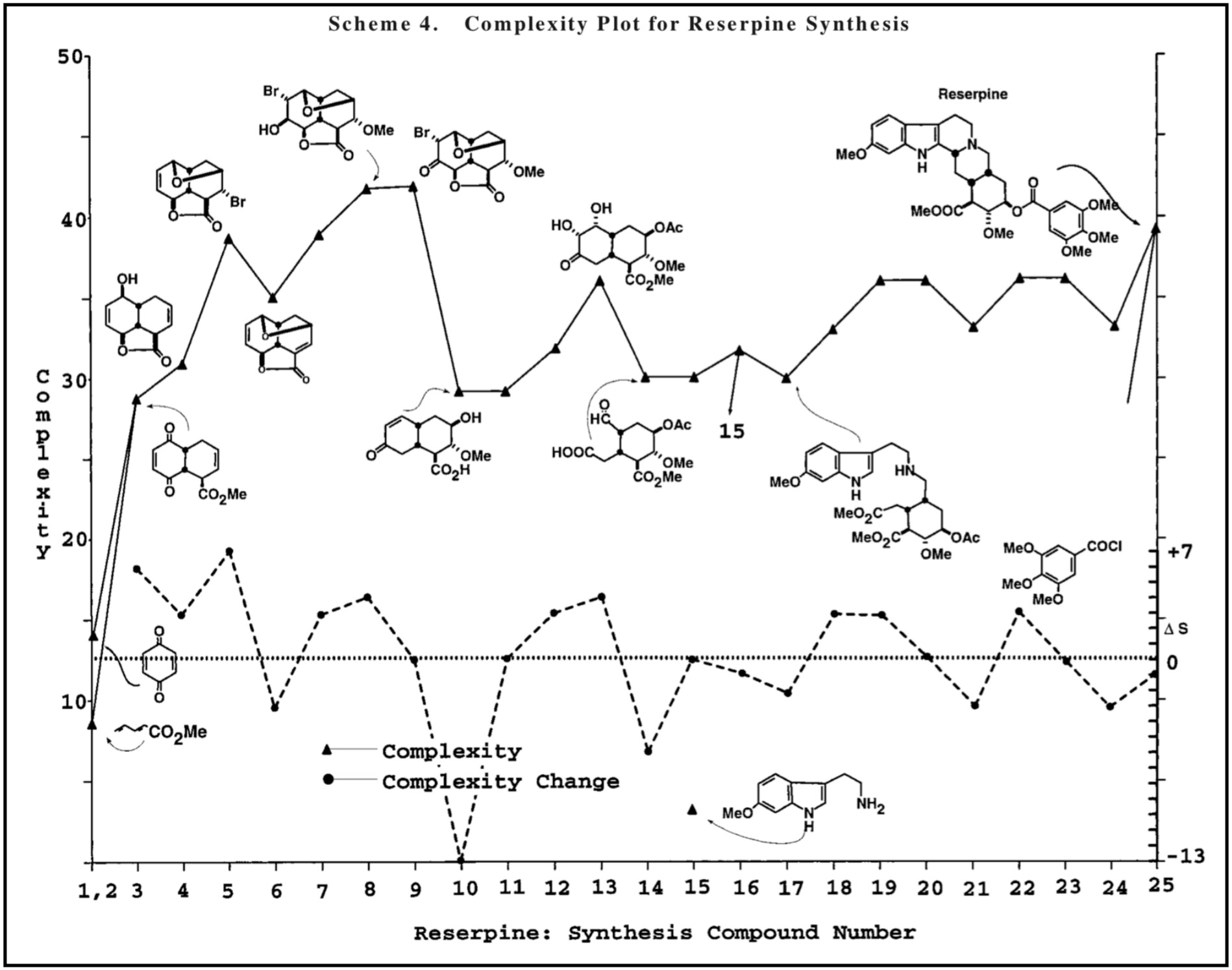
Received July 23, 1998

A chemically intuitive measure (metric) for molecular complexity is described. It is then applied to several syntheses of complex natural products. We conclude that the metric corresponds well with chemical ideas of complexity, and that different syntheses do in fact have appreciably different logical structures as measured by this. Possible implications of this interesting observation are discussed.

- Attempted to quantify chemically intuitive measures for molecular complexity, largely based on the Woodward synthesis of Reserpine
- Divided into two metrics:
  - Size metric, *H*, defined as the number of bonds in a molecule
  - Complexity metric, *S*, defined as follows: *S = 4(# of rings) + 2(# of unsaturations) + (# of heteroatoms) + 2(# of chiral centers)*

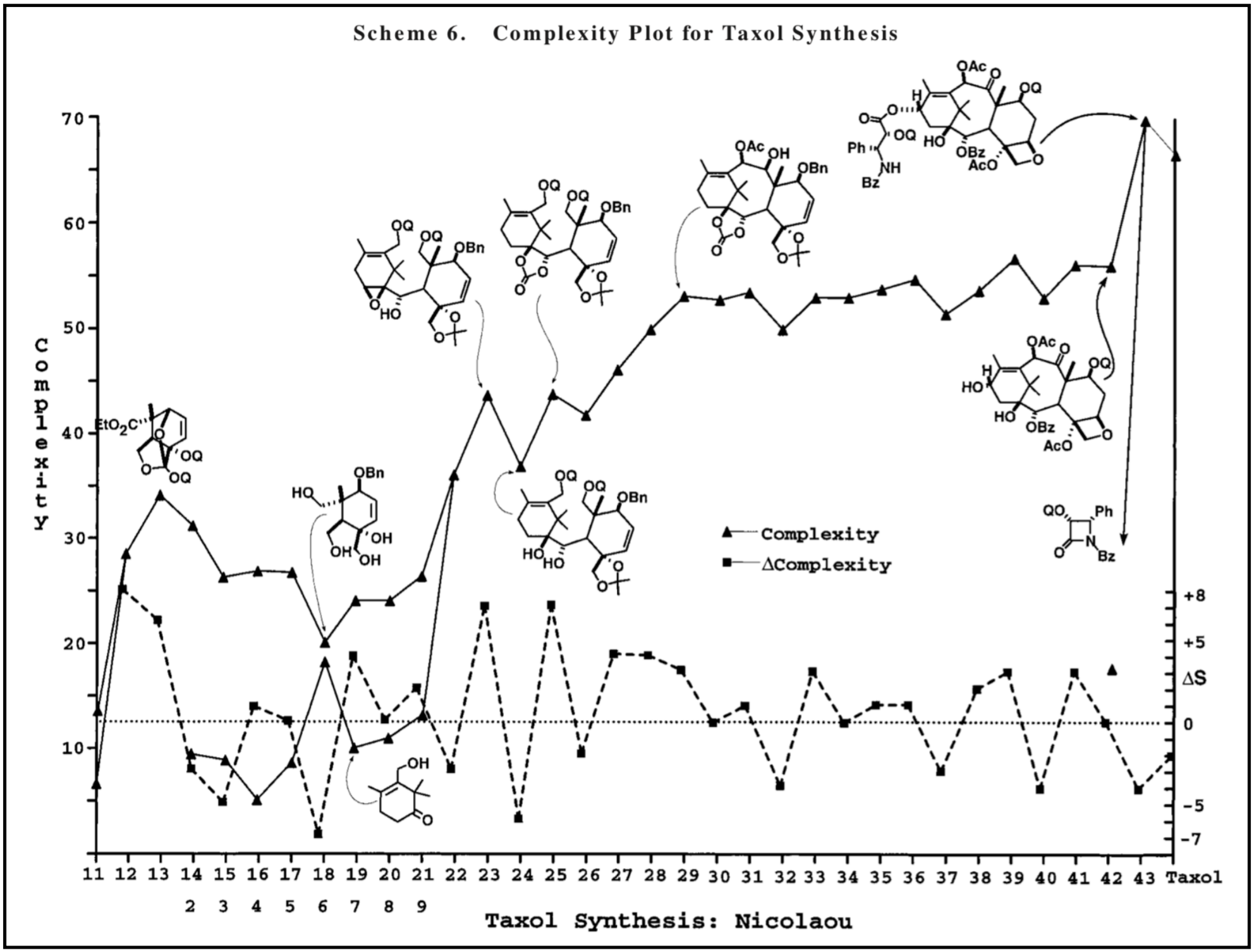
J. Org. Chem. 1998, 63, 22, 7982–7989

Whitlock: On the Structure of Total Synthesis of Complex Natural Products



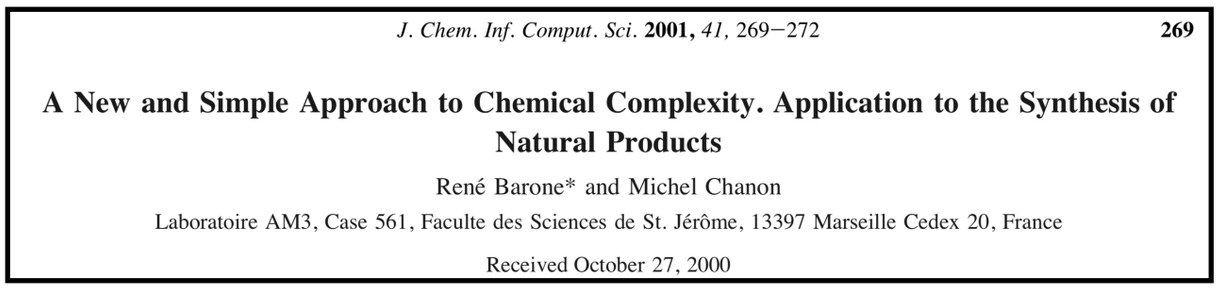
J. Org. Chem. 1998, 63, 22, 7982–7989

Whitlock: On the Structure of Total Synthesis of Complex Natural Products



J. Org. Chem. 1998, 63, 22, 7982–7989

Borone: A New and Simple Approach to Chemical Complexity



J. Chem. Inf. Comput. Sci. **2001,** *41,* 269−272 **269**

A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products

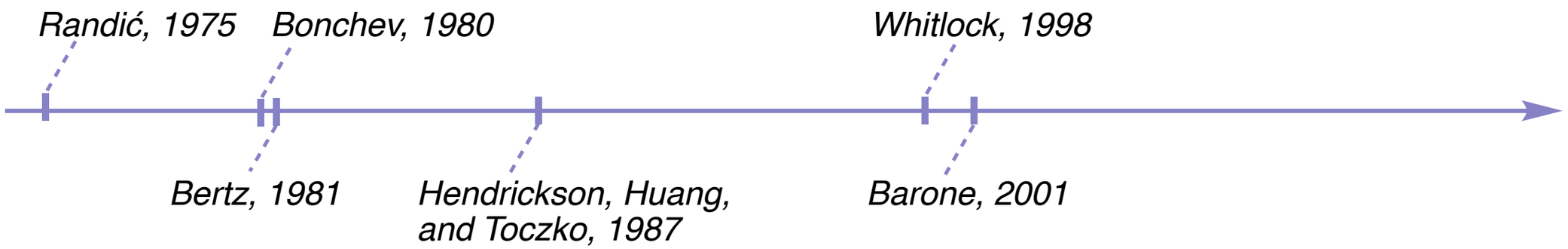
René Barone* and Michel Chanon

Laboratoire AM3, Case 561, Faculte des Sciences de St. Jérôme, 13397 Marseille Cedex 20, France

Received October 27, 2000

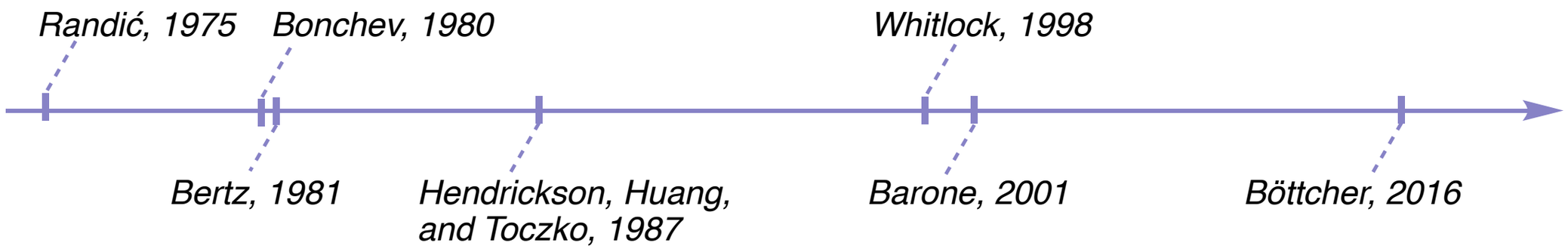
- Developed index to correct for the following critiques of the Whitlock index: substituents not considered, all rings weighted equally
- Index developed weights rings by the ring size and weights substituents by assigning carbons higher values for fewer attached protons

Molecular Complexity: A History



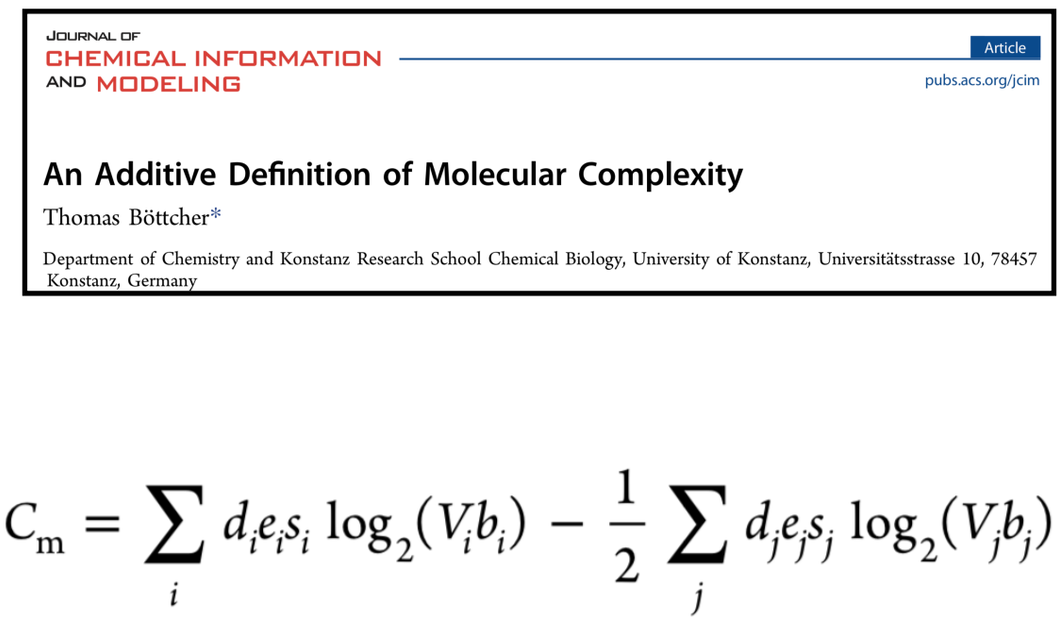
J. Chem. Inf. Model. 2016, 56, 3, 462–470

Molecular Complexity: A History



J. Chem. Inf. Model. 2016, 56, 3, 462–470

Böttcher: An Additive Definition of Molecular Complexity



JOURNAL OF CHEMICAL INFORMATION AND MODELING — Article — pubs.acs.org/jcim

An Additive Definition of Molecular Complexity

Thomas Böttcher*

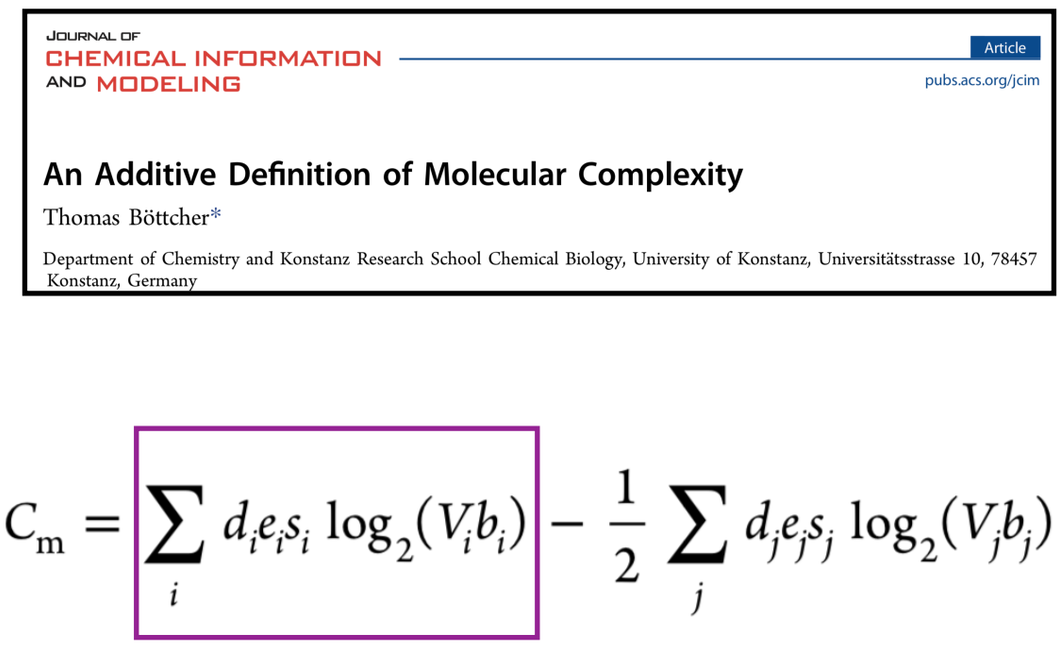
Department of Chemistry and Konstanz Research School Chemical Biology, University of Konstanz, Universitätsstrasse 10, 78457 Konstanz, Germany

$$C_m = \sum_i d_i e_i s_i \log_2(V_i b_i) - \frac{1}{2} \sum_j d_j e_j s_j \log_2(V_j b_j)$$

- Inspired by critiques on graph theory approaches and more intuitive approaches
  - Graph theory approaches: don't adequately address chirality
  - Other approaches: don't adequately address skeletal complexity
- Atom-based approach in order to achieve an additive definition of molecular complexity, based on Shannon entropy
- Uses five variables: valency of each atom, number of bonds, number of chemically distinct bonds, element diversity, and stereochemistry

J. Chem. Inf. Model. 2016, 56, 3, 462–470

Böttcher: An Additive Definition of Molecular Complexity



JOURNAL OF CHEMICAL INFORMATION AND MODELING — Article — pubs.acs.org/jcim

An Additive Definition of Molecular Complexity

Thomas Böttcher*

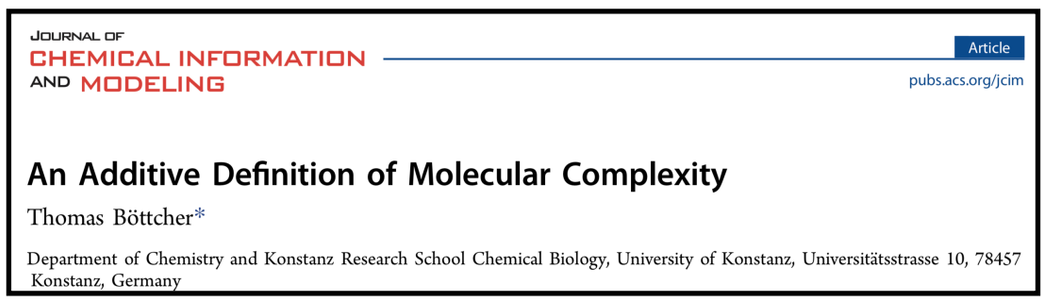
Department of Chemistry and Konstanz Research School Chemical Biology, University of Konstanz, Universitätsstrasse 10, 78457 Konstanz, Germany

$$C_m = \sum_i d_i e_i s_i \log_2(V_i b_i) - \frac{1}{2} \sum_j d_j e_j s_j \log_2(V_j b_j)$$

- Inspired by critiques on graph theory approaches and more intuitive approaches
  - Graph theory approaches: don't adequately address chirality
  - Other approaches: don't adequately address skeletal complexity
- Atom-based approach in order to achieve an additive definition of molecular complexity, based on Shannon entropy
- Uses five variables: valency of each atom, number of bonds, number of chemically distinct bonds, element diversity, and stereochemistry

J. Chem. Inf. Model. 2016, 56, 3, 462–470

Böttcher: An Additive Definition of Molecular Complexity

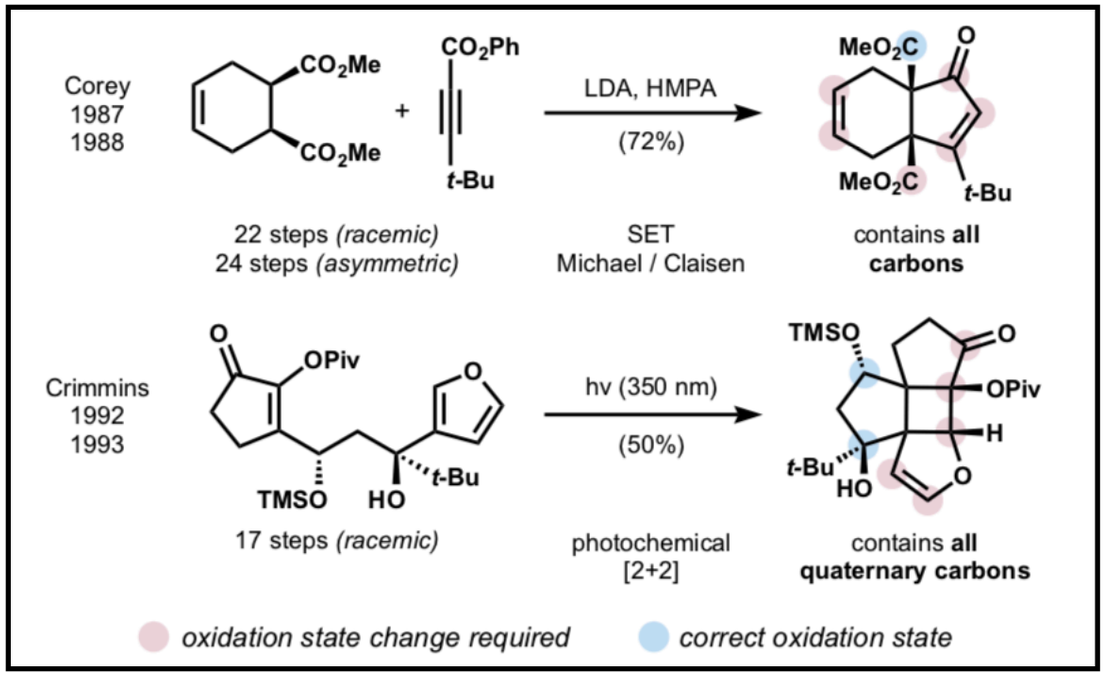


$$C_m = \sum_i d_i e_i s_i \log_2(V_i b_i) - \frac{1}{2} \sum_j d_j e_j s_j \log_2(V_j b_j)$$

- Inspired by critiques on graph theory approaches and more intuitive approaches
  - Graph theory approaches: don't adequately address chirality
  - Other approaches: don't adequately address skeletal complexity
- Atom-based approach in order to achieve an additive definition of molecular complexity, based on Shannon entropy
- Uses five variables: valency of each atom, number of bonds, number of chemically distinct bonds, element diversity, and stereochemistry
- Units of 'mcbits' are given for molecular information

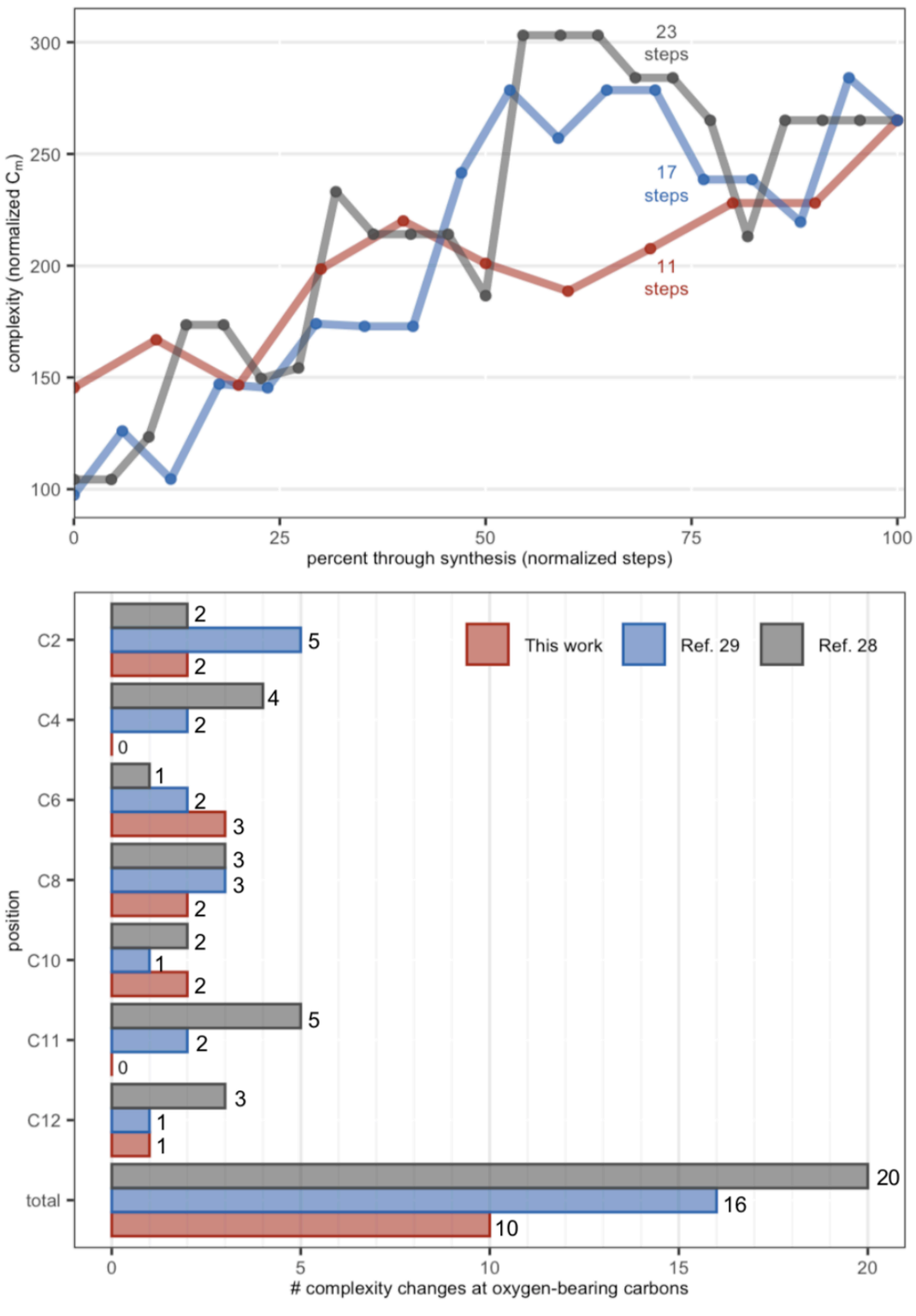
J. Chem. Inf. Model. 2016, 56, 3, 462–470

Bilobalide: A Case Study



- Shenvi and coworkers recognize the benefit of open-shell intermediates to achieving the sterically congested core, but disadvantage of further redox manipulations to efficiency
- Sought to incorporate correct oxidation state of each carbon in starting materials and use open-shell intermediates to form hindered C-C bonds
- Result is a "generally positive increase in average information content" (based on Böttcher index)

Bilobalide: A Case Study



- Percentage of steps which decrease complexity at key carbons is fairly consistent
  - *Corey - 27%*
  - *Crimmins - 35%*
  - *Shenvi - 25%*
- Noted a correlation between step count and complexity changes at heteroatom-bearing carbons
  - *Corey - 20 of 24*
  - *Crimmins - 16 of 17*
  - *Shenvi - 10 of 12*
- "Ideally, a carbon atom is introduced in the synthesis at its final value (oxidation state, hybridization, stereochemistry) or undergoes minimal transformation to reach the target score."

ChemRxiv, 2020, 10.26434/chemrxiv.12132939.v2