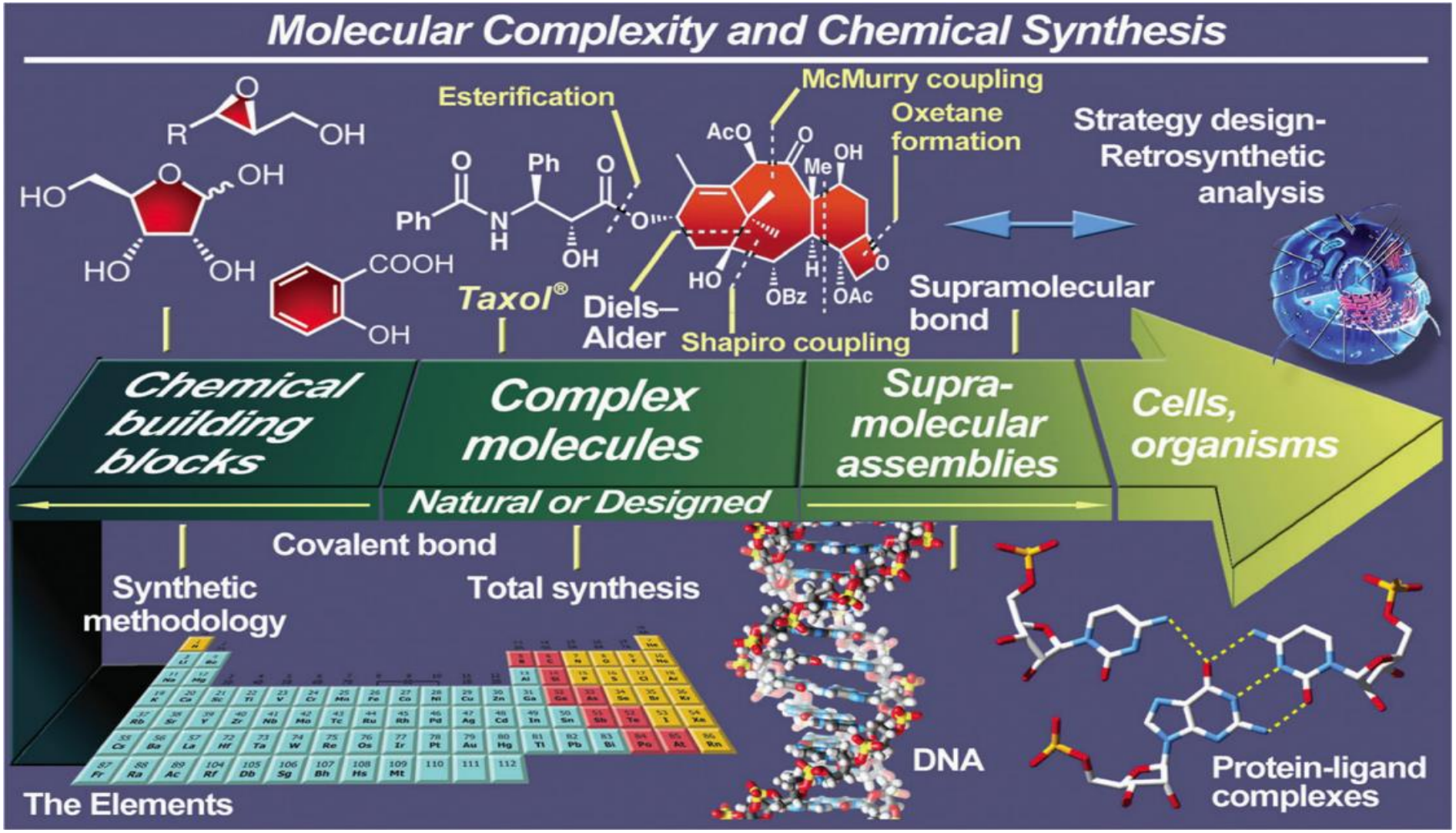
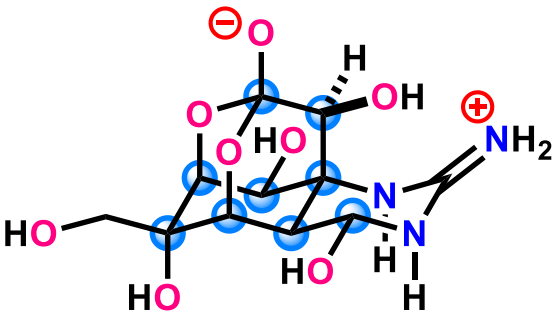
Literature Seminar
2020/11/21 Masanori Nagatomo
Molecular Complexity

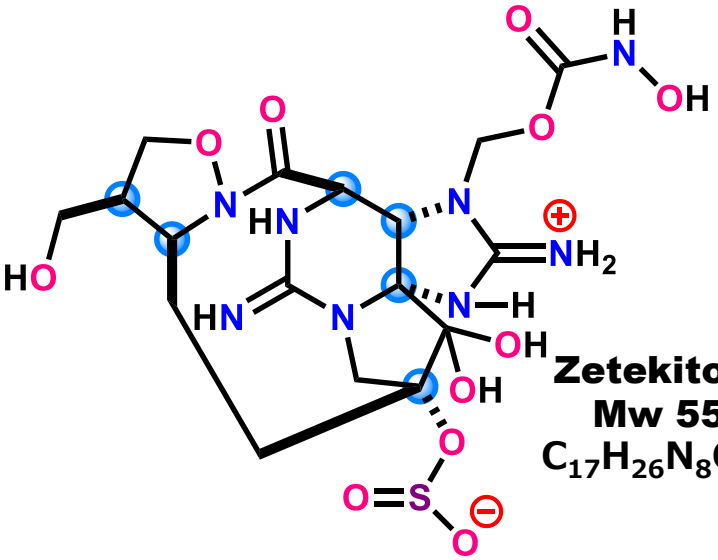


Nicolaou, K. C.; Hale, C. R. H.; Nilewski, C.; Ioannidou, H. A. *Chem. Soc. Rev*. **2012**, *41*, 5185−5238.

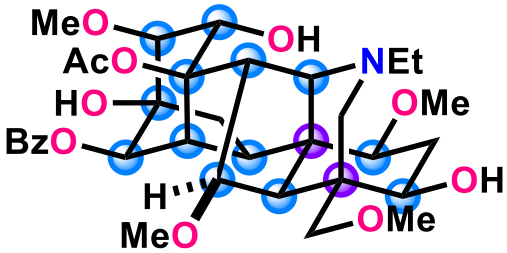
What Is Complexity?



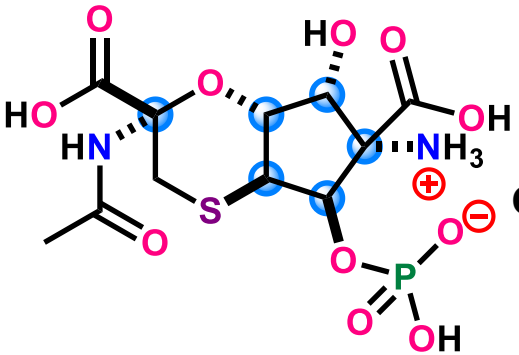
Tetrodotoxin
Mw 319
$C_{11}H_{17}N_3O_8$



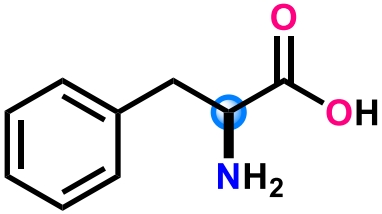
Zetekitoxin
Mw 550
$C_{17}H_{26}N_8O_{11}S$



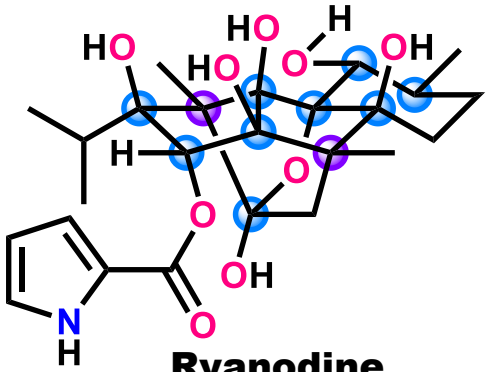
Aconitine
Mw 645
$C_{34}H_{47}NO_{11}$



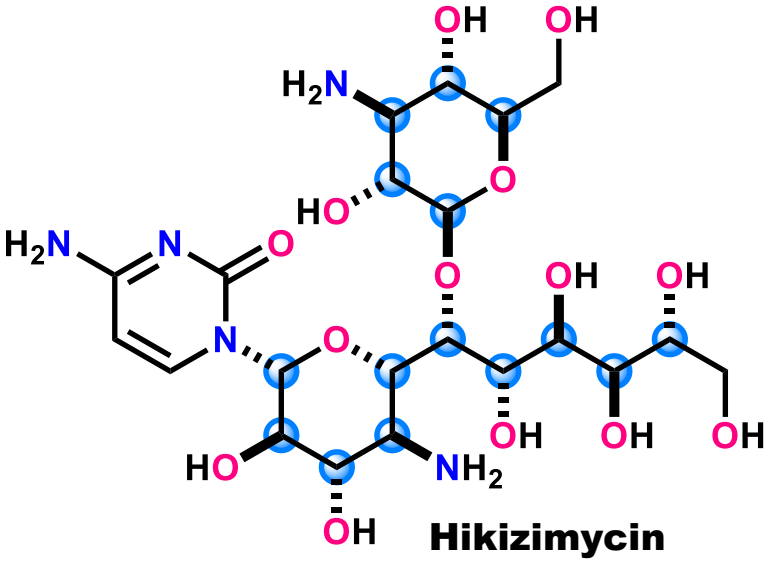
Tagetitoxin
Mw 416
$C_{11}H_{17}N_2O_{11}PS$



***L*-Phenylalanine**
Mw 116
$C_9H_{11}NO_2$



Ryanodine
Mw 493
$C_{25}H_{35}NO_9$



Hikizimycin
Mw 583
$C_{21}H_{37}N_5O_{14}$

How Do We Chemist Define Molecular Complexity?

Complexity

Intrinsic factor (fixed)

objective criteria

determined only by the chemical structure

many attempts for a combination of graph theory and information theory to determine the molecular topology

Böttcher Score

Extrinsic factor (variable)

subjective perception

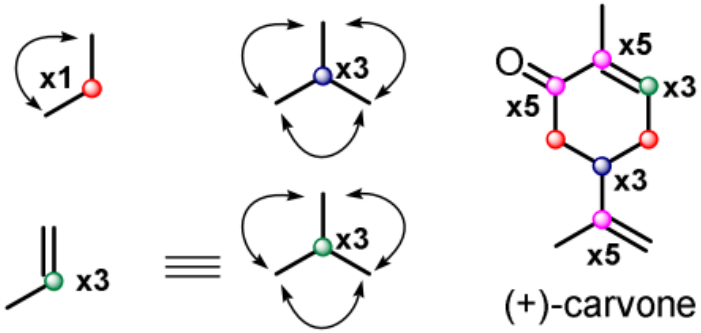
dependent on external conditions, such as temporal factor, including knowledge, experience, and synthetic technology

Peer-based assessment of complexity by letting experienced chemists rank the complexities of compounds has been used to assess the synthetic complexity, accessibility, and drug-likeness of different chemical compounds

Main Paper

Bertz-(Hendrickson) Index

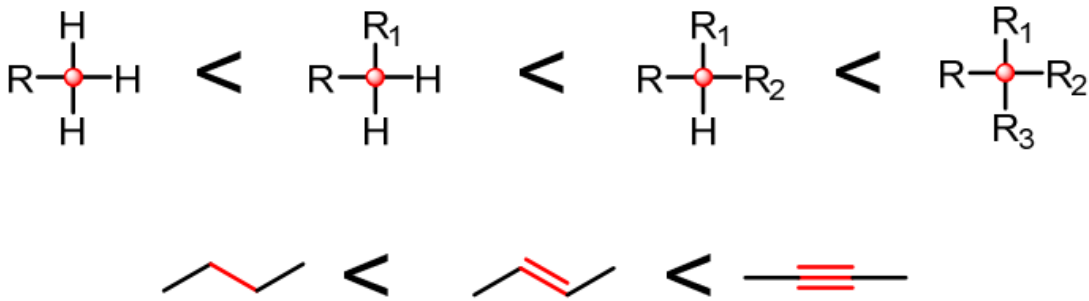
A Bertz's principle



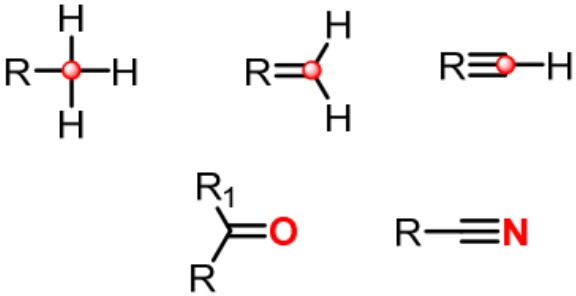
$C_T = C\eta + C\varepsilon$

$C\eta$: bond connectivities - skeletal complexity
$C\varepsilon$: diversity of elements - kinds of atoms

B complexity gain

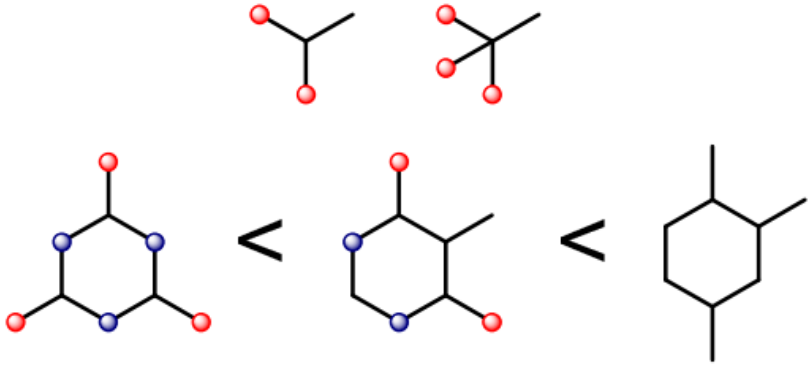


C Zero complexity



terminal atoms does not increment the complexity calculation

D decrease in complexity (symmetry)



1. Bertz, S. H. *J. Am. Chem. Soc.* **1981**, *103*, 3599-3601.
2. Siguara Bastos de Lemos E Silva. Chemistry and biosynthesis of highly complex marine alkaloids from Mediterranean biodiversity. Organic chemistry. Université Paris Saclay (COmUE), 2017.

Applications and Shortcomings

C_T of many chemical compounds is publically available on PubChem.[1)] Molecular complexity also has important implications for organic synthesis planning, in-silico drug design, and pharmaceutical development, including QSPR (Quantitative Structure-Property Relationship) and QSAR (Quantitative Structure-Affinity Relationship) approaches.

Frequently criticized shortcomings are the failure to address chirality in graph-theoretical approaches and missing sensitivity to skeletal structure, branching, and symmetry of other indices.

Proposing a framework for molecular complexity that relies on both mathematical rigor and chemically consistent inherent logic.

Böttcher Score

1. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. Pubchem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

Thomas Böttcher

BIOSPHERE COMPLEXITY:
https://www.youtube.com/watch?v=xFNWVSEuxxc

2009: Ph.D @ LMU, Munich (Prof. Stephan A. Sieber)
2010: Postdoctoral research@ TMU, Munich (Prof. Stephan A. Sieber)
2011-2014: Postdoctoral research@ Harvard Medical School, Department of Biological Chemistry & Molecular Pharmacology (Prof. Jon Clardy)
2014-2020: Independent group leader @ University of Konstanz
2020: An Emmy Noether research group leader @ University of Konstanz, Professor of Microbial Biochemistry, Faculty of Chemistry, Department of Biological Chemistry and Department of Microbiology and Ecosystem Science (DOME), University of Vienna
Research interest: the isolation and identification of natural products that modify and manipulate bacterial population behavior. Aim to inhibit bacterial virulence and discover the chemistry of ecological interactions of microorganisms.

Prof. Stephan A. Sieber

Prof. Jon Clardy

Böttcher Score

The information content is defined by the entropy *H* (Unit is bit).

$$H = -\sum_i p_i \log_x p_i$$

One variable is needed to identify the nature of the element by its valence shell, and four variables are required as descriptors of the bonding environment: the number of bonds, the number of chemically different bonds, the element diversity, and the stereochemistry.

C_m: Molecular Complexity (Unit is mcbit), **Vi**: valence electrons, **Bi**: total number of bonds, **di**: introduced to characterize the number of chemically nonequivalent bonds to atoms with Vibi > 1 at the ith position, **ei**: giving the number of different non-hydrogen elements or isotopes involved in the bonding situation, including atom i and its direct neighbors, to include heteroatoms.

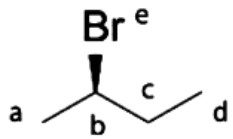
$$C_m = \sum_i d_i e_i s_i \log_2(V_i b_i)$$

To account for symmetries of a molecule, the corresponding atom positions of chemically equivalent sets of atoms for each symmetric position j are subtracted.

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

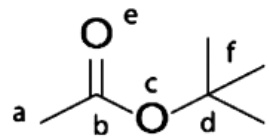
1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

Examples: How to Calculate C_m



i	a	b	c	d	e
d_i	1	3	2	1	1
e_i	1	2	1	1	2
s_i	1	2	1	1	1
V_i	4	4	4	4	7
b_i	1	3	2	1	1

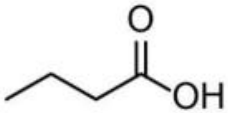
$C_m = \log(4) + 3\cdot2\cdot2\cdot\log(4\cdot3) + 2\cdot\log(4\cdot2) + \log(4) + 2\cdot\log(7) = 58.6$ mcbit



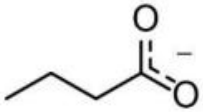
i	a	b	c	d	e	f
d_i	1	3	2	2	1	1
e_i	1	2	2	2	2	1
s_i	1	1	1	1	1	1
V_i	4	4	6	4	6	4
b_i	1	4	2	4	2	1

$C_m = \log(4) + 3\cdot2\cdot\log(4\cdot4) + 2\cdot2\cdot\log(6\cdot2) + 2\cdot2\cdot\log(4\cdot4) + 2\cdot\log(6\cdot2) + 3\cdot\log(4) - 0.5\cdot3\cdot\log(4) = 66.5$ mcbit

Examples illustrating how to calculate Cm on a per-atom basis for a molecule. Different atom positions i are labeled by letters (a−f), and for tert-butyl acetate the symmetry correction for the positions f is included. In the equations, "log" stands for $\log_2$.



$C_m = \log(4) + 2\cdot\log(4\cdot2) + 2\cdot\log(4\cdot2) + 3\cdot2\cdot\log(4\cdot4) + 2\cdot\log(6\cdot2) + 2\cdot\log(6) = 50.34$ mcbit

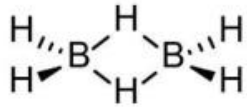


$C_m = \log(4) + 2\cdot\log(4\cdot2) + 2\cdot\log(4\cdot2) + 2\cdot2\cdot\log(4\cdot4) + 2\cdot[2\cdot\log(6\cdot1.5)] - 0.5\cdot2\cdot[2\cdot\log(6\cdot1.5)] = 36.34$ mcbit

$H_3Si—SiH_3$

$C_m = 2\cdot[\log(4)] - 0.5\cdot2\cdot[\log(4)] = \log(4) = 2.00$ mcbit

Special bond situations and non-carbon based compounds. Log stands for $\log_2$.



$C_m = 2\cdot[\log(3\cdot2) + \log(1\cdot2)] - 0.5\cdot2\cdot[\log(3\cdot2) + \log(1\cdot2)] = 3.58$ mcbit

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

Examples: Influence of Isotopes on C_m

A

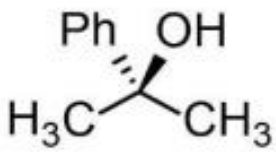
symmetric

$$C_m = 2\cdot\log(4) + 2\cdot2\cdot\log(6\cdot2) + 2\cdot2\cdot\log(4\cdot2) + \log(4\cdot2) = 33.34 \text{ mcbit}$$

asymmetric

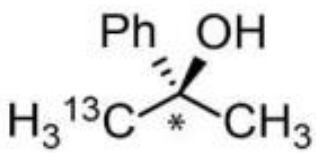
$$C_m = 2\cdot\log(4) + 2\cdot2\cdot\log(6\cdot2) + 2\cdot2\cdot\log(4\cdot2) + 2\cdot\log(4\cdot2) + 2\cdot2\cdot\log(4\cdot2) + 2\cdot3\cdot\log(6\cdot2) + 2\cdot\log(4) = 73.85 \text{ mcbit}$$

B



symmetric achiral

$$C_m = \log(4\cdot3) + 2\cdot\log(4\cdot3) + 2\cdot\log(4\cdot3) + 2\cdot\log(4\cdot4) + 3\cdot2\cdot\log(4\cdot4) + 2\cdot\log(6\cdot2) + \log(4) = 57.09 \text{ mcbit}$$



asymmetric chiral

$$C_m = \log(4\cdot3) + 2\cdot\log(4\cdot3) + 2\cdot\log(4\cdot3) + 2\cdot\log(4\cdot4) + 4\cdot3\cdot2\cdot\log(4\cdot4) + 2\cdot\log(6\cdot2) + \log(4) + \log(4) = 131.09 \text{ mcbit}$$

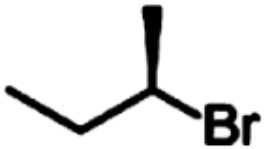
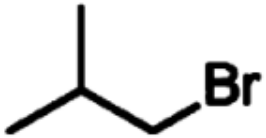
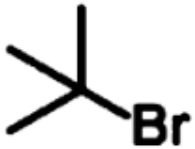
Influence of isotopes on molecular complexity.

A) The 13C isotope of the methyl group disrupts the compound's symmetry, manifesting in changes in di for the central CH_2 group and the omission of symmetry correction. Additionally, the element diversity term ei at one of the oxygens increases with the 13C isotope.

B) In addition to symmetry breaking at the quaternary carbon, stereochemical information (isotope chirality) is introduced, altering si and di and eliminating the symmetry correction term. Changes are highlighted in red. An asterisk labels the stereocenter. Log stands for $\log_2$.

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

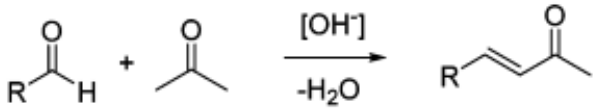
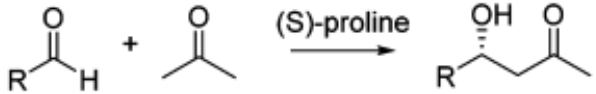
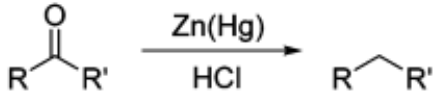
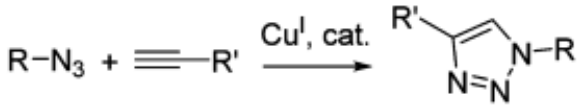
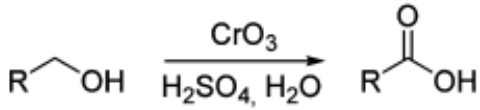
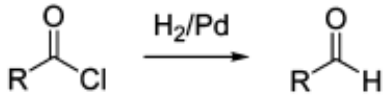
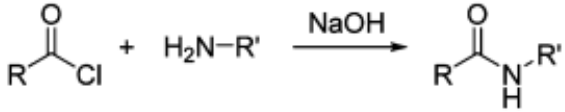
Comparison of C_m with Other Complexity Indices

				
non-equivalent protons (NMR)	5	4	3	1
C_m (mcbit)	58.6	31.6	26.8	24.6
N_S	10	9	9	9
N_T	17	15	9	9
$C(\eta,\varepsilon)$	19.6	13.1	17.6	25.1
S (Whitlock)	3	1	1	1
Barone	45	42	45	54

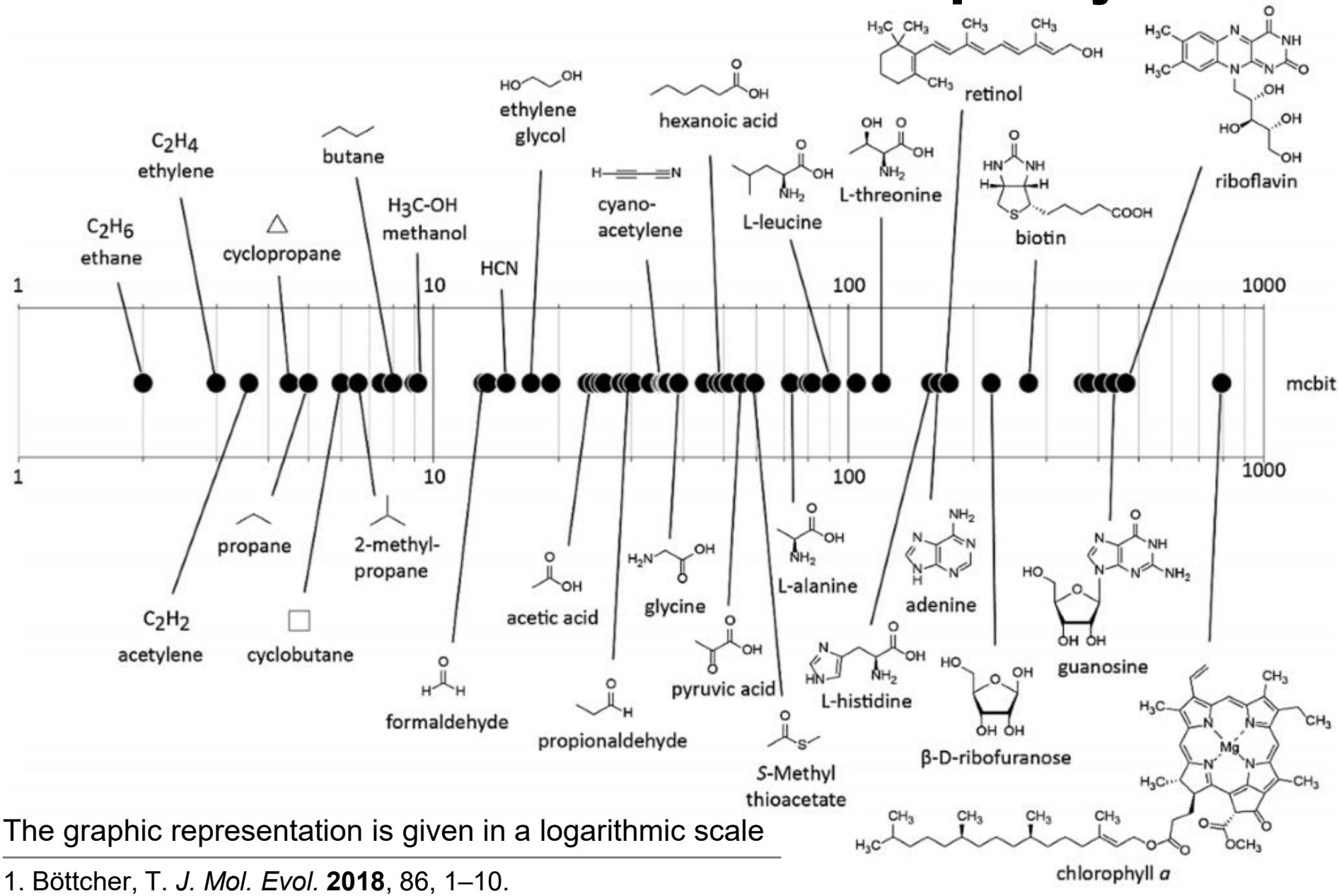
Molecular complexity of bromobutane isomers. The order of the isomers is given from the most complex (left) to the least complex (right) according to the spectral complexity of the 1H NMR spectra as a proxy of molecular complexity. The numbers of nonequivalent protons are in line with the relative complexities given by distinct signals and multiplicities of the corresponding 1H NMR spectra. Complexity values are given for C_m and other indices.

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

Changes in Molecular Momplexity ΔC_m

Name	Reaction	ΔCm
Aldol condensation		8.00
Aldol reaction		40.68
Clemmensen reduction*		-25.17
1,3-dipolar cycloaddition		47.80
Jones oxidation		5.17
Rosenmund reduction		-27.27
Schotten-Baumann reaction		3.03

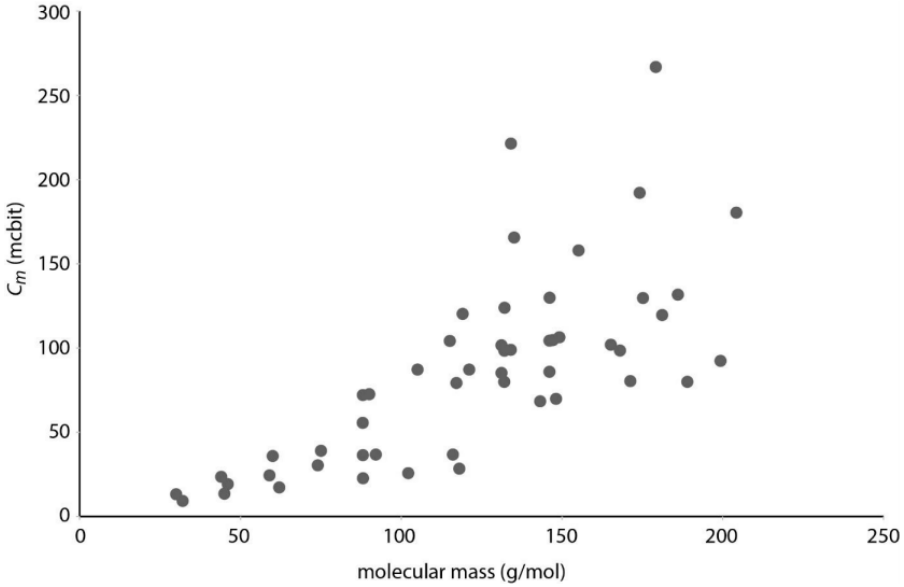
C_m with Calculated for the Structures of Various Small Molecules in a Universal Complexity Scale



The graphic representation is given in a logarithmic scale

1. Böttcher, T. *J. Mol. Evol.* **2018**, 86, 1–10.

Molecular Complexity Per Molecular Mass Unit



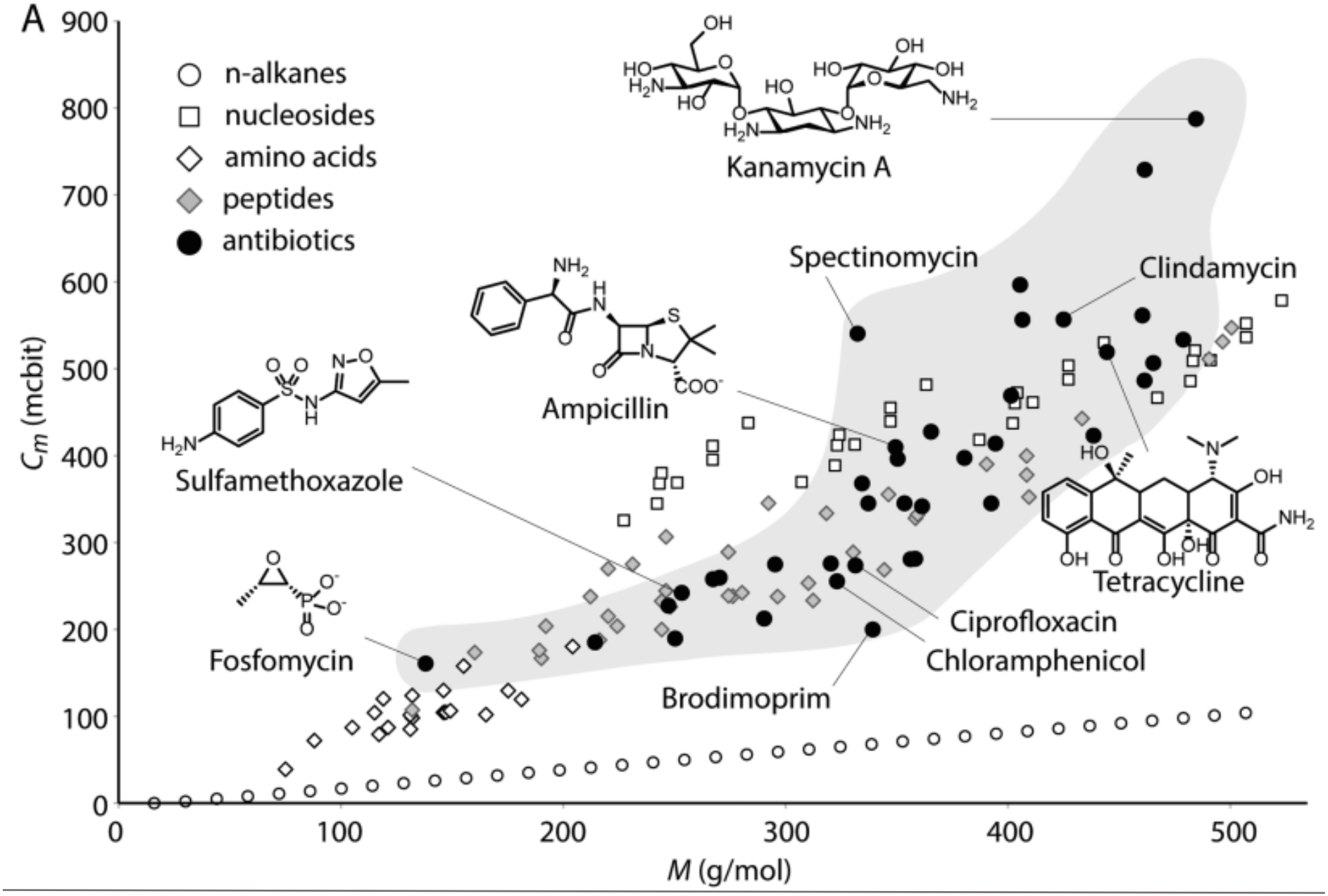
For 51 common metabolites ranging from acetate via amino acids and ribose to phosphoenolpyruvate, the molecular complexity C_m was calculated and plotted against the molecular mass of the compounds, demonstrating that molecular complexity is not directly dependent on molecular mass.

In fact, even small molecules with a comparable mass may differ by 100-200 mcbit.

Compound	C_m/M (mcbit•mol/g)	C_m (mcbit)
amino acids		
Alanine	0.817	72.00
Arginine	0.740	129.65
Asparagine	0.937	123.82
Aspartate	0.744	98.34
Cysteine	0.720	87.17
Glutamine	0.888	129.82
Glutamate	0.714	104.34
Glycine	0.519	38.98
Histidine	1.017	157.83
Isoleucine	0.774	101.51
Leucine	0.692	90.76
Lysine	0.711	104.65
Methionine	0.713	106.34
Phenylalanine	0.617	101.93
Proline	0.905	104.16
Serine	0.829	87.17
Threonine	1.009	120.19
Tryptophane	0.883	180.31
Tyrosine	0.660	119.51
Valine	0.724	84.76
fatty acids		
Hexanoic acid	0.416	48.34
Octanoic acid	0.418	60.34
Decanoic acid	0.420	72.34
Dodecanoic acid	0.421	84.34
Sugars		
Ribose	1.648	221.10
Galactose	1.484	267.29
Glucosamine	1.489	266.76

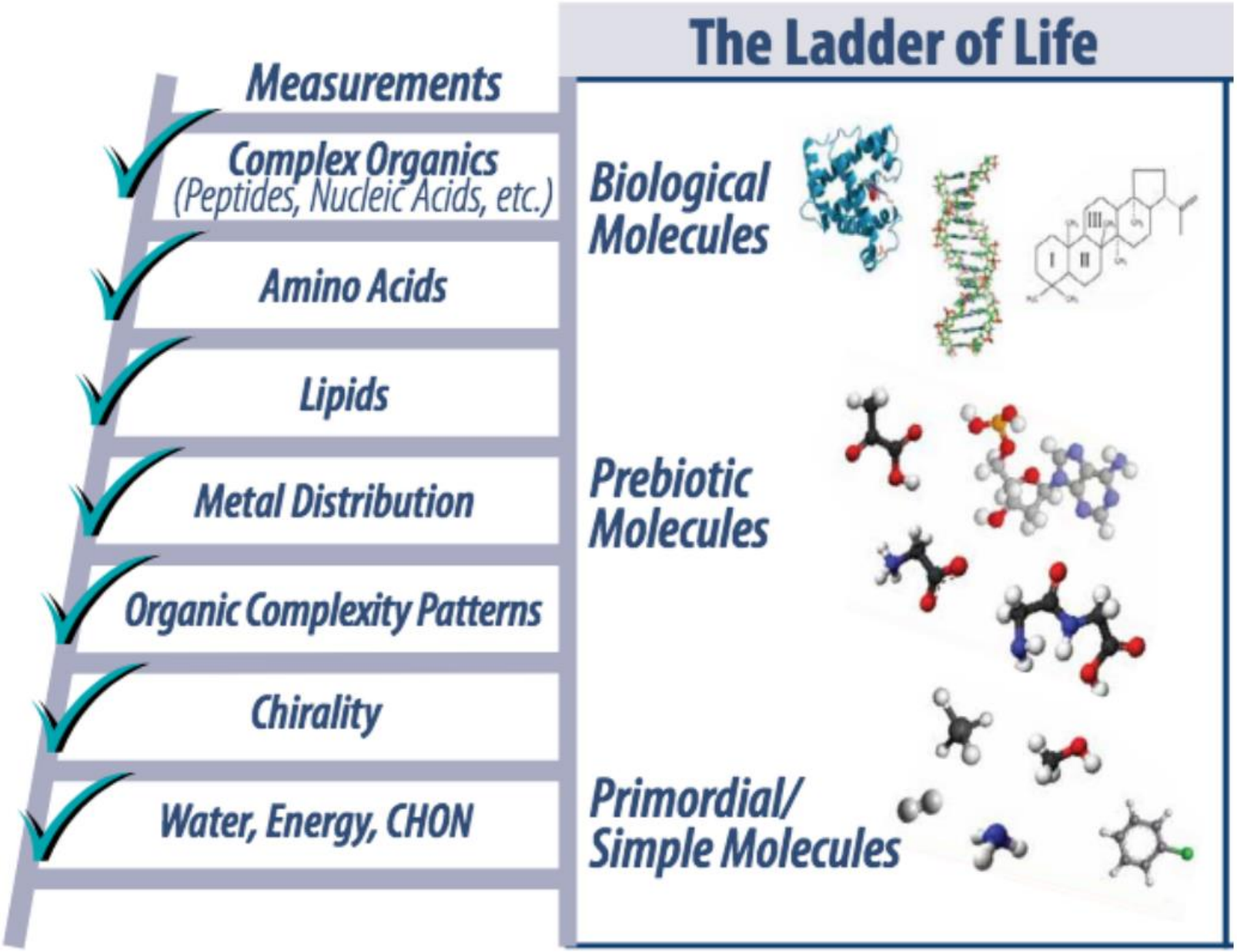
1. Böttcher, T. *J. Mol. Evol.* **2018**, 86, 1–10.

C_m Plotted against Molecular Mass



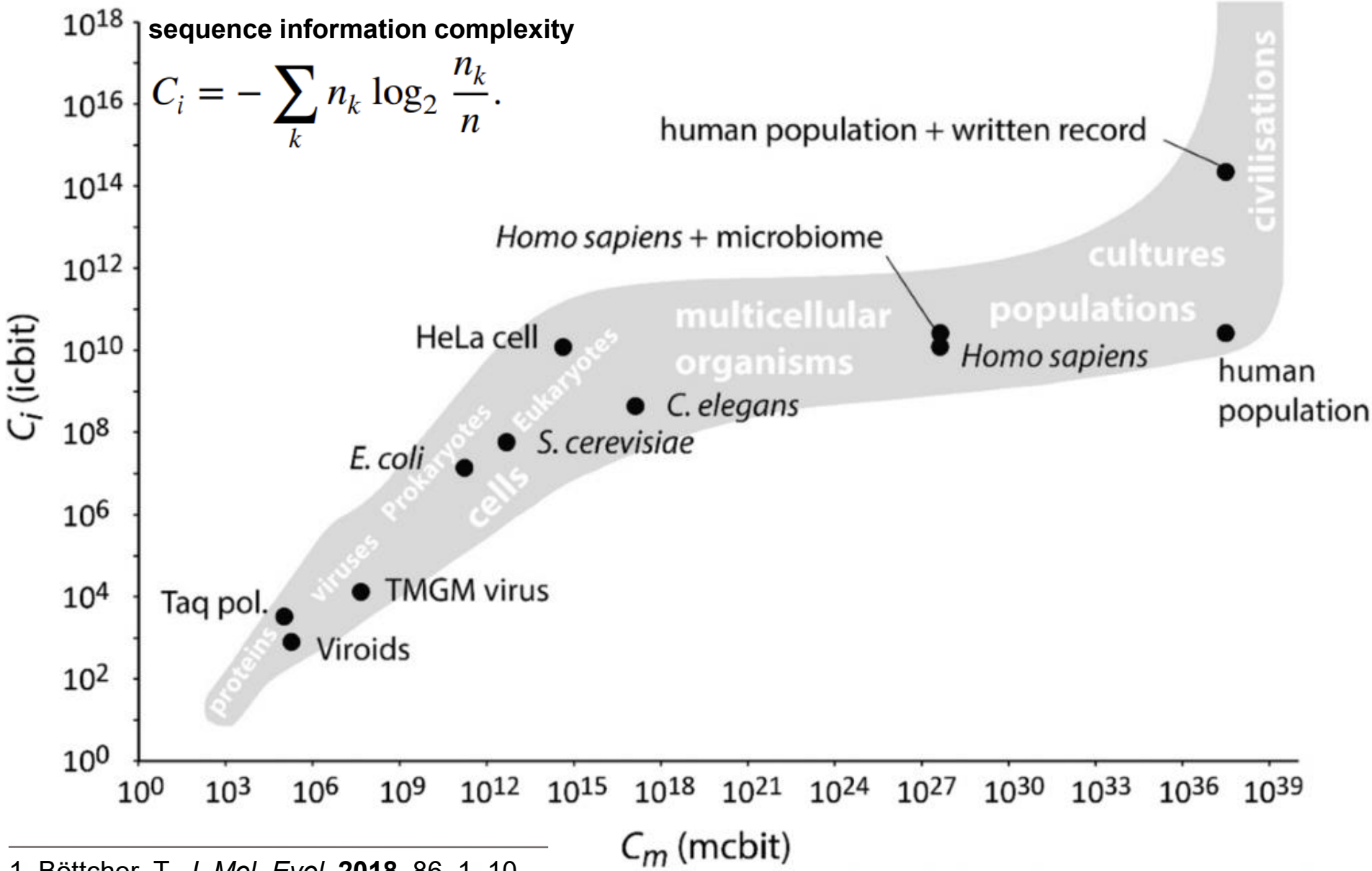
1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

Biosignature (生命存在指標)



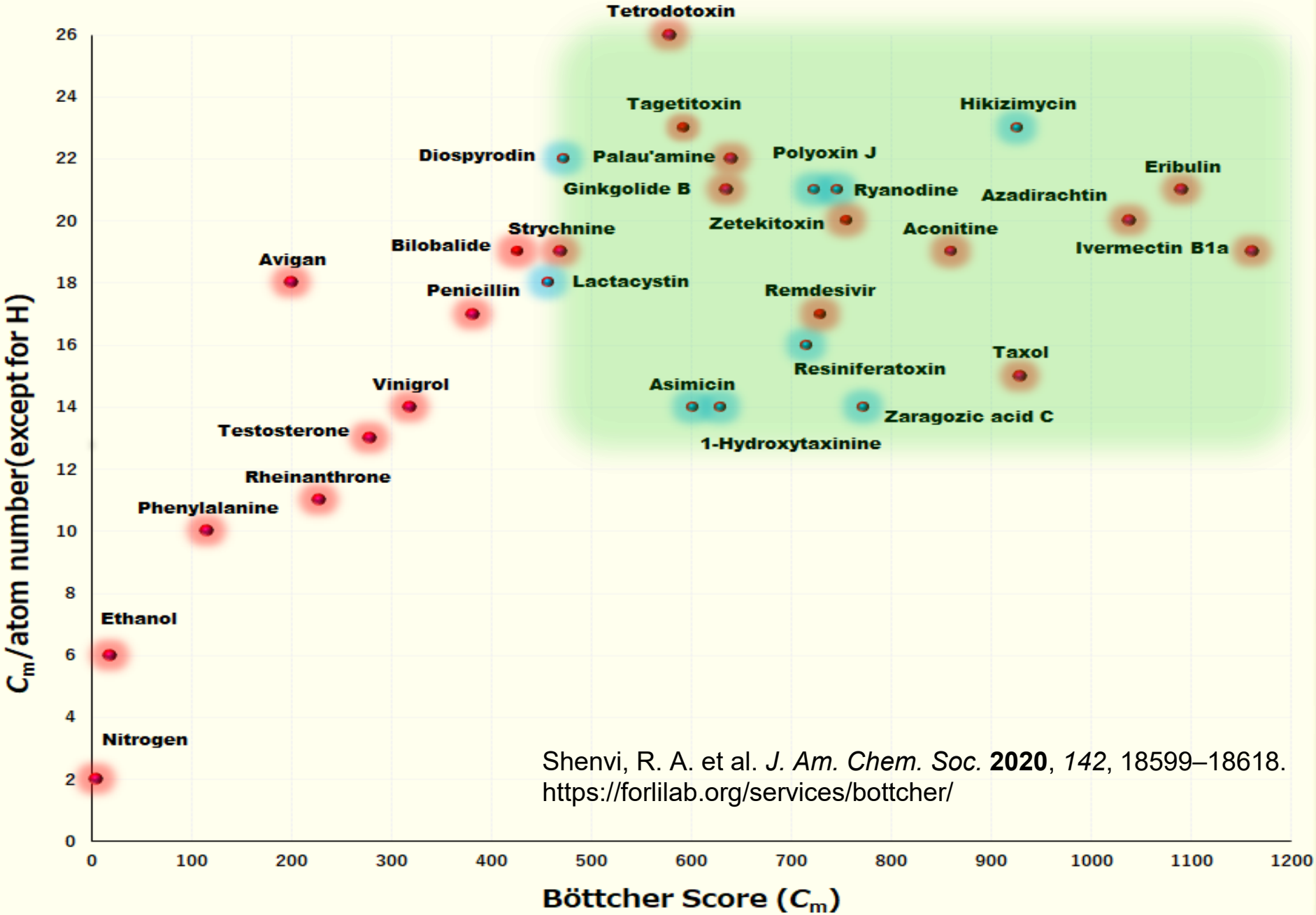
1. https://www.lpi.usra.edu/opag/meetings/sep2017/posters/Graham.pdf

Universal Complexity Scale Plot with Representative Biogenic Units of Earth's Biosphere



1. Böttcher, T. *J. Mol. Evol.* **2018**, 86, 1–10.

$C_m \propto$ Synthesizability?

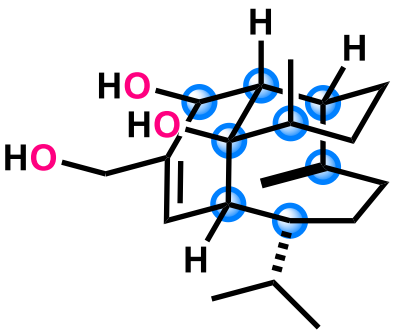


Shenvi, R. A. et al. *J. Am. Chem. Soc.* **2020**, *142*, 18599–18618.
https://forlilab.org/services/bottcher/

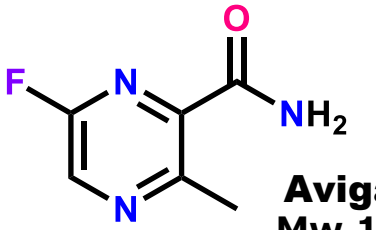
How Do We Chemist Define Molecular Complexity?

In general, the following features of a target structure increase the challenge of chemical synthesis:

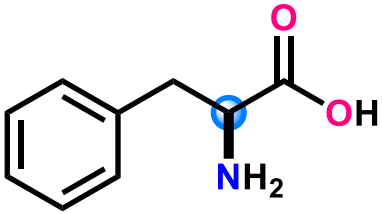
(i) the number and density of functional groups
(ii) the number of stereocenters
(iii) the number and types of rings
(iv) the overall size of the molecule.



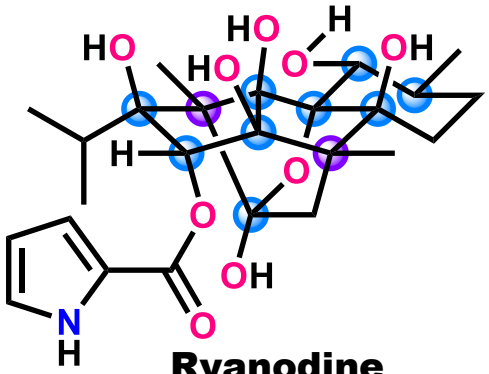
Vinigrol
Mw 322
$C_{20}H_{34}O_3$



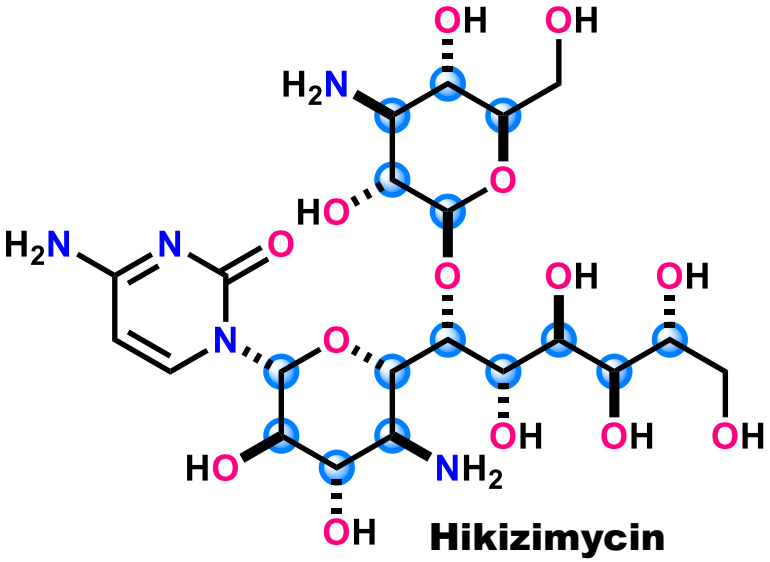
Avigan
Mw 157
$C_5H_4FN_3O_2$



***L*-Phenylalanine**
Mw 116
$C_9H_{11}NO_2$



Ryanodine
Mw 493
$C_{25}H_{35}NO_9$



Hikizimycin
Mw 583
$C_{21}H_{37}N_5O_{14}$

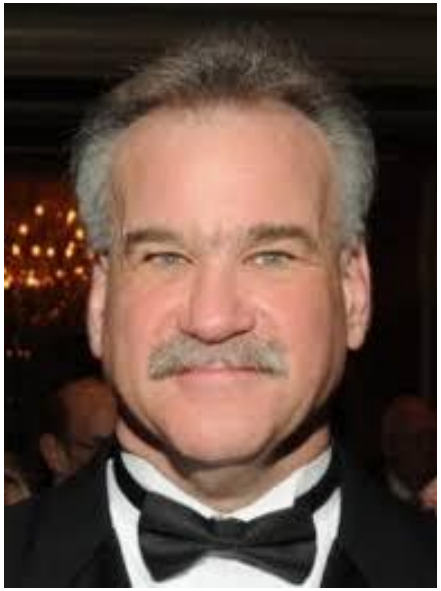
1. a) Peterson, E. A.; Overman, L. E. *Proc. Natl. Acad. Sci. USA*. **2004**, *101*, 11943-11948.
b) Urabe, D.; Asaba, T.; Inoue, M. *Chem. Rev*. **2015**, *115*, 9207–9231.

Martin D. Eastgate

1977: Born in England
1999: B.S in Chemistry @ the University of Surrey, UK
Graduating with first class honors.
2002: Ph.D. in Organic Chemistry @ the University of Cambridge, UK (Prof. Stuart Warren)
sulfur participation chemistry, specifically the generation of thiiranium ions under basic conditions and their use in pyrrolidine synthesis.
2002-2005: Post-doctoral fellow @ the University of Illinois Urbana-Champaign (Prof. Scott E. Denmark)
the Lewis-base activation of Lewis-acids and understanding ligand-field theory in hyper-valent silyl cations.
2005: Bristol-Myers Squibb
Currently: a Director in Chemical and Synthetic Development.

The late Dr. Stuart Warren

Bristol Myers Squibb™

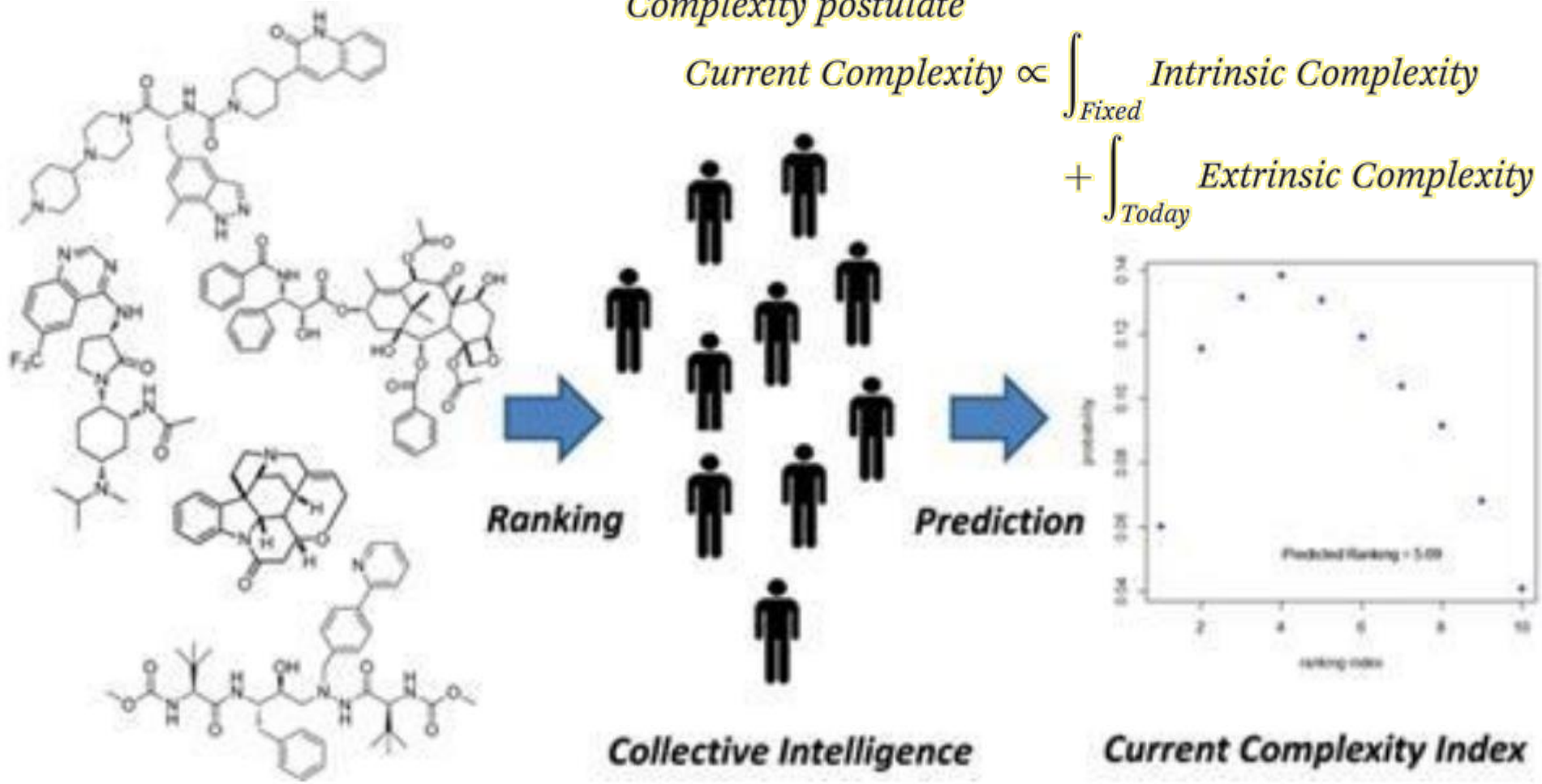


Prof. Scott E. Denmark

"Current" Complexity

Complexity postulate

$$\textit{Current Complexity} \propto \int_{Fixed} \textit{Intrinsic Complexity} + \int_{Today} \textit{Extrinsic Complexity}$$

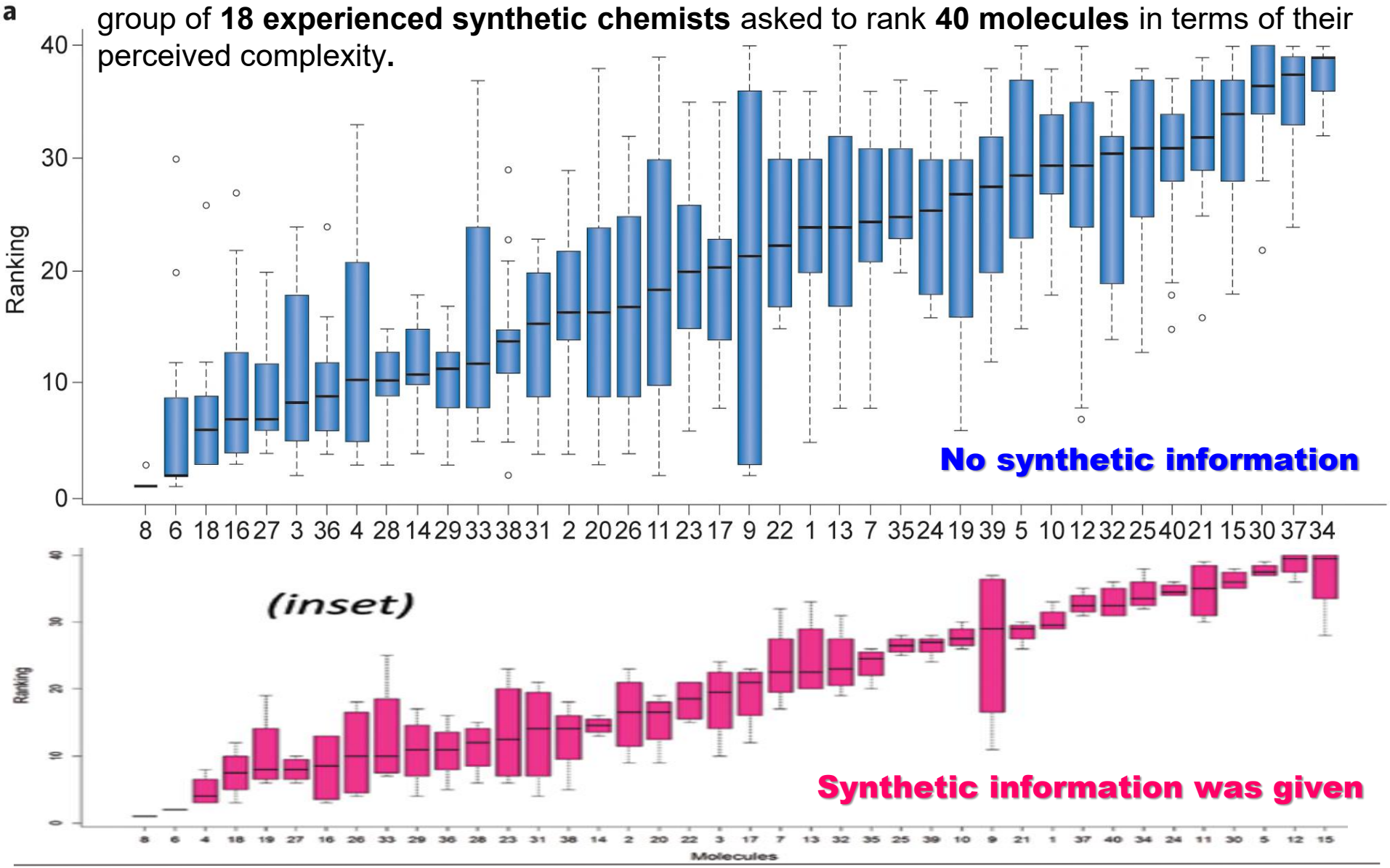


The index is based on a community's perception of complexity, within the context of current technology

1. a) Li, J.; Eastgate, M. D. *Org. Biomol. Chem.* **2015**,*13*, 7164-7176.
b) Woolford, J. *Chemistry World* 22, May, 2015.

Analysis of Collective Intelligence

The current complexity index was based on an analysis of collective intelligence from a group of **18 experienced synthetic chemists** asked to rank **40 molecules** in terms of their perceived complexity.



1. Li, J.; Eastgate, M. D. *Org. Biomol. Chem.* **2015**,*13*, 7164-7176.

Refinement by Intrinsic and Extrinsic Factors

The data obtained from the chemist's intuition was then refined by considering a large series of **intrinsic** and **extrinsic** factors and applying a **Bayesian regression model** to determine the five major factors that impacted the complexity of a structure the most. These are as follows:

(i) the structure's molecular topological (Randic) index[2]
(ii) the number of stereogenic centres established in the synthesis
(iii) the number of heteroatoms on and in aromatic rings
(iv) the number of steps
(v) ideality of the route (as defined by P. S. Baran)[3]

$$\text{ideality} = \frac{(\text{numbers\_of\_construction\_rxns}) + (\text{numbers\_of\_strategic\_redox\_rxns})}{\text{total\_numbers\_of\_steps}}$$

(i) and (iii) are **intrinsic** and unchangeable, whereas the others are **extrinsic** variables reflecting advances that occur over time. From this was established an easily comprehensible 1–10 rating scale, with 1 being the most complex and 10 being least complex.

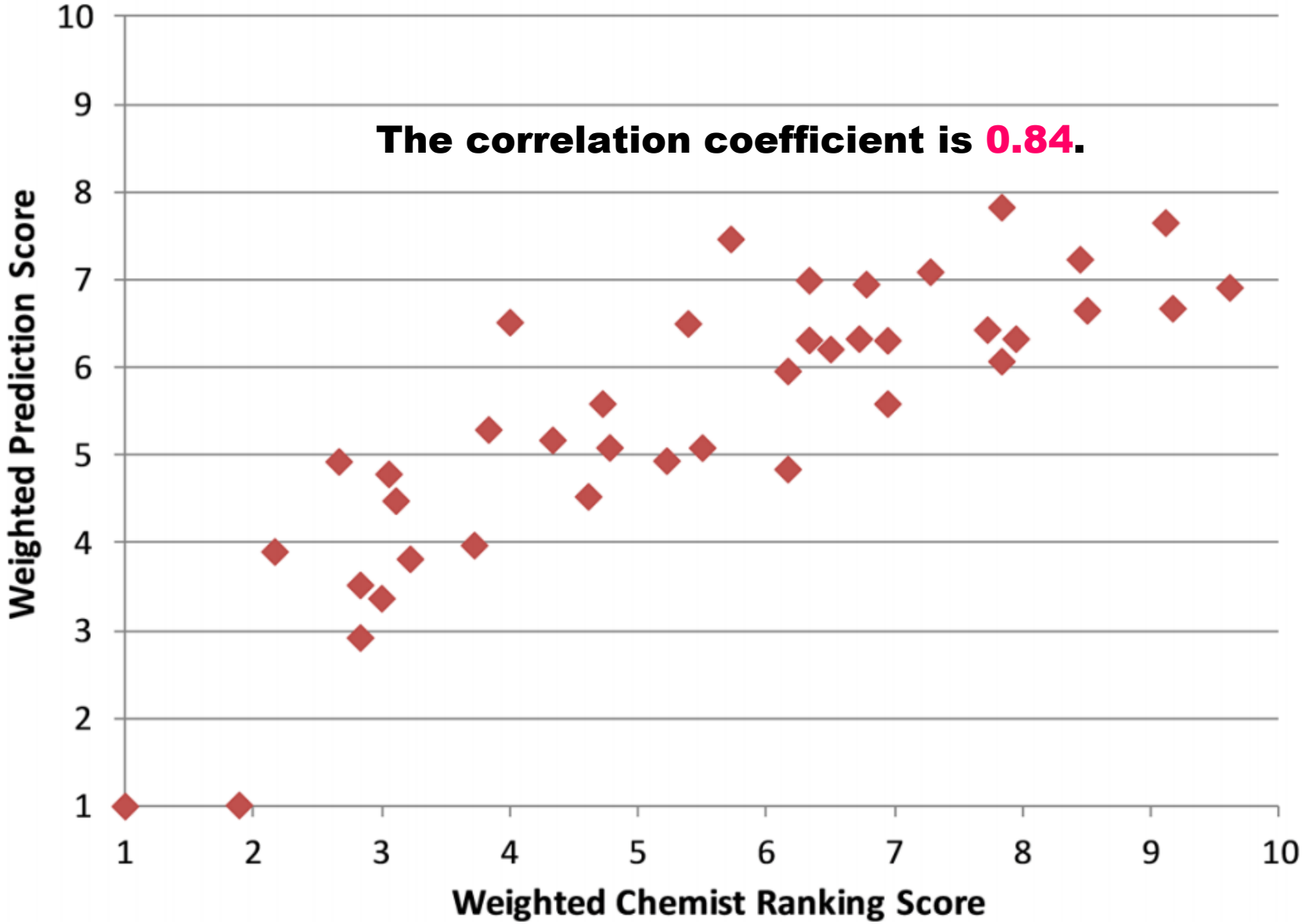
Regression model used in the current complexity index

$$\mu = \beta_0 + \beta^i_{\text{Randic}}\, x_{\text{Randic}} + \beta^e_{\text{SS}}\, x_{\text{SS}} + \beta^i_{\text{HAA}}\, x_{\text{HAA}} + \beta^e_{\text{Steps}}\, x_{\text{Steps}} + \beta^e_{\text{Ideality}}\, x_{\text{Ideality}} + \varepsilon$$

Latent response factor (μ) proportional to five weighted factor coefficients (β). Randic = Randic topology index; SS = number of stereocenters made; HAA = heteroatoms in or on aromatic rings; Steps = longest-linear + 50% of the branching steps; Ideality = ideality score. Intrinsic (i) and extrinsic (e) factors.

1. Li, J.; Eastgate, M. D. *Org. Biomol. Chem.* **2015**,*13*, 7164-7176.
2. Randic, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609-6615.
3. Gaich, T.; Baran, P. S. *J. Org. Chem.* **2010**, *75*, 4657-4673.

Correlation of Weighted Predict Complexity Scores with Chemist Ranking Scores



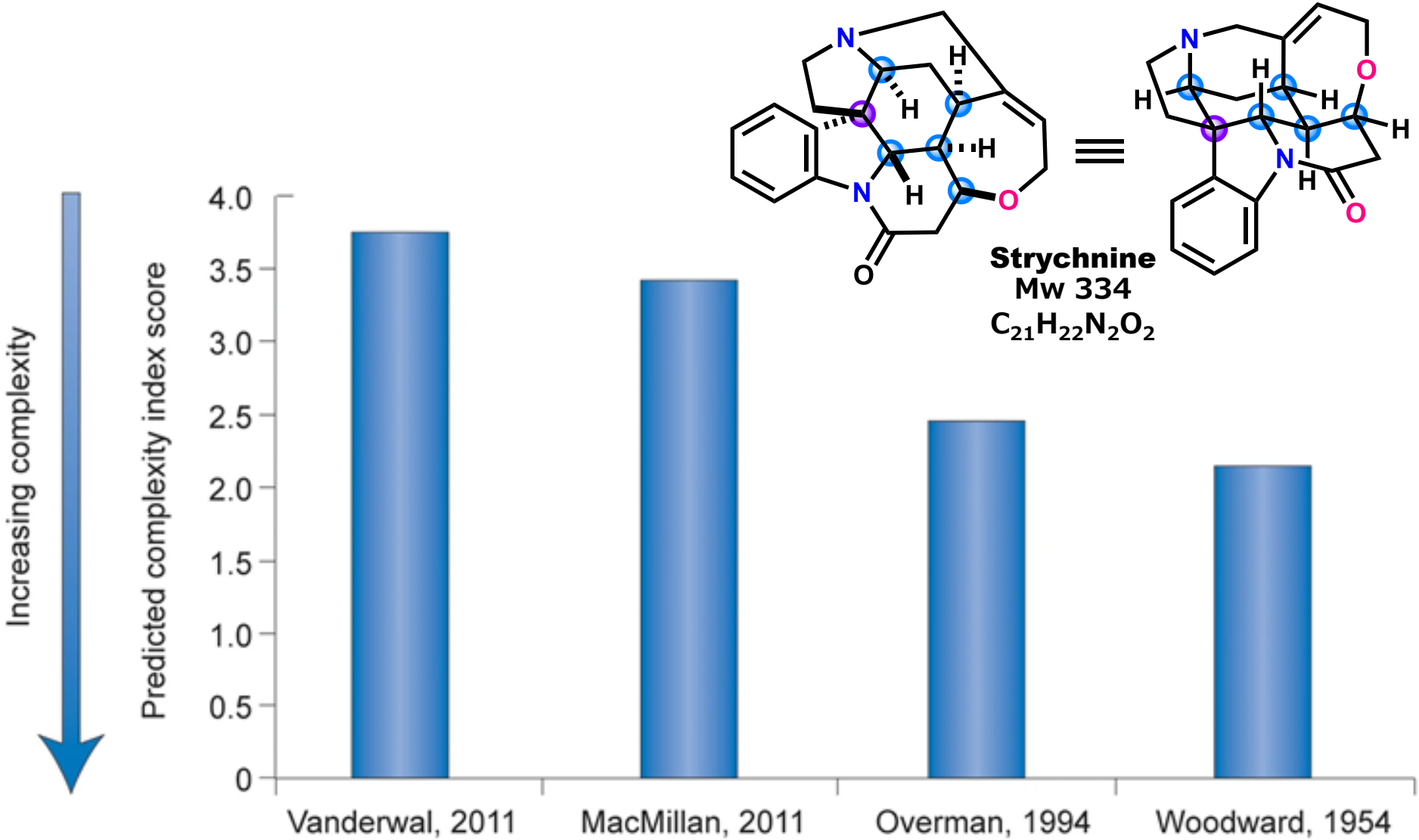
1. Li, J.; Eastgate, M. D. *Org. Biomol. Chem.* **2015**,*13*, 7164-7176.

Input Parameters for Selected Molecules and Weighted Predicted Complexity Score

	Randic	Steps	Ideality	Chiral_made	HAA	predicted complexity score	
Topiramate	10.99264	3	0.67	0	0	7.83	
Sidenafil	15.75559	8	0.89	0	7	6.42	
Atazanavir	25.40878	8.5	0.6	0	1	6.27	
BMS-2	23.16366	11	0.56	1	4	5.16	
BMS-1	18.35124	16	0.56	3	3	4.04	
BILN2061	27.91184	19.5	0.54	3	6	2.88	
EpoA	18.26984	27	0.47	5	2	2.38	Danishefsky
Strychnine	13.72358	8.5	0.60	6	1	3.75	Vanderwal
Strychnine	13.72358	13	0.69	6	1	3.43	MacMillan
Strychnine	13.72358	25	0.56	6	1	2.45	Overman
Strychnine	13.72358	30	0.53	6	1	2.14	Woodward
Discodermolide	23.87016	30	0.59	10	0	1.25	Novartis-Smith-Paterson
Taxol	31.71272	37	0.48	11	0	1.06	Nicolaou
Halaven	31.3614	56	0.38	13	0	1.00	Eisai
Welwit C isonitrile	13.3244	21	0.33	4	1	3.42	Rawal
Welwit C isonitrile	13.3244	23	0.39	4	1	3.30	Garg
Welwit A isonitrile	12.45351	23	0.43	4	1	3.38	Wood
Welwit A isonitrile	12.45351	9	0.78	4	1	4.98	Baran

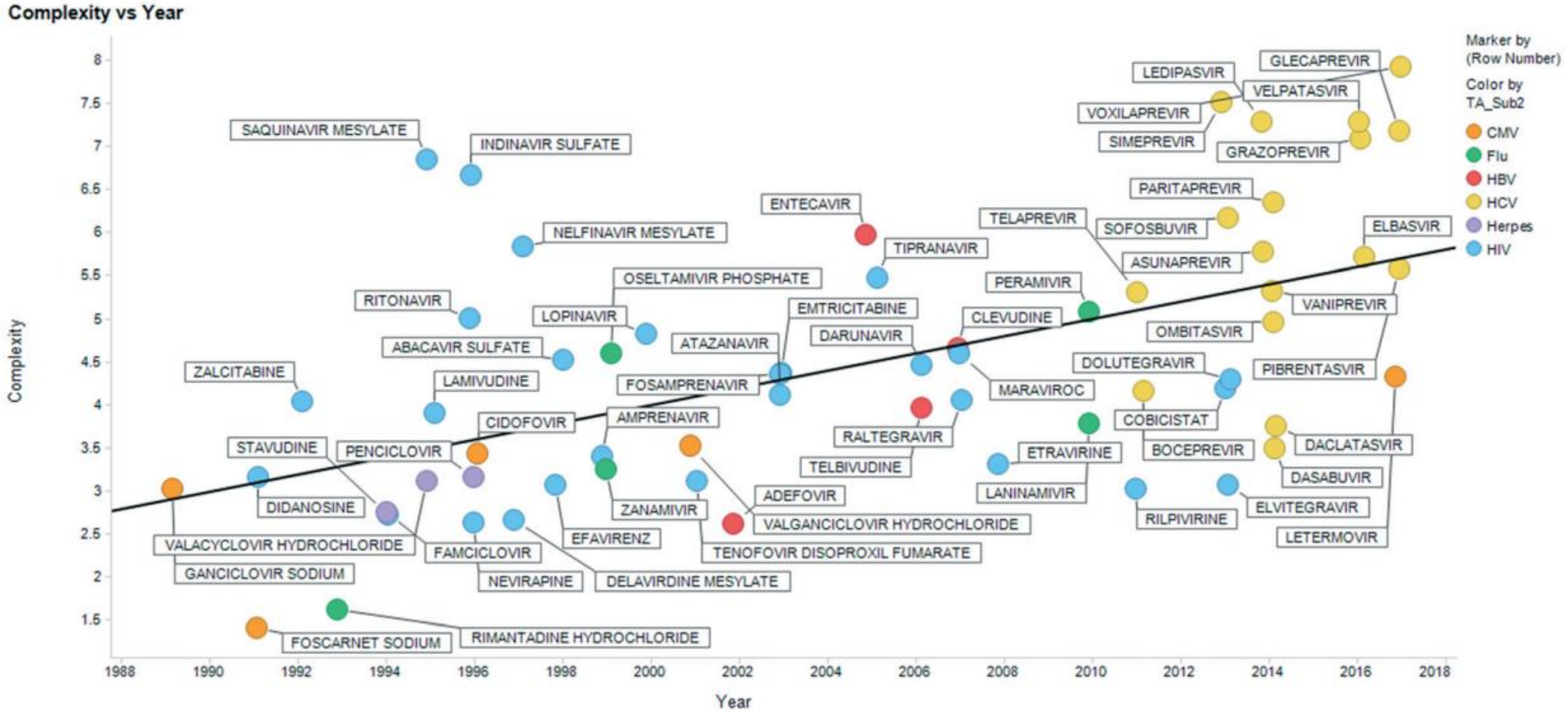
1. Li, J.; Eastgate, M. D. *Org. Biomol. Chem.* **2015**,*13*, 7164-7176.

Predictive Complexity Index Scores for Some of the Strychnine Syntheses



1. a) Li, J.; Eastgate, M. D. *Org. Biomol. Chem.* **2015**,*13*, 7164-7176.
b) Woolford, J. *Chemistry World* 22, May, 2015.

Complexity of Marketed Small-Molecule Antiviral Drugs from 1988 to 2018



Colors represent the disease target.

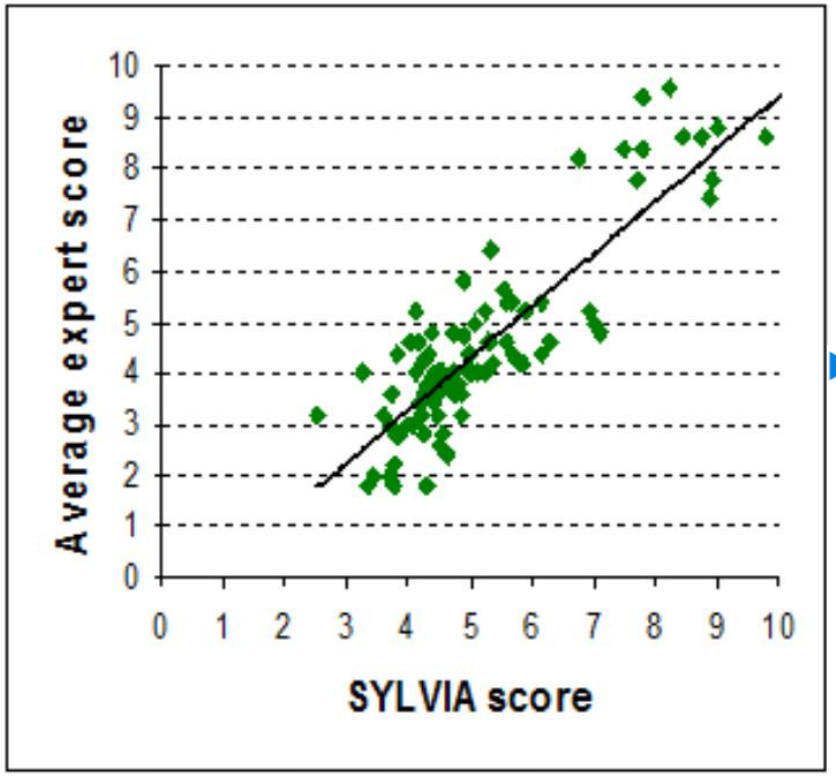
1. Li, J.; Eastgate, M. D. *React. Chem. Eng.* **2019**, *4*, 1595–1607.

SYLVIA

Similarly, **Gasteiger** has defined the **'synthetic accessibility'** of a compound; an estimate to reflect how easily a molecule can be synthesized, based on an analysis of the molecular structure and a comparison with the contents of an organic reaction database.

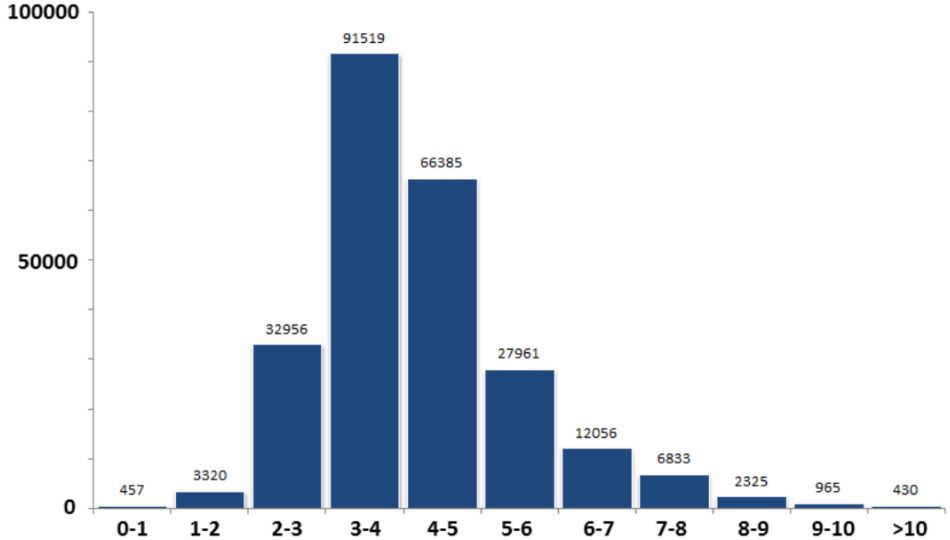
To adjust the synthetic accessibility estimates, **five chemists representing three different pharmaceutical companies** were asked to rank 100 molecules on a ten-point scale.

An accessibility tool called **SYLVIA** that was developed based on these studies is freely available (http://www.molecular-networks.com/products/sylvia).



Correlation: 0.89

Distribution of scores in Open NCI Database with 245,208 records



Summary

Johann Gasteiger, an expert in cheminformatics at the University of Erlangen-Nürnberg, Germany, says:
The idea of quantifying or assigning the complexity of a molecular structure a number has been around for some time without a suitable solution. In other words, **no system has found broad acceptance among the organic community yet**.
The reasons for this are manifold, but not least because of community resistance. Many organic chemists consider synthesis design as an "art" where computers should not have a place.
It seems that organic and process chemists have finally started to recognize the value of computer tools in their work on organic synthesis.

In my opinion, complexity should be defined by the intrinsic factor. The Böttcher Score is considerably better than previous indices but seems to overestimate the degree of complexity reduction due to molecular symmetry. It also seems to lack consideration of fused ring structures (e.g., transannular interaction).

1. a) Gasteiger, J. *Nat. Chem.* **2015**, *7*, 619–620. b) Woolford, J. *Chemistry World* 22, May, 2015.

Appendix

Examples: Various Simple Aliphatic Hydrocarbons

Compound	C_m (mcbit)	No. distinct carbons
Ethane	2.00	1
n-Propane	5.00	2
n-Butane	8.00	2
2-Methylpropane	6.58	2
2-Methylbutane	17.17	4
n-Pentane	11.00	3
2,2-Dimethylpropane	8.00	2
2-Methylpentane	23.17	5
2,2-Dimethylbutane	19.00	4
3-Methylpentane	17.17	4
n-Hexane	14.00	3
2,3-Dimethylbutane	11.17	2
2,3-Dimethylpentane	46.26	7*
3-Methylhexane	45.51	7*
2-Methylhexane	29.17	6
2,2-Dimethylpentane	25.00	5
2,2,3-Trimethylpentane	20.17	4
3,3-Dimethylpentane	18.00	4
n-Heptane	17.00	4
3-Ethylpentane	15.58	3
2,4-Dimethylpentane	14.17	3

*Chiral molecule

Molecular complexity C_m for various simple aliphatic hydrocarbons from ethane to heptane and their isomers correlates with the number of chemically nonequivalent (distinct) carbon atoms.

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

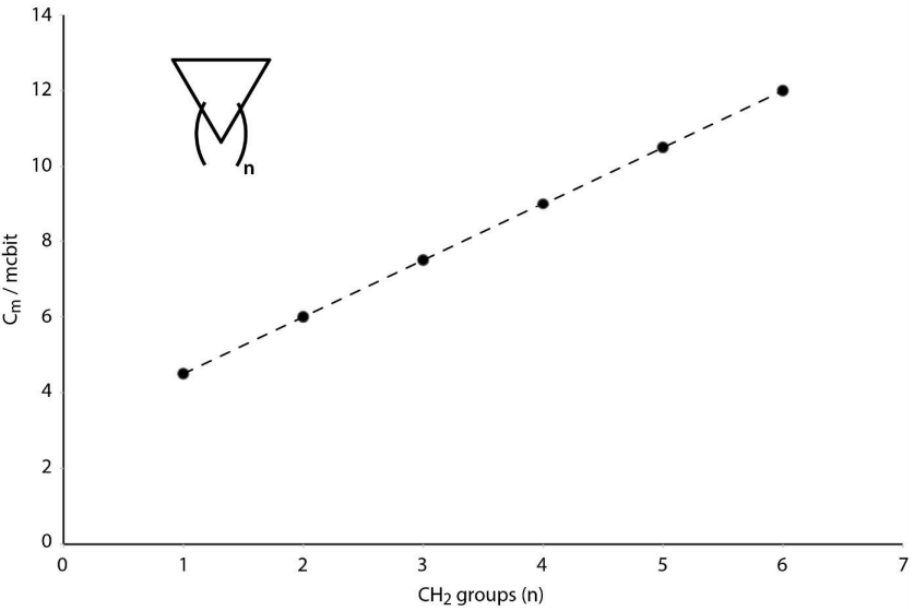
Comparison of C_m with Other Complexity Indices

Compound	C_m (mcbit)	$C(\eta,\varepsilon)$	N_T	N_S
Methane	-*	-*	1	1
Ethane	2.00	-*	3	2
n-Propane	5.00	0.0	6	3
n-Butane	8.00	2.0	10	4
n-Pentane	11.00	7.5	15	5
n-Hexane	14.00	12.0	21	6
Ethylene	3.00	0.0	5	3
Propene	12.17	7.5	10	5
1-Butene	18.17	14.0	16	7
Methanol	9.17	2.0	3	3
Ethanol	19.17	2.8	6	5
1-Propanol	25.17	7.2	10	7
2-Propanol	21.51	10.8	11	7
Acetone	25.17	26.3	19	10
Acetaldehyde	23.51	10.3	10	7
Propanal	29.51	17.2	16	10
Dimethyl ether	11.17	2.8	6	4
Acetylene	3.58	4.8	9	4

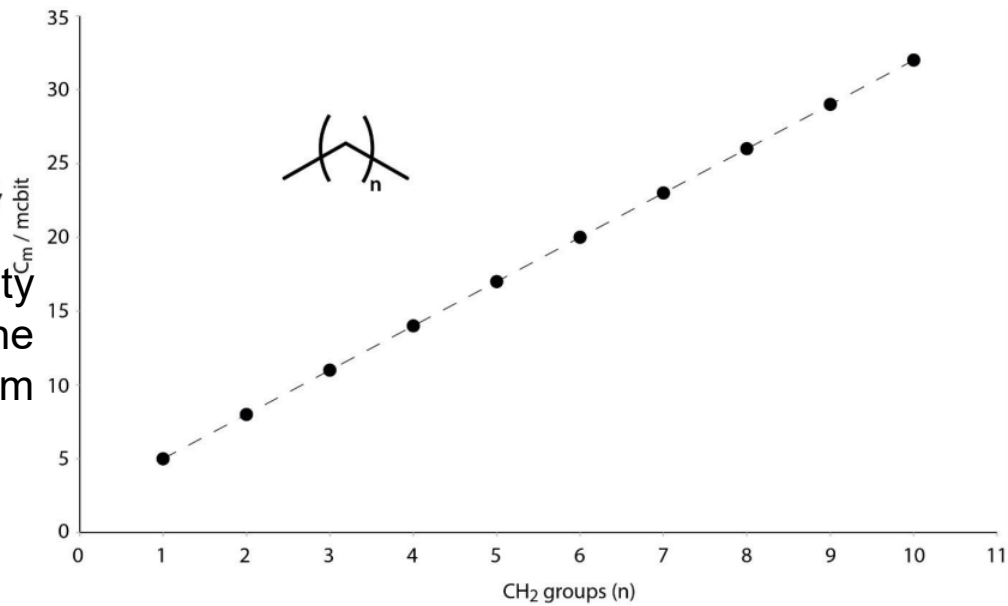
*not defined for these molecules

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

Example: Linear Dependence of C_m



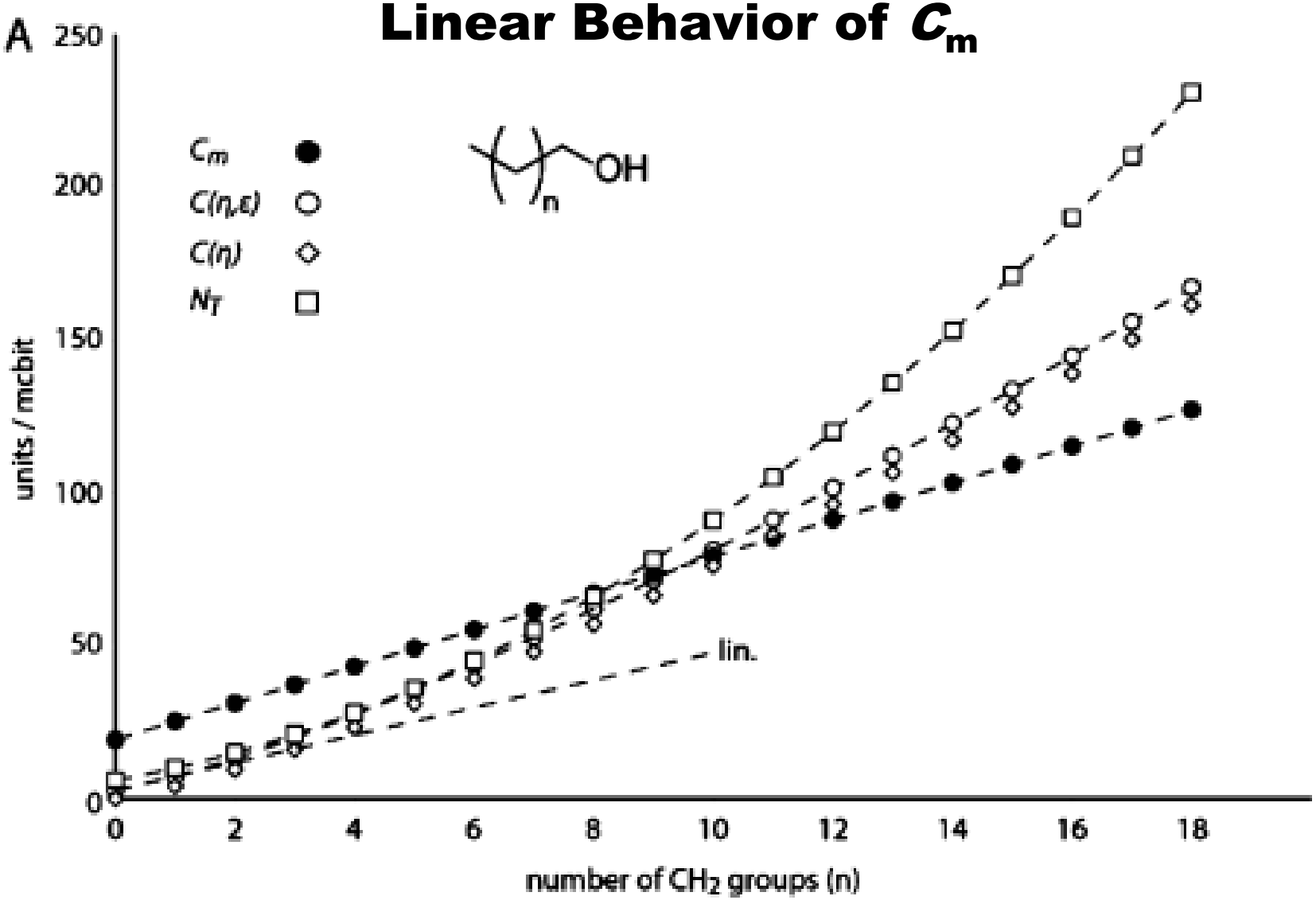
The linear dependence of molecular complexity C_m from the number of methylene groups in the homologous series of cycloalkanes from cyclopropane (n = 1) to cyclooctane (n = 6).



The linear dependence of molecular complexity C_m from the number of methylene groups in the homologous series of alkanes from propane (n = 1) to dodecane (n = 10).

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

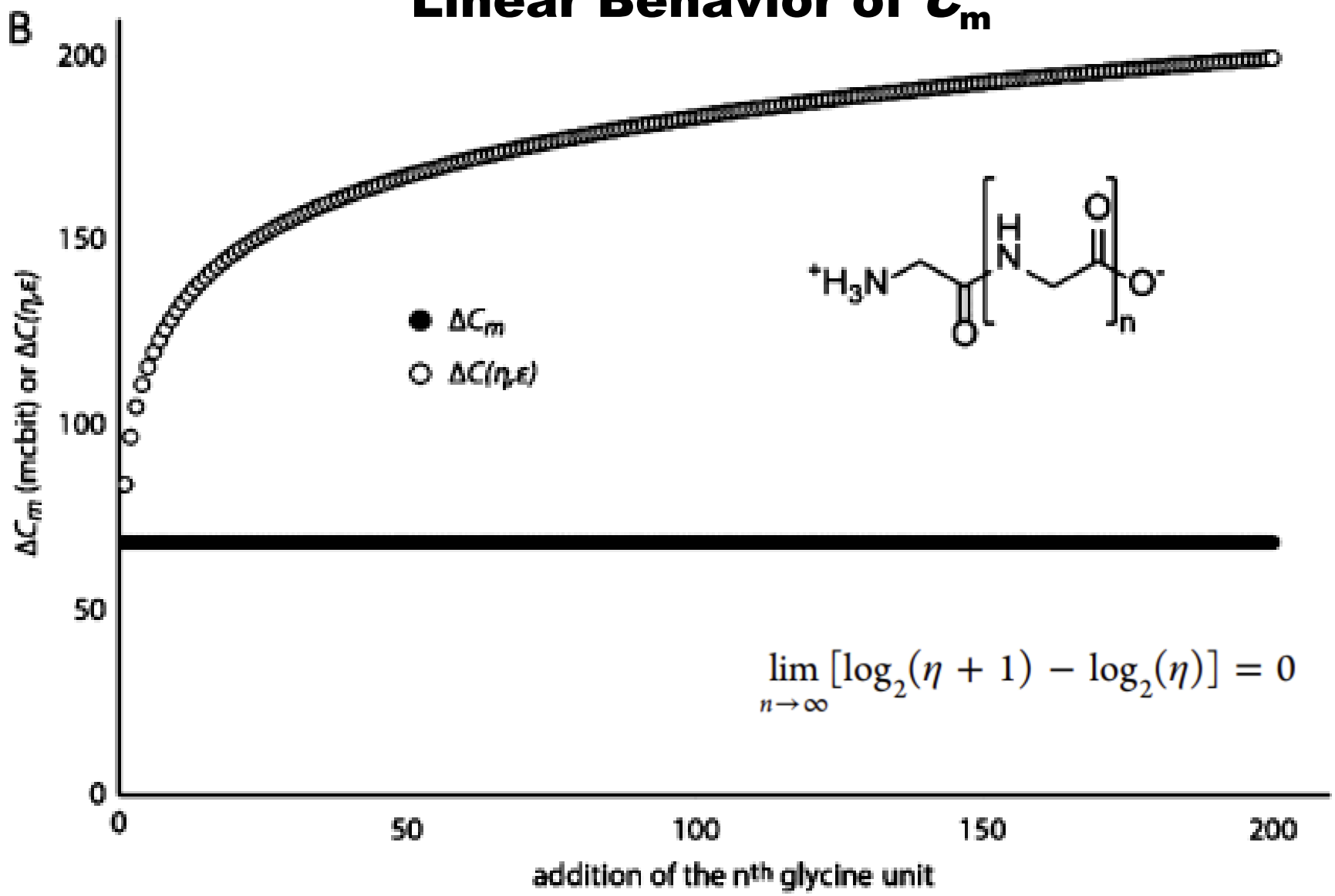
Nonlinear Behavior of $C_{(\eta,\varepsilon)}$ in Comparison to the Linear Behavior of C_m



Homologous series from ethanol to $C_{20}H_{41}OH$, where Cm gives a linear increase and C(η,ε), C(η), and NT give nonlinear increases. A linear extrapolation (lin.) from the first two values of C(η,ε) is given for better visualization.

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

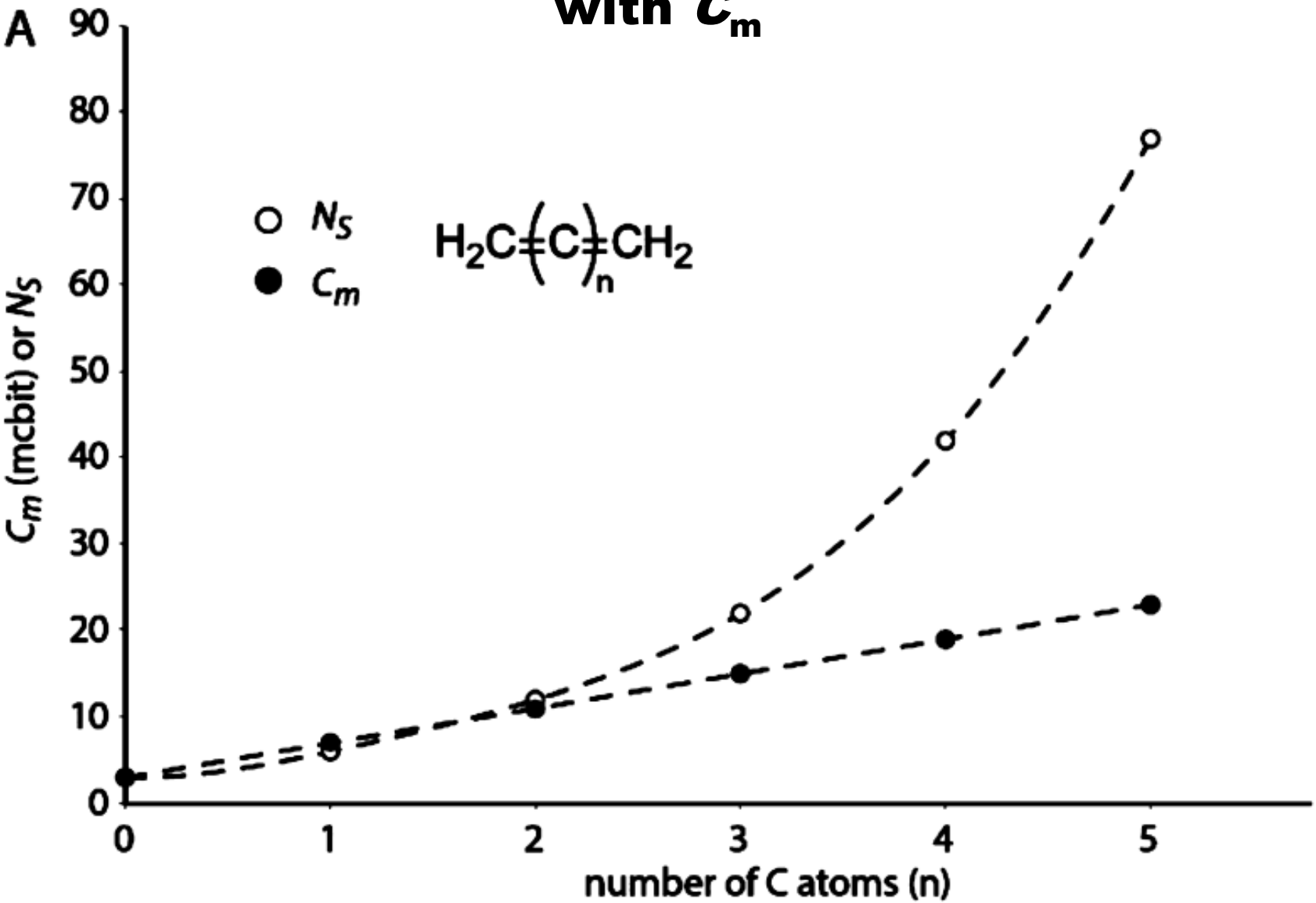
Nonlinear Behavior of $C_{(\eta,\varepsilon)}$ in Comparison to the Linear Behavior of C_m



Changes in complexity $\Delta C_{(\eta,\varepsilon)}$ and ΔC_m for the addition of the nth glycine residue to a polyglycine chain, showing the nonlinear behavior of $C_{(\eta,\varepsilon)}$.

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

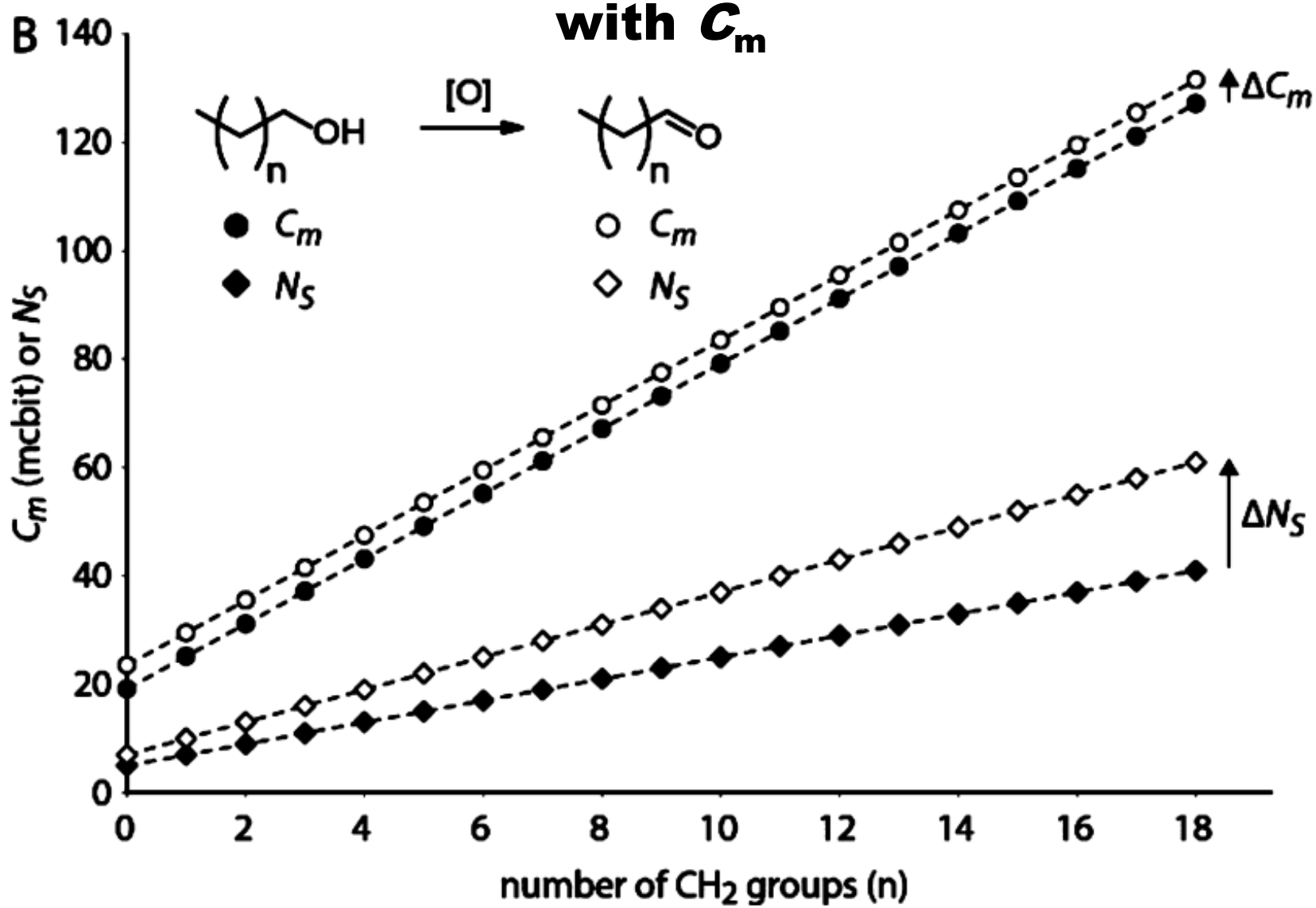
Comparison of the Graph-Theory-Based Index N_S with C_m



Linear and nonlinear increases within the homologous series of cumulenes for C_m and N_S, respectively.

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

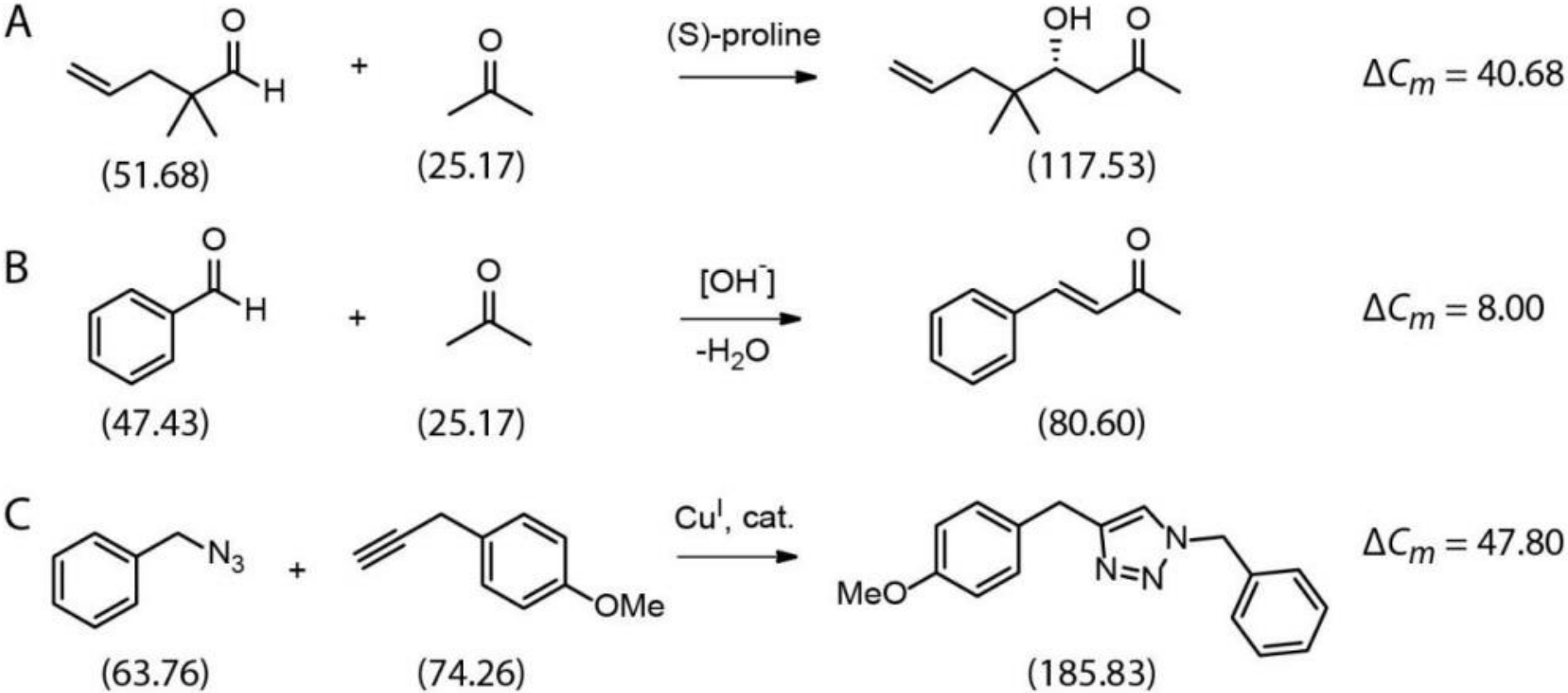
Comparison of the Graph-Theory-Based Index N_S with C_m



Invariance of ΔC_m to the chain length of aliphatic alcohols in the oxidation reaction to give the corresponding aldehydes, in contrast to an increase in ΔN_S for the same reaction with increasing hydrocarbon chain length.

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

Changes in Molecular Momplexity ΔC_m



Changes of molecular complexity ΔC_m for different chemical reactions.

A) Stereoselective aldol reaction
B) aldol condensation
C) 1,3-dipolar azide-alkyne cycloaddition.

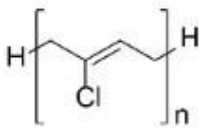
C_m values for all reactants are given in brackets.

1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462−470.

Molecular Complexity for Various Artificial and Biological Polymers without Sequence Information

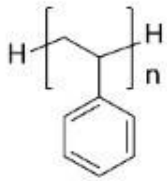
Artificial polymers

Neoprene



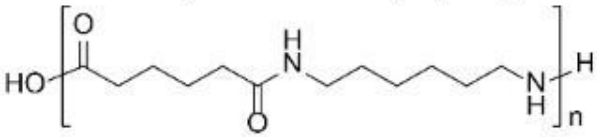
$C_m = (56 \cdot n) - 8$
$C_i = 0$

Polystyrene



$C_m = (61 \cdot n) - 19.5$
$C_i = 0$

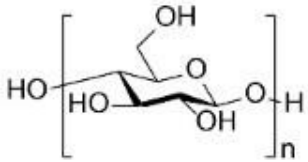
Polyamide 6.6 (Nylon)



$C_m = (172.9 \cdot n) - 15.5$
$C_i = 0$

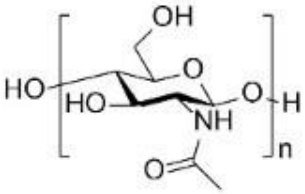
Biological polymers

Cellulose



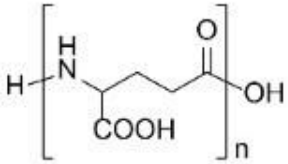
$C_m = (271.3 \cdot n) - 4$
$C_i = 0$

Chitin



$C_m = (324.6 \cdot n) - 4$
$C_i = 0$

Poly-γ-glutamic acid

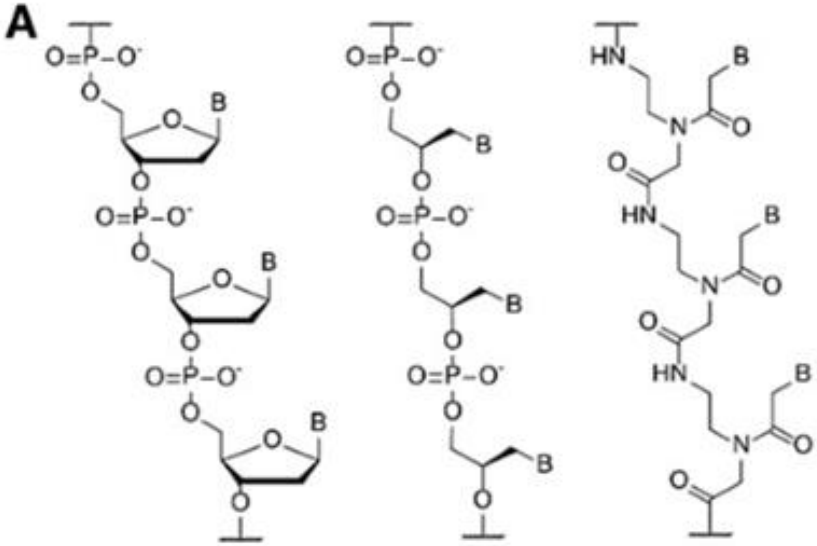


$C_m = (138.1 \cdot n) - 21.5$
$C_i = 0$

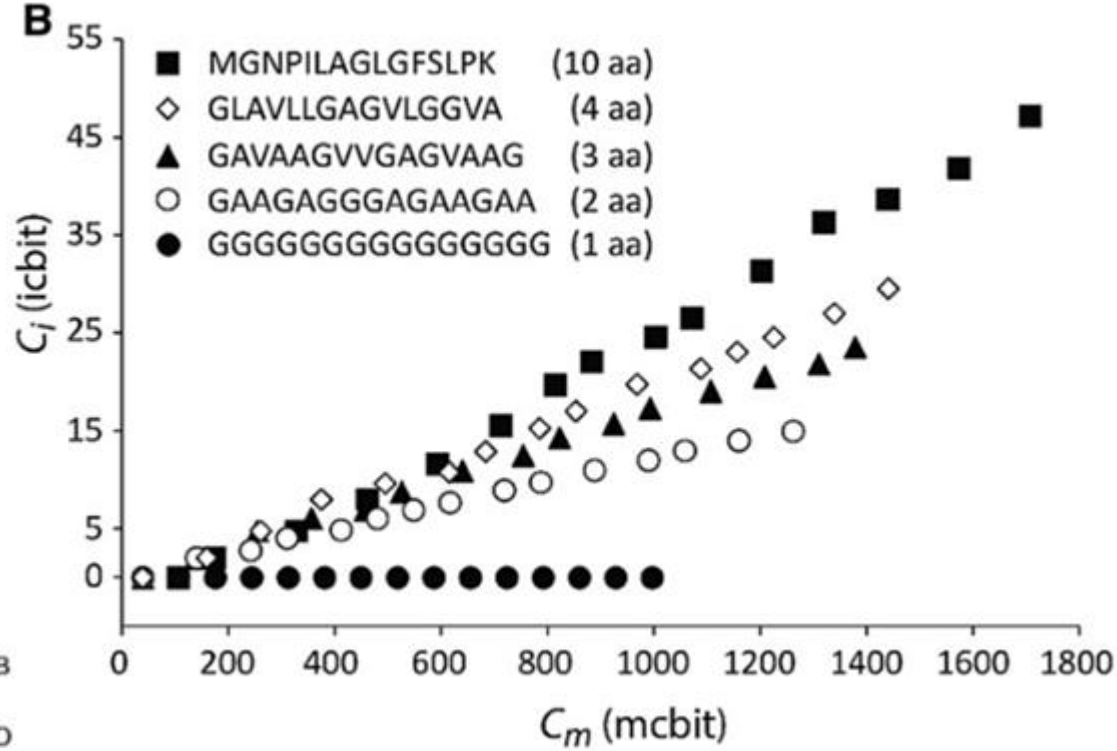
1. Böttcher, T. *J. Mol. Evol.* **2018**, 86, 1–10.

Quantification of Molecular (C_m) and Information Complexity (C_i) for Various Types of Biogenic Units

A) Chemical structures of DNA, (R)-GNA, and PNA and calculated complexity values for an arbitrary model sequence (ATGTGA).



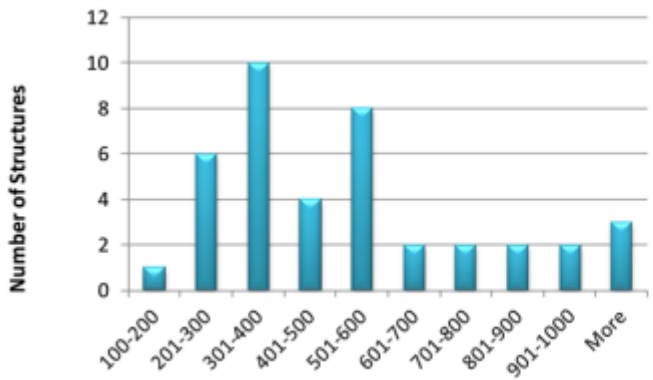
ATGTGA	DNA	(*R*)-GNA	PNA
C_m	2596.2	1831.1	1985.6
C_i	11.5	11.5	11.5

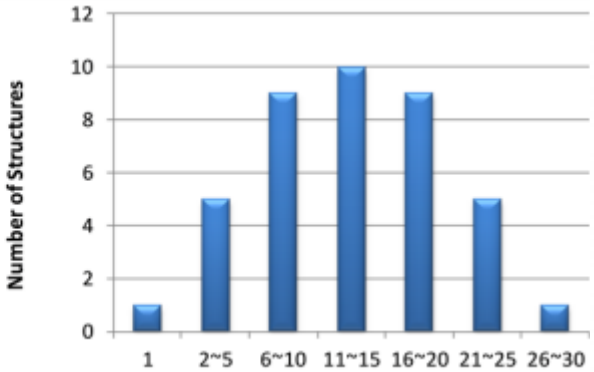


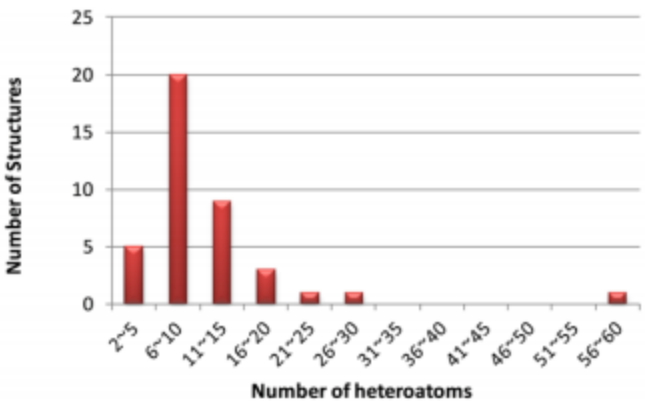
B) Information complexity plotted against molecular complexity for arbitrary protein sequences of different alphabet sizes (1 aa to 10 aa) as a function of length. B: nucleobase, aa: amino acid

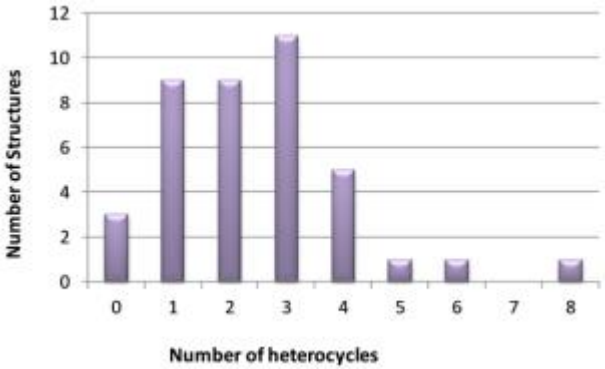
1. Böttcher, T. *J. Mol. Evol.* **2018**, 86, 1–10.

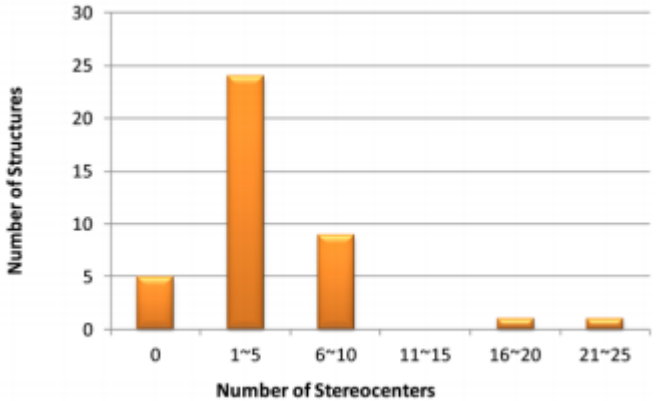
Training Dataset

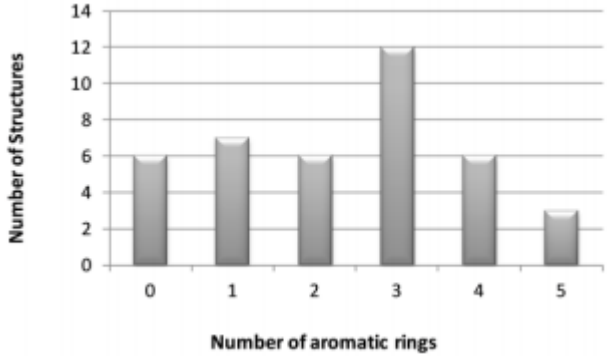




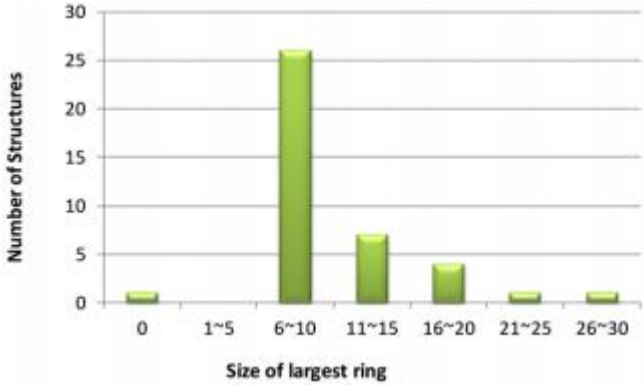








Distributions of molecular properties from MW, double bond equivalent (DBE), heteroatoms, heterocycles, aromatic rings, stereocenters, and the largest ring size in the training dataset.



1. Li, J.; Eastgate, M. D. *Org. Biomol. Chem.* **2015**,*13*, 7164-7176.