Molecular Complexity

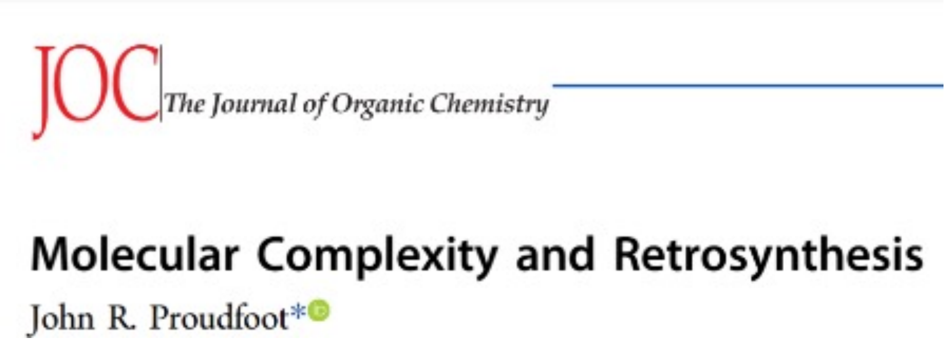
Literature Seminar

2024/06/15

Yuuki Watanabe

Contents

1. Development of indexes



JOC The Journal of Organic Chemistry

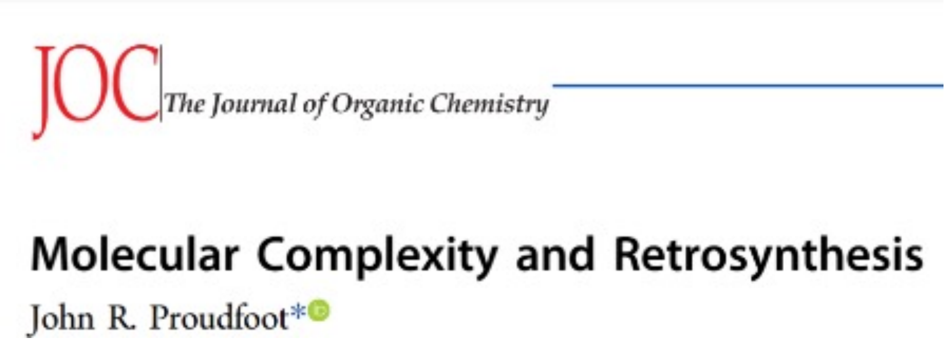
Molecular Complexity and Retrosynthesis

John R. Proudfoot*

2. Application

Contents

1. Development of indexes



JOC *The Journal of Organic Chemistry*

Molecular Complexity and Retrosynthesis

John R. Proudfoot*

2. Application

John R. Proudfoot

John Proudfoot, Ph. D.

Education

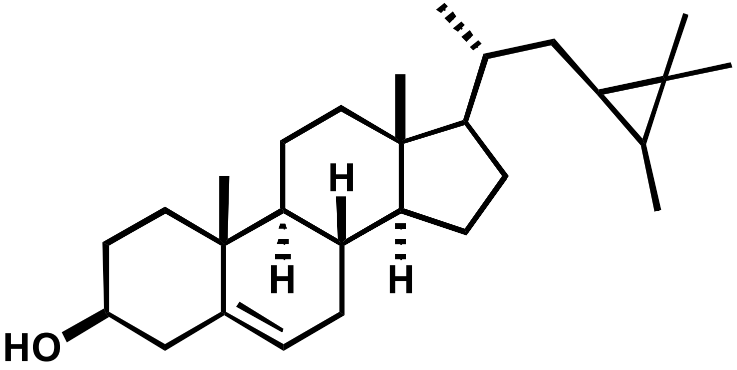
1984?~1987:Ph. D @Stanford University (Prof. Carl Djerassi)
1987~2017: Director of medicinal chemistry
@ Boehringer Ingelheim Pharmaceuticals Inc: Ridgefield, CT, US
2017~: Consultant @Discoverybytes LLC: Newtown, CT, US

Research Area in Stanford University

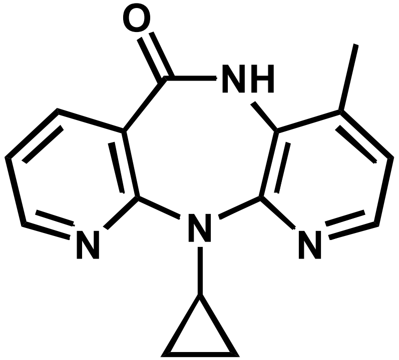
1. Structure determination of steroids
2. (Bio)synthesis of new steroids

Research Area in Pharmaceutical Company

1. Medicinal Chemistry
2. Drug Design
3. Data processing



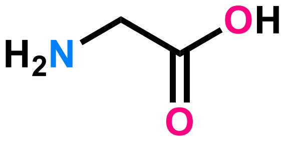
Prof. Carl Djerassi



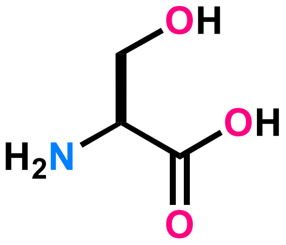
nevirapine
(anti-HIV-1 drug)

1. http://www.sbchem.kyoto-u.ac.jp/suginome-lab/jp/index.php?%E3%83%9B%E3%83%BC%E3%83%A0

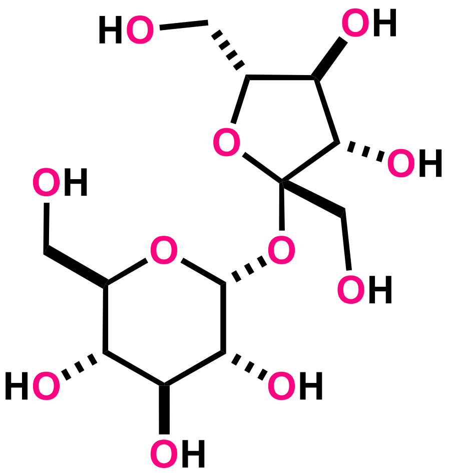
Compound –What do you think?–



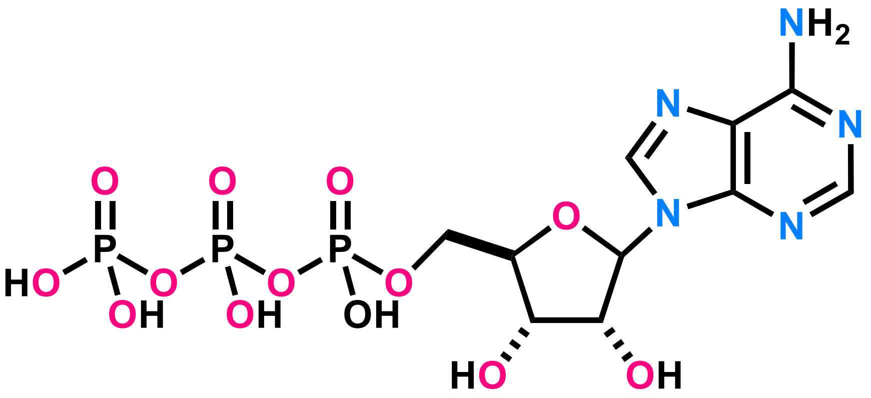
glycine



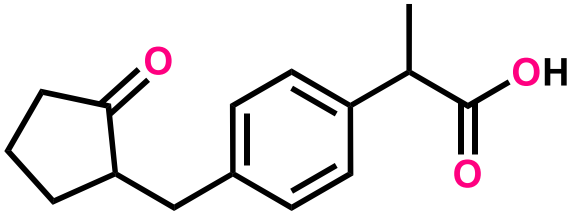
serine



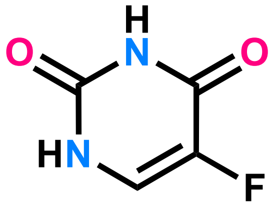
sucrose



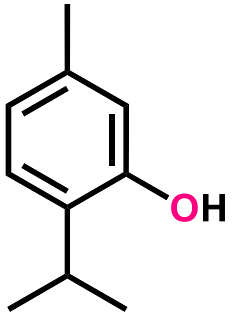
adenosine triphosphate



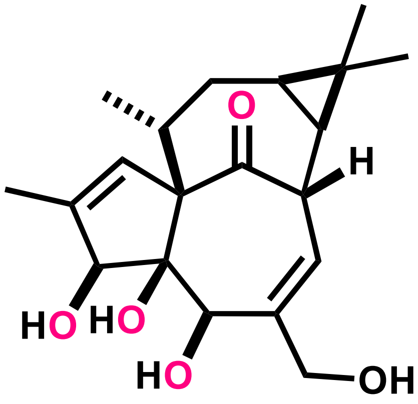
loxoprofen



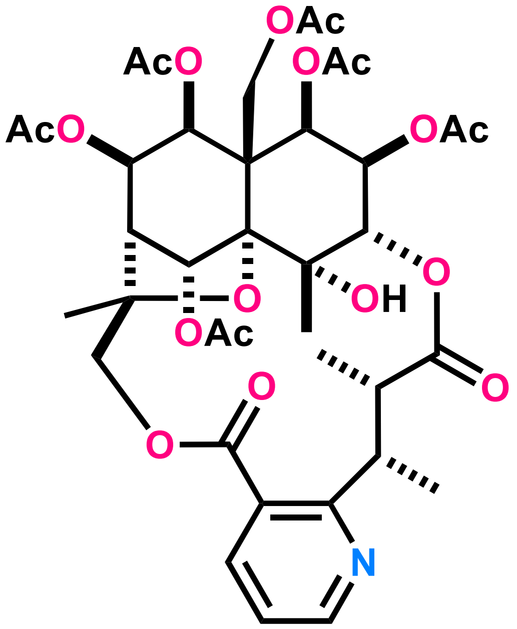
5-FU



thymol



Ingenol



euonymine

What characterizes molecules?

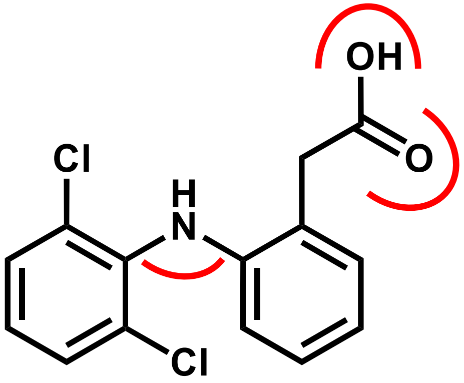
Index for Drug Discovery

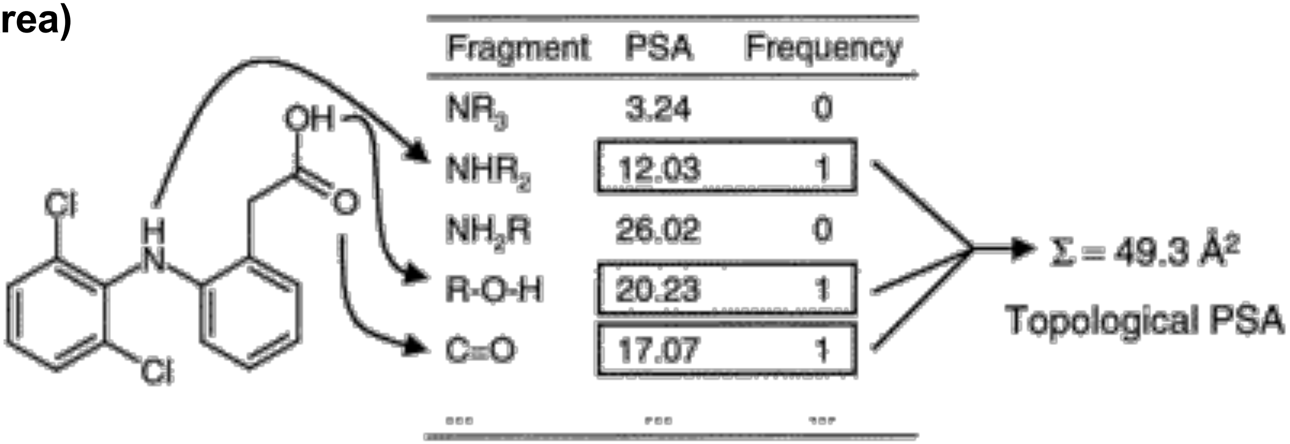
In the drug discovery, the ① polarity, ② specificity, and ③ greenness of synthetic route are important.
→The indexes corresponding to these factors are developed.

① polarity

- Lipinski's rule of 5

- tPSA (topological Polar Surface Area)





Fragment	PSA	Frequency
NR_3	3.24	0
NHR_2	12.03	1
NH_2R	26.02	0
R-O-H	20.23	1
C=O	17.07	1
...	...	...

$\Sigma = 49.3$ Å^2

② specificity

- F_{sp3}, $F_{Cstereo}$

$$F_{\mathrm{sp}^3} = \frac{\text{the number of sp}^3\text{ carbon atom}}{\text{total number of carbon atom}}$$

$$F_{\mathrm{C}_{\mathrm{stereo}}} = \frac{\text{the number of stereocenters}}{\text{total number of carbon atom}}$$

③ greenness of synthetic route

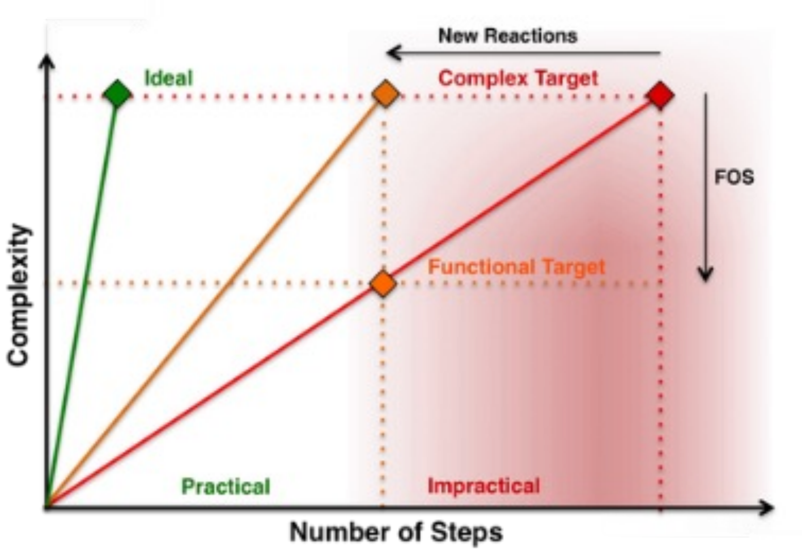
- PMI (Process Mass Intensity)

$$\mathrm{PMI} = \frac{\text{total amounts of materials (reagents, solvents...)}}{\text{amounts of product}}$$

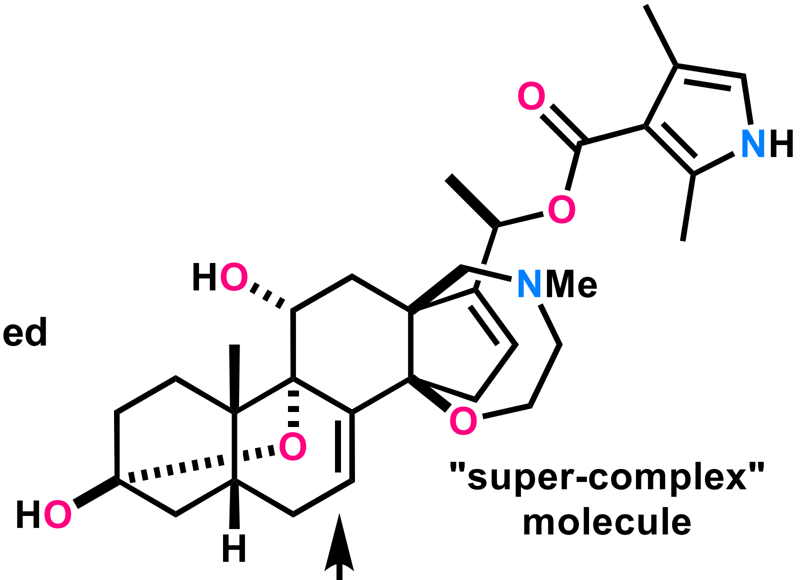
- The key, high-level metric for evaluating and benchmarking progress towards more sustainable manufacturing.

1. Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714.
2. Lovering, F.; Bikker, J.; Humblet, C. *J. Med. Chem.* **2009**, *52*, 6752.
3. Jimenez-Gonzalez, C.; Ponder, C. S.; Broxterman, Q. B.; *Angew. Chem., Int. Ed.* **2021**, *60*, 12819.

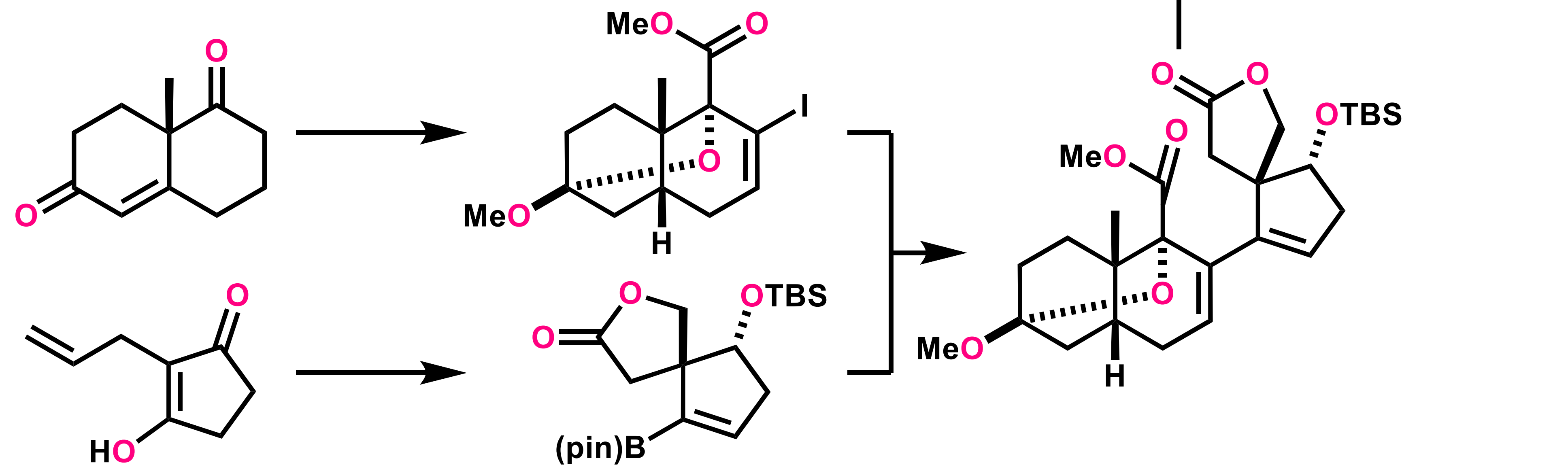
Indexes for chemists



For chemists, complexity was obscurely mentioned



"super-complex" molecule



"simple" molecule

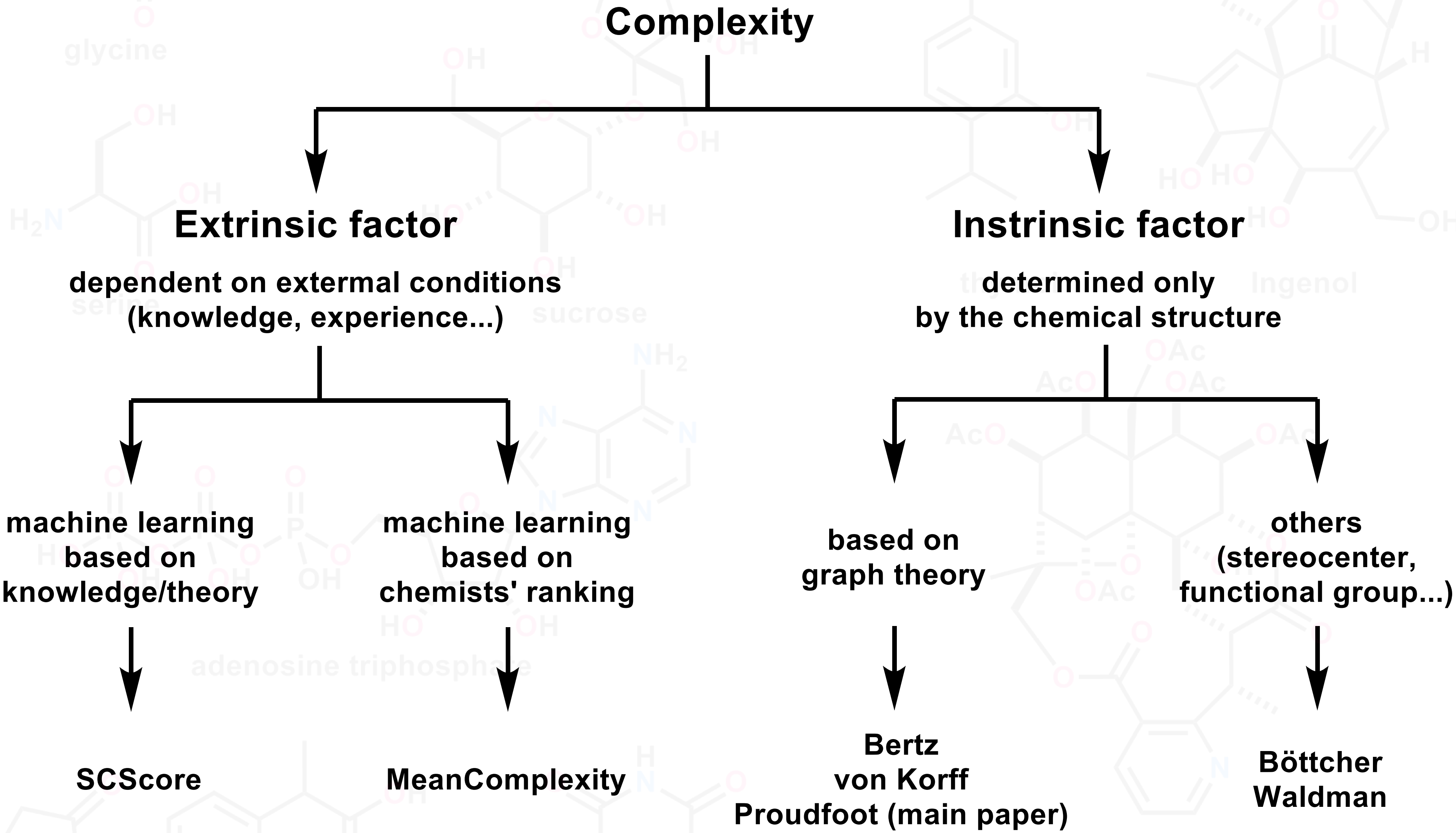
"medium" intermediate

"complex" intermediate

1. https://web.stanford.edu/group/pawender/function-oriented-synthesis.html

2. Watanabe, Y.; Morozumi, H.; Mutoh, K.; Hagiwara, K.; Inoue, M. *Angew. Chem. Int. Ed.* **2023**, *62*, e202309688.

Define Molecular Complexity



Complexity

Extrinsic factor

dependent on extermal conditions
(knowledge, experience...)

machine learning
based on
knowledge/theory

SCScore

machine learning
based on
chemists' ranking

MeanComplexity

Instrinsic factor

determined only
by the chemical structure

based on
graph theory

Bertz
von Korff
Proudfoot (main paper)

others
(stereocenter,
functional group...)

Böttcher
Waldman

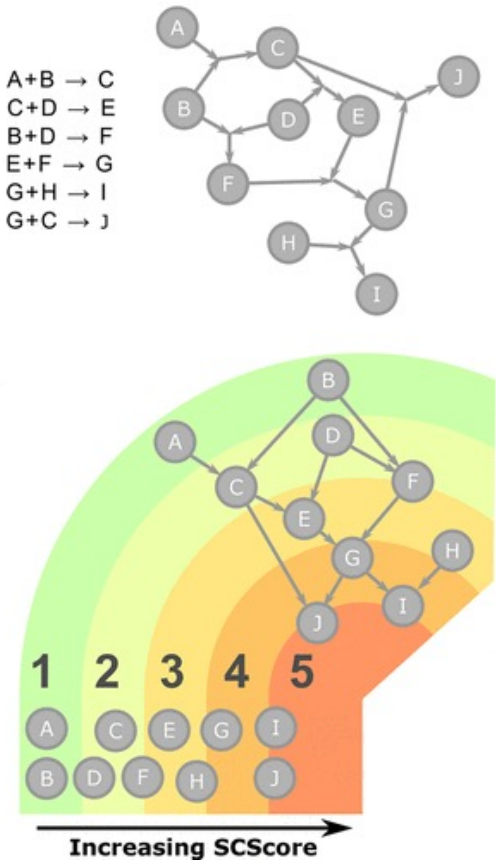
SCScore

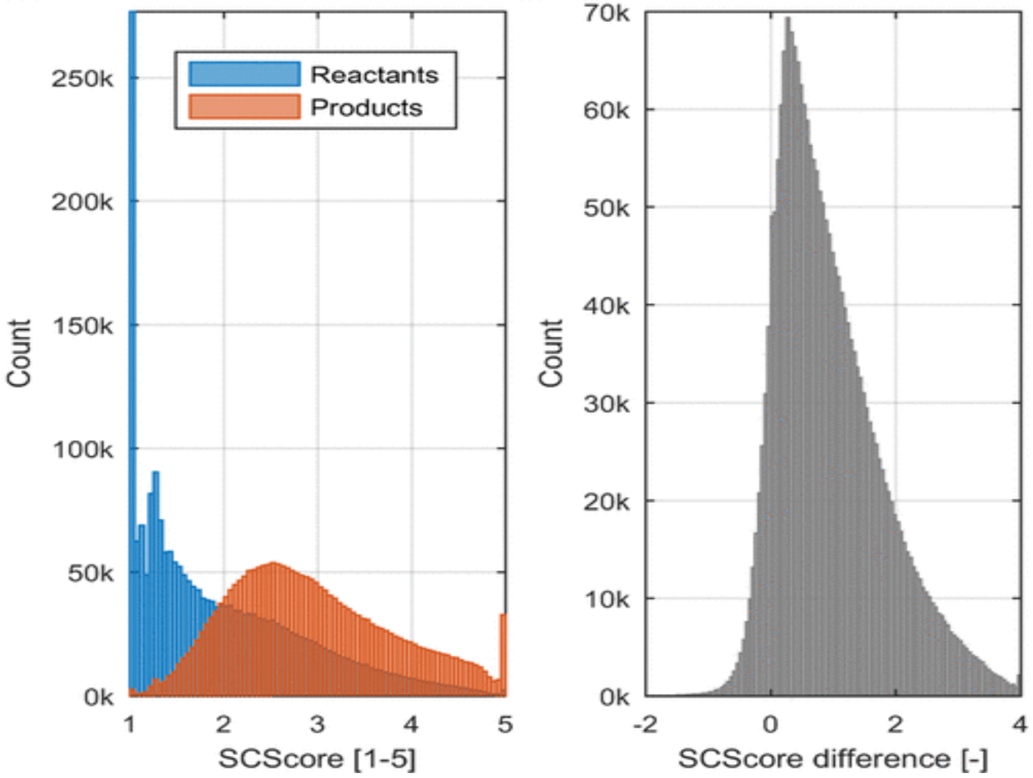
SCScore: Based on precedent reaction knowledge (12 million reactions in Reaxys)
Assuming that the complexity of products should not be less than the complexity of reactants.

$$\mathrm{SCScore}(m) \equiv f(m, \theta)$$

$$f(P, \theta) \geqq \max\{f(R_i, \theta)\}_i \;\forall\; (R_1 + R_2 + \ldots + R_n \rightarrow P)$$

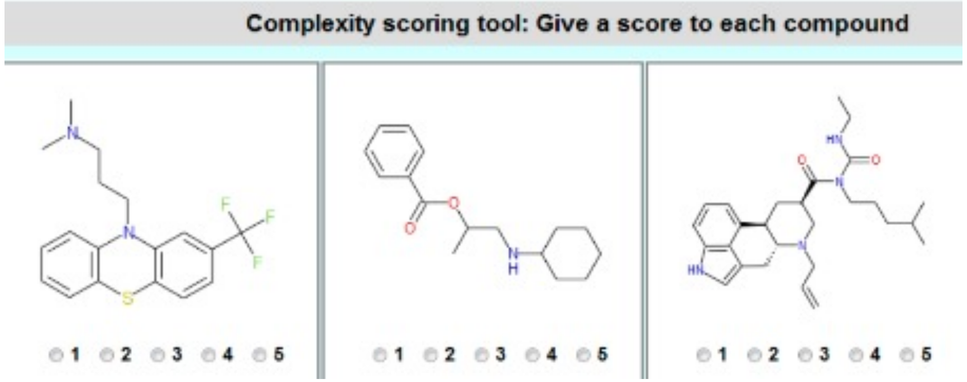
P: Product
R: Reactant





Coley, C. W.; Rpgers, L.; Green, W. H.; Jensen, K. F. *J. Chem. Inf. Model.* **2018**, *58*, 252.

MeanComplexity

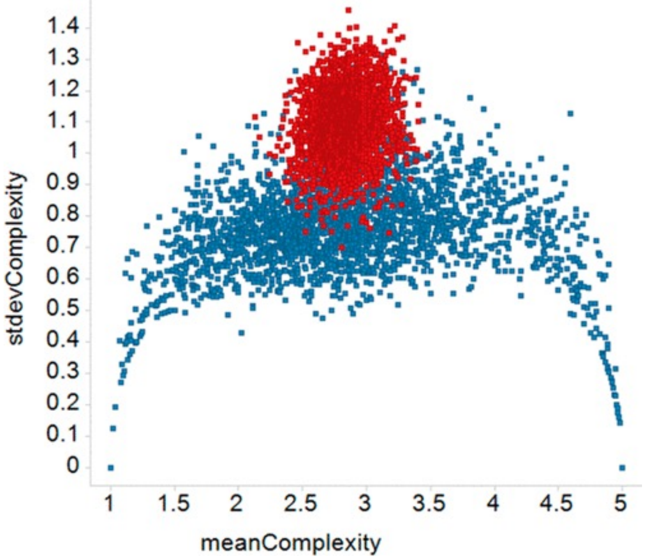


Complexity scoring tool: Give a score to each compound

MeanComplexity: Developed by Merck
- Based on the knowledge of chemists
- Comparing complexity
(not synthesizability)

Chemists voted

108000 votes from
386 chemists for 2681 molecules



Blue: meanComplexity
Red: randomly voted results

Merck developed a system for predicting PMI based on the MW and MeanComplexity to judge the efficiency of synthetic routes.

II

SMART-PMI

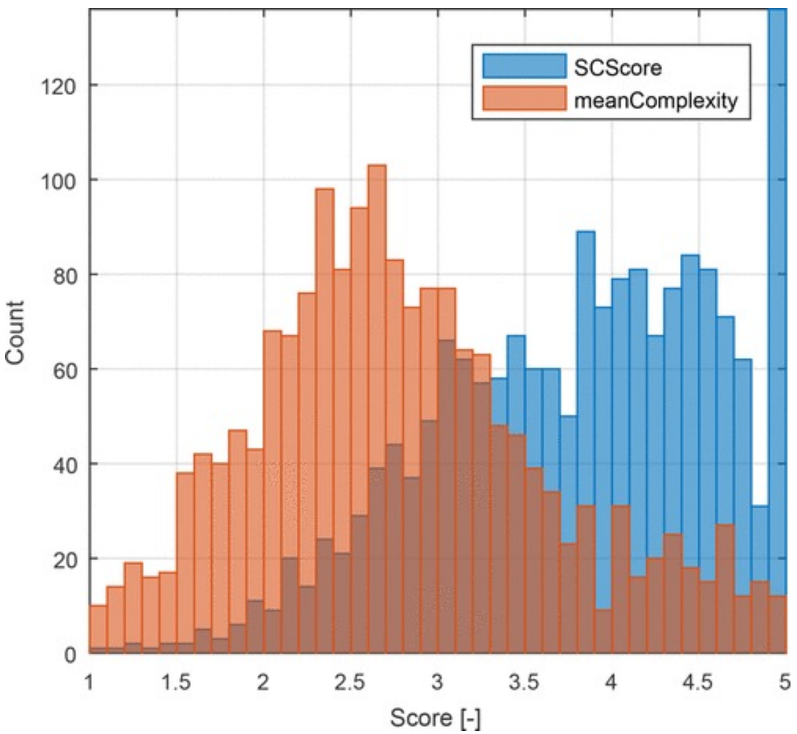
$$\boxed{\begin{aligned}\mathrm{SMART-PMI} &= 0.13 \cdot \mathrm{MW} \\ &+177 \cdot \mathrm{MeanComplexity} - 252\end{aligned}}$$

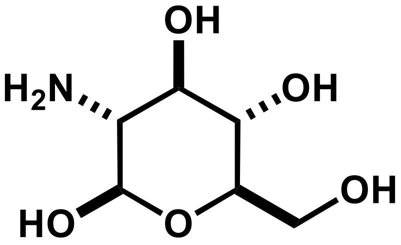
Problem: For the complex molecules,

$$\text{SMART-PMI} \fallingdotseq 0.13 \cdot \mathrm{MW} + 633$$

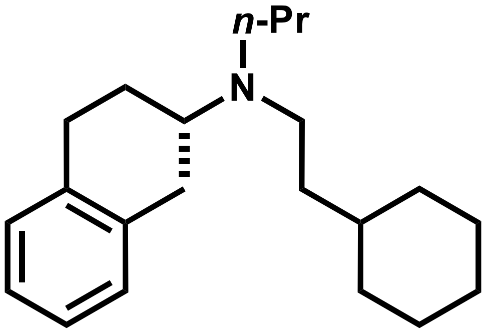
Sheridan, R. P.; Zorn, N.; Sherer, E. C.; Campeau, L.-C.; Chang, C.; Cumming, J.; Maddess, M. L.; Nantermet, P. G.; Sinz, C. J.; O'shea, P. D. *J. Chem. Inf. Model.* **2014**, *54*, 1604.

SCScore v.s. MeanComplexity

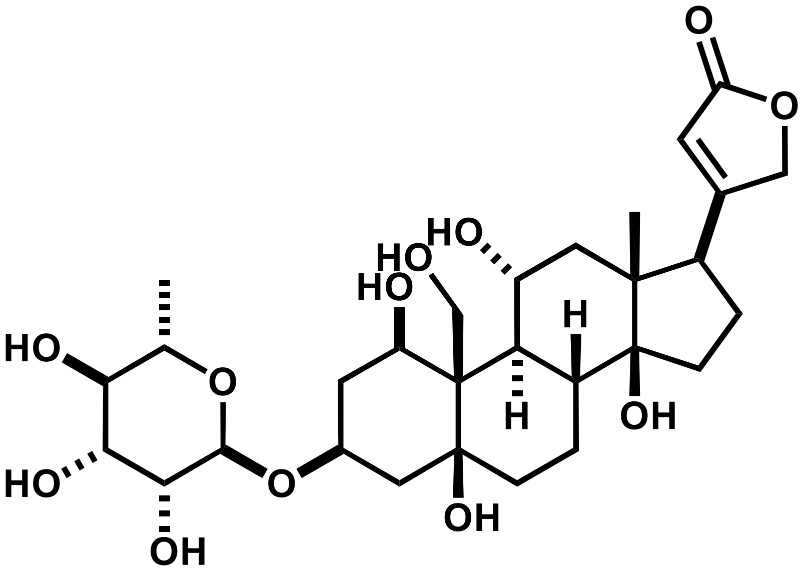




SCScore: 1.249
meanComplexity: 3.105



SCScore: 4.262
meanComplexity: 1.413



SCScore: 1.794
meanComplexity: 4.778

SCScore: the complex natural products might be vey low score because they are isolated and used for the manipulation as starting material.

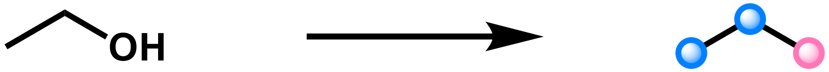
On the other hand, drugs are usually final target molecules, so they have high score.

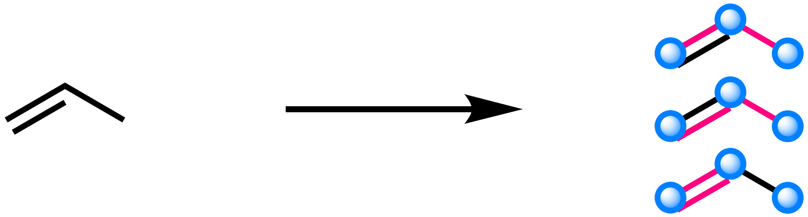
→MeanComplexity is more suitable for the organic synthesis/chemistry.

1. Coley, C. W.; Rpgers, L.; Green, W. H.; Jensen, K. F. *J. Chem. Inf. Model.* **2018**, *58*, 252.
2. Sheridan, R. P.; Zorn, N.; Sherer, E. C.; Campeau, L.-C.; Chang, C.; Cumming, J.; Maddess, M. L.; Nantermet, P. G.; Sinz, C. J.; O'shea, P. D. *J. Chem. Inf. Model.* **2014**, *54*, 1604.

Bertz Complexity C_T (1)

<Graph theory-based complexity>





η : **the total number of "connectivity"**
= the number of propane substructure

$$\eta = \frac{1}{2}\sum_i (4-i)(3-i) - \underline{D} - 3\underline{T}$$

Double bond **Triple bond**

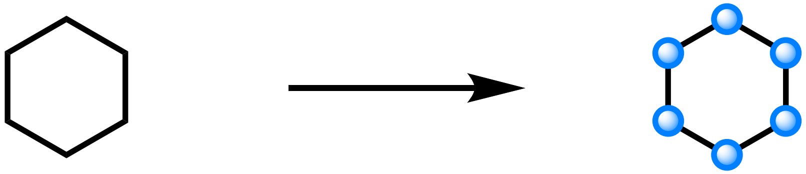
$$\boxed{\begin{aligned} C_T &= C_\eta + C_E \\ C_\eta &= 2\eta \log_2 \eta - \sum_i \eta_i \log_2 \eta_i \\ C_E &= E \log_2 E - \sum_i E_i \log_2 E_i \end{aligned}}$$

η_i : **each number of "same" substructure**

E : **the total number of non-hydrogen atom**

E_i : **each number of "same" non-hydrogen atom**

<Examples>



$$\eta = 6,\ \ E = 6$$

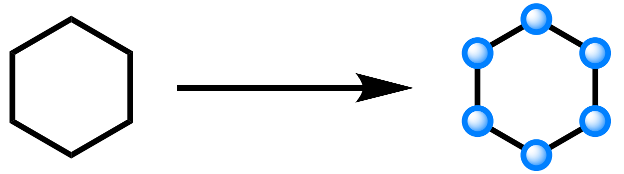
$$C_\eta = 2 \cdot 6 \cdot \log_2 6 - 6 \cdot \log_2 6 = 15.5$$

$$C_E = 6 \cdot \log_2 6 - 6 \cdot \log_2 6 = 0$$

$$C_T = 15.5 + 0 = 15.5$$

1. Bertz, S. H. *J. Am. Chem. Soc.* **1981**, *103*, 3599.
2. Hendrickson, J. B.; Huang, P.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 63.

Bertz Complexity C_T (2)

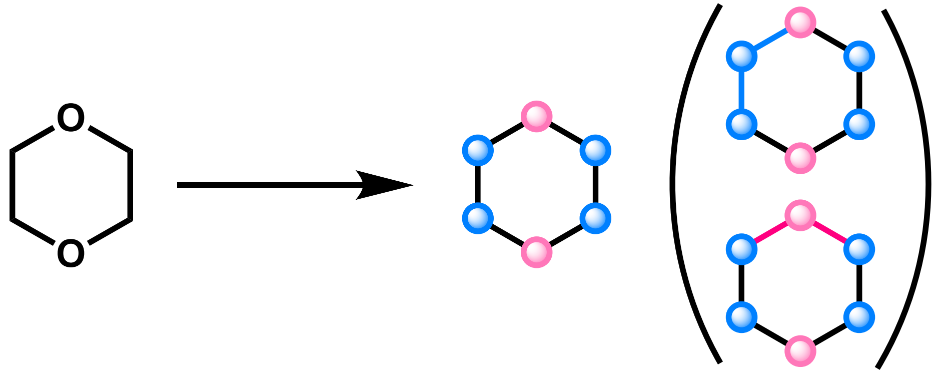


$$\eta = 6,\ \ E = 6$$

$$C_\eta = 2 \cdot 6 \cdot \log_2 6 - 6 \cdot \log_2 6 = 15.5$$

$$C_E = 6 \cdot \log_2 6 - 6 \cdot \log_2 6 = 0$$

$$C_T = 15.5 + 0 = 15.5$$

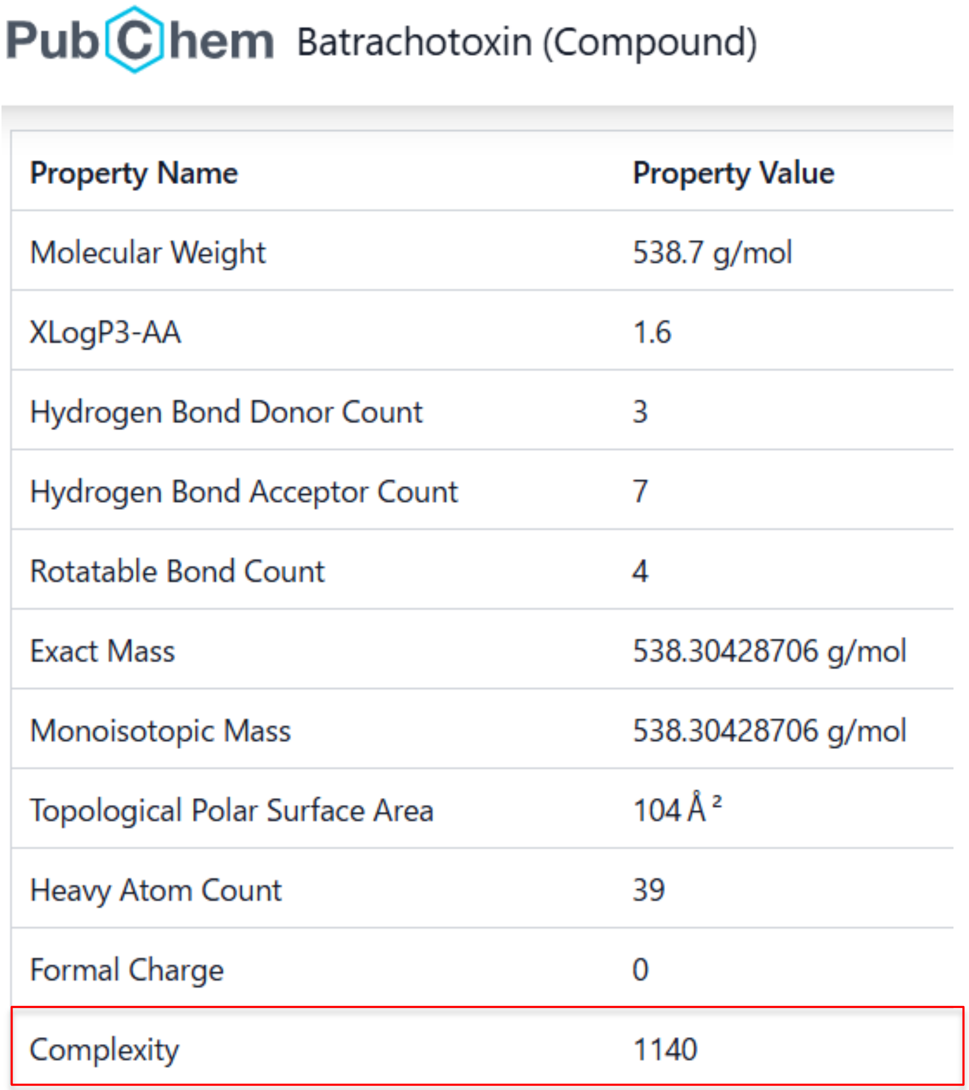


$$\eta = 4 + 2 = 6,\ \ E = 4 + 2 = 6$$

$$C_\eta = 2 \cdot 6 \cdot \log_2 6 - 4 \cdot \log_2 4 - 2 \cdot \log_2 2 = 21.0$$

$$C_E = 6 \cdot \log_2 6 - 4 \cdot \log_2 4 - 2 \cdot \log_2 2 = 5.51$$

$$C_T = 21.0 + 5.51 = 26.5$$



PubChem Batrachotoxin (Compound)

Property Name	Property Value
Molecular Weight	538.7 g/mol
XLogP3-AA	1.6
Hydrogen Bond Donor Count	3
Hydrogen Bond Acceptor Count	7
Rotatable Bond Count	4
Exact Mass	538.30428706 g/mol
Monoisotopic Mass	538.30428706 g/mol
Topological Polar Surface Area	104 Å ²
Heavy Atom Count	39
Formal Charge	0
Complexity	1140

1. Bertz, S. H. *J. Am. Chem. Soc.* **1981**, *103*, 3599.
2. Hendrickson, J. B.; Huang, P.; Toczko, A. G.. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 63.
3. https://pubchem.ncbi.nlm.nih.gov/

Von Korff Complexity

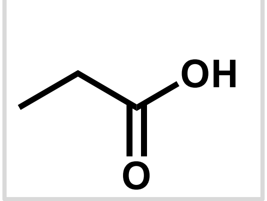
γ : the length of substructure

$N(\gamma)$: the number of substructures whose length is γ.

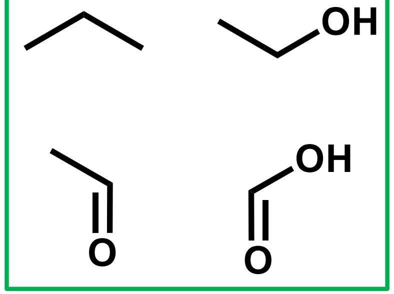
$N(\gamma_{\max})$: the max of $N(\gamma)$.

$$\boxed{\dim(M) = \frac{\log N(\gamma_{\max})}{\log \gamma_{\max}}}$$

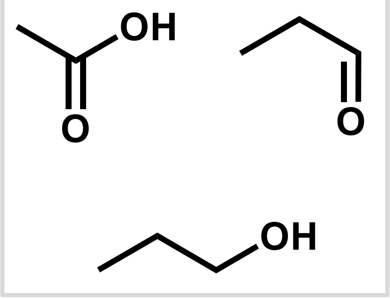
$\gamma = 4$



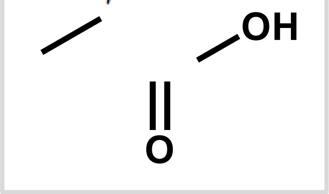
$\gamma = 2$



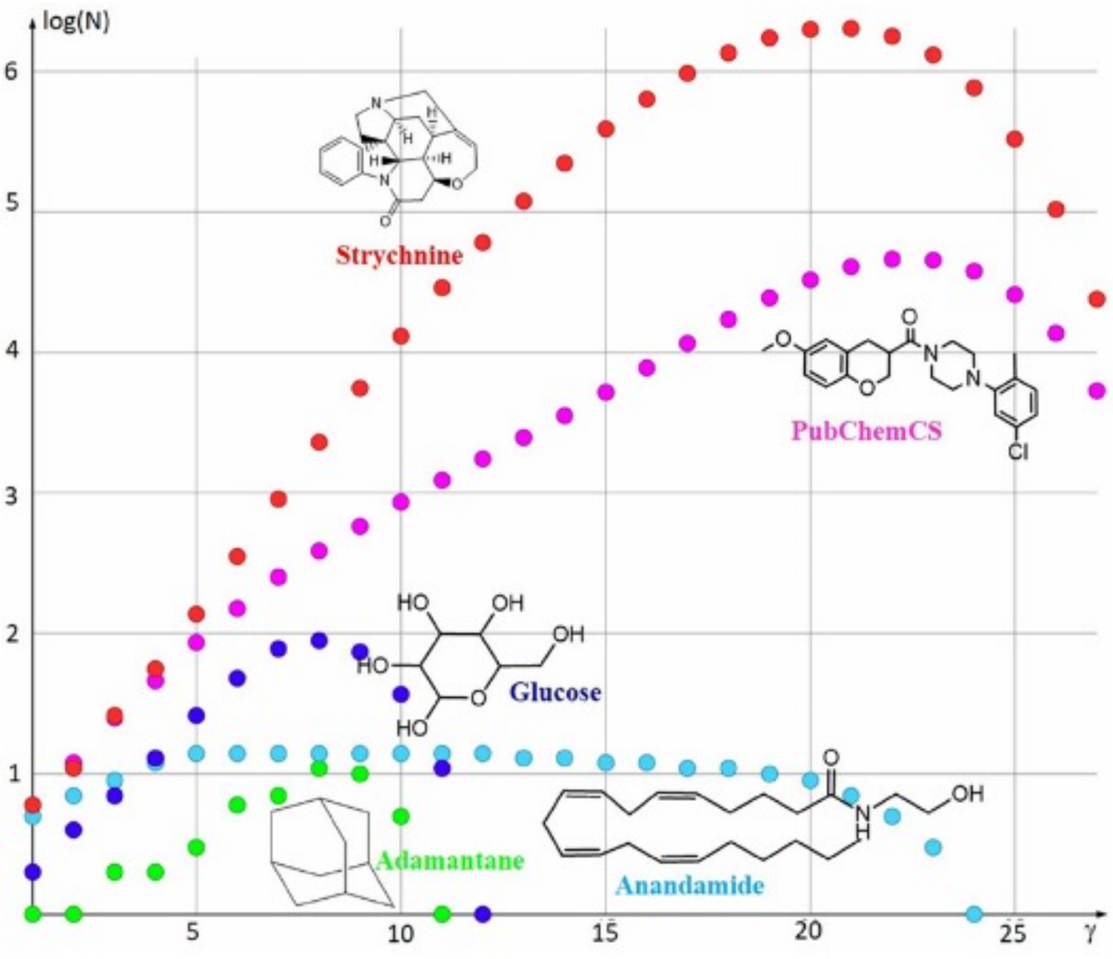
$\gamma = 3$



$\gamma = 1$



$$\frac{\log 4}{\log 2} = 2$$



Name	$\gamma_{\max}$	$N(\gamma_{\max})$	dim
Adamantane	8	11	1.2
PubChemCS	22	45973	3.5
Strychnine	21	2022462	4.8

von Korff, M.; Sander, T. *Scientific Reports* **2019**, *9*, 967.

Proudfoot's Complexity (1)

Proudfoot's approach: Hydrogen including path-based complexity for each atom

→Complexity = the expected value of passing through a certain path

Atom	
#1	H
#6	C
#7	N
#8	O
#9	F
#14	Si
#15	P
#16	S
#17	Cl
#35	Br
#53	I

Connectivity	
X1	1 any type
X2	2 any type
X3	3 any type
X4	4 any type
X5	5 any type
X6	6 any type

non-hydrogen Connectivity	
D1	1 non-H
D2	2 non-H
D3	3 non-H
D4	4 non-H
D5	5 non-H
D6	6 non-H

For each Atom

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

For one molecule

$$C_M = \sum C_A$$

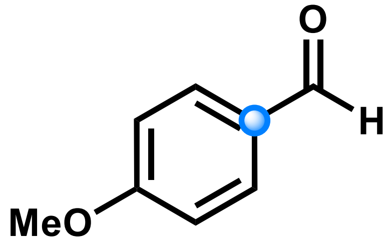
p_i **: the fractional occurrence of each path type**

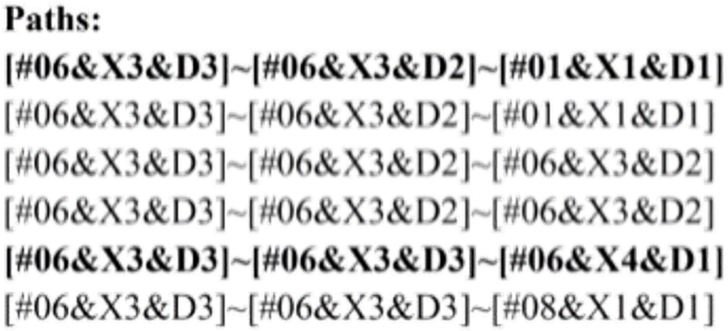
N **: the total number of paths**

Paths: length 2 (in case of terminal elements, length 1)

1. Proudfoot, J. R. *J. Org. Chem*. **2017**, *82*, 6968.
2. Proudfoot, J. R. *Bioorg. Med. Chem. Lett*. **2017**, *82*, 6968.

Proudfoot Complexity (2)





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]

$$C_A = -\left\{\frac{2}{6}\log_2\left(\frac{2}{6}\right)+\frac{2}{6}\log_2\left(\frac{2}{6}\right)+\frac{1}{6}\log_2\left(\frac{1}{6}\right)+\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right\}+\log_2 6$$

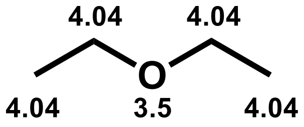
$$= -(-0.53-0.53-0.43-0.43)+2.58$$

$$= 4.5$$

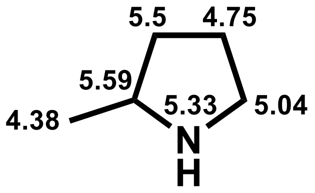
Min = 0, Max = 7.168

<Examples of C_A>

- Compounds

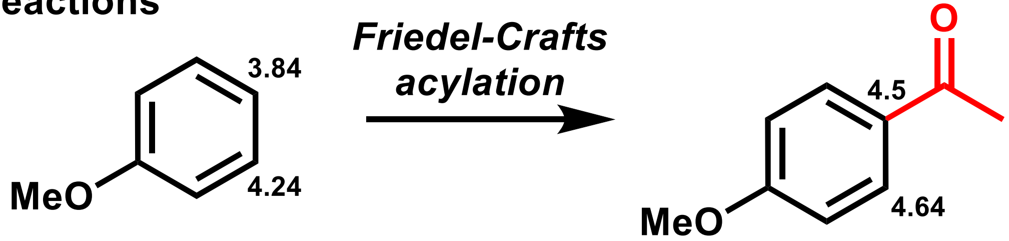


C_M = 19.7

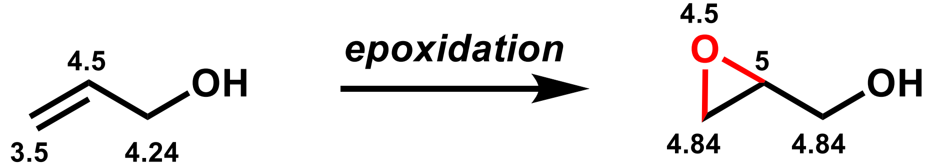


C_M = 30.6

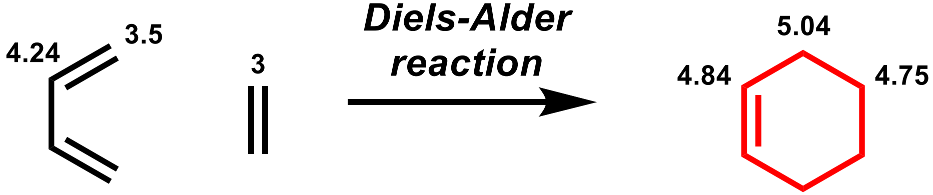
- Reactions



Friedel-Crafts acylation



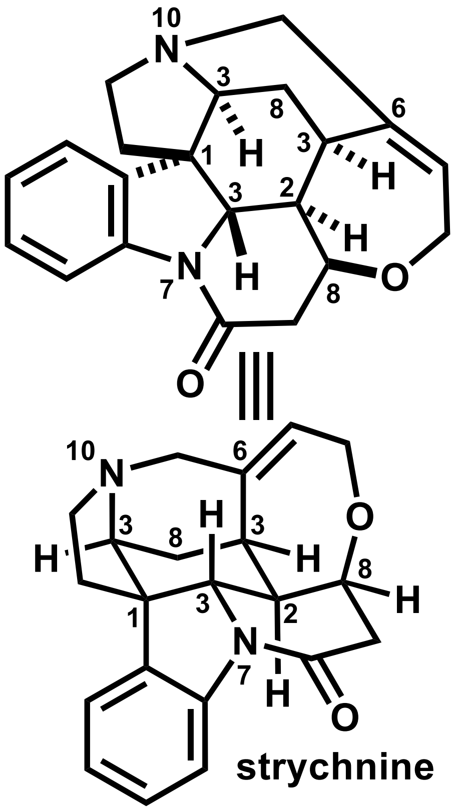
epoxidation



Diels-Alder reaction

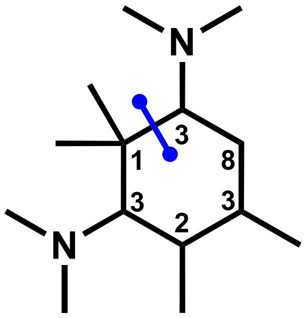
1. Proudfoot, J. R. *J. Org. Chem.* **2017**, *82*, 6968.
2. Proudfoot, J. R. *Bioorg. Med. Chem. Lett.* **2017**, *82*, 6968.

Proudfoot's Complexity (3)

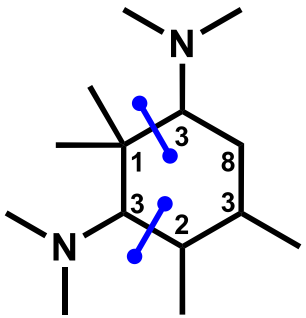


strychnine

rank	C_A
1	6.37
2	5.97
3	5.9
6	5.75
7	5.61
8	5.5
10	5.45



Synthesis	Year	Ring step	Total steps
Magnus, (±)	1992	13	28
Overman, (−)	1993	18	24
Fukuyama, (−)	2004	20	25



Synthesis	Year	Ring step	Total steps
Bodwell, (±)	2002	5	12
Padwa, (±)	2007	4	16
Reissig, (±)	2010	2	9
Vanderwal, (±)	2011	3	6

Due to the path analysis, this score enables to analyze the retrosynthesis based on the complexity of each atom.

1. Proudfoot, J. R. *J. Org. Chem.* **2017**, *82*, 6968.
2. 110625_LS_Kengo_Masuda. 3. 130727_PS_Toshiki_Tabuchi.

Böttcher Score C_m

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.

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

d_i **: the number of chemically nonequivalent bonds**

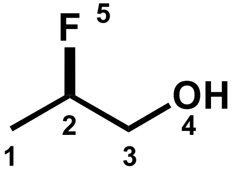
e_i **: the number of different non-hydrogen elements in the bond situation**

s_i **: stereocenters = 2, others = 1**

V_i **: valence electrons**

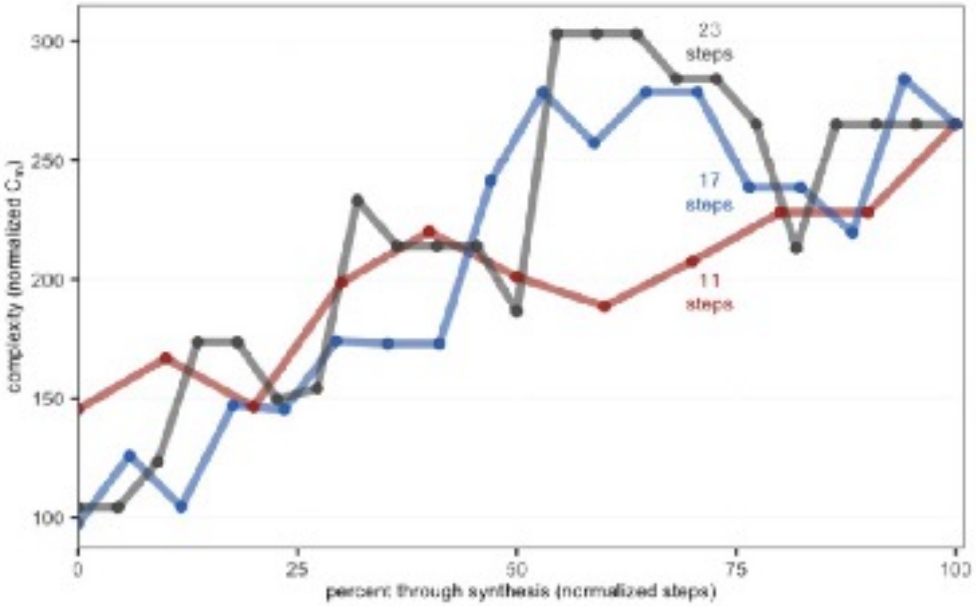
b_i **: total number of bonds**

j **:the corresponding atom positions of chemically equivalent sets of atoms**



i	1	2	3	4	5
d_i	1	3	2	1	1
e_i	1	2	2	1	1
s_i	1	2	1	1	1
V_i	4	4	4	6	7
b_i	1	3	2	1	1
total	2	43.0	12	2.6	2.8

$$C_M = 62.4 \text{ mcbit}$$



1. Böttcher, T. *J. Chem. Inf. Model*. **2016**, *56*, 462. 2. 201121_LS_Masanori_Nagatomo

Waldmann SPS (1)

SPS: Aiming to consider the biologically relevant characteristics of compounds

→The sp^3 richness and the ring system are mainly focused on the factors.

$$\mathrm{SPS} = \sum_i h_i s_i r_i n_i^2$$

$$\mathrm{nSPS} = \frac{1}{a} \sum_i h_i s_i r_i n_i^2$$

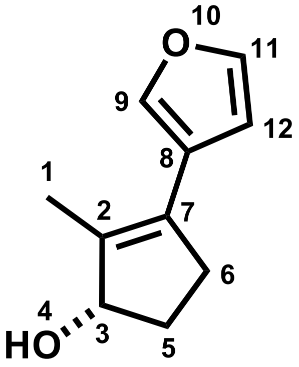
h_i **: equals 3,2, and 1 for sp^3-, sp^2- and sp-hybridized atoms**

s_i **: stereocenters = 2, others = 1**

r_i **: non-aromatic ring = 2, others = 1**

n_i **: the number of non-hydrogen neighbors**

: the number of non-hydrogen atoms



i	h_i	s_i	r_i	n_i	$h_i s_i r_i n_i^2$
1	3	1	1	1	3
2	2	1	2	3	36
3	3	2	2	3	108
4	3	1	1	1	3
5	3	1	2	2	24
6	3	1	2	2	24
7	2	1	2	3	36
8	2	1	1	3	18
9	2	1	1	2	8
10	2	1	1	2	8
11	2	1	1	2	8
12	2	1	1	2	8

$\mathrm{SPS} = 284$

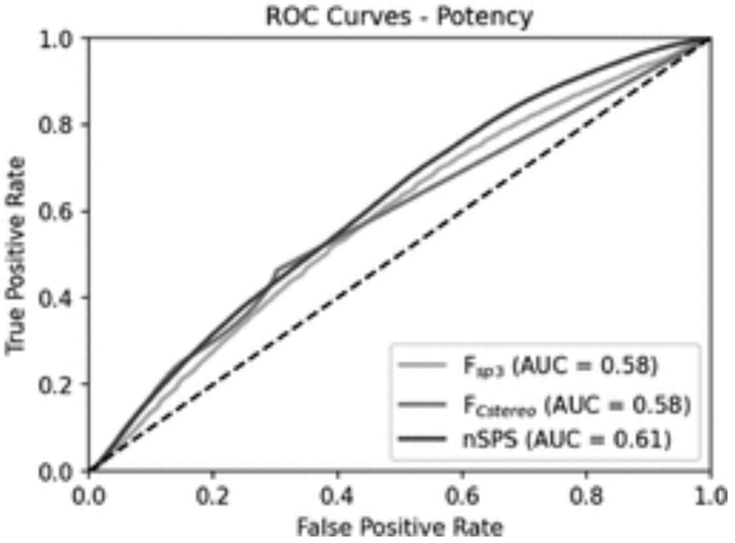
$\mathrm{nSPS} = 23.7$

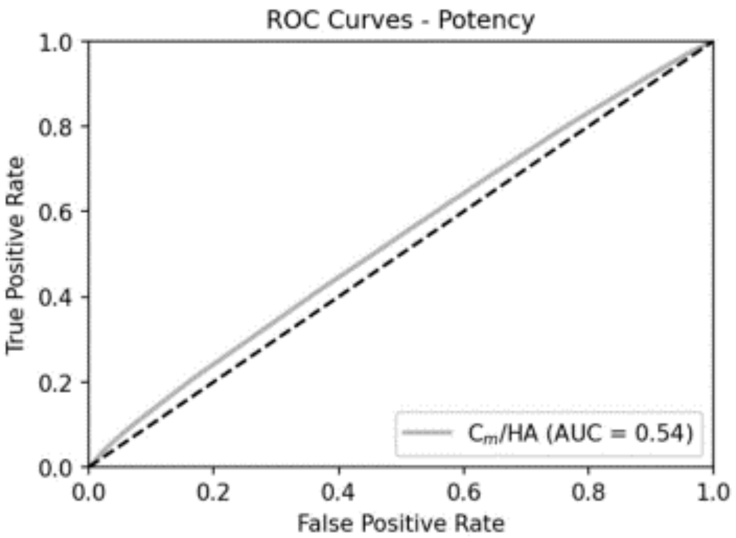
1. Krzyzanowski, A.; Pahl, A.; Grigalunas, M.; Waldmann, H. *J. Med. Chem.* **2023**, *66*, 12739.

SPS (2)

Potency:
the biological activity

If the complexity is correlated to the activity, AUC>0.5.

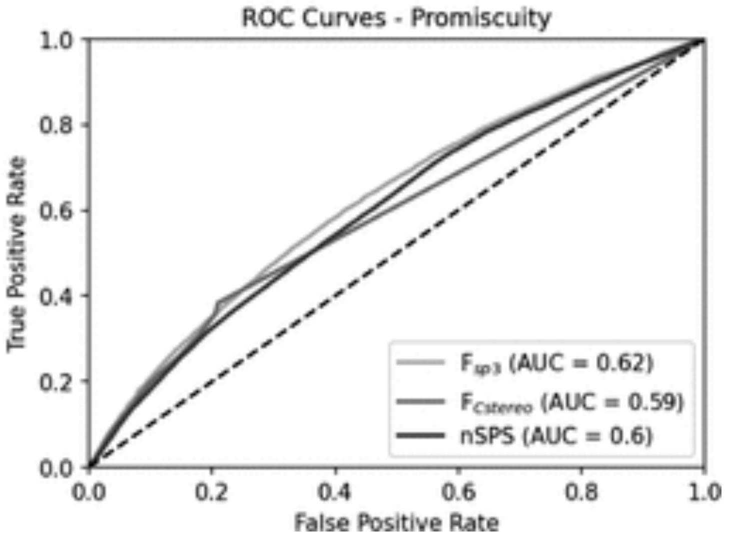


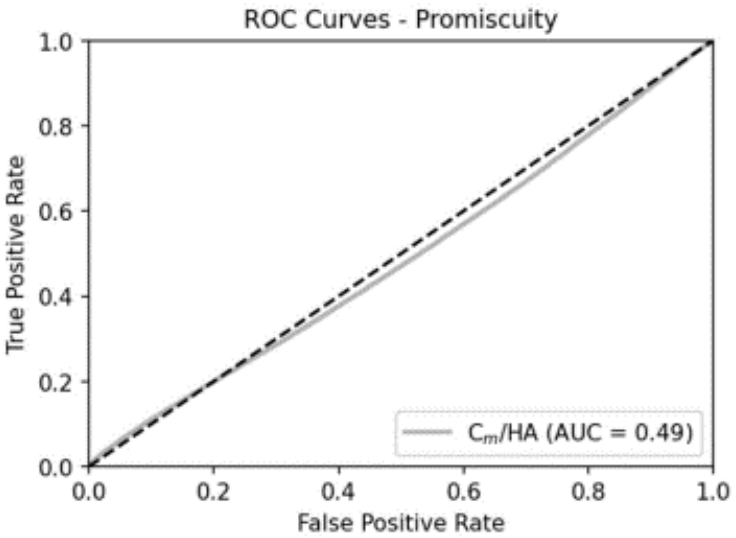


Promiscuity:
the specificity against the target compounds

If the specificity is correlated to the activity, AUC>0.5.

If AUC = 0.5, the index was evaluated randomly.



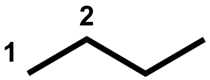
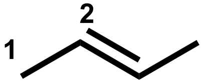


nSPS gave the good potency/promiscuity, while Böttcher score had almost no correlation.

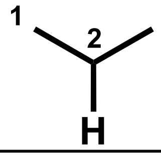
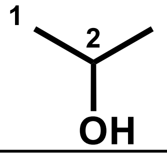
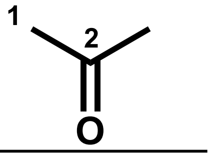
1. Krzyzanowski, A.; Pahl, A.; Grigalunas, M.; Waldmann, H. *J. Med. Chem.* **2023**, *66*, 12739.

Short Summary

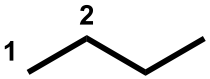
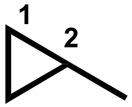
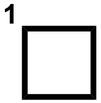
① unsaturation

			
Bertz	2.0	<<	15.2
Böttcher	8.0	<	9.17
SPS	30	>	22
C_M (Proudfoot)	17.9	>	16.1
C_A (Proudfoot)	C1: 4.04 C2: 4.91	> >	C1: 3.69 C2: 4.38

② oxidation/stereochemistry

					
Bertz	0.0	<<	10.8	<<	26.3
Böttcher	5.0	<<	21.5	<	25.2
SPS	18	<<	60	>	24
C_M (Proudfoot)	11.9	<	15.6	>	11.8
C_A (Proudfoot)	C1: 4.04 C2: 3.81	= <	C1: 4.04 C2: 4.06	> >	C1: 3.69 C2: 3.40

③ ring, ④ symmetry

				
Bertz	2.0	<<	19.2	8.0
Böttcher	8.0	<	15.2	6.0
SPS	30	<<	159	72
C_M (Proudfoot)	17.9	<	20.6	18.0
C_A (Proudfoot)	C1: 4.04 C2: 4.91	< <	C1: 5.50 C2: 5.57	C1: 4.50

Cyclobutane: highly symmetric ring
→All scores are decreased.

	Gain	Loss
Bertz	①②	④
Böttcher	②	④
SPS	②③	①
C_M (Proudfoot)	②	①

Contents

1. Development of indexes

JOC The Journal of Organic Chemistry

Molecular Complexity and Retrosynthesis

John R. Proudfoot*

2. Application
    closed to the public