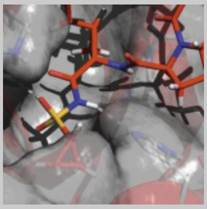
doi:10.2533/chimia.2023.258 Chimia 77 (2023) 258–262 © M. von Korff, T. Sander



Medicinal Chemistry and Chemical Biology Highlights

Division of Medicinal Chemistry and Chemical Biology

A Division of the Swiss Chemical Society

Molecular Complexity for Chemical Reactions

Modest von Korff* and Thomas Sander

*Correspondence: Dr. M. Korff, E-mail: modest.korff@idorsia.com
Idorsia Pharmaceuticals Ltd, CH-4123 Allschwil, Switzerland

Abstract: A new method is presented on how to calculate molecular complexity for chemical reactions by the fractal dimension of reactants and products. Two pathways for the total synthesis of strychnine were compared. Significant differences in the two synthesis pathways were reflected by reaction complexity. These results demonstrate that reaction complexity is a powerful measure to group chemical reactions beyond substructural changes.

Keywords: Molecular complexity · Natural product synthesis · Reactions · Strychnine

Modest von Korff received his MSc in Pharmacy at Julius-Maximilians-Universität, Würzburg, Germany, followed by his PhD in Pharmaceutical Chemistry (supervisor Prof. Siegfried Ebel) at the Department of Pharmaceutical Chemistry, University of Würzburg. He is specialized in software engineering to gather knowledge from big data in life sciences. He has been working as a scientist for drug discovery in several pharmaceutical research companies and since 2017 in Idorsia Pharmaceuticals Ltd.

Thomas Sander earned his diploma and PhD with Prof. R. W. Hoffmann in organic chemistry at Philipps University in Marburg, Germany. He undertook a PostDoc with Prof. J. B. Hendrickson in cheminformatics at Brandeis University in Waltham, Massachusetts. He then joined Roche in 1993 to develop software for drug discovery, moving in 1998 to Actelion as an early start-up to build the drug discovery informatics environment from the ground up. In 2017 he joined Idorsia as spin-off heading drug discovery informatics. Thomas Sander is the creator of OpenChemLib, an open-source Java cheminformatics framework and the main author of DataWarrior, one of the most widely used open-source cheminformatics applications.

Introduction

Molecular complexity has a long history in organic chemistry. With the increase of available molecules for drug discovery and the plethora of reactions for synthesis, calculating molecular complexity becomes an increasingly prominent topic.[1] A standard measure of molecular complexity would allow the analysis, classification, and search for molecules and reactions for complexity. In this contribution, we outline how the idea of molecular complexity evolves up to the point where it will provide useful support in synthetic organic chemistry in the future. The original idea of molecular complexity aimed to understand the living cell. Rashevsky put forward the hypothesis: if we can calculate the complexity of each molecule in a cell, we can calculate the complexity of the cell.[2] Rashevsky was approaching molecular complexity from information theory. Shannon's central formula for information theory (Eqn. (1)) was introduced in 1948 in his manuscript ‘A mathematical theory of communication’.[3]

$$\mathrm{H}(X) = -\sum_{i=1}^{n} P(x_i)\ log_2 P(x_i) \qquad (1)$$

A central concept of Shannon's information theory is the entropy H(X). The entropy is calculated from the characters of a given alphabet with specific probabilities, Eqn. (1), where x is a character in an alphabet with n characters, and P is the probability that this character occurs in a text. Entropy is a familiar term to chemists as a measure of disorder in a system. In information theory, the entropy can be taken as equivalent to the complexity with a negative sign.

Rashevsky proposed to express graph complexity *via* the entropy of an alphabet of graph invariants.[2] Graph invariants are features that are invariant to isomorphisms of the graph, *i.e.,* for a molecular graph, graph invariants are the number of atoms and the number of bonds. Isomorphism of a molecular graph means two different depictions of the same molecule. Without the burden of graph theoretical considerations, chemists used in the same period the term molecular complexity to describe organic molecules.

Robinson wrote the following about strychnine in 1952: “For its molecular size, it is the most complex substance known”.[4] And in 1959, Bradshaw *et al*., wrote that the “Clarification of the biosynthesis of complex organic structures, *e.g.,* terpenoids and alkaloids, is proceeding rapidly”.[5] We state that the concept of molecular complexity was in the minds of organic chemists from the middle of the twentieth century.

In 1954[6] and in 1963,[7] Woodward *et al*. published the first total synthesis of strychnine. Woodward cited the complexity statement from Robinson. A few years later, Corey set the pace for the synthesis of natural products with his manuscript ‘General methods for the construction of complex molecules’.[8] But it took almost two more decades until Bertz described a measurement for molecular complexity.[9] His work relied Rashevsky's approach. The graph invariants chosen by Bertz were developed from the number of ways a pre-defined subgraph could be 'cut-out' of the original molecular graph. Hence, understanding the approach from Bertz is difficult, as was stated by Whitlock who developed his own complexity measurement.[10] Whitlock relied on counting rings, unsaturated bonds, hetero atoms, and stereo centers. In his contribution, Whitlock analyzed and compared the complexity changes for the synthesis of natural products.

Since the publication by Bertz, many approaches for the calculation of molecular complexity have been developed.[11] This included calculating molecular complexity by a model derived from artificial neural network[12] and a crowd-sourcing approach.[13] The importance of molecular complexity for pharmaceutical industry was recently emphasized.[1]

On the Way to our Novel Algorithm

While reviewing the literature about molecular complexity, we remarked that the graph theoretical- and the molecular feature-based approaches possess a similarity. Both approaches need graph invariants. Graph invariants are structural patterns that do not change when a molecule is depicted in different ways. And both approaches have disadvantages. All graph theoretical approaches using Shannon entropy are not size independent. With an increasing number of features the entropy increases. This is also true for the molecular feature-based approaches if they are not normalized, *i.e.*, by the number of non-hydrogen atoms.

Another relevant disadvantage of the pre-defined substructure-based features was already pointed out by Whitlock.[10] He stated that counting substructures is a very insufficient way to calculate molecular complexity, because counting pre-defined substructures neglects the relations between substructures over the topology of the molecule.

Additionally, complexity methods relying on pre-defined substructures will omit new features in molecules, synthesized after the definition of the substructure-based complexity. And many molecular complexity methods apply a weighing scheme to the molecular features. This scheme needs to be calibrated by an arbitrarily chosen dataset. It would be more desirable to have a molecular complexity measure that accounts for the whole molecular topology, without the need of pre-defined substructure patterns and datasets for calibration.

It took us several years to realize that the fractal-dimension from Mandelbrot[14] contained the solution for our quest. Mandelbrot created a mathematical formula to calculate a measurement for the observation that many objects in the real world are made by smaller copies of the object itself. For calculating the fractal dimension, Mandelbrot's term quantifies self-repeating patterns on different scales. His breakthrough example was a map of the coastline of Britain,[14] in which the smallest scale corresponded to the level of highest detail. A central term in Mandelbrot's complexity calculation is 'the highest level of detail', which refers to the state of a self-repeating pattern analysis where the number of distinguishable patterns becomes maximum. For a map, the highest level of detail is the smallest scale.

Our Novel Algorithm

We transferred Mandelbrot's term for the calculation of the fractal dimension of coastlines in two dimensions to graph topology and retrieved a measurement for molecular complexity.[15]

Our algorithm is straightforward. A molecule is decomposed into all possible substructures. Only non-hydrogen atoms are considered. The substructures are grouped by their number of bonds. The group g_{max} with the highest number n_{max} of unique substructures characterizes the highest level of detail in the molecule. For a molecular structure, the scale for the highest level of detail is the bond count γ_{max} where the maximum number of unique substructures n_{max} was extracted from the molecule.

Molecular complexity c results from dividing the logarithm of the group size n_{max} by the logarithm of the bond count γ_{max} for this group. The Java source-code for the complexity calculation is freely available in the OpenChemLib project on GitHub [https://github.com/Actelion/openchemlib]. Eqn. (2) gives the term for calculating molecular complexity by fractal dimension.

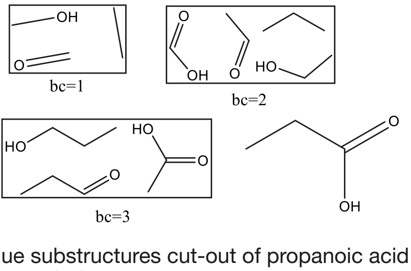
$$c = \frac{log\ n_{max}}{log\ \gamma_{max}} \qquad (2)$$

Propanoic acid may serve as an example (Scheme 1). For one bond, three distinct substructures were cut-out of the molecule. With two bonds, we obtained four distinct subgraphs. Three distinct subgraphs were obtained for three bonds. The highest level of detail, the smallest scale, is two bonds. According to Eqn. (2), the value calculated for propanoic acid is 2, *i.e.*, log(4)/log(2).

Although the calculation is straightforward, the molecule decomposition needs considerable computational resources for complex molecules. If the molecule has a high complexity, according to our definition several million substructures might be extracted. The extracted substructures are graph invariants. The number of unique substructures and, therefore, the number of graph invariants depends on the level of detail in the representation of the molecular graph.

If the graph representation considers stereochemistry, the number of unique substructures is higher with stereochemistry than without. In other words, molecular complexity by fractal dimension considers graph invariants like the majority of approaches. Whereas most current approaches addressing molecular complexity consider graph invariants, our algorithm is the first to include normalization by the number of bonds. The innovation in our algorithm was the observation that the number of bonds for the extracted substructures is the graph-analogue for the scale on a map.

For the complexity given by Mandelbrot, the smallest scale on the map provided the highest level of detail. This was an arbitrary part of the complexity term. In principle, a map scale can go down to the atomic level. In contrast, for subgraph extraction exists a defined highest level of detail. The highest level of detail correlates to the highest number of unique substructures grouped by their bond-counts. For linear alkanes, the highest level of detail is always reached at a bond-count one, with one unique substructure. For bond-count one, the complexity is defined as zero. For strychnine, the bond-count (γ_{max}) at the highest number of unique substructures is twenty-one.



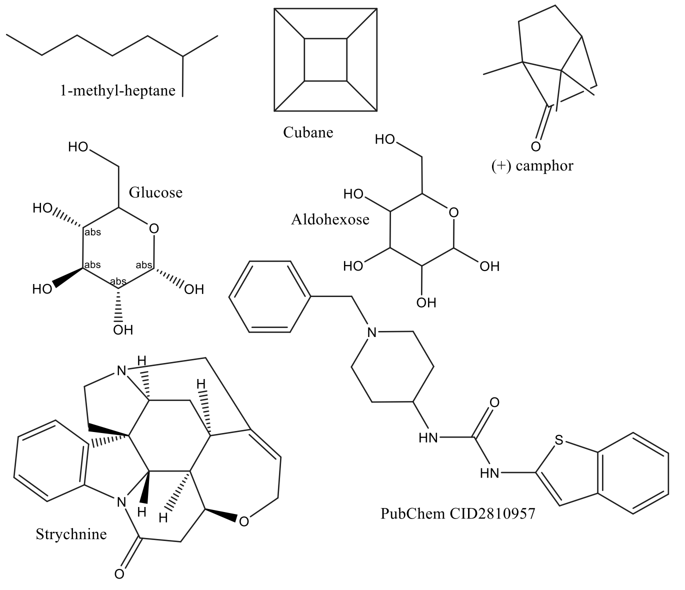
Scheme 1. Unique substructures cut-out of propanoic acid and grouped by their bond counts (bc).

Addressing Molecular Complexity of Molecules Using our Innovative Algorithm

In Table 1 and Scheme. 2, examples are given for molecular complexity by fractal dimension.

Table 1. Example molecules with number of atoms, bonds, number of bonds at highest level of detail (γ_{max}), and corresponding number of unique substructures (n_{max}).

Molecule	Atoms	Bonds	γ_{max}	n_{max}	Complexity
n-octane	8	7	1	1	0.00
1-methyl-heptane	8	7	3	2	0.63
Cubane	8	12	8	14	1.27
Propionic acid	5	4	2	4	2.00
(+) camphor	11	12	8	98	2.21
Aldohexose	12	12	8	89	2.16
Glucose	12	12	8	106	2.24
PubChem CID2810957	26	29	21	3’153	2.65
Strychnine	25	31	21	2’025’643	4.77



Scheme 2. Example molecules for Table 1.

As presented earlier, the simple propionic acid shows a molecular complexity of two. Camphor has a complexity of 2.21, which is slightly higher, even though all organic chemists would agree that camphor is a much more complex molecule than propionic acid. This illustrates the intrinsic relationship between the size of a molecule and its complexity. Fractal dimension yields a complexity measure that normalizes for the size of the molecule.

Aldohexose, the general form for aldehyde containing sugar molecules, has a lower complexity than camphor despite having an additional atom. Glucose, an aldohexose with defined stereochemistry, has a higher complexity than camphor. This shows that stereochemistry is well-considered in the fractal dimension calculation for molecules.

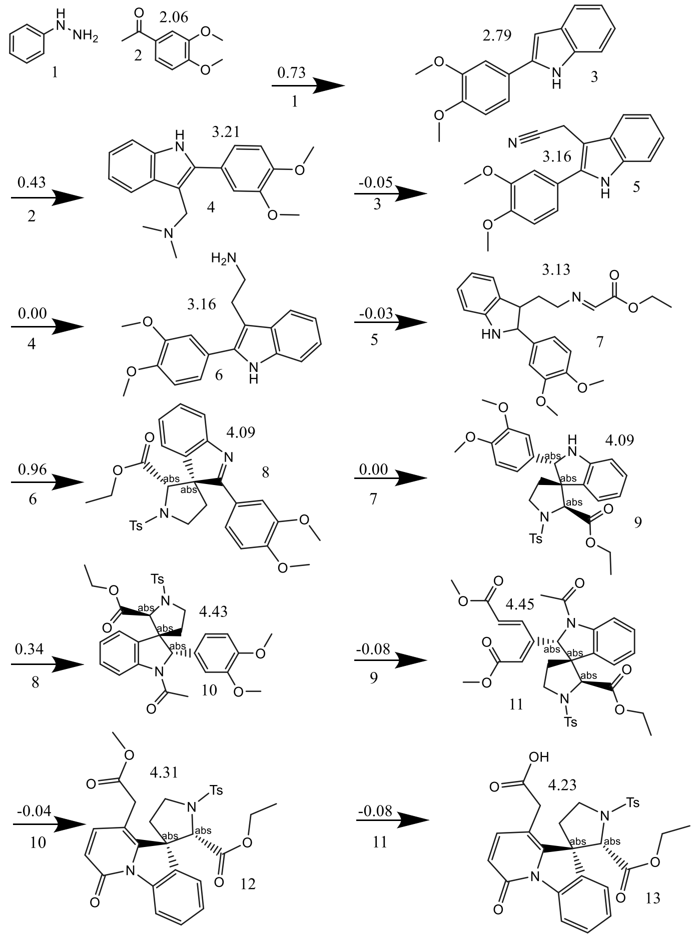
The molecule from PubChem with CID 2810957 was used for high-throughput screening (HTS) and possesses one atom more than strychnine. Both molecules have the highest level of detail at 21 bonds. The difference in the number of substructures for this bond count was striking. The PubChem CID 2810957 molecule has more than three thousand unique substructures, whereas strychnine has over two million substructures. Consequently, the molecular complexity for these two molecules differs a lot, which is reflected in the experience of synthetic organic chemists. Consequently, for molecules of about the same size, the fractal dimension is a valid measure of molecular complexity.

Assessing Molecular Complexity in Synthesis Paths

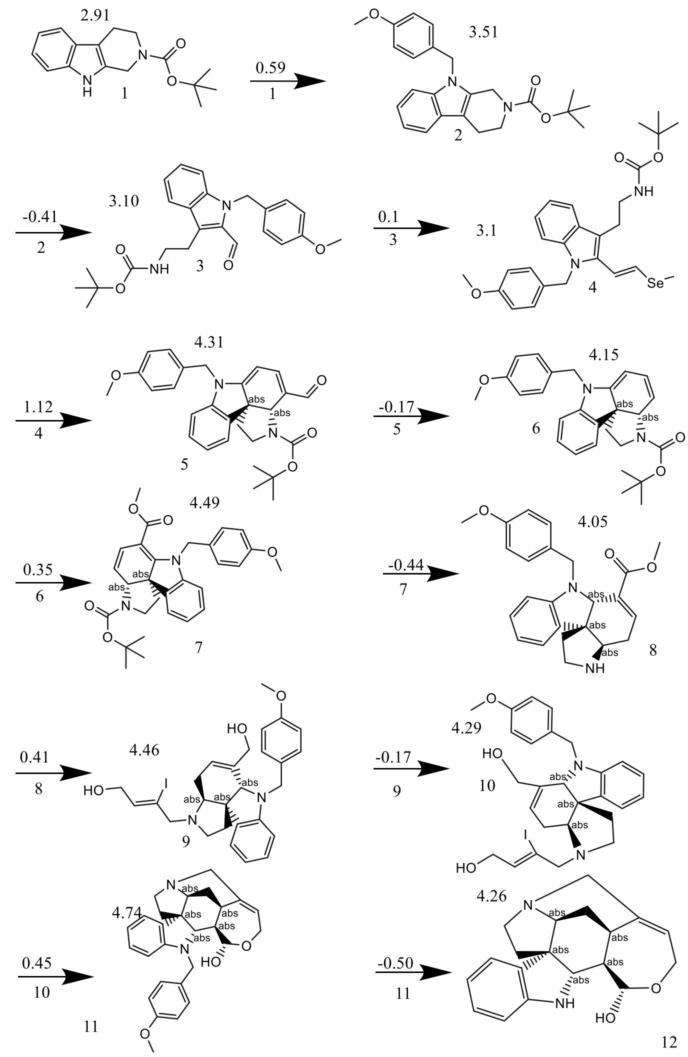
The changes in molecular complexity for two synthesis pathways were calculated using our algorithm for chemical synthesis analysis. Both synthesis pathways depict the total synthesis of strychnine. Woodward *et al.* needed twenty-seven steps to synthesize strychnine[6,7] (Scheme 3). Around fifty years later, MacMillan and coworkers published a manuscript with a much shorter synthesis of thirteen steps (Scheme 4).[16] We looked at the two synthesis routes from a molecular complexity point of view. The change of complexity in a reaction step was calculated by subtracting the complexity of the educt from the complexity of the product. If there were two educts, the complexity of the molecule with higher complexity was taken.

The changes in complexity are summarized in Table 2. MacMillan used a more complex starting material than Woodward (2.91 vs 2.06, respectively). The split into positive and negative complexity changes revealed no large difference between the two syntheses. But it must be considered that for the similar sums of complexity changes quite a different number of synthesis steps were needed. Consequently, the median of absolute complexity changes shows that MacMillan employed reactions that increased or decreased the complexity of the products much more than the reactions applied by Woodward.

Looking at the complexity changes along the synthesis paths provided interesting details (Fig. 1). In all cases except one, MacMillan’s synthesis showed a pattern of increasing complexity followed by a reduction of complexity. Only in reaction steps



Scheme 3. Synthesis route of strychnine by Woodward. The number below the arrow specifies the reaction step. The value above the arrow indicates the difference in molecular complexity in the reaction step between product and educt. The value above the structure depicts the complexity of a molecule. Reaction steps 12–27 are available as supplementary material.



Scheme 4. Synthesis route of strychnine by MacMillan. The number below the arrow specifies the reaction step. The value above the arrow indicates the difference in molecular complexity in the reaction step between product and educt. The value above the structure depicts the complexity of a molecule. The following reaction steps are available as supplementary material.

Table 2. Sums of changes in complexity in strychnine syntheses. Split up by positive and negative complexity changes.

	Strychnine syntheses	
	Woodward (27 steps)	MacMillan (13 steps)
Complexity starting material	2.06	2.91
Sum of positive complexity changes	4.274	3.65
Sum of negative complexity changes	–1.562	–1.794
Total sum (complexity strychnine)	4.77	4.77
Median of changes (absolute values)	0.13	0.41

three and four were there two increasing complexity steps in a row. For Woodward's synthesis, the pattern was similar but more irregular. Looking for reactions with the highest changes in complexity, we find reaction step four for MacMillan and step seven for Woodward. In both cases it was a ring-closing reaction for the central scaffold of strychnine. The large negative changes in complexity were either ring openings or removal of larger groups. A small increase/decrease in complexity indicated the addition/removal of a small group. Complexity changes zero or close to zero occurred through bond order conversion.

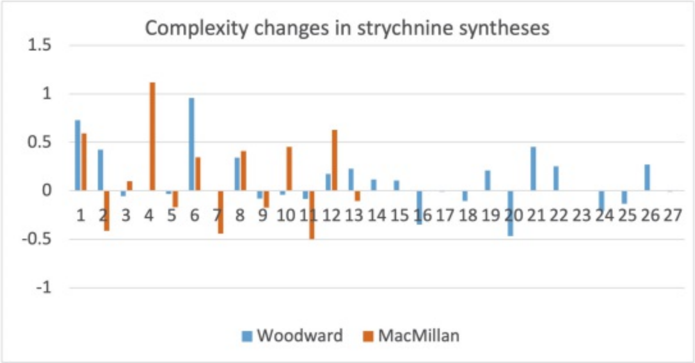


Fig. 1. Complexity changes on in the synthesis paths for strychnine. X-axis = synthesis step as given in Schemes 3 and 4. Y-axis = change in molecular complexity.

Conclusions and Outlook

The work of thousands of chemists in the fifty years between the two strychnine syntheses of Woodward and MacMillan equipped the latter with a toolset of chemical reactions that allowed for a much more concise synthesis pathway than for Woodward. MacMillan exemplified the pathway for modern synthesis with the consecutive pattern of increasing/decreasing complexity reactions.

From the complexity analysis of the two syntheses, and some more not shown here for the sake of brevity, we conclude the following: Molecular complexity by fractal dimension can group chemical reactions by type. Ring closures, ring-openings, fragment addition by cross-coupling reactions, and removal of chemical groups showed significant changes in molecular complexity from educt to product.

Molecular Complexity for Reaction Databases

Adding molecular complexity changes to reaction databases will support the medicinal chemist in the search for specific types of reactions. Molecular complexity allows for a broader search than substructure searches. Searching for reaction types is only possible if the reactions were tagged with appropriate keywords, whereas searching for reaction complexity does not require tagging. This is because the complexity of a chemical reaction is the difference in complexity from product and educt, which is characteristic for a reaction type. So, searching for reaction complexity needs no tagging.

Semiautomatic Synthesis with Molecular Complexity

A more advanced application of our algorithm will be to use reaction complexity for semi-automatic retrosynthesis. Cutting complex molecules automatically to find a possible synthesis route remains a very challenging task for computational algorithms. Employing reaction complexity solves two issues. First, it can find the retrosynthetic cut with the maximum gain of reaction complexity. Second, this reaction complexity can be used to perform a fully automated search for a set of reactions with the same or similar reaction complexity. This selection would be presented to the medicinal chemist, who knows much better than any software if one of the reactions is promising enough to be tried. If the chemist rejects the presented selection, the algorithm simply processes the next best retrosynthetic cut until a feasible synthesis pathway is found.

Supplementary Information

Supplementary information for this publication is available on https://chimia.ch/article/view/2023.258.

Author Contributions

T. Sander designed and implemented most of the source code. M. Korff developed the mathematical term for molecular complexity by fractal dimension and wrote the manuscript.

Acknowledgements

Joel Wahl and Thomas Liphardt for fruitful discussions.

Received: February 16, 2023

[1] S. Caille, S. Cui, M. M. Faul, S. M. Mennen, J. S. Tedrow, S. D. Walker, *J. Org. Chem*. **2019**, *84*, 4583, https://doi.org/10.1021/acs.joc.9b00735.

[2] N. Rashevsky, *Bull. Math. Biophys*. **1955**, *17*, 229, https://doi.org/10.1007/BF02477860.

[3] C. E. Shannon, *Bell Sys. Techn. J.* **1948**, *27*, 379, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[4] R. Robinson, 'Molecular structure of strychnine, brucine and vomicine', Vol. 1, Butterworths, London, **1952**.

[5] W. H. Bradshaw, H. E. Conrad, E. J. Corey, I. C. Gunsalus, D. Lednicer, *J. Am. Chem. Soc*. **1959**, *81*, 5507, https://doi.org/10.1021/ja01529a060.

[6] R. B. Woodward, M. P. Cava, W. D. Ollis, A. Hunger, H. U. Daeniker, K. Schenker, *J. Am. Chem. Soc.* **1954**, *76*, 4749, https://doi.org/10.1021/ja01647a088.

[7] R. B. Woodward, M. P. Cava, W. D. Ollis, A. Hunger, H. U. Daeniker, K. Schenker, *Tetrahedron* **1963**, *19*, 247, https://doi.org/.1016/S0040-4020(01)98529-1.

[8] E. J. Corey, *Pure Appl. Chem*. **1967**, *14*, 19, https://doi.org/10.1351/pac196714010019.

[9] S. H. Bertz, *J. Am. Chem. Soc*. **1981**, *103*, 3599, https://doi.org/10.1021/ja00402a071.

[10] H. W. Whitlock, *J. Org. Chem*. **1998**, *63*, 7982, https://doi.org/10.1021/jo9814546.

[11] O. Mendez-Lucio, J. L. Medina-Franco, *Drug Discov. Today* **2017**, *22*, 120, https://doi.org/10.1016/j.drudis.2016.08.009.

[12] C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, *J. Chem. Inf. Model.* **2018**, *58*, 252, https://doi.org/10.1021/acs.jcim.7b00622.

[13] R. P. Sheridan, N. Zorn, E. C. Sherer, L. C. Campeau, C. Z. Chang, J. Cumming, M. L. Maddess, P. G. Nantermet, C. J. Sinz, P. D. O'Shea, *J. Chem. Inf. Model*. **2014**, *54*, 1604, https://doi.org/10.1021/ci5001778.

[14] B. Mandelbrot, *Science* **1967**, *156*, 636, https://doi.org/10.1126/science.156.3775.636.

[15] M. von Korff, T. Sander, *Sci. Rep*. **2019**, *9*, 967, https://doi.org/10.1038/s41598-018-37253-8.

[16] S. B. Jones, B. Simmons, A. Mastracchio, D. W. MacMillan, *Nature* **2011**, *475*, 183, https://doi.org/10.1038/nature10232.