Current Complexity: A Tool for Assessing the Complexity of Organic Molecules

Journal:	*Organic & Biomolecular Chemistry*
Manuscript ID:	OB-ART-04-2015-000709.R1
Article Type:	Paper
Date Submitted by the Author:	29-Apr-2015
Complete List of Authors:	Li, Jun; Bristol-Myers Squibb, Chemical Development Eastgate, Martin; Bristol-Myers Squibb, Chemical Development

Current Complexity: A Tool for Assessing the Complexity of Organic Molecules

Jun Li and Martin D. Eastgate*

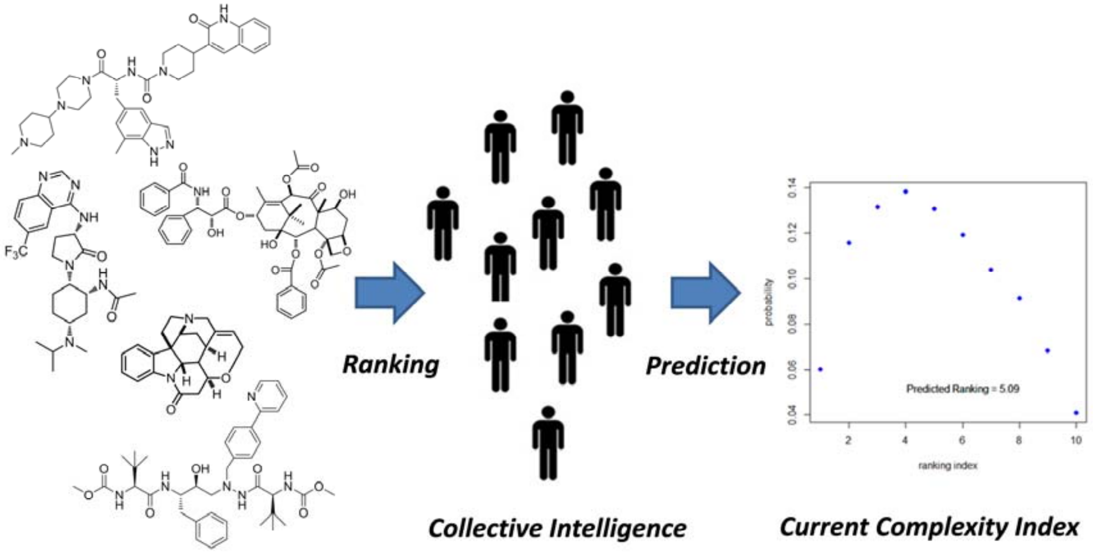
Chemical Development, Bristol-Myers Squibb, 1 Squibb Drive, New Brunswick, NJ, 08903 (USA)

Fax: (+1-732-227-5148)

E-mail: martin.eastgate@bms.com

ABSTRACT: Molecular complexity for a synthetic organic chemist is difficult to define, though intuitively known. Despite the importance of this concept, the quantitative assessment of complexity within organic chemistry has remained a challenge. We report here on the development of an approach for generating a unique complexity index, which is reflective of both intrinsic molecular complexity and extrinsic synthetic complexity. This index is based on a community's perception of complexity, within the context of current technology, calculating a molecules *current complexity*. Our approach allows for a direct comparison between molecules, the analysis of trends within research programs, it also enables an assessment (and comparison) of new synthetic approaches to known molecules and is capable of following a molecules apparent complexity as it changes over time.

TOC Graphic:



A new complexity index is described, leveraging crowd-sourced knowledge to assess a molecules complexity in the context of current technology.

Introduction

With aspects of art and science, organic synthesis is a field focused on the synthesis of molecules, both natural and designed.[1] In the context of *synthetic* organic chemistry, the assessment of a molecules 'complexity' (a paraphrase for the challenge of synthesizing the molecule), is a multi-faceted problem, open for interpretation and challenging to quantify.[2] The comparison of two molecules (in terms of their apparent complexity) is akin to comparing a Van Gogh to a da Vinci, difficult to define and highly dependent on individual bias. For this reason, assessing the 'complexity' of an organic compound is extremely challenging, having components of both rigorous fact and our *perception* of the challenge. A realistic measure of complexity therefore requires an understanding of how the chemical community *perceives* the complexity of organic molecules.

Since the pioneering work of Bertz,[3a] Bonchev,[3c] and Randić,[3d] structure based topology indices have been used for assessing molecular complexity.[3] These approaches leverage graph theory[4] to quantitatively calculate the absolute complexity (via connectivity) of the system, representing an 'architectural' assessment of a molecule. Other complexity measures have attempted to quantify synthetic 'accessibility' as a derivative of complexity.[5] These approaches are often used to compare virtual molecules, such as the output of *in silico* drug discovery, to predict their 'accessibility' or ease of synthesis *before* attempting to prepare them. While this is an important application, the true complexity of a molecule, as experienced by synthetic chemists, is often an unpredictable attribute of the interplay between structure and reactivity. It is this interplay where the true complexity and unpredictability of organic molecules is to be found.

While computational methods to predict a viable synthesis are improving, assessing accessibility without demonstration remains challenging.

Graph theory, while mathematically accurate, fails to account for the practical view of organic chemists; said another way, graph theory alone cannot assess the complexity experienced in reduction to practice (synthesis). For example, if two systems have approximately the same score in a structural index, yet one is much easier to prepare, how can the perception of complexity be similar for those engaged in the synthesis of the compounds? Limitations in calculating 'accessibility' also exist and can be traced to the challenge of predicting the impact of structural, electronic and steric effects, heteroatom substitution and the many other issues found during reduction to practice (synthesis). However, while many of these concepts are intuitively included by experienced chemists when comparing molecules, they are challenging to capture in a mathematical assessment of the structures alone. Some approaches to solving these problems have been explored, but with limited success.[5g]

An additional concept, common in graph theory, is the notion that 'complexity' is fixed. This is true in terms of the structure itself (the molecule never changes), but does this notion apply to those involved in chemical synthesis? In most areas of physical science, the challenges of the field change over time. The coherent study of a system inevitably leads to an improved ability to understand, modify and replicate that system. In synthetic organic chemistry, this study produces improved synthetic strategies, new chemical reactions, better methodologies and innovation in their application, all improving our ability to prepare molecules. The discovery of reactions such as Diels-Alder (1920's), Wittig (1950), Ring-Closing Metathesis (RCM, 1980's) and boron mediated aldols (1990's) significantly impact our capability. These discoveries enable the synthesis of complex systems such as Isochrysohermidin, Manzamine A and Swinholide A, with efficiencies unimaginable before their discovery/invention.[1] Such shifts in the understanding of chemical reactivity change our very reference for what constitutes complexity.[1c] *Thus: the impact of complexity in organic chemistry is time dependent*. Measuring the complexity of an organic molecule needs to be done in the context of current technology. It should be comprised of both *intrinsic* factors (fixed) and *extrinsic* factors (variable), eq 1, and should therefore change over time.[2a]

Strychnine is a classic example of this change: After being isolated in 1818, the structure was finally solved after nearly 130 years (complexity in analysis). Following this advance, Woodwards land-mark synthesis (accomplished in 29 steps) was completed in 1954.[6] In stark contrast, Vanderwalls recent approach required only 6 steps,[7] with new complex intermediates being structurally elucidated in hours, not centuries. The tools to accomplish total synthesis have vastly improved – we now have the capability to prepare compounds such as Halaven, a commercial drug supplied via total synthesis.[8] Strychnine has not changed since 1818, but chemical technology has, and the significant changes must surely impact how we *percieve* the challenge (complexity) of such a molecule *vs* the perception of the challenge that Woodward faced in 1954. Therefore, to be relevant, complexity indices need to reflect the advances in the field and the changes in our capability and perception.

Due to the challenges and concepts outlined above, we considered an alternate approach to developing a complexity index, based on an understanding of how we perceive complexity as a community of chemists. Herein we describe a new protocol for assessing the impact of structural complexity in the context of current technology. This index is empowered by probabilistic modelling, leverages a population's opinion of complexity and incorporates both molecular complexity and synthetic accessibility. This approach can be used in a predictive fashion through combination with route design software and can track changes in a molecule's '*current complexity*' as technology (the synthesis) improves.

$$CurrentComplexity \propto \int_{Fixed} IntrinsicComplexity + \int_{Today} ExtrinsicComplexity$$

Eq 1. *Complexity Postulate*

Results and Discussion

We considered several approaches for developing a quantitative assessment of complexity. As outlined above, we wanted to develop a protocol based on interpreting our perceptions of complexity within the context of both the molecules structure and the technology available. Thus, the method needed to include fixed factors along with variables; however, assessing which factors were important to the chemical community in defining 'complexity' required data.

Pooled group analysis of individual responses has been used in several settings to develop models of complex systems related to human behaviour and perception.[5g-i] The advantages, challenges, methods and biases associated with using human judgement as the ultimate measurement instrument have been well documented.[9] We felt that this approach offered several advantages, such as leveraging expert experience to assess electronic or steric effects, but also had some disadvantages. One such limitation is in using grouped assessments (ie categorized ratings, such as selecting a response on a scale from 1-10), which can result in the loss of subtle differences between molecules and across individuals. We therefore proposed that leveraging real experience (based on personal opinion), along with forced ranking and a Bayesian statistical interpretation of the data, could address many of these concerns. This approach would produce a distribution of *views* (opinions) regarding the complexity of the molecules in the data set. This distribution would reflect the communities varying view of complexity, akin to a collective intelligence of what truly contributes to an assessment of molecular complexity.

In order to test this approach to developing a complexity index, and to provide the initial data-set for analysis, a group of 18 experienced synthetic chemists were asked to force-rank a diverse list of 40 molecules in order of *perceived* complexity (1-40, no duplicates). To enable the proof of concept for this methodology, we limited the initial size of the rating group. We expected that agreement on ‘simple’ and ‘complex’ systems was more likely, with greater variability for compounds in the middle;[5h] this was confirmed (a box plot analysis is shown, figure 1). This variability (distribution of opinions) contains a wealth of information regarding the knowledge of the group, the individual scientists understanding of chemical reactivity and the different opinions as to the weighting given to any individual attribute of perceived complexity.

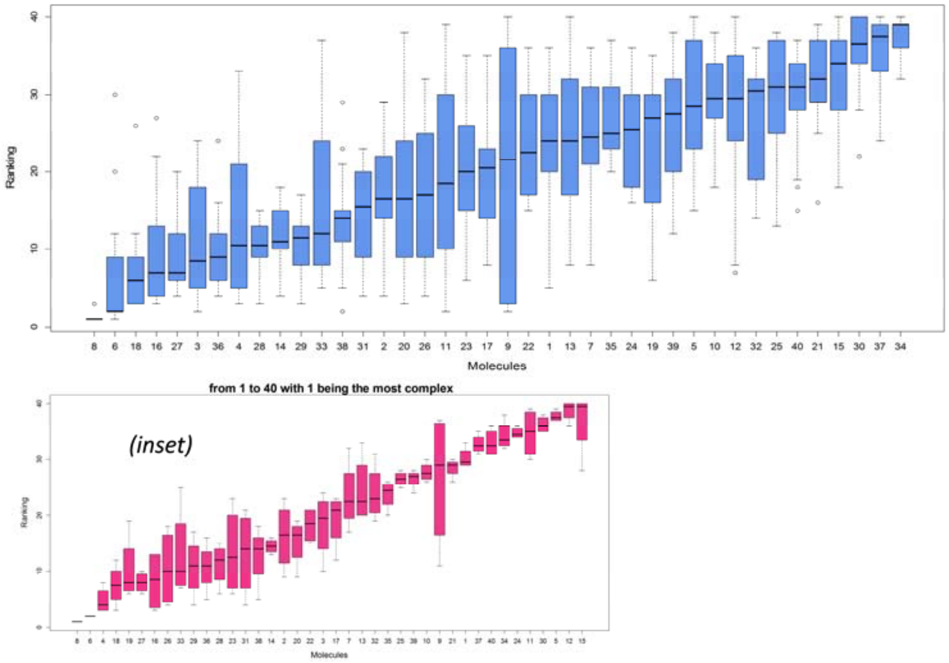


Figure 1: Chemists force ranking of a 40 molecule training set. A score of 1 represented the most complex molecule, 40 the least.

These variations in perceived complexity are important, reflecting the *collective intelligence* of the group. No synthetic information was given to the main rating group. However, all relevant synthetic information was given to a subset of individuals (in-set). This group showed significantly reduced variability in their perception of the molecules complexity – giving credence to our thesis that technology (the synthesis) impacts our perception of complexity.

The variability in the ratings given by various individuals could fall into two main categories: i) differences in perceived significance of an individual factor (ie the importance of a stereogenic center to the complexity of the molecule); ii) aggregated unmeasured and unmeasurable idiosyncrasies based on individual experience (knowledge). This variability is an essential part of our analysis (*vide infra*). *The distribution of views reflects the range of experience of the group* (ie a collective intelligence) and enables the analysis of difficult to predict factors (such as electronic or steric challenges) which may be predicted based on an individual's prior experience.

Although there are many approaches for understanding data sets, a probabilistic approach cognizant of uncertainty can offer a significant advantage.[10] These approaches have been utilized in several applications related to human rating systems.[11] Using the ranking data, we hoped to determine the underlying factors influencing the perceptions of complexity. With the factors identified, we could leverage the data to generate a predicted distribution of the views from within the group (ie a probability distribution would reflect how a group of individuals would respond when asked to rate the complexity of a given molecule).[11,12] Using this probabilistic approach the *intrinsic* and *extrinsic* influences on the cumulative rankings can be balanced with each individuals bias, to build an assessment based on the collective intelligence of the group (figure 2).

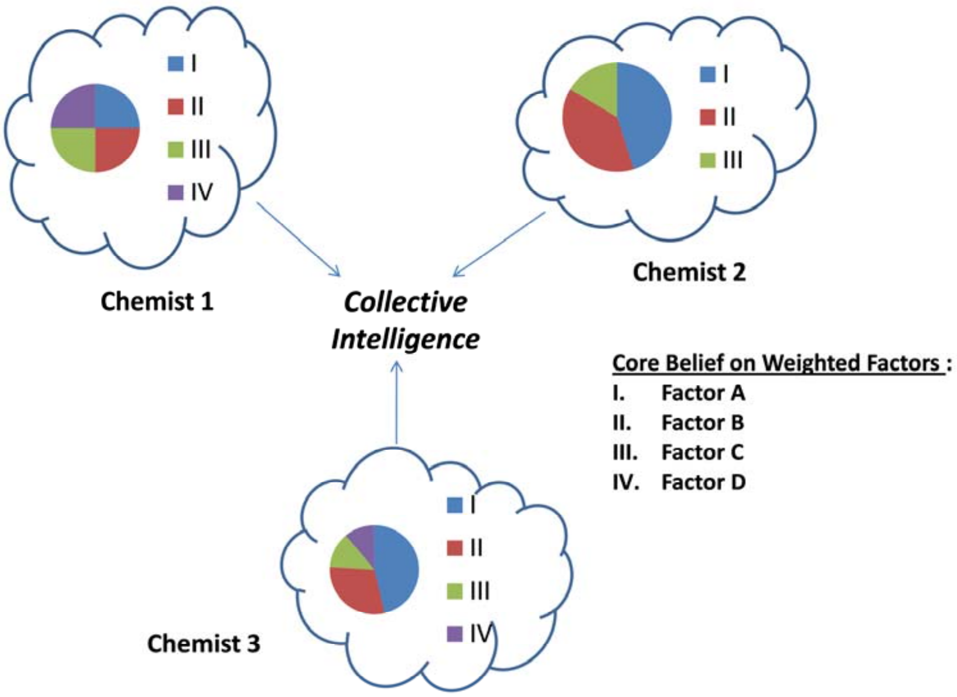


Figure 2: Schematic representation of the ranking process using collective intelligence. Different chemists rate a molecule with different significance given to various influencing factors (I-IV). The probabilistic analysis of this distribution results in collective intelligence.

To develop the model we needed to establish which factors were significant to the data. We considered many *extrinsic* factors reflective of reduction to practice (ie the synthetic route for organic molecules). Some of the factors considered were ideality, developed by Baran (reflective of the ‘complexity’ in the synthesis);[13] step count (how effective current technology is for the given system); and yield. Other factors such as reactivity, stability, and physical characteristics were found to be less important and therefore not included. To quantify *intrinsic* complexity, quantitative structural information from the molecular network (graph theory indices, such as Bertz, Randic, Zagreb etc) were investigated.[14] Since these indices do not include heteroatoms or stereochemical information, a number of additional *intrinsic* variables were considered (ie number of stereogenic centers, heteroatoms, unsaturation, aryl heterocycle fragment prevalence, etc).

These intrinsic and extrinsic complexity factors were evaluated against the training data using a regression subset selection approach;[15] determining the combination of underlying factors which contributed to the observed rankings – using the group data to select which factors were important to the raters perception of complexity. These factors were then verified using a Bayesian regression model. We found that five major factors impacted the complexity ranking (Eq 2): i) a molecular topological index (Randic); ii) the number of stereogenic centers established during the synthesis; iii) number of heteroatoms *on and in* aromatic rings; iv) number of steps; v) ideality of the route. The 1st and 3rd factors are related to *intrinsic* complexity, are unchangeable and reflect the molecule itself. The remaining *extrinsic* factors will vary over time as technology (synthesis) improves.

$$\mu = \beta_0 + \beta^{i}_{Randic} x_{Randic} + \beta^{e}_{SS} x_{ss} + \beta^{i}_{HAA} x_{HAA} + \beta^{e}_{Steps} x_{Steps} + \beta^{e}_{Ideality} x_{Ideality} + \varepsilon$$

Eq 2. Regression model used in the Current Complexity index. 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. The expanded equation is available in the supporting information.

With the main factors identified we developed a probability based model that could reproduce the data from the training set, employing the factors outlined above. The goal was to establish a rating system based on an indexed scale (1-10), we therefore segmented the molecules in the original assessment into 10 groups based on score, but *maintaining* the original rating data. In order to link the ordinal ranking values to the postulated underlying factors, an ordinal probit regression was established where the linear regression output was mapped to the ordinal value via a cumulative normal density link function.[11,12] The Bayesian inference of this regression model then reallocates credibility across the model parameters to be consistent with the observed data. This generates the *probability* of a molecule being in each of the 10 classes of molecular complexity (the indexed complexity grouping), reflecting a predicted response distribution of an actual group of chemists. For example, two molecules from the training list are shown (figure 3).

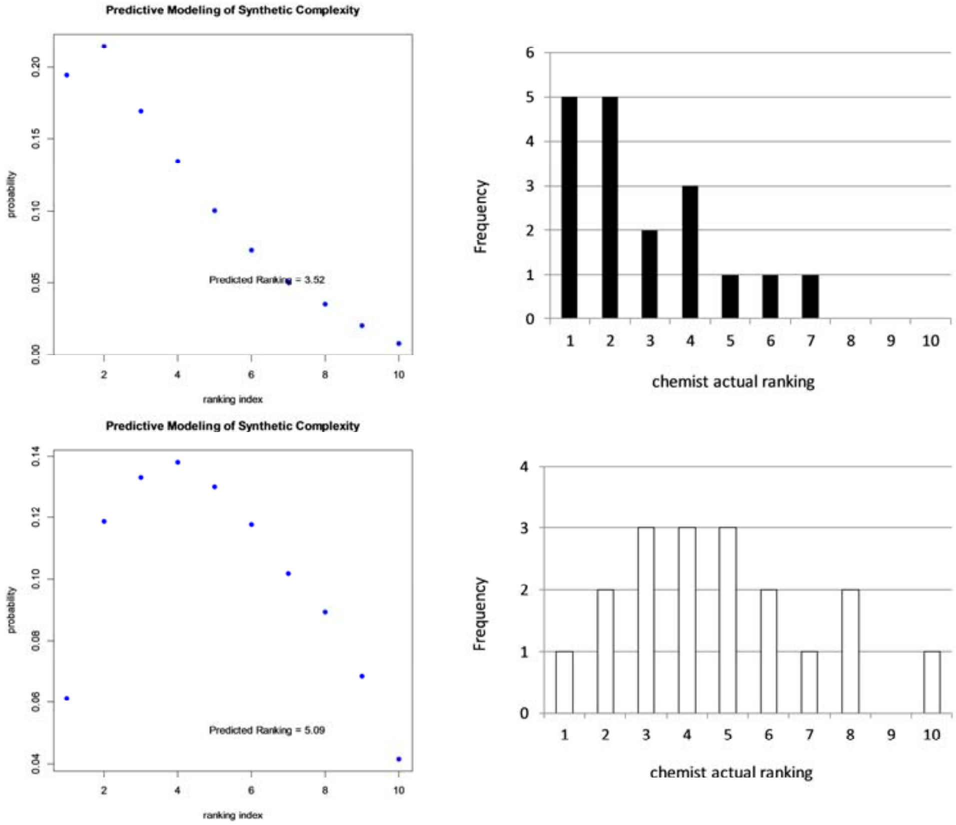


Figure 3. Predictive probability distributions (LHS) versus Chemist actual ranking distributions (RHS) for two representative molecules.

As can be seen the distribution of actual ratings from the group of chemists (RHS bar-charts) compares favourably to the predicted probability distributions from the model (LHS plots). These probability distributions can be considered ‘finger-prints’ of the molecule – they are unique to that system. The summed weighted average of the probability distribution then gives a non-integer index score, 1-10 (1 being the most complex), reflecting the molecules *‘current complexity.’*

In order to compare our model to other approaches, we assessed the fit of the model against the chemist's data. The correlation coefficient (between the models weighted prediction and the actual chemist rankings), is 0.84 (p=1.65E-11, i.e. the probability is less than 1.65E-11 that the correlation score is due solely to chance). This clearly indicates that our model provides an excellent correlation between the calculated view of molecular complexity and the view of a community of trained synthetic organic chemists. With the factors identified, the initial model developed, and a validation of the approach, we could interrogate new molecules in a dispassionate (automated) fashion.

To illustrate the usefulness of this model, we tested its application in a few settings. In characterizing changes in a systems complexity over time, as synthesis improves, we chose to test Strychnine (figure 4).[16] For this molecule, technology has improved in moving from Woodward's landmark synthesis[6] to Overman's,[17] to MacMillan's,[18] and to Vanderwals (the most recent) synthesis.[7] The current complexity of the four approaches were evaluated using the methodology outlined below:

1. Molecular topological index (Randic) of the target molecule is calculated using the encoded SMILES molecular structure notation.
2. Number of stereogenic centers established in the target molecule during the synthesis is enumerated.
3. Number of heteroatoms *on and in* any aromatic rings of the target molecule are enumerated.
4. Number of synthetic steps is calculated as a summation of the number of steps in the longest-linear sequence, with 50% of total number of steps in the branching

sequences. This helps differentiate convergent from linear syntheses. Additionally, any chiral separations are treated as a separate step.

5. Ideality is calculated as follows, using the numerical expression and definitions proposed by Baran:[13]

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

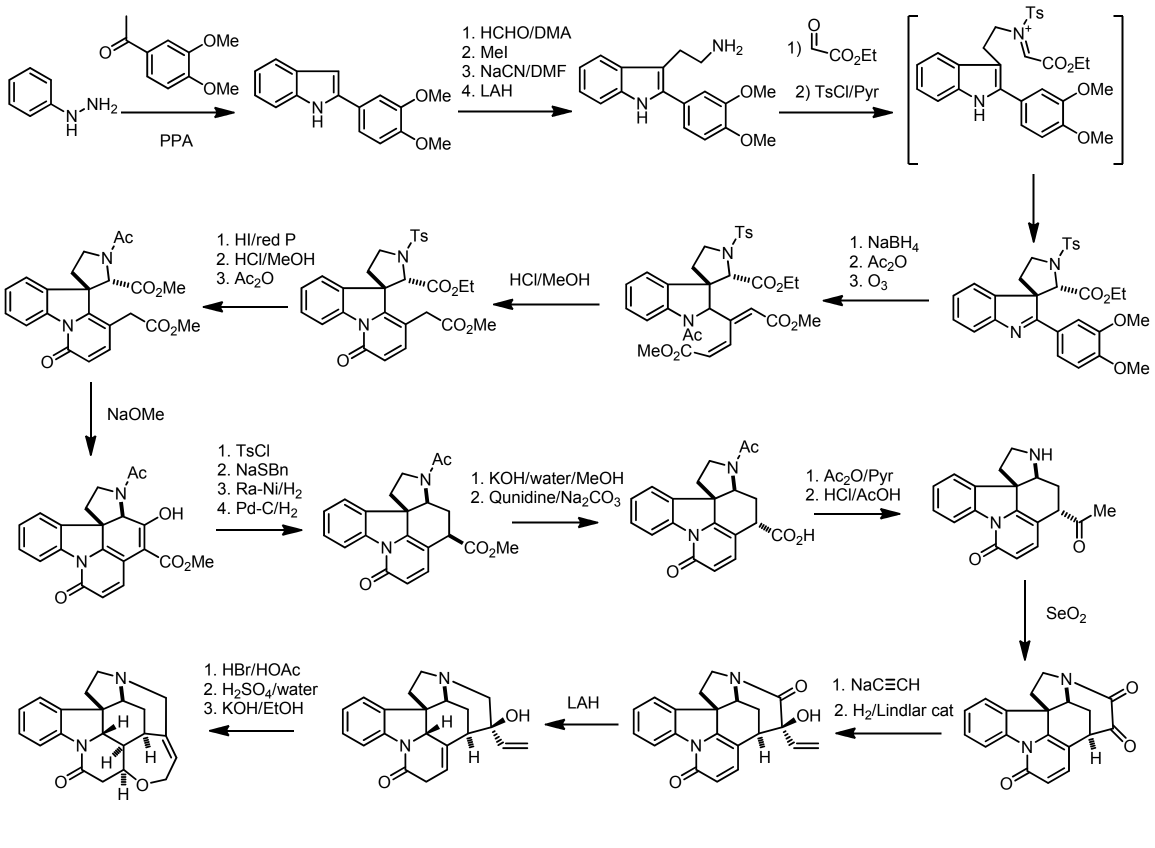
Construction reactions: C-C and C-heteroatom skeletal bonds formation

Strategic Redox: Oxidation and reduction to form correct functionality in the target molecule. All the other steps including non-strategic redox, protecting group manipulation, and functional group inter-conversion were categorized as concession steps. *Since the original definition of steps only considered the chemical reaction steps, we expanded the definition to include non-dynamic kinetic resolution, crystallization-induced chiral resolution, and chiral column separation, to implicitly separate these less efficient processes from direct asymmetric synthesis. For dynamic kinetic resolution or crystallization-induced dynamic resolution (CIDR), the epimerization will be treated as a strategic redox.*

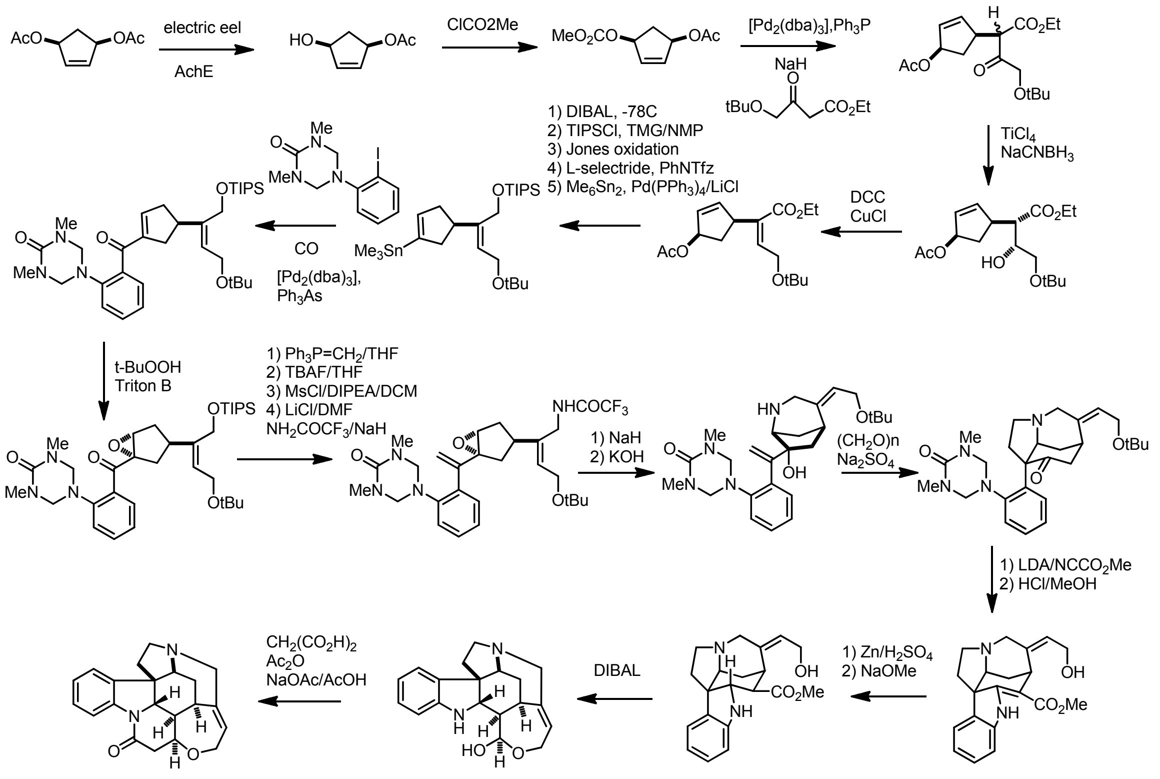
Woodward's total synthesis of strychnine constitutes of 30 steps (Scheme 1).[6] Previous reviews of this route have summarized the total step count of 29 steps to isostrychnine. Here we include the final ethanolic KOH isomerization from isostrychnine to strychnine for direct comparison with other approaches. The quinidine salt resolution was treated as a concession step. The overall route involves 10 construction and 6 strategic redox steps leading to an ideality of 0.53. Overman's enantioselective total synthesis has a total of 25 steps in which 8 construction and 6 strategic redox steps affords an ideality of 0.56 (Scheme 2).[17] MacMillan's enantioselective total synthesis involves a total of 13 steps among which are 7 construction and 2 strategic redox steps, giving rise to a further improved ideality of 0.69 (Scheme 3).[18] Vanderwal's concise racemic synthesis exploiting an intramolecular cycloaddition of a Zincke aldehyde has only 6 steps for the longest linear sequence (Scheme 4).[7] To compare to the approaches preparing enantiomerically enriched material, a separate chiral separation was arbitrarily added. Therefore, a total of 10 steps with 5 construction and one strategic redox results in an ideality of 0.6.

To calculate 'total steps,' we can simply use the total number of steps from Woodward, Overman and MacMillan's linear syntheses in the complexity model. To compare linear and branched routes, 50% of the total steps from the branching sequences are added to the longest linear sequence – this was done to discount the greater efficiency of branched approaches. Thus, for Vanderwal's synthesis, we add the steps of longest linear sequence (including the chiral separation) to 50% of the branching steps [ie 7 + (0.5×3)] giving 8.5 as the model input for total step count.

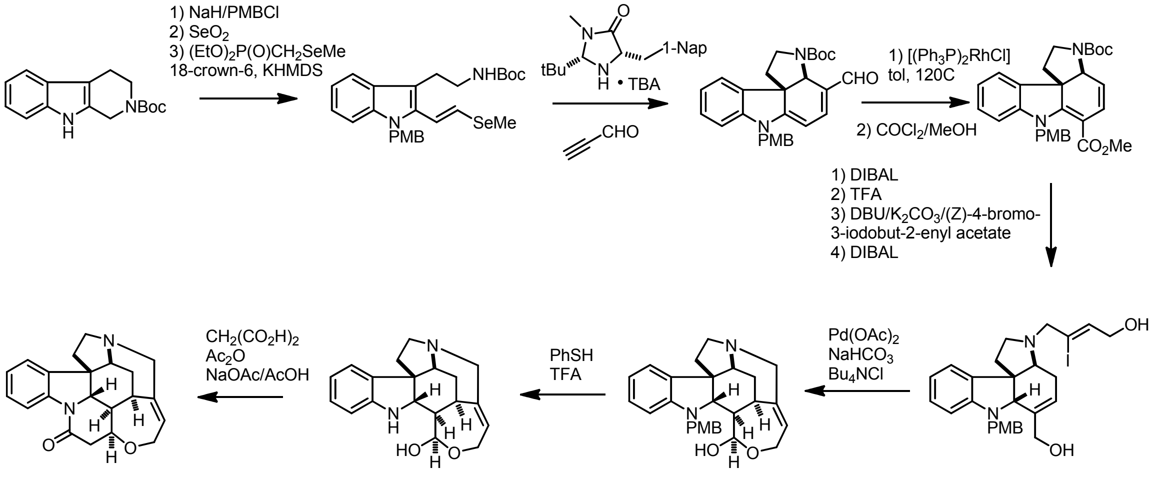
The number of stereogenic centers established in Strychnine and the number of heteroatoms *on and in* aromatic rings are 6 and 1, respectively. With these model parameters available, we can derive the current complexity from the four Strychnine syntheses.



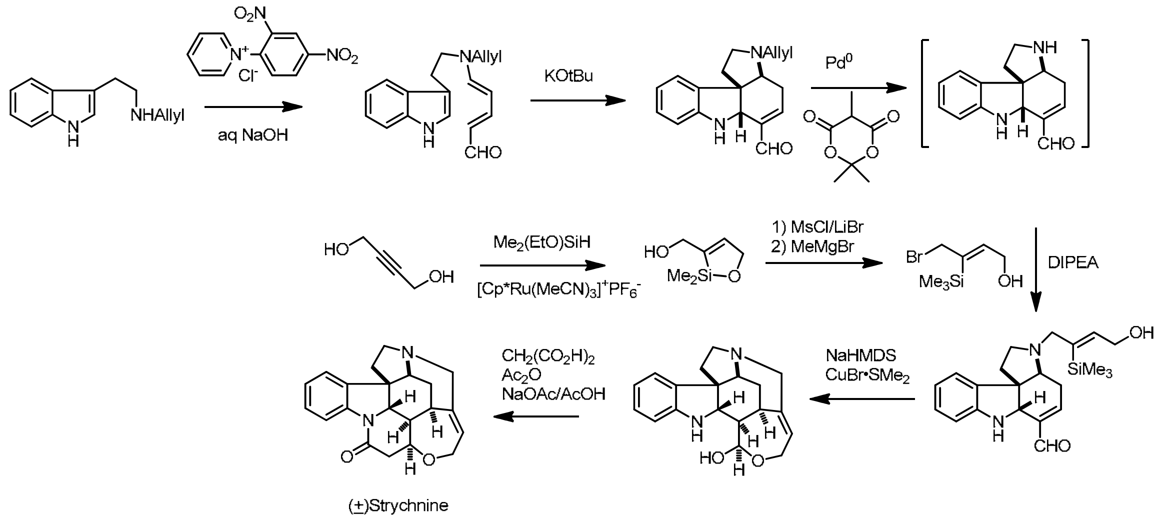
Scheme 1: Woodward's synthesis of strychnine.[6]



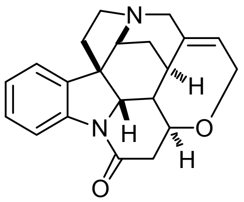
Scheme 2: Overman's Synthesis of strychnine.[17]



Scheme 3: MacMillan's Synthesis of strychnine.[18]



Scheme 4: Vanderwal's Synthesis of strychnine.[7]



Strychnine

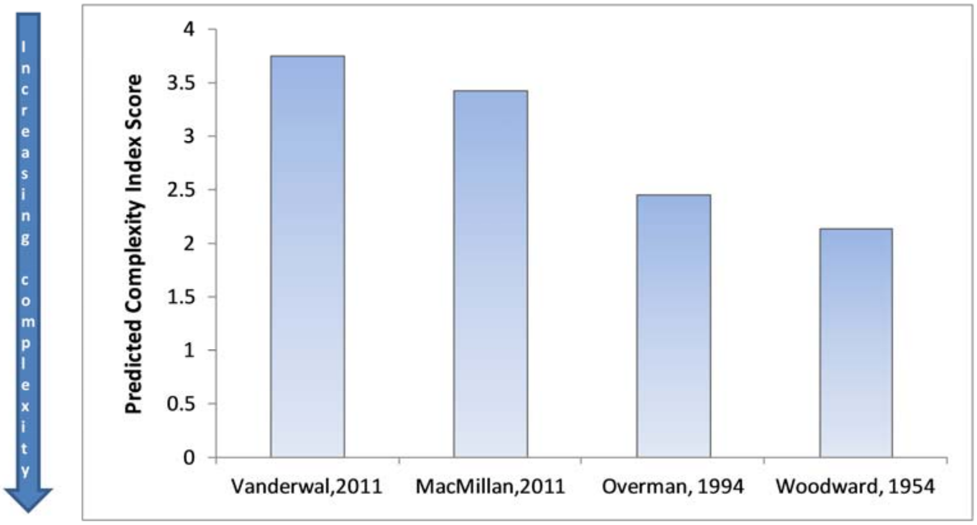


Figure 4: Predictive complexity index scores for some of the Strychnine syntheses – charting the impact of technology on complexity.

Strychnine is a complex system (scoring 2.14 for Woodward's original synthesis), however, the model shows how the *current complexity* has been impacted by improving technology; Vanderwal's synthesis shows a >1.5 unit change in complexity (to 3.75). It is important to note that the improved technology (preparing racemic Strychnine in 6 steps) does not make Strychnine a 'simple' molecule, it is still architecturally complex with a score reflective of that complexity today (3.75 is still a complex molecule). This demonstrates the balance of our model (neither extrinsic nor intrinsic factors dominate).

Next we detailed the complexity analysis on a family of natural products with closely-related intrinsic complexity, but distinct synthetic strategies. For this we chose the welwitindolinone family, containing an intriguing bridged ring system, which has inspired several synthetic approaches to various members of the family (figure 5).[19] Welwitindolinone C isonitrile (Rawal[20] and Garg's syntheses[21]) and A isonitrile (Wood[22] and Baran's syntheses[23]) are analyzed as two closely related systems for illustration.

Taking advantage of a sequence of palladium mediated enolate arylation, electrophilic chlorination and oxime rearrangement, Rawal's total synthesis of (-) N-methyl welwitindolinone C isonitrile involves 21 steps; 5 construction and 2 strategic redox, including the final desulfurization from isothiocyanate (Scheme 5).[20] This approach gives rise to an ideality of 0.33. In comparison, Garg's total synthesis utilizes a ring closure via an indolyne, chlorination via a vinyl stannane and an innovative deuterium isotope effect for controlling regioselectivity (Scheme 6).[21] This route involves 23 steps, with 6 construction and 3 strategic redox producing a final ideality of 0.39.[24] The current complexities of these approaches are 3.42 and 3.30 respectively. It should be noted that a higher ideality does not necessarily result in lower *current complexity*. A longer synthetic route with a higher percentage of productive steps (more 'ideal') would still be considered more complex.

Wood's total synthesis of (±)-Welwitindolinone A isonitrile involves a creative chloronium ion mediated semi-pinacol rearrangement and a stereo-controlled reductive cyclization (Scheme 7).[22] To compare to other enantiospecific routes, a separate chiral separation step was arbitrarily added. Therefore, a total of 23 steps, 8 construction and 2 strategic redox, resulted in an ideality of 0.43 and current complexity of 3.38. Baran's concise enantiospecific synthesis of Welwitindolinone A isonitrile via a redox strategy involves 9 steps, 4 construction and 3 strategic redox, giving an ideality of 0.78 (Scheme 8).[23] The short step count and high ideality resulted in a current complexity of 4.98, indicating a remarkable >1.5 unit change in complexity. Thus the development of this new synthetic technology significantly reduced the *current complexity* of the system, making Welwitindolinone A isonitrile a less complex structure, compared to C (from today's perspective).

In considering the impact of a new route to a molecule, current complexity stands as a multifaceted approached to detailing the impact of that route to our ability to prepare the system in question. This is of great significance in industrial settings, where understanding the impact of a new synthetic approach is more complex than just 'counting steps' and allows for a comparison between various molecules.

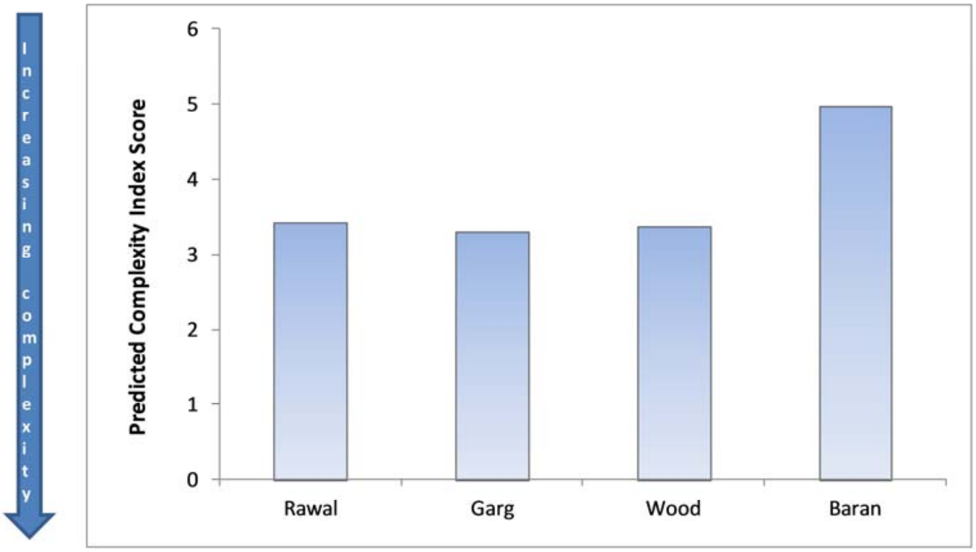
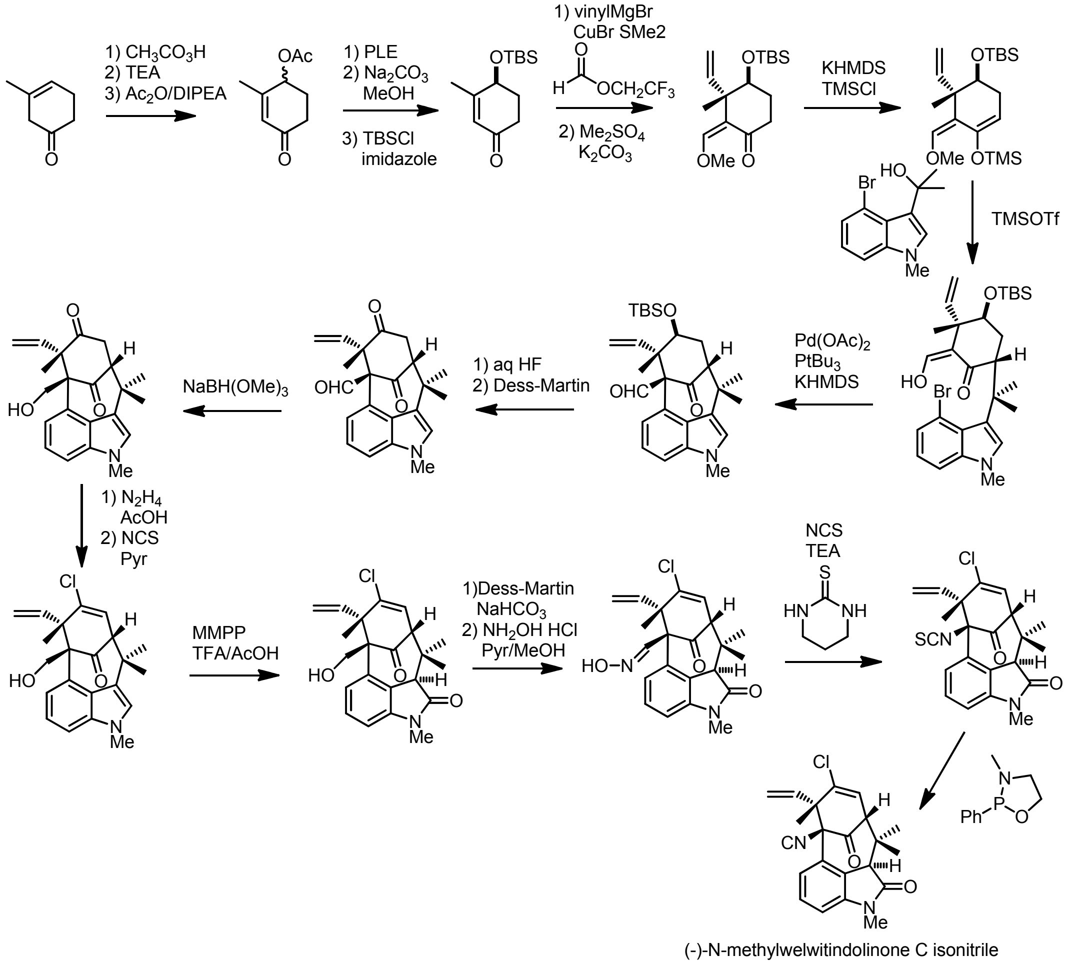


Figure 5: Predicted complexity index scores for some of the Welwitindolinone syntheses.

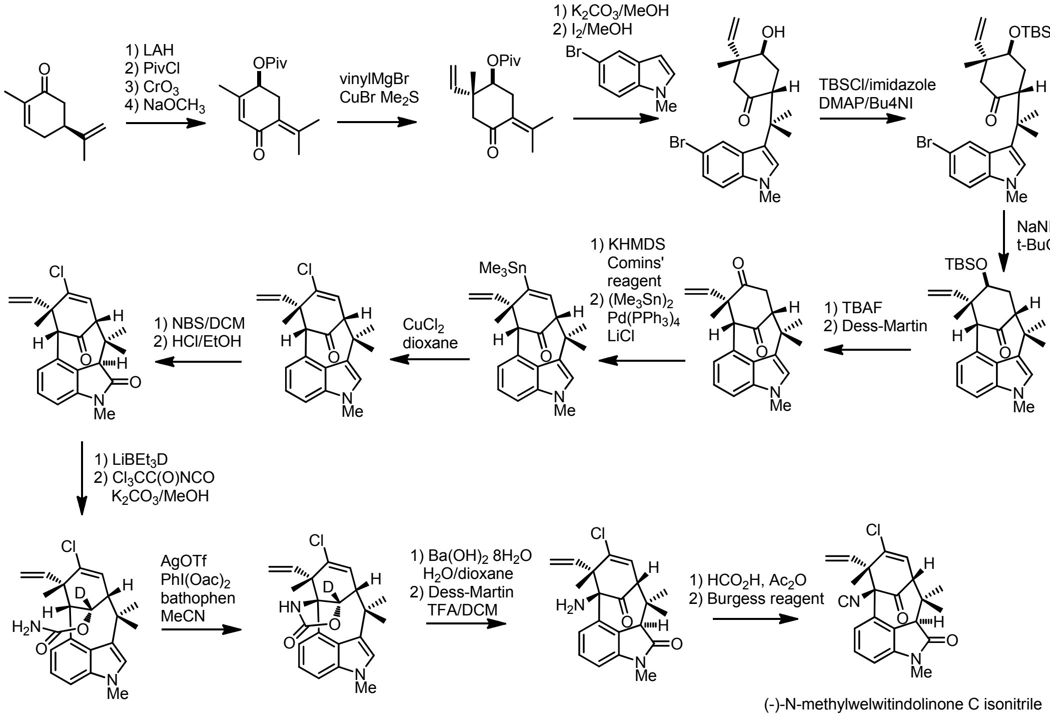
Table 1. Summary of the Strychnine and Welwitindoline Model Parameters for each demonstrated synthesis analyzed.

Target molecule	Total steps	Non-strategic redox	PG manipulation	FGI	Other Concession	Strategic redox	Construction	Ideality
Strychnine (Woodward)	30	1	6	4	1	6	10	0.53
Strychnine (Overman)	25	1	6	4	0	6	8	0.56
Strychnine (MacMillan)	13	1	3	0	0	2	7	0.69
Strychnine[a] (Vanderwal)	10	0	1	2	1	1	5	0.60
N-Welwitindolinone C (Rawal)	21	5	4	4	1	2	5	0.33
N-Welwitindolinone C (Garg)	23	4	4	6	0	3	6	0.39
Welwitindolinone A[a] (Wood)	23	3	6	3	1	2	8	0.43
Welwitindolinone A (Baran)	9	0	0	2	0	3	4	0.78

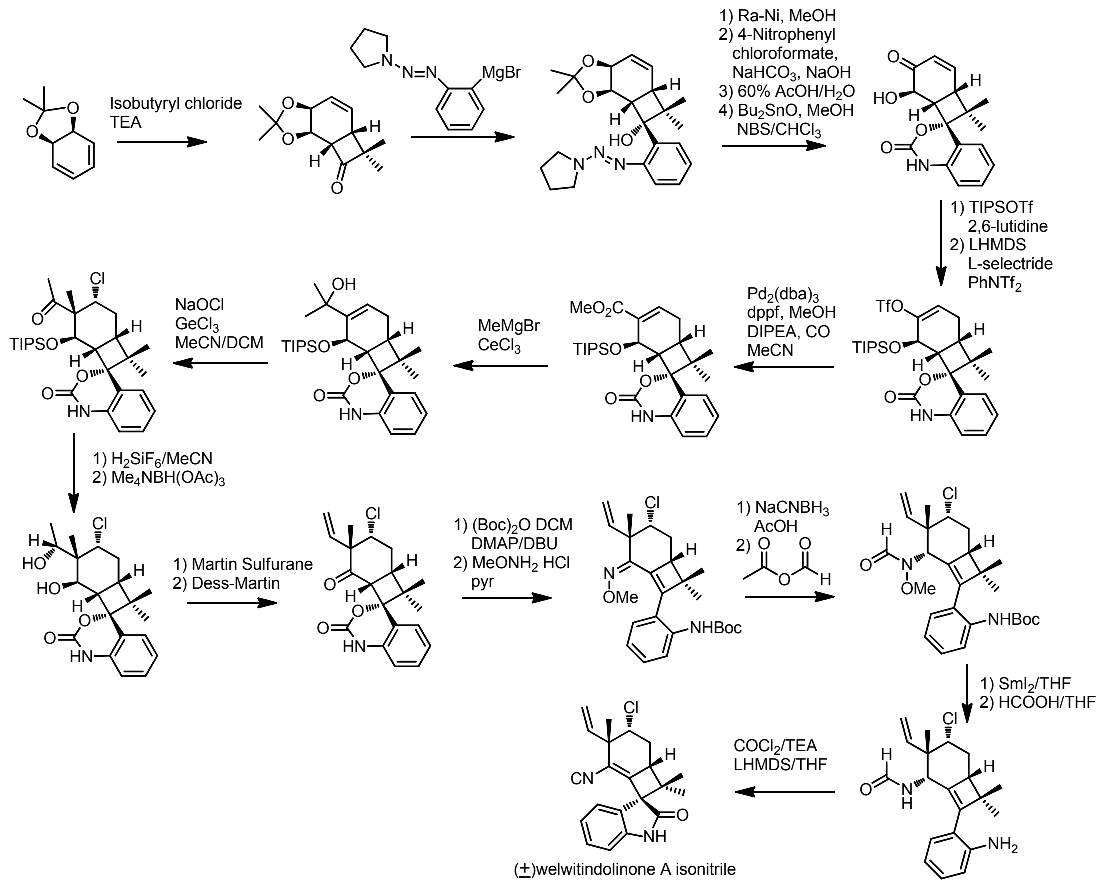
[a] As noted previously, in comparing racemic and enantioselective approaches an additional resolution step was added to the racemic routes.



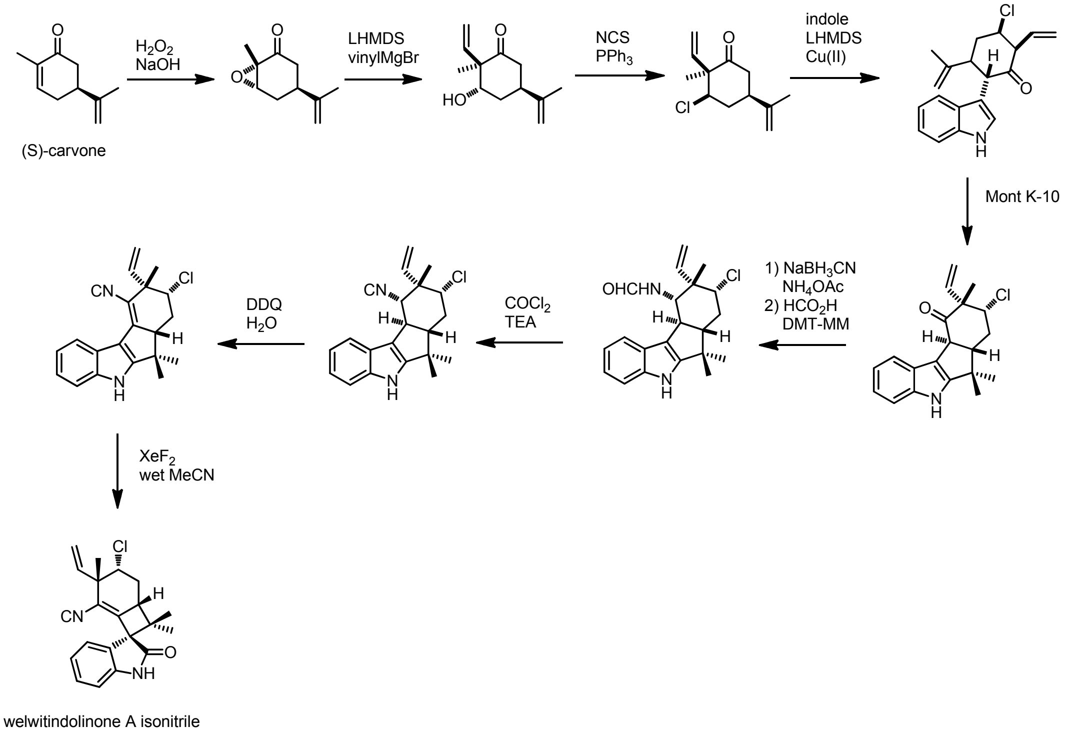
Scheme 5: Rawal's Synthesis of N-methylwelwitindolinone C isonitrile.[20]



Scheme 6: Garg's Synthesis of N-methylwelwitindolinone C isonitrile.[21]

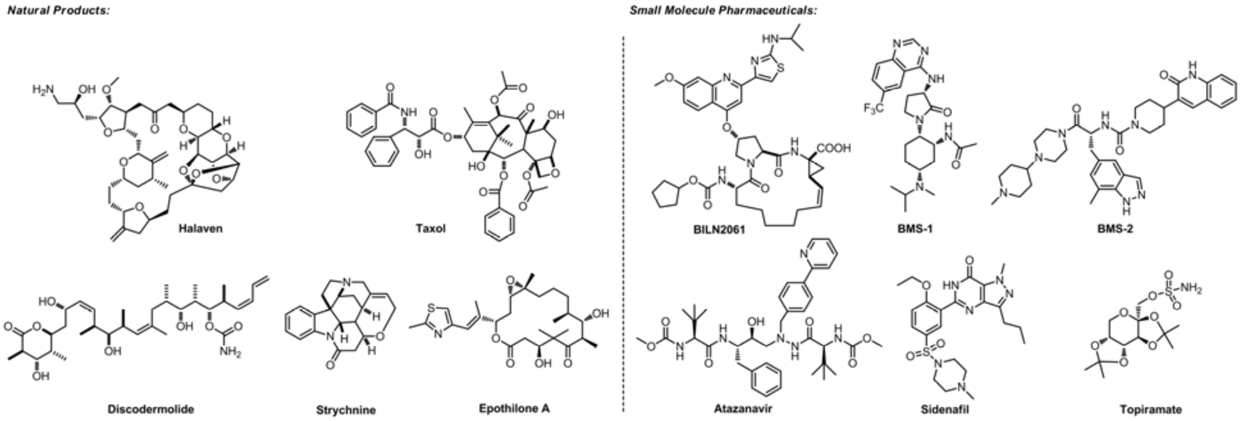


Scheme 7: Wood's Synthesis of welwitindolinone A isonitrile.[22]

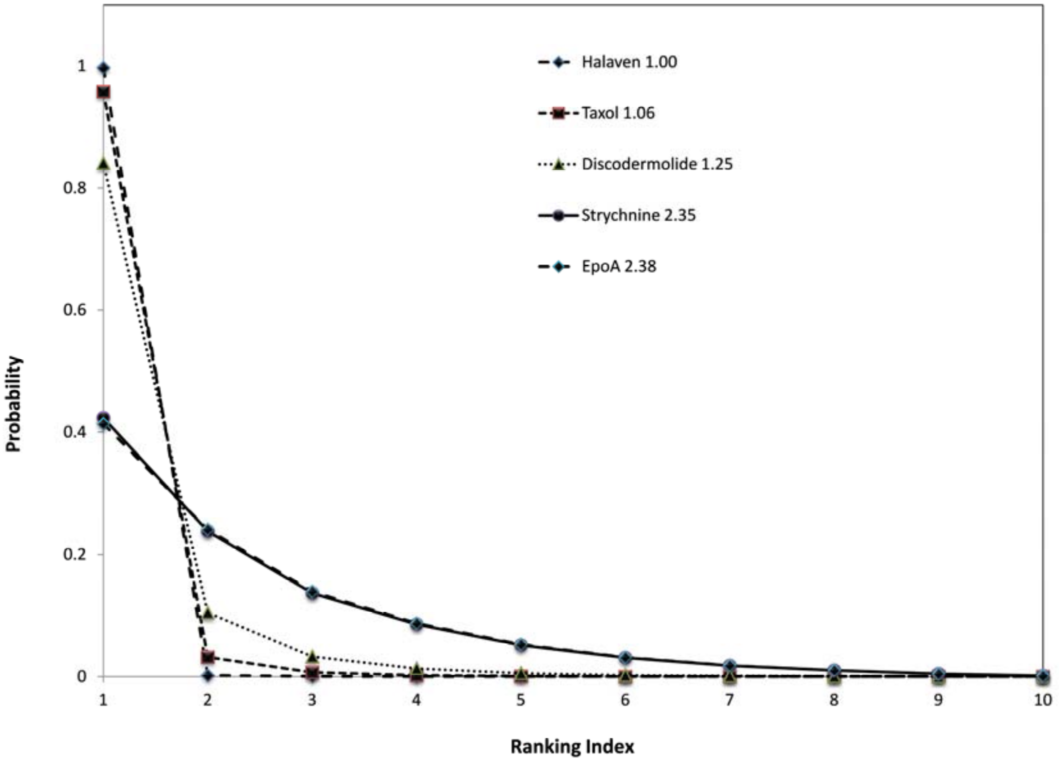


Scheme 8: Baran's Synthesis of welwitindolinone A isonitrile.[23]

We also compared a number of natural products and pharmaceutical drug candidates, to showcase the power of this new index for comparing across systems (figure 6). The results show credible differentiation between molecules. In figure 6a, complexity probability distributions for Eisai's synthesis of Halaven,[8] Nicolaou's Taxol synthesis,[25] Novartis-Smith-Paterson's Discodermolide,[26] Woodward's Strychnine and Danishefsky's Epotholine A[27] are *quantitatively* compared for the first time. Additionally, a number of pharmaceutical compounds with diverse topological features including atazanavir,[28] topiramate,[29] sidenafil,[30] BILN2061[31] and two BMS clinical candidates, BMS-1[32] and BMS-2,[33] and were compared (figure 6b). The data demonstrates how molecules with stereocenters and planar heterocycles compare in terms of complexity, along with the representative scores of complex natural products and pharmaceutical systems. In the natural products studied, EpoA is the least complex, with Halaven and Taxol representing similar levels of complexity. While structurally different, the current total synthesis of Halaven and Taxol build similar numbers of stereogenic centers (13 and 11 respectively) and have similar high step counts (56 vs 37), resulting in similar complexity, with Halaven being more complex. Strychnine, EpoA and BILN2061 are similar in complexity, with the triamine (BMS-1) being a slightly simpler system. This order reflects reality. Strychnine is a tightly packed stereochemically rich system, EpoA has difficult stereochemistry on the macrocyclic ring and BILN2061 has simpler stereochemistry, but contains a challenging macrocycle. In the pharmaceuticals presented, BMS-2 represents a more challenging problem than the poly-peptide Atazanavir, the planar heterocycle Sidenafil has more complex chemistry than the simply derived, though stereochemically rich, sugar Topiramate, both of which are less complex than the nitrogen-rich system BMS-2.


Natural Products:
Small Molecule Pharmaceuticals:
Halaven
Taxol
BILN2061
BMS-1
BMS-2
Discodermolide
Strychnine
Epothilone A
Atazanavir
Sidenafil
Topiramate

(a)


Halaven 1.00
Taxol 1.06
Discodermolide 1.25
Strychnine 2.35
EpoA 2.38
Probability
Ranking Index
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5
6
7
8
9
10

(b)

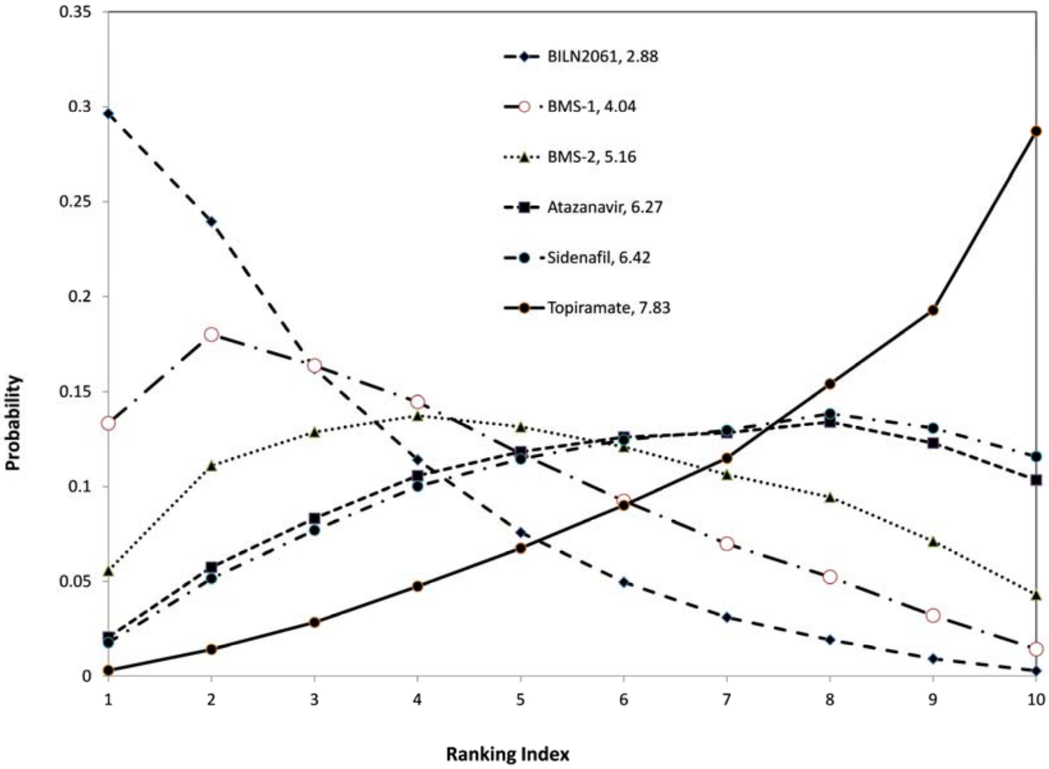


Figure 6. Molecules selected for analysis, and (a) the complexity probability distributions for the natural products Halaven,[8] Taxol,[25] discodermolide,[26] Strychnine[6] and epothilone A;[27] (b) the complexity probability distribution for pharmaceutical compounds atazanavir,[28] sidenafil,[30] topiramate,[29] BILN2061[31] along with some BMS development candidates (BMS-1[32] and BMS-2[32]). Complexity scores are listed next to the legends.

We have conducted this assessment over many BMS internal proprietary compounds (>60) and found that this approach easily passes the ‘common-sense’ check, with molecules falling into the correct locations on the indexed scale. Improvements in routes for a given molecule are captured and the changes quantitatively reflect the impact of the improved technology. The ability of the ranking model to give non-integer results allows differentiation of close systems. While not all results are agreed on by all people, the model generates a distribution reflective of a *community's* predicted response. We should stress that the utility of this model is not restricted to the assessment of reported/completed synthesis. In this application we have focused on demonstrated synthetic approaches. However, merging this analysis with new 'synthetic route design' software, such as ARChem from SimBioSys,[34] or the synthesis prediction tool ICSYNTH,[35] this method could be applied in a purely predictive sense – predicting the perceived complexity of *de novo* molecules prior to a synthetic endeavour. This application may prove especially valuable in comparing molecules derived from *in silico* drug design.

Conclusion

In summary, we have demonstrated a new concept for assessing a molecules complexity. *Current complexity* assesses structure in the context of capability and is variable over time. Our approach uses quantitative data from the molecular network, along with the collective intelligence of a population, to accurately determine the current complexity of a molecule in the context of current technology. We have demonstrated this approach, which leverages probability functions to combine molecular complexity and synthetic complexity, in a diverse series of compounds. This approach can help chemists assess differences in complexity across a portfolio of compounds, analyze changes in pharmaceuticals year on year, track the 'current' complexity of a given molecule over time, or enable the quantitative impact of a new route or process to be determined. Future development of this methodology should focus on expanding the data set used in the analysis (expanding the number of ‘raters’) along with expanding our understanding of the factors which influence our perception of the complexity of organic molecules.

Acknowledgement

The authors would like to thank Drs. Rajendra Deshpande and David Kronenthal for supporting this work. We would also like to thank Drs. Ke Chen, Nicolas Cuniere, Akin Davulcu, Murali Dhar, Yu Fan, Michaël Fenster, Monica Fitzgerald, Sergei Kolotuchin, Sha Lou, Harold Mastalerz, Boguslaw Mudryk, Mark Saulnier, Michael Schmidt, Chris Sfouggatakis, Omid Soltani, Steven Tymonko, Zhongmin Xu, Ian Young, and Bin Zheng for providing their ranking assessments.

Supporting Information

This material is available free of charge via the Internet.

References and Footnotes

[1] (a) E. J. Corey, X.-M. Cheng, *The Logic of Chemical Synthesis*; John Wiley: New York, **1989**. (b) K. C. Nicolaou, E. J. Sorensen, *Classics in Total Synthesis: Targets, Strategies, Methods*; VCH: Weinheim; New York, **1996**; (c) K. C. Nicolaou, S. A. Snyder, *Classics in Total Synthesis II: More Targets, Strategies, Methods*; Wiley-VCH: Weinheim, **2003**; (d) K. C. Nicolaou, *Proc. R. Soc. A* **2014**, 470.

[2] (a) D. H. Rouvray, D. Bonchev, *Complexity in Chemistry: Introduction and Fundamentals*, **2003**, Taylor & Francis.; (b) G. M. Whitesides, R. F. Ismagilov, *Science.* **1999**, *284*, 89; (c) P. Selzer, H.-J. Roth, P. Ertl, Schuffenhauer, A. *Curr. Opin. Chem. Biol*. **2005**, *9*, 310.

[3] (a) S. H. Bertz, *J. Am. Chem. Soc.* **1981***, 103,* 3599*;* (b) S. H. Bertz, *J. Am. Chem. Soc.* **1982***, 104,* 5801*;* (c) D. Bonchev, O. Mekenyan, N. Trinajstić, *Int. J. Quant. Chem.* **1980***, 17,* 845.; (d) M. Randić, *J. Am. chem. Soc.* **1975***, 97,* 6609.; *J. Am. Chem. Soc.,* **1977**, 99, 444; (e) J. B. Hendrickson, P. Huang, A. G. Toczko, *J. Chem. Inf. Comput. Sci.* **1987***, 27,* 63.; (f) I. Gutman, B. Ruscic, N. Trinajstic, C. F. Wilcox, *J. Chem. Phys*, **1975**, 62, 3399.; (g) H. Wiener, *J. Am. Chem. Soc.,* **1947**, 69, 17.

[4] F. Farary, **1969** *Graph Theory,* Addison-Wesley, Reading, MA.

[5] (a) J. Gasteiger, **2003** *Handbook of chemoinformatics,* Wiley-VCH, Weinheim; (b) H. Whitlock, *J. Org. chem.* **1998***, 63,* 7982; (c) R. Barone, R. Chanon, *J. Chem. Inf. Comput. Sci.* **2001***, 41,* 269; (d) A.P. Johnson, C. Marshall, P. N. Judson, *J. Chem. Inf. Comput. Sci*. **1992***, 32,* 411; (e) V. J. Gillet, G. Myatt, Z. Zsoldos, A. P. Johnson, *Perspect. Drug Discovery Des.* **1995,** *3,* 34. (f) K. Boda, A. P. Johnson, *J. Med. Chem.* **2006,** *49,* 5869.*;* (g) P. Ertl, A. Schuffenhauer, *J. Cheminformatics* **2009***, 1,* 8; (h) K. Boda, T. Seidel, J. Gasteiger, *J. Comput. Aided. Mol. Des.* **2007***, 21,* 311. (i) Y. Takaoka, Y. Endo, S. Yamanobe, H. Kakinuma, T. Okubo, Y. Shimazaki, T.Ota, S. Sumiya, K. Yoshikawa, *J. Chem. Inf. Comput. Sci*. **2003***, 43,* 1269. (j) Y. Pdolyan, M. A. Walters, G. Karypis, *J. Chem. Inf. Model*. **2010***, 50,* 979. (k) M. M. Hann, A. R. Leach, G. Harper, *J. Chem. Inf. Model.,* **2001**, *41*, 856. (l) S. H. Nilar, N. L. Mo, T. H. Keller, *J. Comput. Aid. Mol. Des.* **2013**, *27*, 783.

[6] R. B. Woodward, M. P. Cava, W. D. Ollis, A. Hunger, H. U. Daeniker, K. Schenker, *J. Am. Chem. Soc.* **1954**, 4749.

[7] D. B. C. Martin, C. D. Vanderwal, *Chem. Sci*. **2011**, *2*, 649.

[8] W. Zheng, B. M. Seletsky, M. H. Palme, P. J. Lydon, L. A. Singer, C. E. Chase, C. A. Lemelin, Y. Shen, H. Davis, L. Tremblay, M. J. Towle, K. A. Salvato, B. F. Wels, K. K. Aalfs, Y. Kishi, B. A. Littlefield, M. J. Yu, *Bioorg. Med. Chem. Lett*. **2004**, 5551.

[9] D. W. Hubbard, “How to measure anything”, 2010, 2nd Ed. Wiley.

[10] (a) A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A.Vehtari, D. B. Rubin, *Bayesian Data Analysis.* **2014**, Chapman & Hall, 3rd Ed.; (b) J. K. Kruschke, *Doing Bayesian Data Analysis.* **2011**, Elsevier.

[11] W. H. Greene, D. A. Hensher, *Modeling Ordered Choices*, **2010**, Cambridge Univ.

[12] (a) J. H. Albert, S. Chib, *J. Amer. Statist. Assoc*. **1993,** *88*, 669; (b) M. K. Cowles, *Statistics and Computing*. **1996,** *6*, 101.

[13] T. Gaich, P. S. Baran, *J. Org. Chem.* **2010**, *75*, 4657.

[14] (a) F. Emmert-Streib, M. Dehmer, *PLoS one*, **2012**, *7*, e34523. (b) M. Dehmer, K. Varmuza, S. Borgert, F. Emmert-Streib, *J. Chem. Inf. Comput. Sci*. **2009***, 49,* 1655.

[15] A. J. Miller, *Subset selection in Regression* **2002** Chapman & Hall, 2nd Ed.

[16] J. S. Cannon, L. E. Overman, *Angew. Chem. Int. Ed.,* **2012**, *51*, 4288.

[17] S. D. Knight, L. E. Overman, G. Pairaudeau, *J. Am. Chem. Soc.* **1993**, *115*, 9293.

[18] S. B. Jones, B. Simmons, A. Mastracchio, D. W. C. MacMillan, *Nature*, **2011**, *475*, 183.

[19] J. L. Wood, *Nature Chemistry* **2012**, *4*, 341.

[20] K. M. J. Allan, K. Kobayashi, V. H. Rawal, *J. Am. Chem. Soc*. **2012**, *134*, 1392.

[21] K. W. Quasdorf, A. D. Huters, M. W. Lodewyk, D. J. Tantillo, N. K. Garg, *J. Am. Chem. Soc*. **2012**, *134*, 1396.

[22] S. E. Reisman, J. M. Ready, M. M. Weiss, A. Hasuoka, M. Hirata, K. Tamaki, T. V. Ovaska, C. J. Smith, J. L. Wood, *J. Am. Chem. Soc*. **2008**, *130*, 2087.

[23] J. M. Richter, Y. Ishihara, T. Masuda, B. W. Whitefield, T. Llamas, A. Pohjakallio, P. S. Baran, *J. Am. Chem. Soc*. **2008**, *130*, 17938.

[24] We note that the CrO_3 oxidation of derivatized (S)-carvone did not directly introduce the carbonyl ketone functionality present in the target molecule. It was reduced by super deuteride to control the site of nitrene insertion. The resulting alcohol was later re-oxidized by Dess-Martin periodinane. We therefore assigned the initial CrO_3 oxidation a non-strategic redox for purposes of calculating the overall ideality of the sequence.

[25] K. C. Nicolaou, Z.Yang, J. J. Liu, H. Ueno, P. G. Nantermet, R. K. Guy, C. F. Claiborne, J. Renaud, E. A. Couladouros, K. Paulvannan, E. J. Sorensen, *Nature*, **1994**, *367*, 630.

[26] S. J. Mickel, D. Niederer, R. Daeffler, A. Osmani, E. Kuesters, E. Schmid, K. Schaer, R. Gamboni, W. Chen, E. Loeser, F. R. Kinder, Jr., K. Konigsberger, K. Prasad, T.

M. Ramsey, O. Repi, R.-M. Wang, G. Florence, I. Lyothier, I. Paterson, *Org. Proc. Res. Dev.* **2004**, *8*, 122.

[27] A. Balog, D. Meng, T. Kamenecka, P. Bertinato, D.-S. Su, E. J. Sorensen, S. J. Danishefsky, *Angew. Chem. Int. Ed. Engl.,* **1996**, *35*, 2801.

[28] Z. Xu, J. Singh, M. D. Schwinden, B. Zheng, T. P. Kissick, B. Patel, M. J. Humora, F. Quiroz, L. Dong, D.-M. Hsieh, J. E. Heikes, M. Pudipeddi, M. D. Lindrud, S. K. Srivastava, D. R. Kronenthal, R. H. Mueller, *Org. Proc. Res. Dev.* **2002**, *6*, 323.

[29] G. Arvai, S. Garaczi, A. G. Mate, F. Lukacs, Z. Viski, G. Schneider, US20060040874

[30] D. J. Dale, P. J. Dunn, C. Golightly, M. L. Hughes, P. C. Levett, A. K. Pearce, P. M. Searle, G. Ward, A. S. Wood, *Org. Proc. Res. Dev.* **2000**, *4*, 17.

[31] N. K. Yee, V. Farina, I. N. Houpis, N. Haddad, R. P. Frutos, F. Gallou, X.-J. Wang, X. Wei, R. D. Simpson, X. Feng, V. Fuchs, Y. Xu, J. Tan, L. Zhang, J. Xu, L. L. Smith-Keenan, J. Vitous, M. D. Ridges, E. M. Spinelli, M. Johnson, *J. Org. Chem*. **2006**, *71*, 7133.

[32] M. G. Yang, R. J. Cherney, M. D. Eastgate, J. Muslehiddinoglu, S. J. Prasas, Z. Xiao, US20080027083.

[33] R. O. Cann, C.P. H. Chen, Q. Gao, R. L. Hanson, D. Hsieh, J. Li, D. Lin, R. L. Parsons, Y. Pendri, R. B. Nielsen, W. A. Nugent, W. L. Parker, S. Quinlan, N. P. Reising, B. Remy, J. Sausker, X. Wang, *Org. Proc. Res. Dev.* **2012**, *16*, 1953.

[34] (a) J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, H. Y. Ando, *J. Chem. Inf. Model.* **2009**, *49*, 593; (b) http://www.simbiosys.ca/archem/index.html

[35] A. Bøgevig, H. J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Löw, C. Oppawsky, T. Rein, H. Saller, *Org. Proc. Res. Dev*. **2015**, *19*, 357.