nature reviews chemistry

https://doi.org/10.1038/s41570-024-00645-8

Perspective

Check for updates

Molecular complexity as a driving force for the advancement of organic synthesis

Brandon A. Wright & Richmond Sarpong

Abstract

The generation of molecular complexity is a primary goal in the field of synthetic chemistry. In the context of retrosynthetic analysis, the concept of molecular complexity is central to identifying productive disconnections and the development of efficient total syntheses. However, this field-defining concept is frequently invoked on an intuitive basis without precise definition or appreciation of its subtleties. Methods for quantifying molecular complexity could prove useful for characterizing the state of synthesis in a more rigorous, reliable and reproducible fashion. As a first step to evaluating the importance of these methods to the state of the field, here we present our perspective on the development of molecular complexity quantification and its implications for chemical synthesis. The extension and application of these methods beyond computer-aided synthesis planning and medicinal chemistry to the traditional practice of 'complex molecule' synthesis could have the potential to unearth new opportunities and more efficient approaches for synthesis.

Sections

Introduction

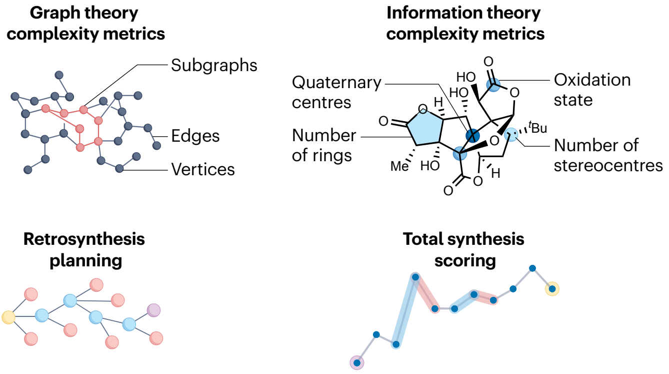
Quantifying structural complexity

Quantifying synthetic complexity

Applications of molecular complexity analysis

Analysis and evaluation of total syntheses

Conclusion and outlook



Department of Chemistry, University of California, Berkeley, USA. e-mail: rsarpong@berkeley.edu

Introduction

Synthetic chemists are fascinated by complex molecules. Densely packaged carbon skeletons decorated with oxygen, nitrogen or other heteroatoms evoke a sense of awe and intense interest from those steeped in the art and science of organic synthesis. Generating molecular complexity, whether in the context of a useful synthetic method or the strategy of a total synthesis, is arguably a focal point for many synthetic chemists at some level. Synthesis is, after all, the enterprise of building new molecules. Many naturally occurring compounds, often isolated in unsustainably small quantities, contain unusual structural motifs that challenge the field to advance new strategies and methods for their preparation. Access to novel structural motifs may lead to desirable biological properties or applications as functional materials. The construction of complex molecules is, therefore, a primary goal of the field of synthetic chemistry.

For as often as the term complexity is discussed colloquially within the field, one must consider: what exactly do we mean when we call a molecule complex? For example, one could argue that molecular complexity is a purely subjective concept which chiefly relies on human perception, much like an analysis of a piece of art or work of literature[1]. Conversely, one might also consider that molecular complexity bears a certain mathematical objectivity, describing inherent features or relationships that render a particular system 'complex'[2]. The complexity of systems is certainly not limited to chemistry, and many have applied this concept in other fields such as in physics[3], biology[4–6], climate science[7] or engineering[8]. However, in particular, the subjective and objective measures of complexity are often held in tension in organic synthesis, a field which cultivates appreciation for both the elegance of a synthetic strategy and the quantitative description of a reaction mechanism. Woodward was awarded the Nobel Prize in Chemistry in 1965 "for his outstanding achievements in the art of organic synthesis"[9], officially recognizing organic synthesis as a field which is both an art and a science. So, too, is molecular complexity.

The concept of molecular complexity was central to Corey's development of retrosynthetic analysis, the systematic method for planning syntheses[10]. In a retrosynthetic analysis, one works backwards from a synthetic target by considering all possible disconnections which correspond to synthetic transformations in the forward direction. At each stage in the analysis, desirable disconnections are selected and carried onwards, a process which "receives direction and selectivity from the all-important goal of reducing molecular complexity"[10].

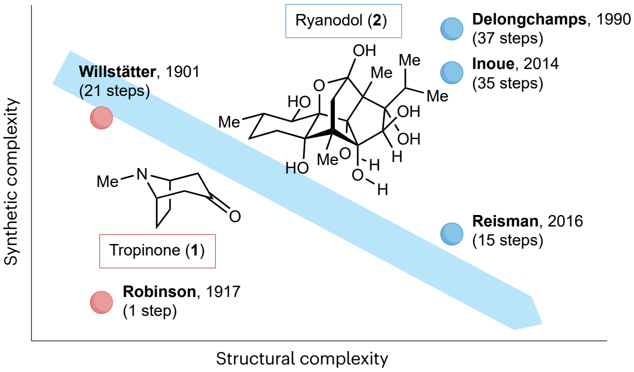


Fig. 1 | Relationship between structural and synthetic complexity. A plot of structural and synthetic complexity space with selected targets and their respective syntheses.

Guided by this singular goal, Corey argues that one can reduce any complex target molecule into simpler and simpler fragments which can be accessed either commercially or according to previously reported methods. This retrosynthesis logic was codified into a series of discrete rules[11] and eventually programmed as a retrosynthesis software named Logic and Heuristics Applied to Synthetic Analysis (LHASA)[10,12], laying the groundwork for the development of many additional automated and semi-automated synthesis planning programs which have recently emerged[13]. In many of these algorithms, a 'scoring function' evaluates which disconnection yields the greatest decrease in complexity and is critical to the success of the algorithm in identifying full-length synthetic pathways. Although there are instances in which brief increases in complexity – such as installation of a protecting group or a generation of a more reactive intermediate – can enable efficient overall syntheses, reducing molecular complexity, the 'all-important goal'[10], occupies a central place in retrosynthetic analysis.

Taken more broadly, molecular complexity can also serve as an aspirational goal for synthetic chemistry as it advances new methods and more efficient synthetic strategies. According to Corey, "Molecular complexity can be used as an indicator of the frontiers of synthesis, since it often causes failures which expose gaps in existing methodology. The realization of such limitations can stimulate the discovery of new chemistry and new ways of thinking about synthesis"[11]. In the context of total synthesis, attempts to construct highly complex molecules reveal the current limitations of existing synthetic methods and pose opportunities to develop new chemistry. In this light, molecular complexity can be best understood in two distinct dimensions: structural complexity and synthetic complexity. Structural complexity refers to inherent structural features of a molecule which contribute to its overall complexity. Factors such as number of rings, stereocentres or heteroatoms, which describe the structural composition of a compound, are often invoked. Synthetic complexity, as defined previously by others[14,15], describes how easily a particular target could be synthesized, for example, the number of steps required to access the molecule. As Eastgate and Li propose[14], this aspect of complexity is extrinsic to the target and largely dependent on currently available methodology. Structural complexity, conversely, is intrinsic to the target, immutable.

These two facets of complexity are distinct, yet related, and the interplay between them provides a useful framework for understanding the progress of the field in synthesizing complex molecular scaffolds over the past two centuries. On a conceptual level, as similarly illustrated by Wender[16], one might consider structural and synthetic complexity as plotted along two axes (Fig. 1). Navigating this 'complexity space' has allowed synthetic chemists to approach increasingly complex molecular architectures in shorter sequences of steps. For example, tropinone (**1**), a target of medium–low structural complexity, was first synthesized by Willstätter in 1901 in a reported 21 steps[17]. Notably, the work of Willstätter on tropine synthesis led to the structure elucidation of cocaine and other tropane alkaloids[18,19]. Despite the broader impact of this work, tropinone (**1**) remained synthetically complex – that is, until 1917, when Robinson reported a one-step synthesis of tropinone featuring a decarboxylative double-Mannich transformation to efficiently construct the 8-azabicyclo[3.2.1]octane core[20]. The precipitous drop in the synthetic complexity of tropinone illustrated the power of the then recently reported (though later named) Mannich reaction to the synthetic community and rendered the approach of Robinson an instant classic. Nevertheless, targets of greater structural complexity than **1** remained largely out of reach until newly developed synthetic methods enabled the synthesis of increasingly complex molecules.

Because structurally complex molecules often demand numerous synthetic transformations to complete their synthesis, structural complexity is, naturally, often correlated with synthetic complexity. Diterpenoids such as ryanodol (**2**) have attracted significant attention for their complex architectures and intriguing insecticidal functions. As one might expect for a target of high structural complexity, two of the first completed syntheses of ryanodol by Delongchamps (37 steps)[21] and Inoue (35 steps)[22] reflect its similarly high synthetic complexity. Informed by these previous syntheses, a recent 15-step synthesis from Reisman and coworkers[23] rendered **2** synthetically much less complex, illustrating how cumulative advances in strategy and methodology can tame structurally complex molecules down to a much more reasonable level of synthetic complexity as measured by step count. In the so-called age of feasibility[24], when total synthesis was first pushed to its limits by tackling targets such as Taxol[25–33], calicheamicin[34–37], halichondrin B[38,39] or palytoxin[40–42], the goal was demonstrating that organic synthesis could produce a few milligrams of any target at any cost. High structural complexity was tethered to high synthetic complexity. In the modern 'age of scalability'[43], however, the field is steadily transitioning towards a new phase of innovation wherein structurally complex targets can be easily accessed in short step sequences and with highly efficient transformations (Fig. 1, indicated by blue arrow). Indeed, the most powerful synthetic methods and strategies are those which minimize the synthetic complexity of structurally complex molecules.

Together, structural and synthetic complexity are useful concepts through which the state of organic synthesis can be evaluated and the field can be driven forward. Given the various applications of molecular complexity within medicinal chemistry, total synthesis and retrosynthetic planning, a more quantitative treatment of this key concept is necessary. In a time when quantitative tools such as machine learning[44–47] and statistical modelling and parametrization[48–50] are being applied to reaction prediction[51–54], catalyst design[55,56] or retrosynthesis planning[57–59], the realm of complex molecule synthesis could benefit from even more of a focus on quantitative metrics for guiding synthetic strategy. This Perspective will survey several pivotal contributions useful for measuring structural and synthetic molecular complexity and will then examine how these efforts relate to a series of applications in organic chemistry.

Quantifying structural complexity

Frameworks for measuring complexity

The structural complexity of molecules can be assessed according to two branches of mathematics: graph theory and information theory. First, a brief examination of these fields of study is necessary to fully appreciate the underlying framework behind existing complexity metrics. Broadly speaking, graph theory deals with the way objects are connected. Adapted to chemistry, chemical graph theory models chemical structures as molecular graphs, which are abstract representations of 'objects' that are 'related' in a network[60]. Instead of atoms and bonds in a chemical structure, chemical graphs feature vertices and edges, respectively (Fig. 2a). Ignoring most chemical or physical considerations, chemical graph theory primarily treats the connectivity of a molecule as an adequate representation of its structure. The consideration of molecular connectivity in the context of retrosynthesis, for example, is reflected in several of the retrosynthetic rules of Corey for choosing disconnections in a complex molecule based on a 'topological strategy'[11]. Disconnections that best reduce molecular complexity, according to Corey, are those that simplify the topology of a molecule by a number of means: reduction in the overall number of rings, division into fragments of equal size, or cleavage of maximally bridged rings. As it relates to methods development[61,62], any given reaction can be analysed with chemical graph theory by looking at changes to the respective molecular graph. In this way, chemical graph theory can be used to examine molecular connectivity as a representation of structural complexity.

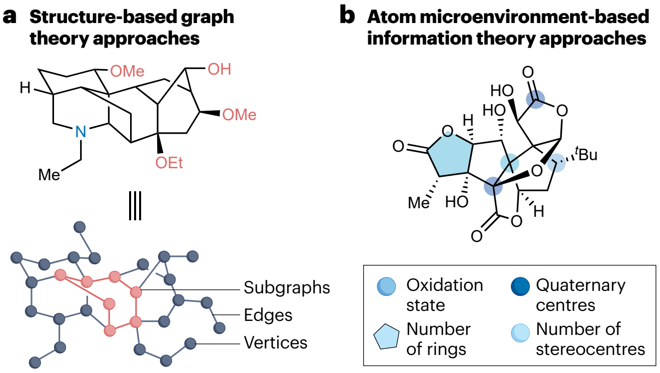


Fig. 2 | Measurements of complexity. a, Graph theory methods for measuring complexity rely on substructures and properties of molecular graphs. **b**, Information theory methods draw from features of atomic microenvironments.

Information theory deals with the way information and uncertainty are quantified, manipulated and represented. Rooted in Shannon's formula[63], information theory states that uncertainty, or information entropy, is related to the sum of all possible states of individual variables. Chemical information theory, by analogy, treats molecules as a series of variables or features that encode information in a particular state[64] (Fig. 2b). A string of binary digits (0 or 1) can contain information interpreted by a computer. Likewise, stereocentres on a molecule (*R* or *S*) can encode molecular properties or molecular shapes that are recognized by cellular machinery or manifested as chemical reactivity. Unlike chemical graph theory, which approximates molecules as a set of nodes and connections, chemical information theory focuses on factors, such as atom identity, atom microenvironment or other molecular features. In an era of 'Big Data', when key relationships can be unearthed by way of large-scale data science, information theory has been a framework for the development of quantitative structure–activity relationship (QSAR) parameters in medicinal chemistry programs[65]. In our view, the basic challenge of selectivity in organic synthesis – as seen through enantioselective catalysis or site-selective functionalization – can be understood fundamentally as a challenge in molecular information encoding[66]. In the context of drug–protein interactions, the chiral information present in small molecules can trigger vastly different biochemical signalling pathways[67]. As such, the installation of stereogenic centres is, essentially, an information-encoding event. In general, chemical information theory assesses structural complexity by accounting for all the variables present in a molecule which, together, constitute its overall structure.

Methods based on chemical graph theory

Structural complexity metrics frequently draw from graph theory and information theory in attempts to capture the inherent complexity of a molecule. One of the most influential methods for evaluating structural complexity was developed by Bertz in 1981 (ref. 68). Termed a 'general index' for molecular complexity, the method is built on principles from both chemical graph theory (in its analysis of molecular connectivity) and chemical information theory (in its analysis of heteroatom content). Bertz proposes that by considering graph theoretical invariants, which are features of the molecular graph solely based on the composition and structure of the graph, a reasonable measure of molecular complexity owing to connectivity $C(\eta)$ can be obtained. Borrowing from the mathematical form of Shannon's formula[63], Bertz defines $C(\eta)$ as the sum of connections (pairs of edges joined by a vertex) in an all-carbon molecular graph when corrected for symmetry by factoring out equivalent connections (Fig. 3a). The term $C(\eta)$ counts the number of nonequivalent 'propane' subgraphs within the parent molecular graph structure. To capture aspects of molecular complexity owing to heteroatom content, the function $C(E)$ is introduced, approaching zero for molecules with little atom diversity such as hydrocarbons and maximized for molecules with a large variety of heteroatoms present (hydrogen atoms are generally ignored in complexity metrics). The total complexity, C_T, is thus a function of both complexity owing to connectivity and heteroatom diversity (Fig. 3b). Overall, the Bertz C_T index accounts for molecular size, degree of branching, local symmetry, bond and ring count, and the presence of heteroatoms through these graph theory-based and information theory-based methods.

Since the seminal work of Bertz on a general approach to assessing molecular complexity, a number of additional methods have emerged. Hendrickson proposed a variant of the connectivity-based analysis of Bertz by counting the number of hydrogens appended to each carbon in the structure (that is, methyls, methylenes, methines and quaternary centres)[69]. Instead of enumerating the number of two-bond connections, as in the approach of Bertz, molecular topology could be described by simply accounting for how hydrogen atoms are distributed on the carbon skeleton. Because of the challenge associated with manually performing the complexity analysis of Bertz, modern computer algorithms have automated and operationally simplified many of these calculations[69,70]. Expanding on his initial work, in which propane subgraphs are counted to capture molecular complexity owing to connectivity, Bertz later proposed considering all possible subgraphs within a given structure[71] (Fig. 3c). These novel graph theoretical invariants, N_S (number of kinds of subgraphs) and N_T (total number of subgraphs), account for subgraphs containing heteroatoms while maintaining an approach fundamentally based in graph theory. Total walk count (twc) has also been proposed by Rücker as an effective index for characterizing molecular connectivity[72]. Beyond these reports, a whole host of approaches to measuring molecular complexity based on chemical graph theory have been reported: Randíc and coworkers proposes assessing molecular branching and self-avoiding paths in the molecular graph[73–75]; the chemically intuitive metric of Whitlock counts molecular features such as rings, chiral centres heteroatoms[76]; Barone and Chanon have expanded on the work of Whitlock by factoring in ring size[77]; Bonchev and coworkers have elaborated the Wiener index to include subgraph considerations[78,79]; Proudfoot considers the paths in the molecular graph emanating from the microenvironment of each atom[80,81]. Together, these applications of graph theory to organic chemistry attempt to represent the extent to which connectivity is an essential component of molecular complexity.

Methods based on chemical information theory

Many complexity metrics influenced by information theory still retain a strong bias towards using molecular graphs to represent chemical structures. Calculating the information content of a molecular graph has led to the development of similarity indices[82] for differentiating molecules on the basis of these graph-theoretical representations[64]. As a departure from solely characterizing molecular graphs, Bonchev proposed an atom-by-atom method for measuring information content based on individual features of atomic microenvironments[83]. This approach by Bonchev advanced the key assumption that atomic microenvironments are independent from one another and, therefore, should be represented as additive information-bearing variables. Recently, this application of information theory was extended by Böttcher to a novel molecular complexity index[66]. In Böttcher's method, each atom in a molecule is characterized by its valency, isomeric possibilities, and diversity of chemical groups or elements in its immediate microenvironment (Fig. 3d). Just like information in a string of binary digits can be measured by the number of bits, the sum of these individual atomic features gives rise to the total information content of the molecule measured in 'molecular complexity bits'. As such, significant weight

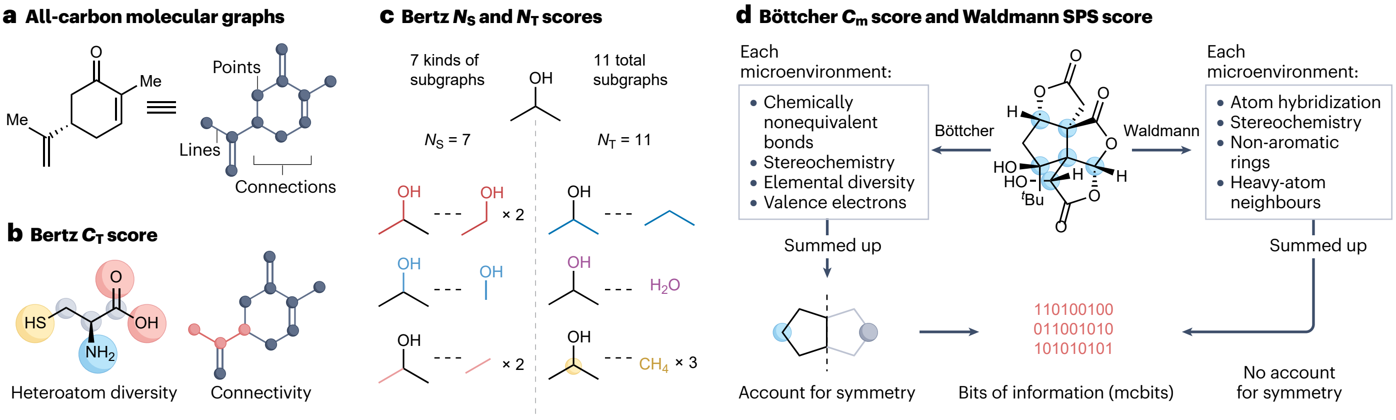


Fig. 3 | Computational representations of molecular structure and their use in complexity analysis. a, Representation of molecular structures as all-carbon molecular graphs, wherein atoms are points, bonds are lines, and connections two-bond units on the graph. **b**, Bertz's C_T score incorporating complexity from heteroatom diversity and connectivity. **c**, Bertz's method for counting subgraphs, N_S and N_T. **d**, The information theory-based methods of Böttcher (complexity metric (C_m)) and Waldmann (spacial score (SPS)) for evaluating structural complexity based on the microenvironment of each atom.

is assigned to factors such as stereochemistry and symmetry, which are often underappreciated elements of many graph theory-based methods. Interestingly, this complexity metric (C_m) scales with the number of peaks present in NMR spectra of alkane isomers. As a practicing synthetic chemist might intuit, symmetry-breaking features in a molecule which give rise to diastereotopicity can wildly complicate structural analysis by NMR spectroscopy – the method of Böttcher aptly captures this phenomenon.

Additive methods such as that of Böttcher quantify larger molecules as more complex than smaller ones. In terms of structural complexity, this principle is fairly intuitive: bimolecular coupling reactions generally increase overall molecular complexity; macromolecules bear more complexity than small molecules. However, for complex small molecules such as natural products, which pose unique synthetic challenges and often have biological signalling capabilities, the 'proximity' of structural features is a key aspect of its overall complexity and synthetic challenge. Shenvi and coworkers recently extended the method of Böttcher for measuring total information content to the concept of information 'density' by factoring in molecular volume[84]. Many biologically relevant natural products are densely functionalized and feature unusual structural motifs, and the manner in which these functional groups are packed in space gives rise to unforeseen emergent properties which complicate their synthesis (that is, increase their synthetic complexity)[85]. Biologically active natural products often engage protein receptors with exquisite selectivity owing to the high information content encoded in their complex structures, and indeed, many natural product analogues have led to approved drugs[86,87]. Therefore, to advance the field of synthetic methodology and facilitate access to biologically active molecular scaffolds, the construction of targets with high information density from fragments with low information density is a fundamental aim of total synthesis.

Building off the substructure analysis of the C_m score of Böttcher, Krzyzanowski, Waldmann and coworkers have recently reported the development of spacial score (SPS)[88], which aims to be more focused on topological complexity as it relates to biological function. In their analysis, although Böttcher's score has generated noteworthy interest in the organic synthesis community for scoring the complexity of chemical transformations, it has proven less useful as a descriptor for predicting biological selectivity and potency, as has been previously shown for scores such as F_{sp^3} and F_{Cstereo} (ref. 89). In the SPS metric, the microenvironment and local complexity of each atom are evaluated for atomic hybridization (penalizing unsaturation), stereoisomerism, presence of non-aromatic rings (rewarding connectivity) and number of heavy atom neighbours (prioritizing branching). These substructure scores are then summed across all atoms to arrive at the total complexity, which can be optionally normalized (nSPS) by number of heavy atoms to allow comparison of molecules of different molecular weights. In addition to using SPS to correlate structural complexity with biological activity and selectivity, the authors illustrate the intuitive nature of the SPS metric by illustrating Shenvi's bilobalide synthesis using an accompanying Python package available for automated scoring[84,90]. Compared to the 'first-principles' approach of Böttcher with C_m, SPS is more intentionally tailored to the kinds of structural features (that is, saturation, rings and so on) which might be relevant for assessing complexity as it relates to biological activity or total synthesis.

Quantifying synthetic complexity

Synthetic complexity captures how easily a particular target could be synthesized. As many synthetic chemists can surely attest, estimating the synthetic complexity of a molecule can be far from trivial. Although the number of synthetic steps required to access a target is often related to its structural features, determining synthetic complexity from structure alone can be incomplete. For example, steroid derivatives, which are structurally complex by virtue of containing several rings, stereocentres and heteroatoms, can often be prepared in a single step from readily available commercial material and, therefore, are synthetically less complex. In any retrosynthetic analysis, one must enumerate possible disconnections and synthetic precursors until either commercially available or previously reported starting materials are identified. In theory, the synthetic complexity of a given target is essentially a function of its retrosynthesis, representing the path length from starting material to final target. Of course, in practice, this estimated synthetic complexity is only validated upon completion of a synthesis when final step count and reaction efficiencies are experimentally determined. Although the exact definitions of a 'synthetic step', such as those described by Guerrero and coworkers[91] and Johnson[92], have varied historically, step count remains an important (albeit an imperfect) parameter for understanding synthetic complexity. There are, of course, many 'soft' practical considerations for encompassing the ease of synthesis of a compound: reaction efficiency (and overall yield), cost of materials, purification of intermediates or supply of starting materials or reagents. In the context of process chemistry, these become essential considerations when approaching manufacturing scales. The synthetic complexity of a target is also subject to change: as the field develops new synthetic methods and a wider range of starting materials and reagents, more efficient synthetic strategies can be realized, making this concept somewhat of a moving target. Of course, the rigorous evaluation of synthetic complexity requires proper experimental validation by completion of a synthetic route, making assessment of the synthetic complexity of a potential target somewhat of a circular task. Therefore, in this Perspective, we estimate synthetic complexity using a combination of existing metrics such as step count or synthetically challenging structural features as simplified approximations for this hard-to-measure concept.

Despite the many challenges associated with definitively measuring synthetic complexity, several methods have been reported which aim to estimate this useful property. In the context of medicinal chemistry, synthetic complexity scores can sort lead candidates by their ease of synthesis, highlighting promising targets which can be quickly prepared while discounting those whose synthesis would require considerable resources. To that end, Ertl and Schuffenhauer developed an approach at Novartis for estimating synthetic accessibility (SAscore) by accounting for structural motifs which have the greatest impact on the synthesizability of a target[93]. A fraction of the PubChem database was analysed to identify common structural features which are well-established in the literature, and these fragments contribute to synthetic accessibility. The method also accounts for structural features – such as rings, stereocentres or macrocycles – which generally complicate synthetic accessibility and are counted as penalties in the SAscore. Whereas estimating synthetic complexity based purely on structural features can overlook certain subtleties, this SAscore not only accounts for literature-precedented substructures but also comes at a considerably lower computational cost than enumerating retrosynthesis for each molecule.

Although assessing synthetic complexity by generating an exhaustive retrosynthesis tree for every target is a computationally demanding task, modern machine learning methods offer a cheaper alternative for measuring these properties. In the context of retrosynthesis algorithms, which are guided from target to starting material through productive reductions in complexity, a rapid method for quickly scoring many intermediates is needed. Coley and coworkers disclosed a learned synthetic complexity metric, SCScore, trained on 12 million reactions from the Reaxys database[94]. In contrast to many previous approaches, which measure structural or synthetic complexity using principles from domain expertise, SCScore models reported literature data to formulate a conception of synthetic complexity. The neural network was trained on reaction data with the key constraint that products have higher scores than reactants, and thus molecules are ranked in a pairwise fashion to maintain this requirement when assigned a score between 1 and 5. Notably, this approach does not contain an explicit database of commercially available compounds such as many retrosynthesis algorithms[95]; instead, SCScore learns from reaction data, implicitly, the types of molecules that tend to be starting materials and those which tend to be products. When SCScore is mapped onto existing drug syntheses, the tool appropriately characterizes each target compound as more synthetically complex than its precursor, a marked improvement over heuristic methods (such as SAscore[93] or length of SMILES[13]) that do not appropriately show synthetic complexity monotonically increasing over the course of a synthesis. Because SCScore is trained on the kinds of compounds published in the Reaxys database, evaluating infrequently appearing structures such as natural products is a fundamental challenge of this approach, and highly complex substrates often 'saturate' the function with scores approaching 5. Nevertheless, SCScore represents a novel method for evaluating synthetic complexity at low computational cost while maintaining high fidelity to the body of published synthetic transformations.

Synthetic complexity is a function of the relative ease of obtaining starting materials and the synthetic tools available at any point in time, evolving as the field advances and various bonds or structural motifs become easier to assemble. Eastgate and Li developed a metric for 'current complexity', a function of intrinsic factors related to structural complexity and extrinsic parameters which vary over time[14]. The authors' recognition that many aspects of complexity, both structural and synthetic, are often intuitively perceived by expert chemists prompted a small-scale survey for assessing human-perceived complexity ranking of a set of molecules. From these data, a regression model was developed using parameters such as Randić's molecular topology index[73], Baran's synthetic 'ideality' definition[96], and a series of heuristic factors such as number of stereocentres constructed and overall step count. Whereas intrinsic parameters such as molecular topology remain constant, extrinsic parameters such as step count are subject to change as the field advances. These principal factors identified by regression analysis were refitted with a probabilistic model which represents the current complexity as a distribution of scores, much like how a panel of chemists might collectively rank molecular complexity with differences in individual perception and biases.

Although measures of complexity – both synthetic and structural – have traditionally been built on the theoretical foundation of graph theory and information theory, another mode for evaluating these properties is human chemical intuition. Of course, an individual's assessment of complexity-defining molecular features is subject to their own perception and human biases, and complexity rankings can vary wildly from chemist to chemist. However, large aggregations of chemist rankings can suppress these biases by averaging individual scores, thereby yielding a 'crowdsourced' definition of complexity which might be more general than individual perception[97]. Sheridan and coworkers applied this crowdsourcing model to a collection of 386 chemists at Merck across several subdivisions within the company[98]. Users of a voting module were tasked with ranking groups of five molecules by their complexity, which was left intentionally undefined to capture unbiased conceptions of complexity. These data were then averaged to generate a meanComplexity score for each molecule, which yielded a series of notable findings: first, the manner in which molecules were drawn or represented (that is, explicit wedges or dashes, molecular orientation and so on) had a measurable, though not overwhelming, impact on the assigned score; second, individual chemists did not agree based on their assigned scores, but averaging over many voters yielded a self-consistent QSAR-based model; last, meanComplexity correlates reasonably well ($R^2 = 0.89$) with Ertl and Schuffenhauer's SAscore[93], revealing that the chemists polled understood complexity in a way that closely resembles synthetic complexity. Although this crowdsourcing approach has its limitations (that is, appropriate molecular representations and ambiguous definitions of complexity), it reveals the degree to which human intuition aligns with many of the theory-based definitions of molecular complexity.

Applications of molecular complexity analysis

Retrosynthesis and computer-aided synthesis planning

Because retrosynthesis is ultimately guided, as Corey states, by "the all-important goal of reducing molecular complexity"[10], the way in which chemists think about complexity has a substantial impact on which disconnections are favoured and which synthetic strategies are pursued in the laboratory. The "rules of retrosynthesis", codified by Corey and coworkers in the LHASA program[99–109], attempts to capture the many facets of molecular complexity with a series of heuristics derived from years of experience in organic synthesis rather than from a first-principles approach. Corey's 'Logic'[11] has prompted further analysis from a graph-theoretical standpoint to validate the now-accepted wisdom of organic synthesis. Bertz and Sommer[71], and later Bertz and Rücker[1,110], questioned whether these retrosynthesis principles hold true under a mathematical lens according to a set of previously discussed structural complexity metrics (N_S, N_T and twc)[111]. Re-examining a set of polycyclic synthetic targets once analysed by Corey[112], the authors rank each bond in the structure by greatest reduction in structural complexity according to the chosen metrics (Fig. 4). Compared to the heuristic-derived rules, many of the top-ranked bonds by the aforementioned complexity metrics are in agreement: bonds that are directly attached to (that is, are *exo* to) another ring or are contained within the maximally bridged ring generally result in the greatest reduction in complexity. The strategy of identifying maximally bridged rings has been used by Sarpong and coworkers to complete the synthesis of several complex diterpenoid alkaloids[113,114]. One notable exception in which the LHASA heuristic is, in fact, contramathematical is the consideration of central fusion bonds which, when disconnected, generate macrocyclic intermediates (ring size > 7). According to Bertz and Sommer[71], breaking these transannular bonds often result in the most complexity-reducing disconnections to the molecular graph (Fig. 4), but owing to the historical paucity of efficient methods to construct carbocyclic macrocycles, transannular disconnections are explicitly discouraged by Corey. Newer developments in synthetic methodology that enable the rapid construction of precursor macrocycles[115–119] reveal new strategic possibilities which, in alignment with the mathematical analysis of Bertz and Rücker, ought to be considered for the efficient construction of complex, polycyclic targets by transannular bond formation.

In the Digital Age, the adaptation of retrosynthesis logic to computer programs and the automation of synthesis planning has been

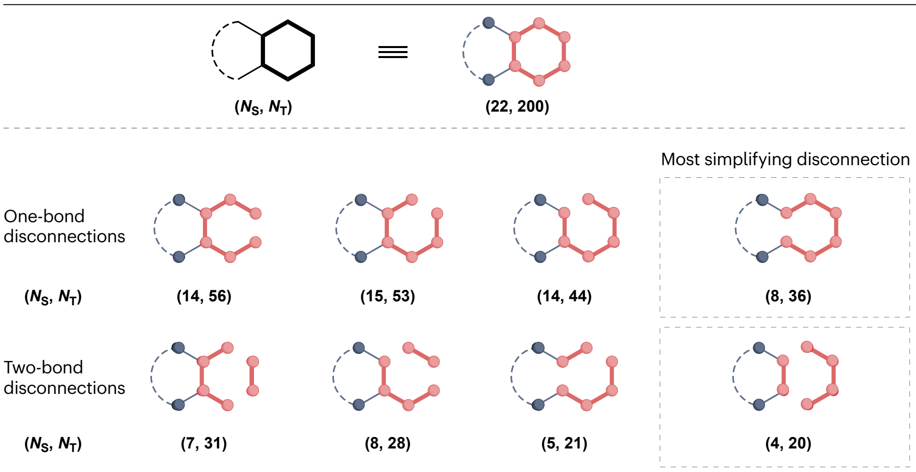


Fig. 4 | Complexity analysis of fused ring systems. Evaluation of various disconnections in a fused ring system according to N_S and N_T metrics, as shown by Bertz and Sommer[71].

a longstanding effort at the interface of chemistry and computer science[13,120]. The LHASA program developed by Corey set the stage for the development of countless additional programs aimed at streamlining computer-aided synthesis planning (CASP): Chematica/SYNTHIA[95], ASKCOS[59], ICSYNTH[121], IBM RXN[122], AiZynthFinder[123], WODCA[124], and many others. Retrosynthesis algorithms typically follow a similar form to human-performed retrosynthetic analysis (Fig. 5a). Currently used retrosynthesis programs make use of expert-coded reaction templates[95] or machine learning techniques[46,57,125] to learn possible chemical transformations and, by recognizing relevant structural motifs or functional groups present in the target molecule, propose a series of disconnections. The number of possible disconnections at any given stage can be extremely large owing to the 'combinatorial explosion'[126] of recursive multistep retrosynthesis searches; efficient navigation of this immense search space is handled by scoring functions, which rank disconnections and highlight those which lead to the shortest achievable path accordingly (Fig. 5b). With SYNTHIA, for example, proposed routes can also be ranked by user-defined criteria such as use of protecting groups, number of steps, or the possibility of reactivity conflicts (for example, a Grignard reaction in the presence of an additional carbonyl functional group). Often, in this structure search, the most 'downhill' disconnections lead to precursor compounds which are structurally dissimilar from the parent molecule. Therefore, one might envision that measures of molecular similarity[127,128] such as Tanimoto coefficients[129] could be used in cheminformatics and indirectly capture some aspects of this analysis. However, structural complexity has a higher tendency to guide a retrosynthesis back to simple building blocks, which are commercially available. Currently, programs such as SYNTHIA use oversimplified complexity measures based on the length of the SMILES string[13]. However, insofar as reducing structural complexity is a primary goal of retrosynthesis, there is significant opportunity to use more sophisticated measures of complexity in these synthesis planning programs to potentially improve algorithm performance. The growing field of machine learning has made ample use of mathematical optimization methods for multidimensional functions[130], but in order for these powerful tools to take hold in the realm of CASP, the complexity landscape must first be quantitatively well-defined. Therefore, metrics for complexity — the 'distance' in both synthetic and chemical space[131,132] from readily available starting materials — are particularly significant in the context of CASP and merit further development from both synthetic and computational chemists.

Recently, Cernak and coworkers applied this concept of 'graph edit distance' to the enantioselective synthesis of the alkaloid stemoamide[133]. Although there are over 30 previous approaches to stemoamide, application of SYNTHIA to the natural product highlighted a novel Mannich transformation which had not been previously reported. However, inefficiencies in the rest of the route made the SYNTHIA proposal less competitive compared to previous reports, and human intervention was required to shorten the sequence. Ultimately, Cernak and coworkers leveraged graph edit distance as a way to prioritize high-impact steps in a machine-readable format and synergize with programs such as SYNTHIA which contain a vast database of known reactions. After developing a first-generation approach to stemoamide via an organocatalyzed Mannich reaction, graph edit analysis highlighted which part of the route could be further improved by cutting out functional group interconversions, resulting in a three-step second-generation approach featuring two key steps: an auxiliary-directed Michael addition and a final Aubé–Schmidt rearrangement.

Analysis and evaluation of total syntheses

If one of the goals of total synthesis is to navigate complexity space from starting material to a target compound through the most efficient path, then measures of structural complexity can be useful in describing how various syntheses traverse this landscape. In addition to the frameworks of 'redox economy',[134] 'atom economy'[135] or 'step economy',[136] which have been used in retrosynthetic analysis, 'complexity economy' has also been used as the basis for comparing synthetic strategies to classic complex molecules such as Taxol, strychnine and longifolene[76,77,81,137]. Plotting the structural complexity of each intermediate in a synthetic route illustrates in two dimensions how each step contributes to reaching the final target. In general, this kind of analysis can provide valuable insight into how molecular complexity is navigated in the total synthesis of natural products. These analyses also raise intriguing fundamental questions about synthetic strategy: for example, in an 'ideal synthesis', should structural complexity be generated at the beginning or end of a route? Can strategies for complexity generation be tailored to the purpose of the synthesis (for example, late-stage diversification for preparing libraries of analogues)? Are there special considerations for 'overbred' intermediates, which contain excess structural complexity relative to the target? Molecular complexity, when quantified, can be a useful analytical framework

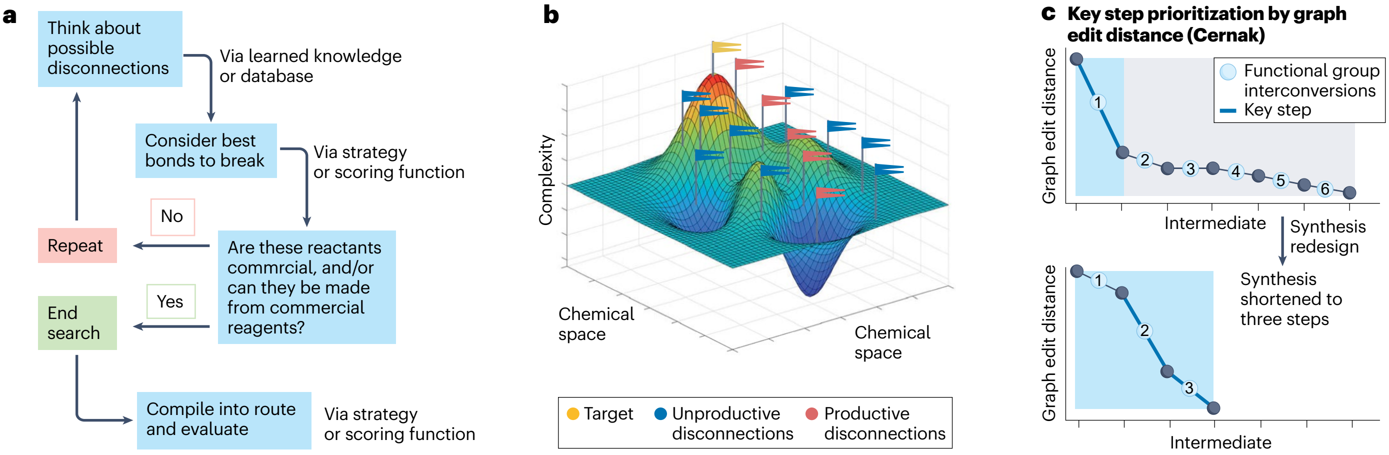

a
Think about possible disconnections
Via learned knowledge or database
Consider best bonds to break
Via strategy or scoring function
No
Repeat
Are these reactants commrcial, and/or can they be made from commercial reagents?
Yes
End search
Compile into route and evaluate
Via strategy or scoring function
b
Complexity
Chemical space
Chemical space
● Target
● Unproductive disconnections
● Productive disconnections
c Key step prioritization by graph edit distance (Cernak)
Graph edit distance
Functional group interconversions
Key step
1
2
3
4
5
6
Intermediate
Synthesis redesign
Synthesis shortened to three steps
Graph edit distance
1
2
3
Intermediate

Fig. 5 | Complexity analysis applied to retrosynthesis. a, General algorithm for retrosynthesis performed by human or computer. **b**, Navigation of complexity space as the primary goal of retrosynthetic analysis. **c**, Prioritization of graph edit distance by Cernak and coworkers[133] led to the development of a three-step synthesis of the alkaloid stemoamide.

for addressing these kinds of issues. To this end, we examined three paradigms for navigating structural complexity (late-stage, early-stage and 'excess' complexity generation) and applied the SPS complexity score of Krzyzanowski and Waldmann[88] to evaluate the paths traversed by total syntheses within these categories. Discussion of how these synthetic efforts take distinct approaches to the target is illustrated by complexity plots, which depict the complexity analysis for each synthetic intermediate.

Building complexity at a late stage

In terms of retrosynthetic analysis, achieving rapid reduction in molecular complexity when working backwards from a molecular target is a primary goal of synthetic planning. This type of approach often entails, in the forward sense, a late-stage introduction of stereocentres, rings or key structural motifs by powerful complexity-building transformations. Strategically, one might opt for this late-stage complexity paradigm to build up the necessary fragments in a reliable fashion (without emergent properties[85] thwarting standard transformations) and attempt a final multi-bond forming process. There is, however, a significant degree of risk in planning the most ambitious transformation for the end of a route, and synthetic flexibility is often a consideration at the outset of a total synthesis. Nevertheless, many such strategies aim to model proposed biosynthetic pathways and involve nature-inspired cascades to assemble polycyclic architectures from linear precursors. For instance, in the late 1960s, Johnson investigated the proposed cationic polyene cyclization for the biosynthesis of steroids such as (±)-16,17-dehydroprogesterone (**3**)[138] (Fig. 6a). Treatment of tertiary alcohol **4** with tin(IV) chloride as a Lewis acid resulted in the stereospecific cationic cascade to forge three bonds and three rings and set five stereogenic centres (+106% increase in SPS score[88] in going, for example, from **4** to **5**; Fig. 6a, right). Following ozonolytic C–C double bond cleavage on **5** and subsequent aldol cyclization, the final steroid target (**3**) was assembled. Since these pioneering studies in cationic polyene cyclizations, others have extended this approach to additional terpene scaffolds[139–141] and demonstrated analogous radical-mediated cyclizations[142–145] to access high-complexity sp^3-rich intermediates from simpler linear precursors, as illustrated in the complexity analysis (Fig. 6a, right).

This late-stage complexity logic is also evident in Nicolaou's synthesis of the endiandric acids[146–149], in which polyene-yne **6** (Fig. 6b) is subjected to Lindlar reduction conditions followed by a remarkable cascade of thermally promoted pericyclic reactions to access both endiandric acid B (**7**) and C (**8**) in their respective methyl ester forms. This series of pericyclic processes – an 8*π* electrocyclization followed by a 6*π* electrocyclization and final [4 + 2] Diels–Alder cycloaddition – not only validates a proposed biosynthesis but also constitutes a rapid rise in molecular complexity (+238% increase in SPS score in going, for example, from **6** to **7**; Fig. 6b, right) as appreciated by the newly forged rings and stereocentres.

In this same vein, Heathcock and Piettre conducted a series of studies on the proposed biosynthesis for the *Daphniphyllum* family of alkaloids[150,151] (Fig. 6c). Starting from polyene **9** (prepared in six steps), treatment with methylamine followed by acetic acid facilitated amine condensation, two aza-Prins-type cyclizations, and amine-mediated hydride transfer to yield dihydro-*proto*-daphniphylline (**10**) in 65% yield. These studies have since served as the foundation for understanding the biosynthesis of newly isolated *Daphniphyllum* alkaloids in the yuzurimine, calyciphylline and daphnilactone subfamilies[152]. Similar to the work of Johnson[138] and Nicolaou[146–149], as shown in the accompanying complexity analysis plot, Heathcock's synthesis generates significantly more structural complexity (+233% increase in SPS score) in the final biomimetic polycyclization step (that is, **9**→**10**, Fig. 6c, right).

As another paradigm for late-stage generation of structural complexity, transannular bond-forming processes have emerged as powerful transformations for the efficient synthesis of polycyclic natural products. Construction of macrocyclic intermediates from linear precursors can enable the concomitant formation of several bonds in a transannular fashion, providing access to complex scaffolds that otherwise might be challenging to assemble by discrete ring-forming operations. One notable example of this strategy is the parallel efforts of Evans[153] and Sorensen[154] towards (+)-FR-182877 (**11**) and (−)-FR-182877 (**12**) (Fig. 6d), in which **13**, constructed through a series of palladium-mediated cross couplings and alkylation steps by Evans, is allowed to undergo cyclization to macrocycle **14**. Arriving at the desired pattern of unsaturation, both approaches report a tandem transannular Diels–Alder and hetero-Diels–Alder cycloadditions from macrocycle **14** to forge all four ring fusion bonds and afford the pentacyclic core of the natural product, representing a significant spike in calculated structural complexity (Fig. 6d, right). A short sequence of steps from pentacycle **15** achieved the transannular lactonization to complete the synthesis of both enantiomers of FR-182877 ((+), Sorensen; (−), Evans). As shown in the complexity analysis in Fig. 6d,

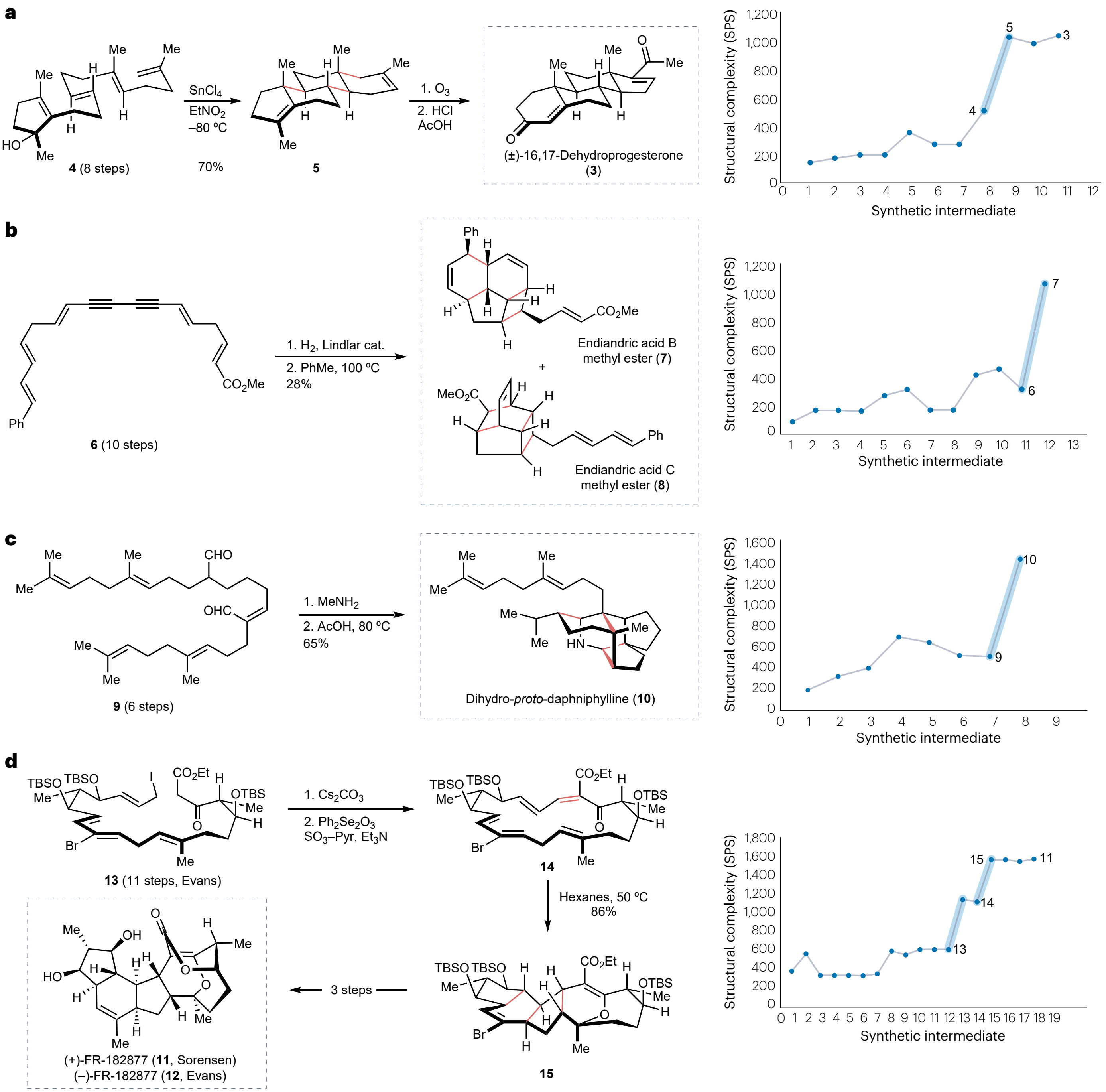


Fig. 6 | Analysing the late-stage generation of molecular complexity. Selected example of late-stage molecular complexity generation in total synthesis with accompanying complexity analysis, as scored by SPS (ignoring protecting groups). Key bond-forming steps (shown in the graphs) are highlighted in blue. **a**, The key cationic polyene cyclization in Johnson's synthesis of (±)-16,17-dehydroprogesterone (**3**). **b**, Nicolaou's late-stage pericyclic cascade to access endiandric acids B (**7**) and C (**8**). **c**, Heathcock's evaluation of a biomimetic cascade to establish the dihydro-proto-daphnphylline (**10**) scaffold. **d**, Evans' and Sorenson's parallel approaches to FR-12877 (**11**, **12**) through macrocyclization and transannular [4+2] cycloadditions. cat., catalyst.

both the macrocyclization step (+102% increase) and transannular [4 + 2] cycloadditions (+43% increase) contribute significantly to overall SPS score. However, 11 steps (LLS) were required to build the precursor for macrocyclization, highlighting the need for more expedient approaches to macrocyclic precursors if such approaches are to be more widely adopted.

Building complexity at an early stage

The second paradigm for understanding the role of molecular complexity in synthesis planning features a rapid rise in structural complexity in the early stages of a synthetic route. Carefully planned couplings of easily accessible building blocks can give rise to a remarkable degree of structural complexity in a very short sequence of steps. One of the inherent challenges of this strategy is that, upon achieving a high level of complexity, emergent properties and unusual reactivity can arise[85]. These unforeseen factors can significantly complicate the rest of the synthesis and generally require additional steps to circumvent unproductive or deleterious reactivity[155]. One notable demonstration of this early-stage paradigm is Crimmins' synthesis of ginkgolide B (**16**)[156], in which a seven-step sequence from 3-furaldehyde to **17** (Fig. 7a) enabled a [2 + 2] photocycloaddition to yield tetracycle **18**, which contains the four quaternary centres present in the natural product. After this rapid rise in complexity, 14 additional steps were required – comprising a series of ring expansions, redox manipulations, and installation of the final gamma lactone – to complete the 22-step synthesis of ginkgolide B (**16**). Interestingly, plotting the SPS score of this route reveals several insights. First, setting the stage for the key step requires some up-front investment; however, modest gains in complexity in the first seven steps (+371 SPS) are rewarded with a significant spike in complexity (+730 SPS) at the [2 + 2] stage (that is, **17** to **18**). Second, although there is a clear jump in skeletal complexity at this stage, many redox manipulations are necessary to traverse the complexity landscape from **18** to the target (**16**), highlighting the challenge of late-stage oxidations on scaffolds which are already densely functionalized.

More recently, highly complex taxoids such as canataxpropellane (**19**, Fig. 7b) have been targeted by Gaich and coworkers[157]. In their synthetic strategy, furan dienophile **20** was treated with dienone **21** to initiate a thermal Diels–Alder cycloaddition followed by UV irradiation to promote an intramolecular [2 + 2] cycloaddition to yield highly caged [4.4.2] propellane **22** in an efficient two-step sequence. This extremely rapid generation of molecular complexity as reflected in the SPS score plot (see Fig. 7b, right) sets the requisite cyclobutene core and provides a platform for a series of scaffold rearrangements, oxidations and functional group manipulations that advanced **22** to canataxpropellane (**19**) in an additional 24 steps. However, although the rapid generation of complexity (+687% increase for **20**→**22**, Fig. 7b, right) to access the rigid, densely functionalized cyclobutane is impressive, this strategic move probably added an extra constraint for navigating downstream reactivity and performing the necessary rearrangements, oxidations and functional group manipulations over the ensuing 24 steps, wherein there is a more modest +51% complexity increase (an average of 2% per step). Nonetheless, the early entry into the cantaxpropellane scaffold nicely demonstrates how sequential cycloadditions can give rise to remarkable levels of structural complexity.

In many syntheses, mimicry of synthetic strategy of nature (that is, biosynthesis) can provide inspiration for an early-complexity strategy. The two-phase approach of Baran to several terpenoids[33,158–163] is designed to model the synthetic strategy of nature, in which the topologically complex carbon skeleton is first rapidly assembled in a 'cyclase phase' and the oxidation pattern is introduced in a subsequent 'oxidase phase'. This two-phase logic was notably applied to ingenol (sold commercially as the mebutate ester, Picato; **22**, Fig. 7c), an especially challenging diterpene with a high degree of both structural and synthetic complexity. Starting from enantiomerically enriched (+)-carene (**23**), a seven-step sequence of alkylations and an allenyl-Pauson–Khand reaction produced tetracycle **24** – a net increase of six C–C bonds formed in the cyclase phase. In the subsequent oxidase phase, peripheral oxidation of the five-membered and seven-membered rings, along with the key vinylogous pinacol rearrangement, yielded ingenol (**22**) in 14 total steps. As depicted in the complexity analysis (Fig. 7c, right), the 'cyclase phase' clearly contributes skeletal complexity, as scored by the SPS metric (an average +26% increase per C–C bond-forming step, highlighted in blue). In the oxidase phase, redox manipulations (highlighted in pink), have a measurable but attenuated impact on SPS complexity (average +8.8% complexity per oxidation step), with several steps in this phase accompanied by functional group manipulations, which are often necessary for managing complex oxidation patterns.

Many molecules which contain a high level of structural complexity may, in fact, have very low synthetic complexity or are commercially available. The so-called chiral pool – the collection of naturally occurring building blocks which already contain one or more stereogenic centres – has long been a rich source of purchasable molecular complexity and, thus, a convenient starting point for many synthetic routes[164]. One notable use of the chiral pool as a source of structural complexity is Maimone's synthesis of *Illicium* sesquiterpenes (−)-majucin (**25**, Fig. 7d) and (−)-jiadifenoxolane A (**26**) from terpene feedstock (+)-cedrol (**27**)[165,166]. Making use of the existing complexity in the carbon framework of **27**, site-selective C(sp^3)–H oxidations and ring fragmentations enabled the synthesis of enol lactone **28**, from which the majucin and anisatin frameworks were established. Not only did lactone transpositions and late-stage oxidations completed the syntheses of (−)-majucin (**25**) and (−)-jiadifenoxolane A (**26**), but also the reported oxidative sequence constituted a formal synthesis of three additional *Illicium* sesquiterpenes. Given the growing development of the field and the application of site-selective C–H functionalization[167,168], readily available compounds such as (+)-cedrol (**27**), which contain high levels of structural complexity (SPS = 939, see plot at Fig. 7d, right), can serve as synthetic platforms for accessing a wide array of highly oxidized terpenoids.

Leveraging excess complexity

Chemical transformations that enable the rapid generation of complexity are prized in the synthetic toolkit. In many cases, the amount of complexity generated can exceed that of the synthetic target, that is, the intermediate produced is 'more complex' than the final goal. In many cases, this is observed with the use of protecting groups, wherein the fully protected penultimate intermediate is structurally more complex than the final target. Although this excess complexity can be seen as somewhat misleading (because protecting groups are not target-relevant complexity[169]), this can be important to consider in the context of retrosynthesis planning algorithms, wherein allowing for temporary hikes in complexity can lead to two-step disconnections that are overall simplifying[170]. Other instances are more nuanced: such compounds, termed 'overbred intermediates' by Hoffmann[171] and others[172], often contain excess C–C bonds, rings or stereocentres relative to the target and require bond cleavage processes to complete the synthesis. Insofar as additional synthetic steps are required to achieve this 'excess complexity', more direct routes through complexity space

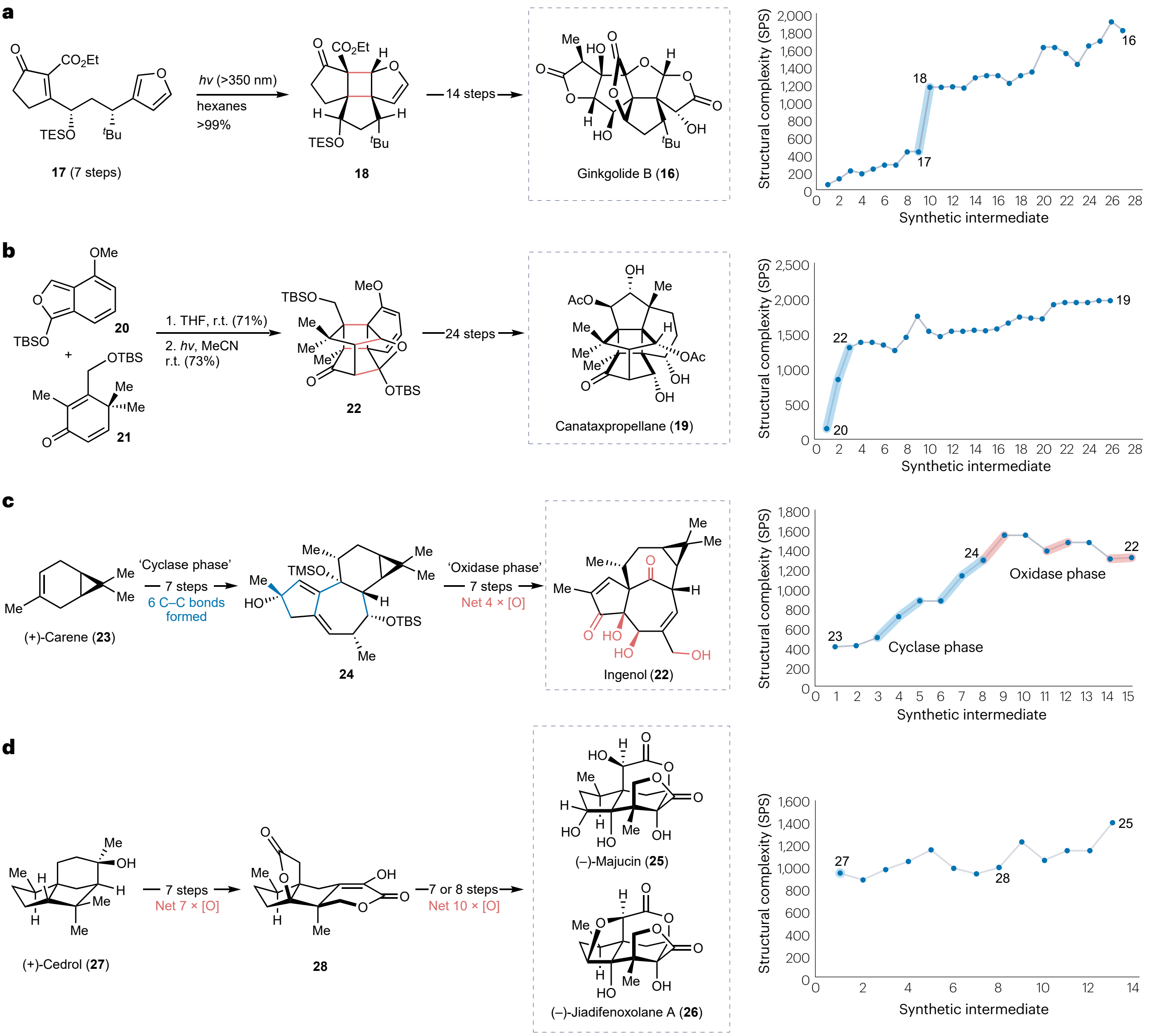


Fig. 7 | Analysing the early-stage generation of molecular complexity. Selected examples of early-stage complexity generation in total synthesis with accompanying complexity analysis, as scored by SPS (ignoring protecting groups). In the graphs (right), key bond-forming steps are highlighted in blue and redox manipulations in pink. **a**, Early-stage [2+2] photocycloaddition in Crimmins' synthesis of ginkgolide B (**16**). **b**, Gaich's Diels–Alder-[2+2] photocycloaddition sequence to rapidly build structural complexity en route to canataxpropellane (**19**). **c**, Baran's two-phase approach to ingenol (**22**), where C–C bonds are first constructed to build structural complexity (cyclase phase) before redox manipulations complete the synthesis (oxidase phase). **d**, Maimone's approach to (–)-majucin (**25**) and (–)-jiadifenoxolane (**26**) starting from (+)-cedrol (**27**). r.t., room temperature.

are seen as ideal[96]. However, if high-complexity structures can be easily prepared, C–C bond cleavage strategies[173] that proceed downhill from a point of maximum complexity can be used advantageously in accessing challenging structural motifs.

Wender's synthesis of α-cedrene[174] (**29**, Fig. 8a) is an example in which generation of excess complexity sets the stage for an exceedingly short synthesis. It is also a synthesis that illustrates the power of photochemical reactions: an arene–olefin cycloaddition was performed from anisole **30** to yield an equimolar mixture of isomers (**31** and **32**), each containing the *endo*-fused cyclopropane motif. Treatment of this mixture with bromine selectively cleaved the desired C–C bond (Fig. 8a, red) in both isomers to complete the cedrene carbon framework. Reductive dehalogenation and Wolff–Kishner reduction afforded α-cedrene (**29**) in a remarkable five steps from previously

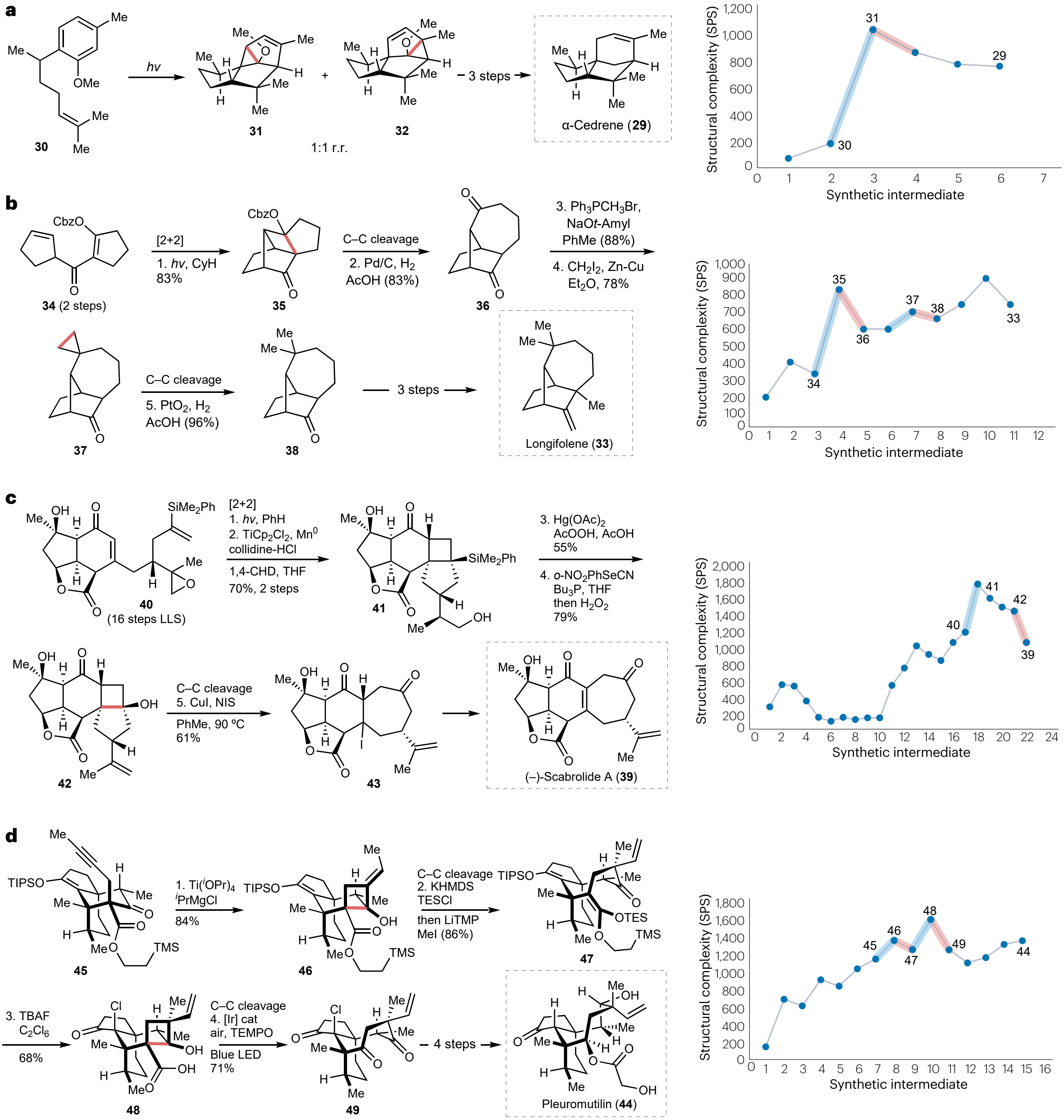


Fig. 8 | Excess complexity generation and the utility of C–C cleavage steps. Selected examples of excess complexity and C–C cleavage in total synthesis with accompanying complexity analysis with SPS score (ignoring protecting groups). In the graphs (right), bond-forming steps are highlighted in blue and bond-cleaving steps in pink. **a**, Wender's synthesis of α-cedrene (**29**) with an arene–olefin photocycloaddition followed by C–C cleavage. **b**, Oppolzer's total synthesis of longifolene (**33**) involving several C–C bond-breaking transformations. **c**, Total synthesis of scabrolide A (**39**) from Stoltz featuring a late-stage [2+2]-C–C cleavage sequence. **d**, Pronin's route to pleuromutilin (**44**) involving the strategic formation of cyclobutanol intermediates for subsequent C–C cleavage. LED, light-emitting diode.

reported materials. As modelled with the complexity analysis (Fig. 8a, right), the arene–olefin cycloaddition from **30** to **31** represents a massive increase in structural complexity (blue, +406% increase), after which a series of bond-cleaving steps (pink, 35% decrease over three steps) establish a concise path to **29**.

The approach of Oppolzer to longifolene (**33**, Fig. 8b) similarly couples photochemical bond-forming steps with subsequent bond-breaking events to access complex bicyclic scaffolds[175]. Photoirradiation of enone **34** yielded cyclobutane **35**, a complex tetracyclic intermediate which contains the bicyclo[2.2.1]heptane core present in longifolene and, according to complexity analysis (+145% SPS increase), is the point of maximum complexity in this approach (Fig. 8b, right). Palladium-mediated hydrogenolytic removal of the Cbz group resulted in spontaneous cleavage of the endocyclic C–C bond in **35** (Fig. 7b, red), resulting in the characteristic seven-membered ring found in **36** after an accompanying reduction in SPS complexity (28% decrease). To install the *gem*-dimethyl unit, Wittig olefination and Simmons–Smith cyclopropanation afforded spirocyclopropane **37**. In a second C–C cleavage transformation, hydrogenolysis of the cyclopropane with Adams' catalyst furnished the *gem*-dimethyl group in **38**, enabling completion of the synthesis of **33** in an additional three steps.

More recently, Stoltz and coworkers reported a total synthesis of (−)-scabrolide A (**39**, Fig. 8c), a norcembranoid diterpenoid, with a key photocycloaddition/C–C cleavage sequence[176]. In their approach, irradiation of enone **40**, followed by reductive epoxide opening, gave cyclobutane **41** as a single isomer. At this stage, oxidation of the tertiary alkyl silane with $Hg(OAc)_2$/AcOOH and elimination of the hydroxy group to the isopropenyl unit afforded cyclobutanol **42**, setting the stage for a transannular C–C cleavage. Treatment of cyclobutanol **42** with CuI/NIS facilitated in situ formation of the corresponding hypoiodite, which underwent homolytic fragmentation, recombination and elimination of the intermediate tertiary iodide (see **43**) to give (−)-scabrolide A (**39**), thus, completing the total synthesis. Interestingly, complexity analysis (Fig. 8c, right) illustrates how intermediate **40** is at the same level of structural complexity as the target, **39**, and yet traversing through a photocycloaddition/C–C bond-cleaving sequence (+49% SPS increase, then 27% decrease) was necessary to access the natural product.

The Pronin synthesis of pleuromutilin (**44**, Fig. 8d), a terpenoid with promising antibacterial properties, stands as a powerful final example of how navigating excess complexity can enable the assembly of highly complex natural product scaffolds[177]. Alkyne **45**, prepared in a rapid six-step sequence, was subjected to Ti-mediated reductive cyclization conditions to afford cyclobutanol **46**. Fragmentation of the endocyclic C–C bond by treatment with excess strong base allowed for tandem functionalization of the resulting extended enolate, giving **47** after methylation. Remarkably, fluoride-mediated cleavage of the silyl ketene acetal re-established the cyclobutanol motif to yield **48**, allowing for further C–C functionalization at that position in a subsequent oxidation. Complexity analysis shows this phase of the route (Fig. 8d, right) as an iterative complexity generation–reorganization sequence, wherein formation of the strained cyclobutanol provided opportunities for key bond-forming events in subsequent steps, ultimately enabling the synthesis of pleuromutilin (**44**) in a total of 12 steps.

In general, because of the requirement for subsequent C–C bond cleaving reactions after introducing excess complexity, the majority of approaches that have used this strategy rely on forming strained rings (three-membered and four-membered). In these cases, C–C bonds are much more easily cleaved using existing technology. It is anticipated that with advances in methods for C–C bond cleavage in less strained systems, there will be significantly more opportunities to use the excess complexity generation strategy in synthesis, especially of terpenoids.

Cheminformatics and medicinal chemistry

In the realm of medicinal chemistry, molecular complexity has been proposed as a key metric for understanding drug–target interactions[89]. Developing chemical descriptors and molecular representations for small molecules in large virtual libraries has been proposed as a useful data-driven approach for predicting drug success or filtering out intractable compounds that possess undesirable properties for drug development[178–180]. In the era of Big Data, cheminformatics and QSAR analysis have emerged as some of the primary pillars of modern drug discovery, and the descriptor of molecular complexity — defined according to a number of previously described methods — has found its place in these disciplines[181,182].

Simple proxies for molecular complexity have driven powerful observations in drug development. Notably, Lovering and coworkers proposed that more complex small molecule clinical candidates — those with higher numbers of sp^3-hybridized carbons and stereogenic centres — tend to advance further in the drug approval process[183]. Much in the same way as Lipinski proposed the 'rule of five' for characterizing the physical properties of successful drug candidates[184], Lovering claims that F_{sp^3} (number of sp^3-hybridized carbons/total carbon count) and stereocentre content are directly correlated with clinical success. The central hypothesis for this relationship is the ability of sp^3-rich compounds to access more diverse chemical space, better position functional groups for improved receptor–drug complementarity, and yield more favourable solubility profiles. Furthermore, high sp^3 carbon and chiral centre count can impart greater selectivity to drugs owing to the three-dimensionality and high information content. In a subsequent report, Lovering further interrogated this selectivity hypothesis by examining assay data for roughly 7,000 compounds against a panel of proteins (Cerep) and cytochrome P450 (CYP) enzymes routinely screened in discovery chemistry at Pfizer[185]. Sorting these compounds according to their molecular complexity (F_{sp^3} and number of chiral centres) revealed an inverse relationship between complexity and assay promiscuity. This confirmed the original hypothesis that more complex molecules, as approximated by F_{sp^3} and stereocentre count, have greater specificity and selectivity in their interactions with biological targets and even higher tendency to evade CYP enzymes, which can facilitate metabolic degradation.

The proposed relationship between molecular complexity and biological function has also found relevance in diversity-oriented synthesis (DOS) and combinatorial chemistry for generating performance-diverse compound libraries[186,187]. To further probe the promiscuity of various subcollections of molecules (natural products, DOS-generated compounds and commercial compounds), Clemons and coworkers systematically carried out a 100-protein assay on a large compound collection[188]. By comparing molecular complexity (fraction of stereogenic carbons) to biological data, the authors described how natural products exhibit high selectivity but have less tendency to yield hits; conversely, commercial compounds display high promiscuity but have higher tendency to result in hits, that is, show activity against several proteins. Designed to have low synthetic complexity, DOS-derived compounds showed reasonable selectivity and activity, making them a potentially attractive alternative to 'simple' commercial compounds or synthetically complex natural products. Indeed, many groups have reported diversity-oriented synthetic strategies to structurally complex 'natural product-like' compounds in a modular and concise number of steps, some of which show promising biological activity[117,189,190]. In medicinal chemistry, molecular structure determines biological function. In this context, the concept of structural complexity – as a particularly important structural descriptor – can prove useful as an indicator of small molecule clinical success or biological target selectivity.

Conclusion and outlook

The generation and manipulation of molecular complexity is a primary goal of the field of synthetic chemistry. However, this field-defining concept is frequently invoked on an intuitive basis without precise definition or appreciation of its subtleties. Although human intuition about complexity can be roughly useful for retrosynthetic analysis, intuition among groups of chemists can remain highly variable and is often riddled with biases and imprecisions[98]. However, current objective implementations of complexity scores in retrosynthesis programs, such as the length of a SMILES string[13], remain somewhat crude, overlooking many subtleties in planning efficient disconnections. Although programs such as SYNTHIA have demonstrated proficiency in planning routes to natural products[170], there is much room for growth for tackling more complex targets. If computer-assisted synthesis planning is to reach greater maturity in the near future, analysing molecular structures and developing synthetic routes much in the same way as expert-trained chemists will require further refinement of this cornerstone concept into a rigorous analytical tool. In this process, complexity analysis might highlight new desirable disconnections that are not yet achievable with the current synthetic toolkit, inspiring the community to develop new methods to close the gap.

In this Perspective, the definitions, methods for quantification and applications of molecular complexity were reviewed. Despite the multitude of analytical methods that attempt to rigorously quantify and measure molecular complexity, however, it remains a somewhat elusive subject. One might reasonably wonder if these complexity analyses – with their own subjective formulations and approaches – are any different than mere human intuition. Each method, with its own assumptions, theoretical bases and computational tradeoffs, invariably fails to capture the whole picture with a single lens. However, complexity analyses do not seek to be holistic or objective for all scenarios. The 'ideal' complexity metric is context-dependent, and their value is demonstrated in their ability to aid in planning successful synthetic routes or inspire new innovations in the field, as shown by Cernak and coworkers. Although each approach indeed bears its own assumptions, they draw from conceptual frameworks such as graph theory or information theory which individually crystallize our understanding of molecular structure, with each analytical framework being a different way to model the ensemble of molecular features. These metrics are, in aggregate, useful for characterizing the state of synthesis in a more rigorous, reliable and reproducible fashion, and they can teach the community how to navigate this landscape more efficiently.

Where do new opportunities remain? Many previous methods for quantifying complexity start with well-understood variables or theoretical frameworks to define complexity from the bottom up. However, with large datasets of composite or crowdsourced complexity scores, such as that from Sheridan and coworkers[98], one might envision machine learning workflows capturing subtle elements of complexity that no single method could report. Such methods could aid in proposing new disconnections that break conventional heuristics but result in efficient syntheses. Furthermore, analysis of human-defined complexity, resulting from years of organic chemistry training, experience and time-tested intuition, might uncover new discrete principles or 'rules' that could simplify retrosynthetic analysis beyond the original LHASA work of Corey. Re-training retrosynthesis algorithms with these data-driven heuristics, in turn, has the potential to render computer-assisted synthesis planning even more powerful. Finally, the development of standardized benchmarks for complexity metrics (for example, by evaluation of a standard set of total syntheses or test compounds) can be used to assess the value of new approaches and allow for more rigorous comparison of these molecular descriptors for future applications.

Synthetic chemists are drawn to building complexity through synthesis: can we be more rigorous with how we measure our progress? Quantifying and characterizing molecular complexity with analytical methods will not only bring needed clarity to an oft-nebulous term, but this can also point the field in new directions for improving how complex structures are assembled. For total synthesis, characterizing the complexity landscape can aid in understanding the successes (and failures) of various synthetic strategies, leading to a better grasp of how one might maximize structural complexity at every step while minimizing the synthetic complexity of medicinally valuable targets. The advent of computer-assisted synthesis planning invites more systematic, reproducible methods for the analysis of complex molecules. Just as the development of LHASA prompted Corey to codify a set of general rules for retrosynthetic analysis, developing new ways to describe molecular complexity has the potential to yield similarly useful applications. Perhaps there are new rules for synthesis that are yet to be formulated, a logic based on algorithm-calculated complexity as the guiding principle. Further development of molecular complexity analysis and its applications can refine the synthetic organic chemist's understanding of their craft and scout new directions for advancing the field.

Published online: 09 September 2024

References

1. Rücker, C., Rücker, G. & Bertz, S. H. Organic synthesis — art or science? *J. Chem. Inf. Comput. Sci.* **44**, 378–386 (2004).
2. Whitesides, G. M. & Ismagilov, R. F. Complexity in chemistry. *Science* **284**, 89–92 (1999).
3. Goldenfeld, N. & Kadanoff, L. P. Simple lessons from complexity. *Science* **284**, 87–89 (1999).
4. Böttcher, T. From molecules to life: quantifying the complexity of chemical and biological systems in the universe. *J. Mol. Evol.* **86**, 1–10 (2018).
5. Parrish, J. K. & Edelstein-Keshet, L. Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science* **284**, 99–101 (1999).
6. Weng, G., Bhalla, U. S. & Iyengar, R. Complexity in biological signaling systems. *Science* **284**, 92–96 (1999).
7. Rind, D. Complexity and climate. *Science* **284**, 105–107 (1999).
8. Marshall, S. M., Murray, A. R. G. & Cronin, L. A probabilistic framework for identifying biosignatures using pathway complexity. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **375**, 20160342 (2017).
9. The Nobel Prize in Chemistry 1965. *Nobel Media AB* https://www.nobelprize.org/prizes/chemistry/1965/summary/ (2020).
10. Corey, E. J. & Todd Wipke, W. Computer-assisted design of complex organic syntheses. *Science* **166**, 178–192 (1969).
11. Corey, E. J. & Cheng, X.-M. *The Logic of Chemical Synthesis* (Wiley, 1996).
12. Corey, E. J., Long, A. K. & Rubenstein, S. D. Computer-assisted analysis in organic synthesis. *Science* **228**, 408–418 (1985).
13. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed. Engl.* **55**, 5904–5937 (2016).
14. Li, J. & Eastgate, M. D. Current complexity: a tool for assessing the complexity of organic molecules. *Org. Biomol. Chem.* **13**, 7164–7176 (2015).
15. Gao, W. & Coley, C. W. The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.0c00174 (2020).
16. Wender, P. A., Verma, V. A., Paxton, T. J. & Pillow, T. H. Function-oriented synthesis, step economy, and drug design. *Acc. Chem. Res.* **41**, 40–49 (2008).
17. Willstätter, R. Synthesen in der tTropingruppe. I. Synthese des tropilidens. *Justus Liebigs Ann. der Chem.* **317**, 204–265 (1901).

18. Humphrey, A. J. & O'Hagan, D. Tropane alkaloid biosynthesis. A century old problem unresolved. *Nat. Prod. Rep.* **18**, 494–502 (2001).
19. Medley, J. W. & Movassaghi, M. Robinson's landmark synthesis of tropinone. *Chem. Commun.* **49**, 10775–10777 (2013).
20. Robinson, R. LXIII. — a synthesis of tropinone. *J. Chem. Soc. Trans.* **111**, 762–768 (1917).
21. Bélanger, A. et al. Total synthesis of ryanodol. *Can. J. Chem.* **57**, 3348–3354 (1979).
22. Nagatomo, M. et al. Total synthesis of ryanodol. *J. Am. Chem. Soc.* **136**, 5916–5919 (2014).
23. Chuang, K. V., Xu, C. & Reisman, S. E. A 15-step synthesis of (+)-ryanodol. *Science* **353**, 912–915 (2016).
24. Baran, P. S. Natural product total synthesis: as exciting as ever and here to stay. *J. Am. Chem. Soc.* **140**, 4751–4755 (2018).
25. Holton, R. A. et al. First total synthesis of taxol. 1. Functionalization of the B ring. *J. Am. Chem. Soc.* **116**, 1597–1598 (1994).
26. Holton, R. A. et al. First total synthesis of taxol. 2. Completion of the C and D rings. *J. Am. Chem. Soc.* **116**, 1599–1600 (1994).
27. Nicolaou, K. C. et al. Total synthesis of taxol. *Nature* **367**, 630–634 (1994).
28. Wender, P. A. et al. The pinene path to taxanes. 5. Stereocontrolled synthesis of a versatile taxane precursor. *J. Am. Chem. Soc.* **119**, 2755–2756 (1997).
29. Wender, P. A. et al. The pinene path to taxanes. 6. A concise stereocontrolled synthesis of taxol. *J. Am. Chem. Soc.* **119**, 2757–2758 (1997).
30. Masters, J. J., Link, J. T., Snyder, L. B., Young, W. B. & Danishefsky, S. J. A total synthesis of taxol. *Angew. Chem. Int. Ed. Engl.* **34**, 1723–1726 (1995).
31. Mukaiyama, T. et al. Asymmetric total synthesis of taxol$^{\backslash\mathrm{R}}$. *Chem. A Eur. J.* **5**, 121–161 (1999).
32. Morihira, K. et al. Enantioselective total synthesis of taxol. *J. Am. Chem. Soc.* **120**, 12980–12981 (1998).
33. Kanda, Y. et al. Two-phase synthesis of taxol. *J. Am. Chem. Soc.* **142**, 10526–10533 (2020).
34. Nicolaou, K. C. et al. Total synthesis of calicheamicin γ_1^{I}. *J. Am. Chem. Soc.* **114**, 10082–10084 (1992).
35. Groneberg, R. D. et al. Total synthesis of calicheamicin γ_1^{I}. 1. Synthesis of the oligosaccharide fragment. *J. Am. Chem. Soc.* **115**, 7593–7611 (1993).
36. Smith, A. L. et al. Total synthesis of calicheamicin γ_1^{I}. 2. Development of an enantioselective route to (−)-calicheamicinone. *J. Am. Chem. Soc.* **115**, 7612–7624 (1993).
37. Nicolaou, K. C. et al. Total synthesis of calicheamicin γ_1^{I}. 3. The final stages. *J. Am. Chem. Soc.* **115**, 7625–7635 (1993).
38. Aicher, T. D. et al. Total synthesis of halichondrin B and norhalichondrin B. *J. Am. Chem. Soc.* **114**, 3162–3164 (1992).
39. Jackson, K. L., Henderson, J. A., Motoyoshi, H. & Phillips, A. J. A total synthesis of norhalichondrin B. *Angew. Chem. Int. Ed. Engl.* **48**, 2346–2350 (2009).
40. Armstrong, R. W. et al. Total synthesis of a fully protected palytoxin carboxylic acid. *J. Am. Chem. Soc.* **111**, 7525–7530 (1989).
41. Armstrong, R. W. et al. Total synthesis of palytoxin carboxylic acid and palytoxin amide. *J. Am. Chem. Soc.* **111**, 7530–7533 (1989).
42. Suh, E. M. & Kishi, Y. Synthesis of palytoxin from palytoxin carboxylic acid. *J. Am. Chem. Soc.* **116**, 11205–11206 (1994).
43. Kuttruff, C. A., Eastgate, M. D. & Baran, P. S. Natural product synthesis in the age of scalability. *Nat. Prod. Rep.* **31**, 419–432 (2014).
44. Pflüger, P. M. & Glorius, F. Molecular machine learning: the future of synthetic chemistry? *Angew. Chem. Int. Ed. Engl.* **59**, 18860–18865 (2020).
45. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2**, 725–732 (2016).
46. Liu, B. et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
47. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
48. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).
49. Reid, J. P. & Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nat. Rev. Chem.* **2**, 290–305 (2018).
50. Zhao, S. et al. Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization. *Science* **362**, 670–674 (2018).
51. Nielsen, M. K., Ahneman, D. T., Riera, O. & Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *J. Am. Chem. Soc.* **140**, 5004–5008 (2018).
52. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
53. Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **6**, 1379–1390 (2020).
54. Coley, C. W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
55. Zahrt, A. F. et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).
56. Burai Patrascu, M. et al. From desktop to benchtop with automated computational workflows for computer-aided design in asymmetric catalysis. *Nat. Catal.* **3**, 574–584 (2020).
57. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
58. Schreck, J. S., Coley, C. W. & Bishop, K. J. M. Learning retrosynthetic planning through simulated experience. *ACS Cent. Sci.* **5**, 970–981 (2019).
59. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
60. Trinajstić, N. *Chemical Graph Theory* (CRC, 1992).
61. Gerry, C. J. et al. Real-time biological annotation of synthetic compounds. *J. Am. Chem. Soc.* **138**, 8920–8927 (2016).
62. Singh, M., Gaskins, B., Johnson, D. R., Elles, C. G. & Boskovic, Z. Synthesis of cycloheptatriene-containing azetidine lactones. *J. Org. Chem.* **87**, 15001–15010 (2022).
63. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
64. Bonchev, D. & Trinajstić, N. Chemical information theory: structural aspects. *Int. J. Quantum Chem.* **22**, 463–480 (1982).
65. Cherkasov, A. et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).
66. Böttcher, T. An additive definition of molecular complexity. *J. Chem. Inf. Model.* **56**, 462–470 (2016).
67. Smith, S. W. Chiral toxicology: it's the same thing...only different. *Toxicol. Sci.* **110**, 4–30 (2009).
68. Bertz, S. H. The first general index of molecular complexity. *J. Am. Chem. Soc.* **103**, 3599–3601 (1981).
69. Hendrickson, J. B., Huang, P. & Toczko, A. G. Molecular complexity: a simplified formula adapted to individual atoms. *J. Chem. Inf. Comput. Sci.* **27**, 63–67 (1987).
70. Rücker, G. & Rücker, C. On finding nonisomorphic connected subgraphs and distinct molecular substructures. *J. Chem. Inf. Comput. Sci.* **41**, 314–320 (2001).
71. Bertz, S. H. & Sommer, T. J. Rigorous mathematical approaches to strategic bonds and synthetic analysis based on conceptually simple new complexity indices. *Chem. Commun.* **16**, 2409–2410 (1997).
72. Ruecker, G. & Ruecker, C. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.* **33**, 683–695 (1993).
73. Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **97**, 6609–6615 (1975).
74. Randić, M., Brissey, G. M., Spencer, R. B. & Wilkins, C. L. Search for all self-avoiding paths for molecular graphs. *Comput. Chem.* **3**, 5–13 (1979).
75. Randić, M. & Plavšić, D. Characterization of molecular complexity. *Int. J. Quantum Chem.* **91**, 20–31 (2002).
76. Whitlock, H. W. On the structure of total synthesis of complex natural products. *J. Org. Chem.* **63**, 7982–7989 (1998).
77. Barone, R. & Chanon, M. A new and simple approach to chemical complexity. Application to the synthesis of natural products. *J. Chem. Inf. Comput. Sci.* **41**, 269–272 (2001).
78. Bonchev, D. The overall wiener index — a new tool for characterization of molecular topology. *J. Chem. Inf. Comput. Sci.* **41**, 582–592 (2001).
79. Bonchev, D., Mekenyan, O. & Trinajstić, N. Topological characterization of cyclic structures. *Int. J. Quantum Chem.* **17**, 845–893 (1980).
80. Proudfoot, J. R. A path based approach to assessing molecular complexity. *Bioorg. Med. Chem. Lett.* **27**, 2014–2017 (2017).
81. Proudfoot, J. R. Molecular complexity and retrosynthesis. *J. Org. Chem.* **82**, 6968–6971 (2017).
82. Bender, A. & Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2**, 3204–3218 (2004).
83. Bonchev, D. & Peev, T. Information theoretic study of chemical elements. Mean information content of a chemical element. *Jahresber. Hochsch. Chem. Tech. Burgas.* **10**, 561 (1973).
84. Demoret, R. M. et al. Synthetic, mechanistic, and biological interrogation of ginkgo biloba chemical space en route to (−)-bilobalide. *J. Am. Chem. Soc.* **142**, 18599–18618 (2020).
85. Herzon, S. B. Emergent properties of natural products. *Synlett* **29**, 1823–1835 (2018).
86. Huffman, B. J. & Shenvi, R. A. Natural products in the 'marketplace': interfacing synthesis and biology. *J. Am. Chem. Soc.* **141**, 3332–3346 (2019).
87. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75**, 311–335 (2012).
88. Krzyzanowski, A., Pahl, A., Grigalunas, M. & Waldmann, H. Spacial score — a comprehensive topological indicator for small-molecule complexity. *J. Med. Chem.* **66**, 12739–12750 (2023).
89. Méndez-Lucio, O. & Medina-Franco, J. L. The many roles of molecular complexity in drug discovery. *Drug Discov. Today* **22**, 120–126 (2017).
90. Baker, M. A., Demoret, R. M., Ohtawa, M. & Shenvi, R. A. Concise asymmetric synthesis of (−)-bilobalide. *Nature* **575**, 643–646 (2019).
91. Del Bel, M., Abela, A. R., Ng, J. D. & Guerrero, C. A. Enantioselective chemical syntheses of the furanosteroids (−)-viridin and (−)-viridiol. *J. Am. Chem. Soc.* **139**, 6819–6822 (2017).
92. Johnson, J. S. Counting steps. *Nat. Synth.* **2**, 6–8 (2023).
93. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 8 (2009).
94. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
95. Klucznik, T. et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522–532 (2018).
96. Gaich, T. & Baran, P. S. Aiming for the ideal synthesis. *J. Org. Chem.* **75**, 4657–4673 (2010).

97. Bonnet, P. Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists. *Eur. J. Med. Chem.* **54**, 679–689 (2012).
98. Sheridan, R. P. et al. Modeling a crowdsourced definition of molecular complexity. *J. Chem. Inf. Model.* **54**, 1604–1616 (2014).
99. Corey, E. J., Iii, R. D. C. & Howe, W. J. Computer-assisted synthetic analysis for complex molecules. Methods and procedures for machine generation of synthetic intermediates. *J. Am. Chem. Soc.* **94**, 440–459 (1972).
100. Corey, E. J. & Frank Feiner, N. Computer-assisted synthetic analysis. A rapid computer method for the semiquantitative assignment of conformation of six-membered ring systems. 1. Derivation of a preliminary conformational description of the six-membered ring. *J. Org. Chem.* **45**, 757–764 (1980).
101. Corey, E. J., Long, A. K., Greene, T. W. & Miller, J. W. Computer-assisted synthetic analysis. Selection of protective groups for multistep organic syntheses. *J. Org. Chem.* **50**, 1920–1927 (1985).
102. Corey, E. J., Wipke, W. T., Iii, R. D. C. & Howe, W. J. Techniques for perception by a computer of synthetically significant structural features in complex molecules. *J. Am. Chem. Soc.* **94**, 431–439 (1972).
103. Corey, E. J., Johnson, A. P. & Long, A. K. Computer-assisted synthetic analysis. Techniques for efficient long-range retrosynthetic searches applied to the Robinson annulation process. *J. Org. Chem.* **45**, 2051–2057 (1980).
104. Corey, E. J., Howe, W. J. & Pensak, D. A. Computer-assisted synthetic analysis. Methods for machine generation of synthetic intermediates involving multistep look-ahead. *J. Am. Chem. Soc.* **96**, 7724–7737 (1974).
105. Corey, E. J. & Frank Feiner, N. Computer-assisted synthetic analysis. A rapid computer method for the semiquantitative assignment of conformation of six-membered ring systems. 2. Assessment of conformational energies. *J. Org. Chem.* **45**, 765–780 (1980).
106. Corey, E. J. & Jorgensen, W. L. Computer-assisted synthetic analysis. Generation of synthetic sequences involving sequential functional group interchanges. *J. Am. Chem. Soc.* **98**, 203–209 (1976).
107. Corey, E. J. et al. Computer-assisted synthetic analysis. long-range search procedures for antithetic simplification of complex targets by application of the halolactonization transform. *J. Chem. Inf. Comput. Sci.* **20**, 221–230 (1980).
108. Corey, E. J., Orf, H. W. & Pensak, D. A. Computer-assisted synthetic analysis. The identification and protection of interfering functionality in machine-generated synthetic intermediates. *J. Am. Chem. Soc.* **98**, 210–221 (1976).
109. Corey, E. J., Long, A. K., Lotto, G. I. & Rubenstein, S. D. Computer-assisted synthetic analysis. Quantitative assessment of transform utilities. *Recl. Trav. Chim. Pays-Bas* **111**, 304–309 (1992).
110. Bertz, S. H. & Rücker, C. In search of simplification: the use of topological complexity indices to guide retrosynthetic analysis. *Croat. Chem. Acta* **77**, 221–235 (2004).
111. Rücker, G. & Rücker, C. Substructure, subgraph, and walk counts as measures of the complexity of graphs and molecules. *J. Chem. Inf. Comput. Sci.* **41**, 1457–1462 (2001).
112. Corey, E. J., Howe, W. J., Orf, H. W., Pensak, D. A. & Petersson, G. General methods of synthetic analysis. Strategic bond disconnections for bridged polycyclic structures. *J. Am. Chem. Soc.* **97**, 6116–6124 (1975).
113. Marth, C. J. et al. Network-analysis-guided synthesis of weisaconitine D and liljestrandinine. *Nature* **528**, 493–498 (2015).
114. Kou, K. G. M. et al. A unifying synthesis approach to the C18-, C19-, and C20-diterpenoid alkaloids. *J. Am. Chem. Soc.* **139**, 13882–13896 (2017).
115. Yudin, A. K. Macrocycles: lessons from the distant past, recent developments, and future directions. *Chem. Sci.* **6**, 30–49 (2015).
116. Martí-Centelles, V., Pandey, M. D., Burguete, M. I. & Luis, S. V. Macrocyclization reactions: the importance of conformational, configurational, and template-induced preorganization. *Chem. Rev.* **115**, 8736–8834 (2015).
117. Mortensen, K. T., Osberger, T. J., King, T. A., Sore, H. F. & Spring, D. R. Strategies for the diversity-oriented synthesis of macrocycles. *Chem. Rev.* **119**, 10288–10317 (2019).
118. Saridakis, I., Kaiser, D. & Maulide, N. Unconventional macrocyclizations in natural product synthesis. *ACS Cent. Sci.* **6**, 1869–1889 (2020).
119. Fürstner, A. Lessons from natural product total synthesis: macrocyclization and postcyclization strategies. *Acc. Chem. Res.* **54**, 861–874 (2021).
120. Hendrickson, J. B. Organic synthesis in the age of computers. *Angew. Chem. Int. Ed. Engl.* **29**, 1286–1295 (1990).
121. Bøgevig, A. et al. Route design in the 21st century: the IC SYNTH software tool as an idea generator for synthesis prediction. *Org. Process Res. Dev.* **19**, 357–368 (2015).
122. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
123. Genheden, S. et al. AiZynthFinder: a fast robust and flexible open-source software for retrosynthetic planning. *J. Cheminform.* **12**, 70 (2020).
124. Gasteiger, J., Ihlenfeldt, W. D. & Röse, P. A collection of computer methods for synthesis design and reaction prediction. *Recl. Trav. Chim. Pays-Bas* **111**, 270–290 (1992).
125. Somnath, V. R., Bunne, C., Coley, C. W., Krause, A. & Barzilay, R. Learning graph models for template-free retrosynthesis. Preprint at https://doi.org/10.48550/arXiv.2006.07038 (2020).
126. Mo, Y. et al. Evaluating and clustering retrosynthesis pathways with learned strategy. *Chem. Sci.* **12**, 1469–1478 (2021).
127. Gasteiger, J., Ihlenfeldt, W. D., Fick, R. & Rose, J. R. Similarity concepts for the planning of organic reactions and syntheses. *J. Chem. Inf. Comput. Sci.* **32**, 700–712 (1992).
128. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**, 1046–1053 (2006).
129. Tanimoto, T. T. *IBM Internal Report* (IBM, 1957).
130. Bottou, L., Curtis, F. E. & Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Rev.* **60**, 223–311 (2018).
131. Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).
132. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
133. Lin, Y., Zhang, R., Wang, D. & Cernak, T. Computer-aided key step generation in alkaloid total synthesis. *Science* **379**, 453–457 (2023).
134. Burns, N. Z., Baran, P. S. & Hoffmann, R. W. Redox economy in organic synthesis. *Angew. Chem. Int. Ed. Engl.* **48**, 2854–2867 (2009).
135. Trost, B. The atom economy — a search for synthetic efficiency. *Science* **254**, 1471–1477 (1991).
136. Newhouse, T., Baran, P. S. & Hoffmann, R. W. The economies of synthesis. *Chem. Soc. Rev.* **38**, 3010–3021 (2009).
137. Barone, R., Petitjean, M., Baralotto, C., Piras, P. & Chanon, M. Information theory description of synthetic strategies. A new similarity index. *J. Phys. Org. Chem.* **16**, 9–15 (2003).
138. Johnson, W. S., Semmelhack, M. F., Sultanbawa, M. U. S. & Dolak, L. A. A new approach to steroid total synthesis. A nonenzymic biogenetic-like olefinic cyclization involving the stereospecific formation of five asymmetric centers. *J. Am. Chem. Soc.* **90**, 2994–2996 (1968).
139. Yoder, R. A. & Johnston, J. N. A case study in biomimetic total synthesis: polyolefin carbocyclizations to terpenes and steroids. *Chem. Rev.* **105**, 4730–4756 (2005).
140. Johnson, W. S. Nonenzymic biogenetic-like olefinic cyclizations. *Acc. Chem. Res.* **1**, 1–8 (1968).
141. Abe, I., Rohmer, M. & Prestwich, G. D. Enzymatic cyclization of squalene and oxidosqualene to sterols and triterpenes. *Chem. Rev.* **93**, 2189–2206 (1993).
142. George, D. T., Kuenstner, E. J. & Pronin, S. V. A concise approach to paxilline indole diterpenes. *J. Am. Chem. Soc.* **137**, 15410–15413 (2015).
143. Sun, Y. et al. Bioinspired total synthesis of sespenine. *Angew. Chem. Int. Ed. Engl.* **53**, 9012–9016 (2014).
144. Brill, Z. G., Grover, H. K. & Maimone, T. J. Enantioselective synthesis of an ophiobolin sesterterpene via a programmed radical cascade. *Science* **352**, 1078–1082 (2016).
145. Hung, K., Hu, X. & Maimone, T. J. Total synthesis of complex terpenoids employing radical cascade processes. *Nat. Prod. Rep.* **35**, 174–202 (2018).
146. Nicolaou, K. C., Petasis, N. A., Zipkin, R. E. & Uenishi, J. The endiandric acid cascade. Electrocyclizations in organic synthesis. I. Stepwise, stereocontrolled total synthesis of endiandric acids A and B. *J. Am. Chem. Soc.* **104**, 5555–5557 (1982).
147. Nicolaou, K. C., Petasis, N. A., Uenishi, J. & Zipkin, R. E. The endiandric acid cascade. Electrocyclizations in organic synthesis. 2. Stepwise, stereocontrolled total synthesis of endiandric acids C-G. *J. Am. Chem. Soc.* **104**, 5557–5558 (1982).
148. Nicolaou, K. C., Zipkin, R. E. & Petasis, N. A. The endiandric acid cascade. Electrocyclizations in organic synthesis. 3. 'Biomimetic' approach to endiandric acids A-G. Synthesis of precursors. *J. Am. Chem. Soc.* **104**, 5558–5560 (1982).
149. Nicolaou, K. C., Petasis, N. A. & Zipkin, R. E. The endiandric acid cascade. Electrocyclizations in organic synthesis. 4. "Biomimetic" approach to endiandric acids A-G. Total synthesis and thermal studies. *J. Am. Chem. Soc.* **104**, 5560–5562 (1982).
150. Piettre, S. & Heathcock, C. H. Biomimetic total synthesis of *proto*-daphniphylline. *Science* **248**, 1532–1534 (1990).
151. Heathcock, C. H. Nature knows best: an amazing reaction cascade is uncovered by design and discovery. *Proc. Natl Acad. Sci. USA* **93**, 14323–14327 (1996).
152. Chattopadhyay, A. K. & Hanessian, S. Recent progress in the chemistry of daphniphyllum alkaloids. *Chem. Rev.* **117**, 4104–4146 (2017).
153. Evans, D. A. & Starr, J. T. A cascade cycloaddition strategy leading to the total synthesis of (−)-FR182877. *Angew. Chem. Int. Ed. Engl.* **41**, 1787–1790 (2002).
154. Vosburg, D. A., Vanderwal, C. D. & Sorensen, E. J. A synthesis of (+)-FR182877, featuring tandem transannular Diels–Alder reactions inspired by a postulated biogenesis. *J. Am. Chem. Soc.* **124**, 4552–4553 (2002).
155. Sierra, M. A. & de la Torre, M. C. Dead ends and detours en route to total syntheses of the 1990s. *Angew. Chem. Int. Ed. Engl.* **39**, 1538–1559 (2000).
156. Crimmins, M. T. et al. The total synthesis of (±)-ginkgolide B. *J. Am. Chem. Soc.* **122**, 8453–8463 (2000).
157. Schneider, F., Samarin, K., Zanella, S. & Gaich, T. Total synthesis of the complex taxane diterpene canataxpropellane. *Science* **367**, 676–681 (2020).
158. McKerrall, S. J., Jørgensen, L., Kuttruff, C. A., Ungeheuer, F. & Baran, P. S. Development of a concise synthesis of (+)-ingenol. *J. Am. Chem. Soc.* **136**, 5799–5810 (2014).
159. Jørgensen, L. et al. 14-step synthesis of (+)-ingenol from (+)-3-carene. *Science* **341**, 878–882 (2013).
160. Kawamura, S., Chu, H., Felding, J. & Baran, P. S. Nineteen-step total synthesis of (+)-phorbol. *Nature* **532**, 90–93 (2016).
161. Mendoza, A., Ishihara, Y. & Baran, P. S. Scalable enantioselective total synthesis of taxanes. *Nat. Chem.* **4**, 21–25 (2012).
162. Wilde, N. C., Isomura, M., Mendoza, A. & Baran, P. S. Two-phase synthesis of (−)-taxuyunnanine D. *J. Am. Chem. Soc.* **136**, 4909–4912 (2014).
163. Yuan, C., Jin, Y., Wilde, N. C. & Baran, P. S. Short, enantioselective total synthesis of highly oxidized taxanes. *Angew. Chem. Int. Ed. Engl.* **55**, 8280–8284 (2016).

164. Brill, Z. G., Condakes, M. L., Ting, C. P. & Maimone, T. J. Navigating the chiral pool in the total synthesis of complex terpene natural products. *Chem. Rev.* **117**, 11753–11795 (2017).
165. Condakes, M. L., Hung, K., Harwood, S. J. & Maimone, T. J. Total syntheses of (−)-majucin and (−)-jiadifenoxolane A, complex majucin-type illicium sesquiterpenes. *J. Am. Chem. Soc.* **139**, 17783–17786 (2017).
166. Hung, K. et al. Development of a terpene feedstock-based oxidative synthetic approach to the illicium sesquiterpenes. *J. Am. Chem. Soc.* **141**, 3083–3099 (2019).
167. Abrams, D. J., Provencher, P. A. & Sorensen, E. J. Recent applications of C–H functionalization in complex natural product synthesis. *Chem. Soc. Rev.* **47**, 8925–8967 (2018).
168. Davies, H. M. L. & Morton, D. Recent advances in C–H functionalization. *J. Org. Chem.* **81**, 343–350 (2016).
169. Wender, P. A. Toward the ideal synthesis and molecular function through synthesis-informed design. *Nat. Prod. Rep.* **31**, 433–440 (2014).
170. Mikulak-Klucznik, B. et al. Computational planning of the synthesis of complex natural products. *Nature* **588**, 83–88 (2020).
171. Hoffmann, R. W. *Elements of Synthesis Planning* (Springer, 2009).
172. Cherney, E. C., Green, J. C. & Baran, P. S. Synthesis of ent-kaurane and beyerane diterpenoids by controlled fragmentations of overbred intermediates. *Angew. Chem. Int. Ed. Engl.* **52**, 9019–9022 (2013).
173. Sarpong, R., Wang, B. & Perea, M. A. Transition metal-mediated C–C single bond cleavage: making the cut in total synthesis. *Angew. Chem. Int. Ed. Engl.* **59**, 18898–18919 (2020).
174. Wender, P. A. & Howbert, J. J. Synthetic studies on arene-olefin cycloadditions: total synthesis of (±)-α-cedrene. *J. Am. Chem. Soc.* **103**, 688–690 (1981).
175. Oppolzer, W. & Godel, T. A new and efficient total synthesis of (±)-longifolene. *J. Am. Chem. Soc.* **100**, 2583–2584 (1978).
176. Hafeman, N. J. et al. The total synthesis of (−)-scabrolide A. *J. Am. Chem. Soc.* **142**, 8585–8590 (2020).
177. Foy, N. J. & Pronin, S. V. Synthesis of pleuromutilin. *J. Am. Chem. Soc.* **144**, 10174–10179 (2022).
178. Chuang, K. V., Gunsalus, L. M. & Keiser, M. J. Learning molecular representations for medicinal chemistry. *J. Med. Chem.* https://doi.org/10.1021/acs.jmedchem.0c00385 (2020).
179. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
180. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
181. Selzer, P., Roth, H. J., Ertl, P. & Schuffenhauer, A. Complex molecules: do they add value? *Curr. Opin. Chem. Biol.* **9**, 310–316 (2005).
182. Hann, M. M., Leach, A. R. & Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **41**, 856–864 (2001).
183. Lovering, F., Bikker, J. & Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **52**, 6752–6756 (2009).
184. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
185. Lovering, F. Escape from flatland 2: complexity and promiscuity. *MedChemComm* **4**, 515–519 (2013).
186. Galloway, W. R. J. D., Isidro-Llobet, A. & Spring, D. R. Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nat. Commun.* **1**, 80 (2010).
187. Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**, 1964–1969 (2000).
188. Clemons, P. A. et al. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl Acad. Sci. USA* **107**, 18787–18792 (2010).
189. Beckmann, H. S. G. et al. A strategy for the diversity-oriented synthesis of macrocyclic scaffolds using multidimensional coupling. *Nat. Chem.* **5**, 861–867 (2013).
190. Kato, N. et al. Diversity-oriented synthesis yields novel multistage antimalarial inhibitors. *Nature* **538**, 344–349 (2016).

Acknowledgements

The authors thank P. A. Wender (Stanford University), R. S. Paton (Colorado State University) and members of the Sarpong Group (UC Berkeley) for stimulating discussions that inspired a part of the discussion presented herein. The authors are also grateful to the US National Institute of General Medical Sciences (NIGMS; R35 GM130345) and the National Science Foundation Center for Computer Assisted Synthesis (CHE 2202693) for the financial support of the primary research that set the stage for development of ideas expressed here.

Author contributions

B.W. researched data for the article and performed complexity analyses. Both authors then discussed content, wrote the article and contributed to review and/or editing of the manuscript before submission.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Chemistry* thanks the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.