Engines of discovery: *Computers in advanced synthesis planning and identification of drug candidates*

After over five decades of efforts, computers have recently begun to plan chemical syntheses of complex targets at a level comparable to human experts. With this milestone achieved, it is now time to ponder not only how the machines will accelerate and multiplex synthetic design, but also how they will guide discovery of new targets having desired properties.

Bartosz A. Grzybowski

Bartosz A. Grzybowski (ORCID 0000-0001-6613-4261; E-mail: grzybor72@unist.ac.kr) is a Distinguished Professor in the Chemistry Department at UNIST, a group leader at the Institute for Basic Science (both South Korea), and a Professor at the Institute of Organic Chemistry, Polish Academy of Sciences in Warsaw. His current research interest center on computer-assisted synthesis and discovery of new reactions, and on the experimental implementation of reaction networks and systems.

PLANNING SYNTHESES OF complex organic molecules is, arguably, the pinnacle of chemical research, sometimes compared to an art requiring not only knowledge but also inspiration. Since syntheses of natural products such as vinigrol, perseanol, or daphlongamine H are extremely nuanced and challenging even to the world's top-level synthetic chemists, it is perhaps not surprising that these chemists have long sought computer's help, trying to codify the discipline's knowledge and strategic thinking, and casting "inspiration" in the form of rigid algorithmic rules. The first ideas[1] and actual programs[2] for computer-aided synthesis emerged already in the 1960s and – at least at that time – it seemed that machines would be conquering the "art of synthetic chemistry" any day. Yet, years have passed and the programs kept faltering – their synthetic predictions failed in the laboratory[3,4] or were applicable only to some simple targets[5] for which a trained chemist does not really need machine's help. Meanwhile, computers managed to conquer even the most intricate games of strategy – in 1997, IBM's DeepBlue defeated the reigning world champion, Garry Kasparov, in chess, and in 2016, Google's AlphaGO[6] meted out a crushing defeat to GO's weltmeister, Lee Sedol. Given that Google has recently demonstrated similar feats of AI and deep learning in natural sciences – for instance, in protein folding[7]—it might be puzzling and frustrating why no news of similar successes have been forthcoming for organic syntheses. To be sure, it was not for the lack of trying as various AI methods have been unleashed on the problem[8,9,10]– it is just that advanced synthetic design turned out to be a tougher nut to crack than either chess or GO! The breakthrough came only recently when a program called Chematica (a.k.a. Synthia™) designed syntheses of complex natural products (**Figure 1**) with precision and elegance the world's leading chemists judged to be indicative of human, expert-level planning.[11] The first part of this article aims to narrate how this was achieved and why the problem turned out to be so difficult to tackle, taking us some 20 years of concerted effort.[12-19] The second part goes further and, inspired by Chematica's success in retrosynthesis, outlines other areas of synthetic chemistry in which computer-driven approaches – in particular forward-synthesis combined with property prediction – can make a profound and lasting impact: In the Origins of Life, in green chemistry, in the generation of molecular diversity, or in the prediction of synthesizable drug candidates. These applications are already taking shape and they are changing the face of modern organic chemistry, boosting the creativity of individual chemists with analyses at scales available only to the machines. Immanuel Kant's famous critique of (synthetic) chemistry as lacking mathematical rigor no longer applies, and

by Bartosz A. Grzybowski

https://doi.org/10.51167/acm00010

the art of making molecules is finally becoming an algorithmic science.

The challenge of retrosynthesis

Let us begin with retrosynthesis – that is, a process in which a desired, often very complex organic molecule of interest is disconnected into smaller fragments, which are then disconnected further and further until reaching some simple and preferably commercially available substrates. The rules for retrosynthetic analysis by humans were codified half a century ago by E.J. Corey[20] who then attempted to apply them to automatic synthetic planning[21] – alas, as mentioned above, with little success[3,4] and only in a semi-automatic fashion whereby the machine provided all possible reactions for each retron while the user had to make his/her choices and construct the pathway. When we started working on the problem some 20 years ago, we were pondering how the process could be fully automated. We identified three interrelated components of what later was to become known as Chematica (**Figure 2**): (1) The rules describing chemical reactions; (2) The algorithms that would iteratively apply these rules to the retrons to generate the synthons and, ultimately, the networks of synthetic possibilities; and (3) the so-called scoring function(s) that would guide navigation of this network, preventing "combinatorial explosion" and concatenating individual reactions into complete pathways.

The rules of the game: reactions

Without repeating our recent reviews on the subject,[15,17] we note that the number of rules required for versatile chemical planning turned out to the be rather high, on the order of 100,000. Each of these rules describes a reaction class (or variant) by the "core" atoms that change during the reaction as well as some flanking atoms (the "environment"). The reaction "template" is written in the so-called SMILES/SMARTS notation and must carefully delineate the scope of admissible substituents – this information can, conceivably, be retrieved from large databases of published reaction examples, and for this purpose it is very tempting to use automated template extraction methods which were available already in the early 2000s.[22] Unfortunately, literature-extracted rules do not *a priori* know how widely to define the "environments" (so that they are appropriate to a given reaction type), or how to account accurately for incompatible groups (by default, not present in published, successful syntheses). If the rules are subsequently applied to molecules featuring such groups, the machine does not recognize them as problematic and can suggest reactions that, in reality, would fail (for detailed discussion, see [17]).

Mindful of such considerations, we decided to code the reaction rules "by hand," inspecting the underlying mechanisms, determining suitable reaction conditions, and then determining which of as many as 400 possible functional groups should be marked as potentially incompatible. In doing so, we focused on reactions that were really useful in synthetic practice, validated by many (and preferably reputable) groups, and transferrable to different scaffolds (i.e., we generally avoided "one trick pony" reactions that might work only for a very specific scaffold but not for other molecules). In some sense, we decided to make our Chematica a conservative planner. This assumption was more than a scientific choice – it was, in large part, a "political" decision addressing a rather widespread disbelief in the prospects of computer-assisted synthesis. Over the years, we have seen many times how a single mistake in Chematica's planning would trigger a triumphant "see, it does not work!" reaction (often from colleagues whose own human-designed syntheses failed multiple times). We realized that for Chematica to become an adopted child of the synthetic community, its reaction rules and synthetic suggestions must be very robust, even at the expense of occasionally missing some inspired but risky solutions.

It took us about a decade to code as many as 100,000+ high-quality chemical rules and then to further fine tune them to predict subtler chemical effects. To this end, we combined expert-coded knowledge base with quantum mechanical calculations (e.g., to determine electron densities at proposed reaction sites[17]) and also with machine learning models which provided more accurate predictions of site-, regio- or diastereoselectivity.[23] Importantly, we developed such models for separate reaction classes, for which there were adequate numbers of literature examples (thousands). Also, as descriptors we used physical-organic measures such as Hammett constants and steric crowding indices, such that the models were taking into account chemically relevant effects (as opposed to only arbitrarily defined structural motifs, as in the so-called fingerprints; for discussion see [23]). With these additions, the rules become a combination of expert-coded knowledge, advanced theory and modern AI – we began to refer to this mix as a "hybrid" approach.

The game itself: network navigation

Of course, the rules themselves were not producing any syntheses. As in chess, the knowledge of how to move a pawn, a rook, or a bishop does not translate into the ability to *play* the game. In chemistry, the "game" of retrosynthesis is how to choose chemically plausible pathways from the giant network of synthetic options. "Giant" here is not an exaggeration – with 100,000 rules, every retron can produce, in one step, on the order of 100 synthons,[11,15] each of these synthons can then produce ~100 progenies, and so on, until commercially available starting materials are reached. In *n* steps, this branching translates into 100^n possible routes one can trace on the network. Even for simple drugs these numbers are very large (e.g., 100^5 or ten billion options for a five-step synthesis), and for the syntheses of natural products they are just exorbitant, as *n* is typically in tens. Clearly, exhaustive exploration of such networks is not feasible and one must devise means for smart navigation – that is, functions that score the synthons and decide which synthetic moves are promising to take.

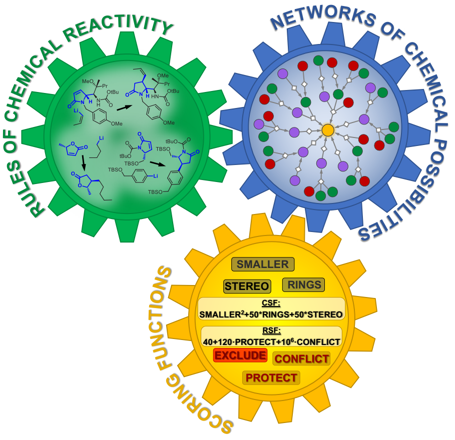

RULES OF CHEMICAL REACTIVITY
NETWORKS OF CHEMICAL POSSIBILITIES
SCORING FUNCTIONS
SMALLER
STEREO
RINGS
CSF:
SMALLER²+50*RINGS+50*STEREO
RSF:
40+120·PROTECT+10⁶·CONFLICT
EXCLUDE
CONFLICT
PROTECT

Figure 2. Main components of a synthesis planning machine: (green) rules describing chemical reactions, (blue) algorithms generating synthons and networks of synthetic possibilities, (c) scoring functions guiding network exploration and preventing "combinatorial explosion".

We started studying reaction networks in the early 2000s, even before we had any reaction rules ready. These early studies[12] used static databases of published reactions turned into a network representation – that is, they were not networks of retrosynthetic options created dynamically for a new target. However, even such mock-up models were useful as they allowed us to learn about network topology, the optimal means of representing the reactions (in the so-called bipartite format, see inset to **Figure 3**), and about the search algorithms and rudimentary scoring functions. By ca. 2010, we had first such algorithms and functions implemented[13,14,24] and showed how they could very rapidly traverse the network of published reactions (a.k.a. Network of Organic Chemistry, NOC) to construct pathways to targets as complex as zaragozic acid (**Figure 3**). Again, these pathways were just a patchwork of reaction steps published by different groups, but they were constructed by the machine without any human guidance.

In the 2010s, we finally combined this knowledge of networks with the reaction rules and began to automatically plan new syntheses to arbitrary – i.e., known or unknown – molecule targets.[15] The rules were applied to the retrons and expanded them into synthons. The scoring functions then evaluated the options and ranked the synthons, marking those that were most promising and merited further expansion. In this way, the scoring functions guided the growth of the network (**Figure 4**) such as to avoid unproductive dead ends and trace plausible syntheses as rapidly as possible. Over the years, the scoring functions evolved and were either (1) based on variables quantifying molecular complexity (lengths of SMILES strings, number of rings, number of stereocenters),[15] or (2) used neural-network hybrids trained on examples of literature syntheses matched onto Chematica's rules.[19] Referring the reader to ref [19] for detailed discussion, we note that functions of type (2) were somewhat better in searching for synthetic routes resembling published approaches, whereas functions of type (1) were better in unbiased design, often suggesting more elegant and unprecedented solutions.

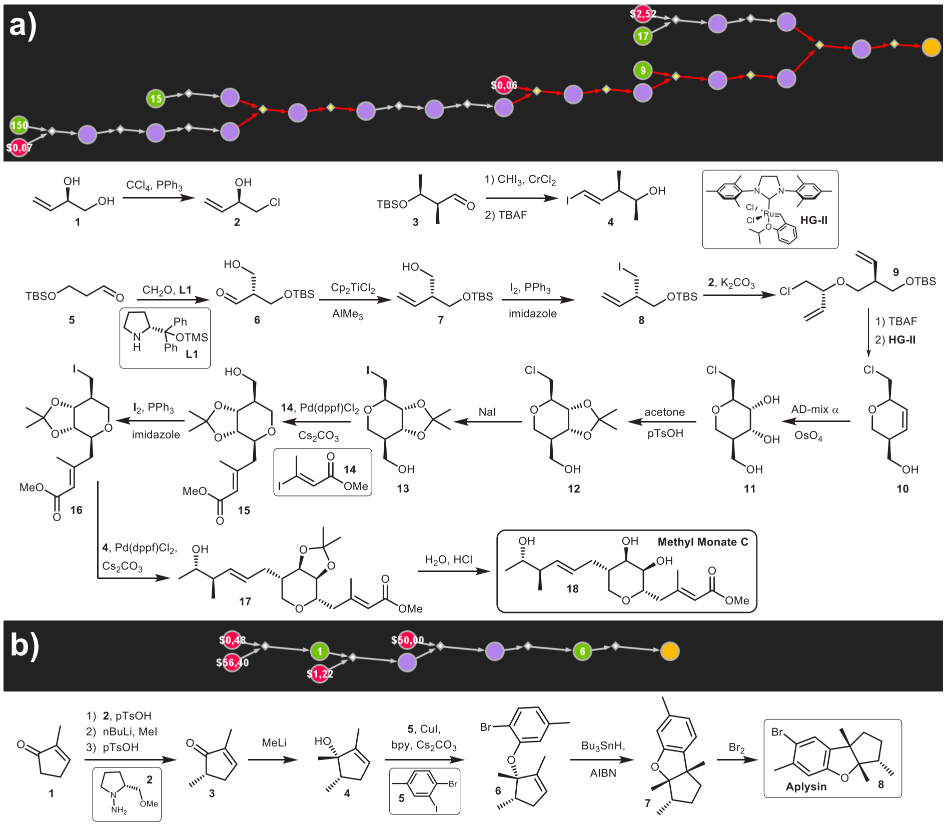


Figure 1. Chematica-designed syntheses leading to **(a)** Methyl Monate C and **(b)** Aplysin. These syntheses were evaluated by human experts in the Synthesis Turing Test described in ref. [11] from which the figure is adapted.

Simple games: syntheses of drugs

In 2017, we finally put all these algorithms to work. At that time, Sigma-Aldrich became interested in long-term sustenance of Chematica but they wanted to check if the program really works – that is, whether its predictions are verifiable in the laboratory. Accordingly, Sigma provided us with six molecules that presented a challenge to their own chemists. Could Chematica design more effective synthetic plans? We agreed to the challenge, added two molecules of our own choosing, and the entire set was committed to synthesis. Somewhat to everyone's surprise, all of these syntheses worked in very good yields and without the need for tedious optimization (see examples in **Figure 5**). We were jubilant, Sigma was convinced (and took over Chematica and began its worldwide marketing as Synthia), and we jointly published a paper describing the results.[16] Yet, the synthetic community remained lukewarm. Even though this work was the first-ever successful validation of computer-designed plans, the targets were deemed too simple. Everyone still waited for the machine to compete at a real expert level – that is, plan syntheses of complex natural products. And, of course, we took up this challenge as well.

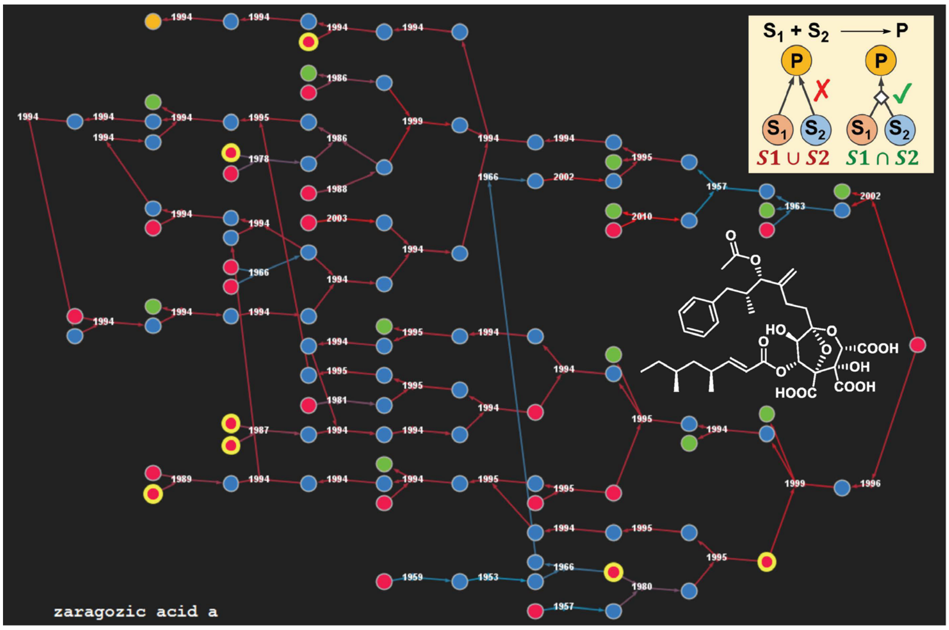


Figure 3. The Network of Organic Chemistry, NOC. The inset on the upper-right illustrates two ways of representing a chemical reaction in a graph form. With only one type of molecule nodes (circles), one would have to draw arrows from both S_1 and S_2 substrates to the product, P, which is inaccurate. To reflect the fact that *both* substrates are needed for the reaction to occur, we use the so-called bipartite representation in which S_1 *and* S_2 "enter" a diamond-shaped node signifying a reaction operation, which then leads to P. The main image has a bipartite graph corresponding to a cost-optimized synthesis of Zaragozic acid A (yellow node in upper left part) found in the NOC. The NOC is a static, predefined network and the NOC-searching algorithm does not plan de novo synthesis – instead, it concatenates reactions already reported in the literature (at years indicated over reaction arrows).

Advanced plays: natural products

From 2017 to 2020, we more than doubled the knowledge-base of reaction rules (to the abovementioned 100,000+), including a large proportion of stereoselective reactions so important in advanced synthesis. We improved the scoring functions, and implemented search algorithms that now used multiple search strategies simultaneously – for instance, one scoring function preferring diversity of approaches ("searching wide"), one putting premium on finishing the routes as rapidly as possible ("searching deep"), and the two exchanging and learning from each other's results. Unfortunately, even with these and other improvements reviewed in[11], the critics seemed to had been right – the program was not robustly identifying routes to complex targets.

Inspecting the results, we noted that Chematica – as all synthesis programs before it – was somewhat short-sighted. If it encountered highly unpromising synthons and could not find a worthy continuation in just one step, it simply withdrew from this branch of the network and did not try to strategize around the problem. What the program was obviously lacking was the ability to think several steps ahead, like the chess masters. Inspired by many classic syntheses designed by the masters of synthesis, we identified and implemented four types of multistep strategies: (1) sequences of steps that allowed the program to overcome local maxima of molecular complexity[18] – that is, to complexify the synthons in one step but, by doing so, open up avenues for elegant, structure simplifying steps later on, (2) sequences that converted highly reactive to less reactive groups (a.k.a., functional group interconversions, FGIs) and thus reduced the numbers of potential chemical incompatibilities in the synthons; (3) "Bypasses" that first removed a conflicting group before trying a step that the program otherwise saw as very promising; and (4) "supersteps" in which certain reactions could be performed simultaneously, under the same reaction conditions. With these four types of strategies, Chematica finally become an expert-level planner, not only using these multistep strategies in separation, but also combining them into even longer, highly logical sequences, sometimes to the depth of five-six synthetic steps.

The improvement was immediately manifest in the program's ability to plan syntheses to complex natural products[11] as illustrated by the synthesis of Methyl Monate C in **Figure 1a** or the synthesis of Aplysin in **Figure 1b**, both of which are hardly discernible from routes that a human expert might design. In fact, in a recent paper on the topic,[11] we assembled a collection of 20 Chematica-planned and 20-literature published (in journals like *J. Org. Chem.*, *Org. Lett.*, *Angew. Chem.*, or *JACS*) syntheses, redrew them in the same format, arranged in no particular order, and asked world-leading experts to guess which ones were machine's creations and which ones were human designs. The experts could no longer tell, indirectly validating Chematica's design. Of course, we also demonstrated direct validation by successfully executing Chematica's synthetic plans to three natural products shown in **Figure 6**. Although there are still some very complex targets Chematica cannot tackle (e.g., CJ – 16,264, Ryanodol or Taxol, see [11]), it is generally for the lack of suitable reaction rules that still need to be coded and improvements in the computer architecture that are still required to handle extremely large reaction networks. Still, these improvements are no longer a question of "if" but rather of "when" and there can be little doubt now that the machine reached a level comparable to human experts.

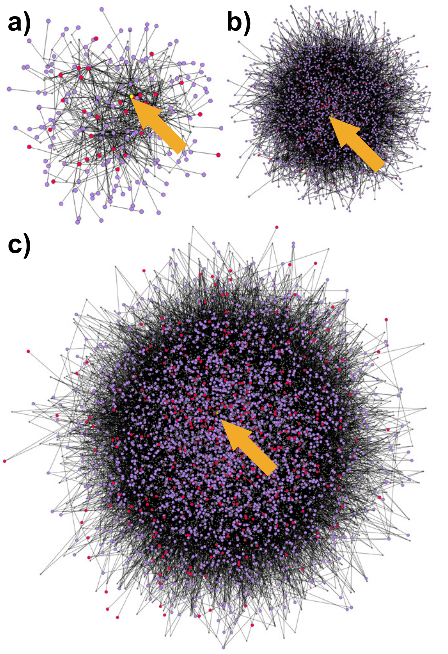


Figure 4. A growing network of synthetic options considered – after **(a)** 15, **(b)** 123 and **(c)** 541 iterations – during retrosynthetic analysis of a simple triarylamine (node indicated by yellow arrow). Figure reproduced from ref. [35]

At this point, it is perhaps wise to pause and ask a provocative question: Who should care about this accomplishment? For the sake of argument, one might say that no matter how intricate, awe-inspiring and even beautiful total syntheses of natural products might be, they are pursued by maybe few hundred research groups worldwide and, compared to the hey-day of the discipline some decades ago, are no longer at the center stage of modern chemistry. Paradoxically, after having spent some twenty years on teaching the machine how to plan such syntheses, we are inclined to agree with this argument. While we believe that demonstrating total synthesis by computer was essential for convincing the community, the machine will likely have more impact when applied to slightly different—though still synthesis-oriented – problems.

Synthesis with multiple constraints

Pondering such problems, we note that computer's major advantage over human brain, beyond sheer speed of performing arithmetic operations, is the number of logical conditions it can handle simultaneously. Imagine a situation in which one seeks not just a viable synthesis of some non-trivial drug molecule, but also a route that is economical, does not involve any toxic intermediates or solvents (i.e., "green"), does not use heavy-metal catalysts (usually undesired in pharmaceutical synthesis, especially in the last steps[25]), and ideally does not infringe upon existing patents. A human performing such planning would have to consult quite a few catalogs of available starting materials, lists of toxic substances, and patent literature – for a computer, "memorizing" such lists and keeping track of these additional, multiple constraints during synthesis planning is straightforward. In fact, in ref [26], we showed how these capabilities can be used to navigate around patented routes and how they can be used to design economical and green routes leading, for instance, to several blockbuster drugs. In a similar genre, in very recent work by us[27] and by Cernak's team,[28] the imposed constraint was to avoid the key intermediates used in the production of antivirals potentially relevant in the context of the current COVID-19 pandemic. In this task, Chematica designed as many as 17 alternative routes to one target whereas Cernak used the program not only to plan alternative syntheses but also carried many of them in the laboratory, validating Chematica's plans once again. In a broader context, the motivation for such analyses is that the known synthetic routes use the same key ingredients and these ingredients might rapidly become unavailable should a given drug candidate prove effective, triggering high worldwide demand. Therefore, creation of "synthetic contingency plans," as we called Chematica's alternative routes, fits well into the strategic efforts of many organizations and governments to secure stable, risk-free supply chains of key pharmaceuticals.[29] Of course, one might argue that skilled medicinal chemists might have come up with such alternative syntheses themselves – but to perform these analyses for thousands of other FDA approved medications would be an extremely tedious task. Only computers have the power to perform such strategic analyses at requisite scales.

The second type of a problem in which computers may outclass humans is planning the syntheses of many targets simultaneously, for example, in the synthesis of a library of compounds around a scaffold of interest. We are not taking here about syntheses planned one-by-one but about "global plans" that make use of intermediates and starting materials common to multiple individual pathways (the use of such common intermediates/substrates may lower the overall cost of the process). As described in ref [30], Chematica is quite adept in constructing such plans within minutes to hours; in the example in **Figure 7** the software's task was to design a global synthetic plan leading to the synthetically most accessible M+6, ^{13}C isotopically labelled derivatives of ten anticoagulant rodenticides. Note how intricate this plan is – it looks like a small network, not just a synthetic path. The problem

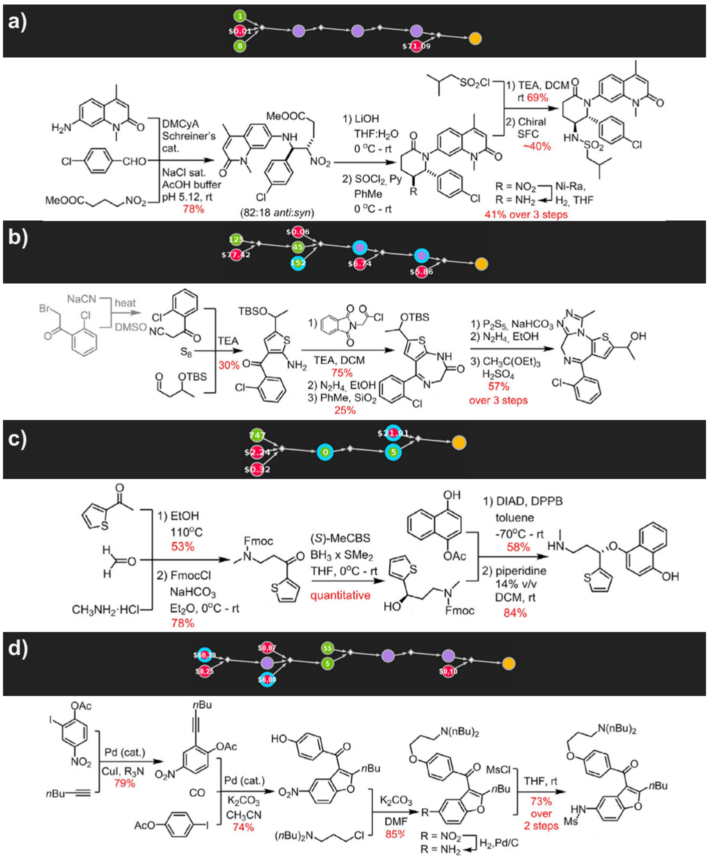


Figure 5. Syntheses of high-value, medicinally relevant targets designed by Chematica and validated by experiment. Syntheses of **(a)** BRD7/9 inhibitor; **(b)** hydroxyetizolam; **(c)** hydroxyduloxetine; and **(d)** dronedarone. Experimental yields are in red font. In Chematica pathway miniatures: yellow nodes = targets; violet = unknown molecules; green = known molecules; red = commercially available chemicals; blue halos = protection needed. Figure reproduced with permission from ref. [16]

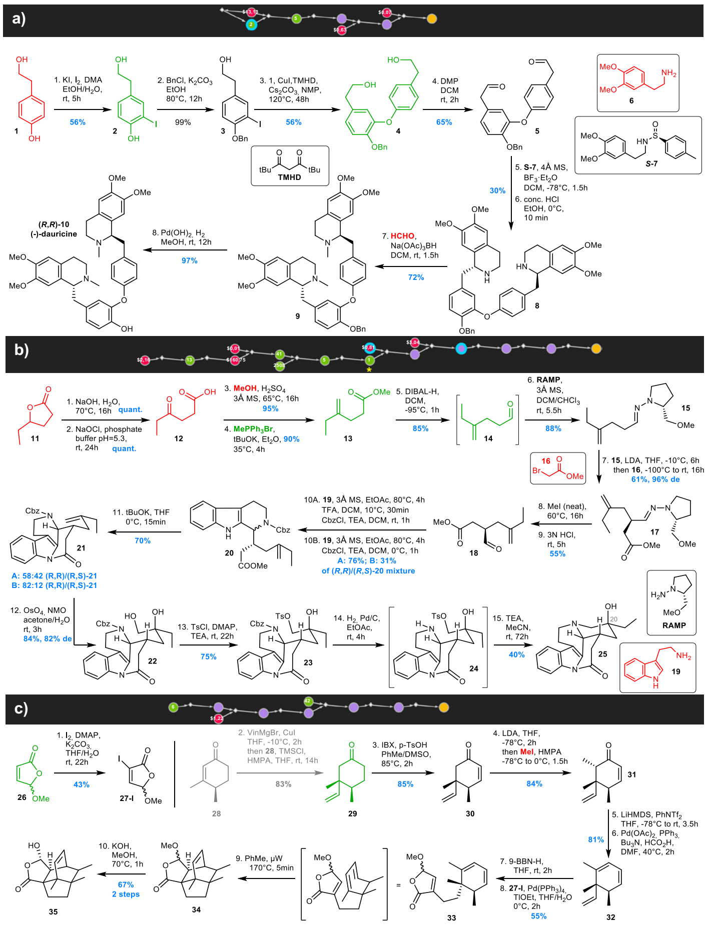

a)
1. KI, I2, DMA EtOH/H2O, rt, 5h 56%
2. BnCl, K2CO3 EtOH 80°C, 12h 99%
3. 1, CuI,TMHD, Cs2CO3, NMP, 120°C, 48h 56%
TMHD
4. DMP DCM rt, 2h 65%
5. S-7, 4Å MS, BF3·Et2O DCM, -78°C, 1.5h
6. conc. HCl EtOH, 0°C, 10 min
30%
7. HCHO, Na(OAc)3BH DCM, rt, 1.5h 72%
8. Pd(OH)2, H2 MeOH, rt, 12h 97%
(R,R)-10 (-)-dauricine
b)
1. NaOH, H2O, 70°C, 16h quant.
2. NaOCl, phosphate buffer pH=5.3, rt, 24h quant.
3. MeOH, H2SO4 3Å MS, 65°C, 16h 95%
4. MePPh3Br, tBuOK, Et2O, 90% 35°C, 4h
5. DIBAL-H, DCM, -95°C, 1h 85%
6. RAMP, 3Å MS, DCM/CHCl3 rt, 5.5h 88%
7. 15, LDA, THF, -10°C, 6h then 16, -100°C to rt, 16h 61%, 96% de
8. MeI (neat), 60°C, 16h
9. 3N HCl, rt, 5h 55%
10A. 19, 3Å MS, EtOAc, 80°C, 4h TFA, DCM, 10°C, 30min CbzCl, TEA, DCM, rt, 1h
10B. 19, 3Å MS, EtOAc, 80°C, 4h CbzCl, TEA, DCM, 0°C, 1h
A: 76%; B: 31% of (R,R)/(R,S)-20 mixture
11. tBuOK, THF 0°C, 15min 70%
A: 58:42 (R,R)/(R,S)-21
B: 82:12 (R,R)/(R,S)-21
12. OsO4, NMO acetone/H2O rt, 3h 84%, 82% de
13. TsCl, DMAP, TEA, rt, 22h 75%
14. H2, Pd/C, EtOAc, rt, 4h
15. TEA, MeCN, rt, 72h 40%
RAMP
c)
1. I2, DMAP, K2CO3 THF/H2O rt, 22h 43%
2. VinMgBr, CuI THF, -10°C, 2h then 28, TMSCl, HMPA, THF, rt, 14h 83%
3. IBX, p-TsOH PhMe/DMSO, 85°C, 2h 85%
4. LDA, THF, -78°C, 2h then MeI, HMPA -78°C to 0°C, 1.5h 84%
5. LiHMDS, PhNTf2 THF, -78°C to rt, 3.5h
6. Pd(OAc)2, PPh3, Bu3N, HCO2H, DMF, 40°C, 2h 81%
7. 9-BBN-H, THF, rt, 2h
8. 27-I, Pd(PPh3)4, TlOEt, THF/H2O 0°C, 2h 55%
9. PhMe, μW 170°C, 5min
10. KOH, MeOH, 70°C, 1h 67% 2 steps

Figure 6. Total syntheses of natural products planned by Chematica and validated in the laboratory: (a) (–)-Dauricine; (b) Tacamonidine, (c) Lammelodysidine A. Figure reproduced from ref. [11]

of library-wide design has many more ramifications (e.g., in ranking library members for the ease of synthesis, or in the selection of the most synthetically accessible isotopomers, for details see [30]). These tasks definitely require computer's assistance when the numbers of library members or the ways in which a compound can be multiply labelled become large (for details, see [30]). Altogether, we are quite excited about computer-assisted synthesis with constraints, as it can really make an impact on green chemistry, economical use and reuse (in multiple syntheses) of the same substrates, or even IP considerations, unlocking new process routes to, e.g., generic drugs.

Forward, not backward!

Still, the examples discussed so far are within the realm of retrosynthesis and are limited in one fundamental aspect – namely, they presuppose the knowledge of the target(s). Can computer-designed syntheses still be of help at the stage od *discovering* new targets with desirable pharmacological or other properties? The answer is in the affirmative provided we reverse the problem and instead of retro- start thinking in the "forward" direction.

Synthesizable molecular spaces

In the forward synthesis process, one starts from a collection of some basic substrates ("generation" G_0) and asks the machine to apply its reaction rules to generate the products of reactions between these substrates. This creates synthetic generation G_1. Subsequently, the molecules from G_0 and G_1 are combined, and their possible reactions are considered, yielding generation G_2. The process is then iteratively applied up to some user-specified generation G_n. As could be expected, the numbers of virtual molecules thus created increase rapidly. In one recently published work,[31] we showed that when ~600 rules describing prebiotically plausible reactions are applied to only six very basic substrates – water, ammonia, hydrogen cyanide, hydrogen sulfide, methane and nitrogen, all assumed to be present on primitive earth – they generate, within just few steps, a network comprised of tens of thousands of structurally diverse molecules, each of which is synthesizable (by definition of our construction) along one and usually many synthetic routes (**Figure 8**).

Property mapping and new pharmaceutical leads. Once created, this "synthesizable molecular space" can be mapped according to some property of interest – in the context of prebiotic chemistry, an obvious choice is to mark molecules that are known as the building blocks of life (amino acids, nucleobases, nucleosides, carbohydrates, and metabolites found in living organisms; red nodes in **Figure 8a**). This simple operation immediately prompts a set of interesting questions: What distinguishes these molecules from other, unmarked ones (i.e., from those that were "not chosen" to become life's components)? In how many ways can the life-like molecules be synthesized? Can they be made along unknown synthetic routes? Do some reaction sequences close into cycles, maybe even autocatalytic ones? For answers to these and other questions – and yes, for new and experimentally validated synthetic routes and cycles – the reader is referred to ref. [31] (and also **Figure 8b,c**). What concerns us here, however, is the uniquely enabling power of combining forward synthesis with property mapping – there is simply no way a human could generate such a complex maze of synthetic options, or inspect its contents for some property (or properties) of interest in a realistic time.

Naturally, this concept has broader implications than just prebiotic analyses. The starting materials can be any substrates one wishes to use, the database of reactions can encompass thousands of reactions relevant to medicinal chemistry, and the properties mapped onto the network can relate to pharmacological properties. This is illustrated in the screenshot from our Allchemy platform in **Figure 9**. Within just n = 3 steps, eight popular building (**Figure 9a**) blocks create a synthesizable space of 3,104 molecules (**Figure 9b,c**), which are then scrutinized by various AI "filters" (**Figure 10**). First, neural networks, NN, pre-trained on the set of >2,000 FDA approved drugs vs. random small molecules (in **Figure 10a,** menu panel circled in green) are used to determine which of the molecules within the space are "drug-like"[32] – that is, have general structural features characteristic of drugs. Second, other NNs are used to filter out molecules that have features indicative of specific toxicity modalities (**Figure 10a**, panel circled in yellow). After application of these two filters, our molecular synthesizable space is reduced – within just seconds of analysis – by ca. 65%, to 1,103 molecules that "look" like drugs and are predicted to be non-toxic. Some of these molecules are shown in the main "window" in **Figure 10a**. At this stage, we may become interested is specifics – for instance, which of these molecules might bind to particular protein targets of interest. Another neural network comes into play – this time, the network is trained on ca. 2 million binding assays describing binding of various small

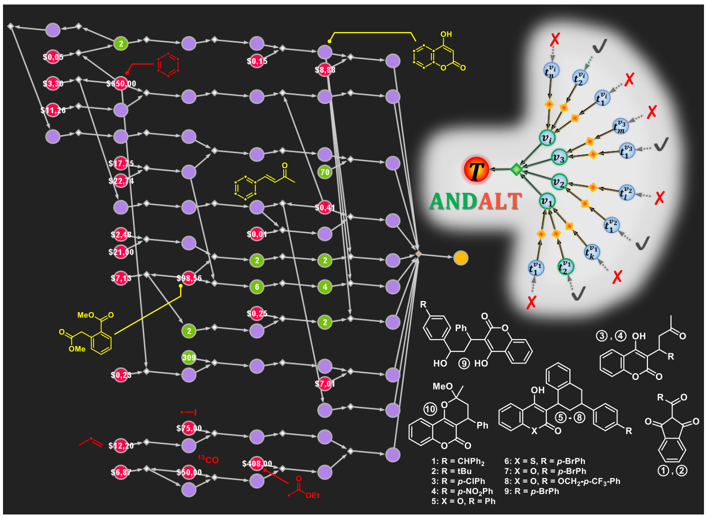


Figure 7. Example of multiple-target design. In this problem, Chematica was presented with ten anticoagulant rodenticides. In addition, each of these parent compounds (drawn in white on the bottom right) was supposed to be isotopically labelled with six ^{13}C carbons – note that there are potentially many options for such labelling. The program was asked to find the most synthetically accessible isotopomer ("ALT" condition meaning one of many alternatives) for each ("AND") parent class and "globally" optimize the synthetic plan to be the most economical. Note that this plan is no longer just "a pathway" – instead, it is a small network of pathways sharing common intermediates, some of which are drawn in yellow. The inexpensive sources of ^{13}C are drawn in red (^{13}C atoms are denoted by small dots). Details of this study will be published separately.

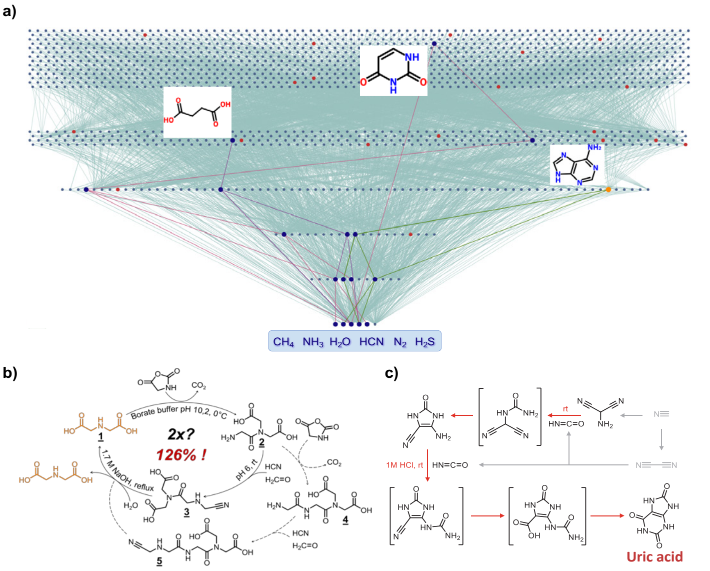


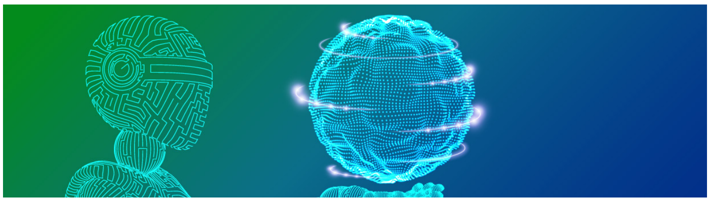
Figure 8. The network of prebiotic chemistry simulated with Allchemy. **(a)** The first five synthetic generations of a network of compounds synthesizable from CH_4, NH_3, H_20, HCN, N_2, and H_2S. Nodes colored red correspond to biotic molecules. Three such biotic molecules (succinic acid, uracil and adenine) are shown along with some of their syntheses traced over the network. **(b)** Allchemy-generated and experimentally validated self-regenerating, prebiotic cycle. After execution of the cycle, iminodiacetic acid – the template molecule – is self-regenerated in 126% yield. **(c)** An example of a new prebiotic synthesis (here, of uric acid) predicted by the software and validated by experiment. For details, see ref.[31] from which the panels **(b,c)** are reproduced.

molecules to several thousand protein targets. This network is taught which structural features in a molecule are indicative of binding (or its lack) to specific proteins. In **Figure 10b**, panel circled in violet is used to specify our targets of interest and also those to which our candidate molecules should not bind; in addition, we can set the threshold of certainty with which these predictions are to be made (the more certainty, the more stringent the network's criteria and the smaller the number of molecules that will pass filtering). Say, we decided to narrow our molecular space – with high certainty – to molecules the network predicts to bind to serotonin 5-HT1D receptor but not to opioid receptors μ or δ. This is a realistic scenario if we were looking for potential anti-migraine drugs that would not be addictive via interaction to the opioid receptor. After few seconds, the network examines our molecular space, and finds 11 molecules that meet our criteria, ranking them according to the predicted binding to serotonin 5-HT1D receptor (main panel of **Figure 10b**). Among these top top-ranking candidates, there is one already known and approved migraine medication Rizatriptan (which is reassuring) but there are also many completely new structures. Clicking on any of them, provides a synthetic route (and sometimes many routes) by which Allchemy generated this molecule (**Figure 10c**). All in all, the full cycle of *in silico* synthesis and property prediction for thousands of candidate molecules took less than five minutes yielding plausible leads worthy of further scrutiny and perhaps even wet lab synthesis and assaying. There is simply no way a human chemist or even a group of chemists could beat such timelines.

As in the case of retrosynthesis with constraints, additional conditions for forward synthesis and/or subsequent filtering can easily be envisioned and applied. In Allchemy, synthetic constraints can be, for instance, to use only green reaction conditions, or to start with a certain molecular fragment and perform only those reactions that make molecules increasingly similar to a given target of interest. By the very nature of our forward-synthesis approach, "similars" thus created are always synthesizable which is not the case for many AI methods that can create molecules that are similar but hard or impossible to make.[33] In terms of filtering, for reasons beyond this short article, we are mostly interested in heats of formation and some optical properties, but any property that can be calculated on the basis of molecular structure can be added.

The new brave world (of computerized synthesis)!

The limited space of this article does not allow us to narrate all these applications in detail. But, we hope, that even this short discussion will serve to convince the readers that synthetic chemistry has entered a new era. The computer-generated syntheses can be now trusted in terms of quality and can be generated on scales previously not thought possible. This newly acquired capability will have tremendous impact on the way we make molecules – before we synthesize them in the laboratory, we will be able to scrutinize many diverse synthetic plans generated on short times. Computers will provide us with suggestions for more economical and more green pathways. We will be able to create synthesizable molecular spaces of breathtaking sizes and will find in them readily-makeable molecules that are likely to have properties we desire. Synthetic planning will be accelerated and more property-oriented. Of course, we the humans will still be needed to execute these synthesis plans (but watch



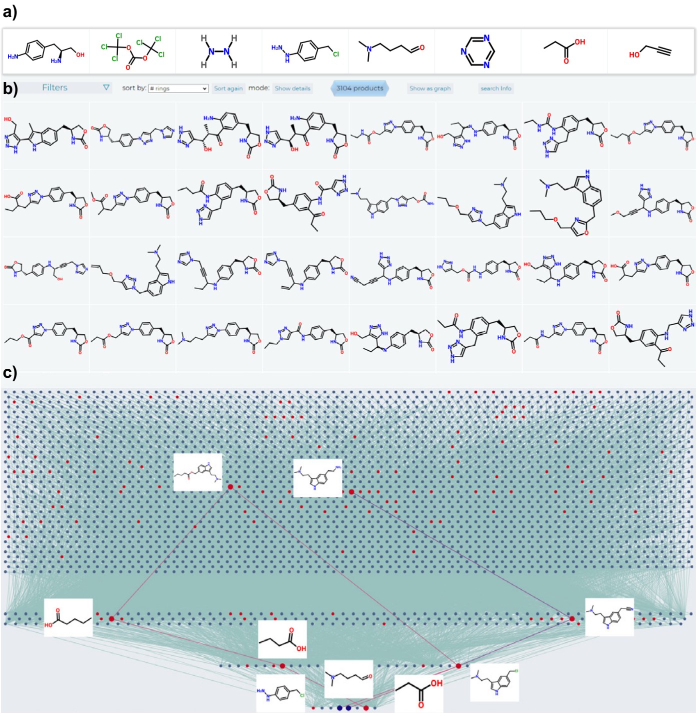
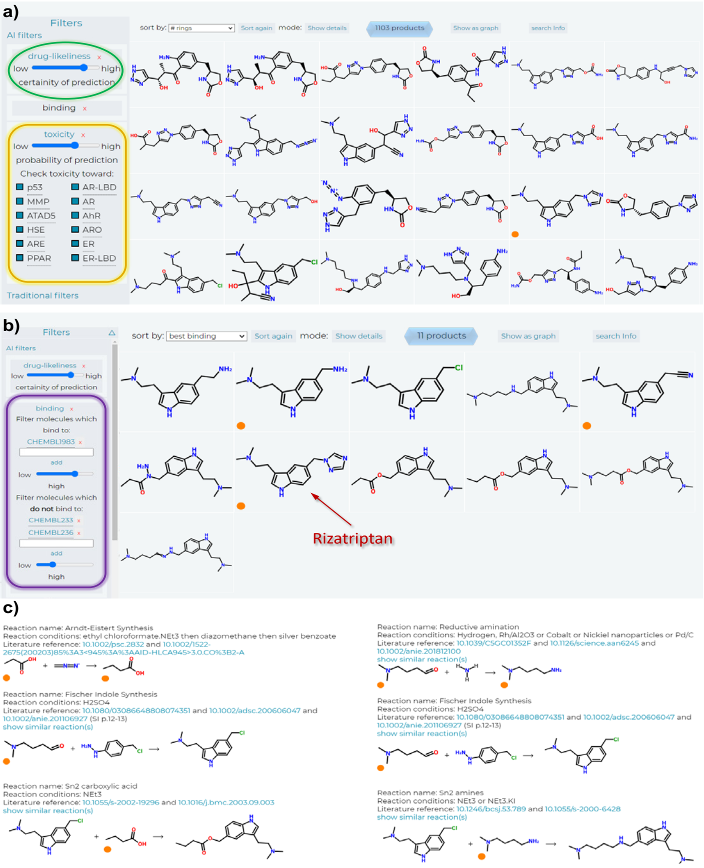


Figure 9. Creation of synthesizable spaces. Eight simple starting materials shown in **(a)** produce a space of 3104 molecules synthesizable within three synthetic steps. The entire process took 4 min on a standard multicore desktop. In **(b)**, fraction of this space is visualized as a list. In **(c)**, it is shown as a network. Red nodes correspond to molecules that are either known drugs or are similar to these known drugs (at a certain, user-specified level). The connections highlighted trace syntheses to two of these drug-similars (see also **Figure 10**). Some intermediates along the synthetic route are also shown. Images are screenshots from the Allchemy platform.



Reaction name: Arndt-Eistert Synthesis
Reaction conditions: ethyl chloroformate.NEt3 then diazomethane then silver benzoate
Literature reference: 10.1002/psc.2832 and 10.1002/1522-2675(200203)85%3A3<945%3A%3AAID-HLCA945>3.0.CO%3B2-A

Reaction name: Fischer Indole Synthesis
Reaction conditions: H2SO4
Literature reference: 10.1080/00397918808074351 and 10.1002/adsc.200606047 and 10.1002/anie.201106927 (SI p.12-13)
show similar reaction(s)

Reaction name: Sn2 carboxylic acid
Reaction conditions: NEt3
Literature reference: 10.1055/s-2002-19296 and 10.1016/j.bmc.2003.09.003
show similar reaction(s)

Reaction name: Reductive amination
Reaction conditions: Hydrogen, Rh/Al2O3 or Cobalt or Nickiel nanoparticles or Pd/C
Literature reference: 10.1039/C5GC01352F and 10.1126/science.aan6245 and 10.1002/anie.201812100
show similar reaction(s)

Reaction name: Fischer Indole Synthesis
Reaction conditions: H2SO4
Literature reference: 10.1080/00397918808074351 and 10.1002/adsc.200606047 and 10.1002/anie.201106927 (SI p.12-13)
show similar reaction(s)

Reaction name: Sn2 amines
Reaction conditions: NEt3 or NEt3.KI
Literature reference: 10.1246/bcsj.53.789 and 10.1055/s-2000-6428
show similar reaction(s)

Figure 10. Evaluation of synthesizable drug candidates. In **(a,b)**, the panels on the left highlight menus for various AI filters – general drug-likeness (in green halo), toxicity (yellow), and binding or not binding to some protein target(s) of interest (violet). Molecules surviving after each modality of filtering are shown in the main window to the right. Orange dots indicate molecules already reported in patents or high-impact publications. In **(b)**, the eleven molecules shown are predicted to bind to serotonin receptor 5-HT1D but not to μ or δ opioid receptors. Molecule indicated by the crimson red arrow is Rizatriptan, an approved medication for acute migraine. Syntheses for this or any other molecule analyzed are available by clicking on the structure of interest. Upon doing so, plans such as those in **(c)** are displayed. Images are screenshots from the Allchemy platform.

out, Chemputers[34] might be coming!) and will definitely be needed to create new synthetic methodologies the machines will then learn and incorporate into their planning. One thing is for sure: Chemistry at large can only benefit from these exciting human-machine synergies that are now emerging.

Funding

Development of Chematica was supported by US DARPA under the Make-It Award, 69461-CH-DRP #W911NF1610384. Some of the computer-planned syntheses were executed under the Symfonia (National Science Center, NCN, Poland, #2014/12/W/ST5/00592) and Maestro (NCN, Poland, #2018/30/A/ST5/00529) Awards. B.A.G. gratefully acknowledges support from the Institute for Basic Science Korea, project code IBS-R020-D1.

Conflict of Interest

Although Chematica was originally developed and owned by B.A.G.'s Grzybowski Scientific Inventions, LLC, he no longer holds any stock in this company, which is now property of Merck KGaA, Darmstadt, Germany. B.A.G. continues to collaborate with Merck KGaA, Darmstadt. All queries about access options to Chematica/Synthia™, including academic collaborations, should be directed to Sarah Trice (sarah.trice@sial.com). B.A.G. currently holds stock in Allchemy, Inc. Some of the Allchemy's modules are commercial (Drug Discovery) and some are publicly available (e.g., Origins of Life, available at https://life.allchemy.net) ◆

References:

1. Vléduts, G. É.; Finn, V. K. Creating a Machine Language for Organic Chemistry. *Inf. Storage Retr.* **1963**, 1, 101–116. https://doi.org/10.1016/0020-0271(63)90012-3; an expanded English version of the original 1957 report to the U.S.S.R. Academy of Sciences.
2. Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, 166, 178–192. https://doi.org/10.1126/science.166.3902.178.
3. van Rozendaal, E. L. M.; Ott, M. A.; Scheeren, H. W. A LHASA Analysis of Taxol. *Recl. des Trav. Chim. des Pays-Bas* **1994**, 113, 297–303. https://doi.org/10.1002/recl.19941130507.
4. van Rozendaal, E. L. M. *Some approaches to the synthesis of taxol and its derivatives: total-synthesis based on a LHASA analysis and semi-synthesis starting from taxine B*. Radboud University Nijmegen, Nijmegen, Netherlands, 1994. https://repository.ubn.ru.nl/bitstream/handle/2066/30052/mmubn000001_181282496.pdf; accessed October 27, 2020.
5. Tanaka, A.; Kawai, T.; Takabatake, T.; Oka, N.; Okamoto, H.; Bersohn, M. Synthesis of an Azaspirane via Birch Reduction Alkylation Prompted by Suggestions from a Computer Program. *Tetrahedron Lett.* **2006**, 47, 6733–6737. https://doi.org/10.1016/j.tetlet.2006.07.100.
6. Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; Hassabis, D. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, 529 (7587), 484–489. https://doi.org/10.1038/nature16961.
7. Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, 577 (7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7.
8. Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, 555, 604–610. https://doi.org/10.1038/nature25978.
9. Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; et al. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, 365, eaax1566. https://doi.org/10.1126/science.aax1566.
10. Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways Using Transformer-Based Models and a Hyper-Graph Exploration Strategy. *Chem. Sci.* **2020**, 11 (12), 3316–3325. https://doi.org/10.1039/C9SC05704H.
11. Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; Molga, K.; Młynarski, J.; Mrksich, M.; Grzybowski, B. A. Computational Planning of the Synthesis of Complex Natural Products. *Nature* **2020**. https://doi.org/10.1038/s41586-020-2855-y.
12. Fialkowski, M.; Bishop, K. J. M.; Chubukov, V. A.; Campbell, C. J.; Grzybowski, B. A. Architecture and Evolution of Organic Chemistry. *Angew. Chem. Int. Ed Engl.* **2005**, 117 (44), 7429–7435. https://doi.org/ 10.1002/anie.200502272.
13. Grzybowski, B. A.; Bishop, K. J. M.; Kowalczyk, B.; Wilmer, C. E. The "wired" Universe of Organic Chemistry. *Nat. Chem.* **2009**, 1, 31–36. https://doi.org/10.1038/nchem.136.
14. Gothard, C. M.; Soh, S.; Gothard, N. A.; Kowalczyk, B.; Wei, Y.; Baytekin, B.; Grzybowski, B. A. Rewiring Chemistry: Algorithmic Discovery and Experimental Validation of One-Pot Reactions in the Network of Organic Chemistry. *Angew. Chem. Int. Ed.* **2012**, 51, 7922–7927. https://doi.org/10.1002/anie.201202155.
15. Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed.* **2016**, 55, 5904–5937. https://doi.org/10.1002/anie.201506101.
16. Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, 4, 522–532. https://doi.org/10.1016/j.chempr.2018.02.002.
17. Molga, K.; Gajewska, E. P.; Szymkuć, S.; Grzybowski, B. A. The Logic of Translating Chemical Knowledge into Machine-Processable Forms: A Modern Playground for Physical-Organic Chemistry. *React. Chem. Eng.* **2019**, 4, 1506–1521. https://doi.org/10.1039/C9RE00076C.
18. Gajewska, E. P.; Szymkuć, S.; Dittwald, P.; Startek, M.; Popik, O.; Mlynarski, J.; Grzybowski, B. A. Algorithmic Discovery of Tactical Combinations for Advanced Organic Syntheses. *Chem* **2020**, 6, 280–293. https://doi.org/10.1016/j.chempr.2019.11.016.
19. Badowski, T.; Gajewska, E. P.; Molga, K.; Grzybowski, B. A. Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angew. Chem. Int. Ed.* **2020**, 59, 725–730. https://doi.org/10.1002/anie.201912083.
20. Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, 14 (1), 19–38. https://doi.org/10.1351/pac196714010019.
21. Corey, E. J.; Long, A. K.; Rubenstein, S. K. Computer-Assisted Analysis in Organic Synthesis.*Science* **1985**, 228, 408–418. https://doi.org/10.1126/science.3838594.
22. Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. Computer-aided Synthesis Design: 40 Years On. *WIREs Comput. Mol. Sci.* **2012**, 2, 79–107. https://doi.org/10.1002/wcms.61.
23. Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels-Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem. Int. Ed.* **2019**, 58, 4515–4519. https://doi.org/10.1002/anie.201806920.
24. Kowalik, M.; Gothard, C. M.; Drews, A. M.; Gothard, N. A.; Weckiewicz, A.; Fuller, P. E.; Grzybowski, B. A.; Bishop, K. J. M. Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry. *Angew. Chem. Int. Ed.* **2012**, 51, 7928–7932. https://doi.org/10.1002/anie.201202209.
25. Patel, R. N. Biocatalysis for Synthesis of Pharmaceuticals. *Bioorg. Med. Chem.* **2018**, 26 (7), 1252–1274. https://doi.org/10.1016/j.bmc.2017.05.023.
26. Molga, K.; Dittwald, P.; Grzybowski, B. A. Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways. *Chem* **2019**, 5, 460–473. https://doi.org/10.1016/j.chempr.2018.12.004.
27. Szymkuć, S.; Gajewska, E. P.; Molga, K.; Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski, B. A. Computer-Generated "Synthetic Contingency" Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, 11, 6736–6744. https://doi.org/10.1039/D0SC01799J.
28. Lin, Y.; Zhang, Z.; Mahjour, B.; Wang, D.; Zhang, R.; Shim, E.; McGrath, A.; Shen, Y.; Brugger, N.; Turnbull, R.; et al. Reinforcing the Supply Chain of COVID-19 Therapeutics with Expert-Coded Retrosynthetic Software. **2020**. https://doi.org/10.26434/chemrxiv.12765410.v1.
29. AI invents new "recipes" for potential COVID-19 drugs https://www.sciencemag.org/news/2020/08/ai-invents-new-recipes-potential-covid-19-drugs (accessed Oct 30, 2020).
30. Molga, K.; Dittwald, P.; Grzybowski, B. A. Computational Design of Syntheses Leading to Compound Libraries or Isotopically Labelled Targets. *Chem. Sci.* **2019**, 10, 9219–9232. https://doi.org/10.1039/C9SC02678A.
31. Wołos, A.; Roszak, R.; Żądło-Dobrowolska, A.; Beker, W.; Mikulak-Klucznik, B.; Spólnik, G.; Dygas, M.; Szymkuć, S.; Grzybowski, B. A. Synthetic Connectivity, Emergence, and Self-Regeneration in the Network of Prebiotic Chemistry. *Science* **2020**, 369 (6511), eaaw1955. https://doi.org/10.1126/science.aaw1955.
32. Beker, W.; Wołos, A.; Szymkuć, S.; Grzybowski, B. A. Minimal-Uncertainty Prediction of General Drug-Likeness Based on Bayesian Neural Networks. *Nature Machine Intelligence* **2020**, 2 (8), 457–465. https://doi.org/10.1038/s42256-020-0209-y.
33. Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. J. Chem. Inf. Model. 2020. https://doi.org/10.1021/acs.jcim.0c00174.
34. Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. *Science* **2019**, 363 (6423), eaav2211. https://doi.org/10.1126/science.aav2211.
35. Badowski, T.; Molga, K.; Grzybowski, B. A. Selection of Cost-Effective yet Chemically Diverse Pathways from the Networks of Computer-Generated Retrosynthetic Plans. *Chem. Sci.* **2019**, 10 (17), 4640–4651. https://doi.org/10.1039/C8SC05611K.