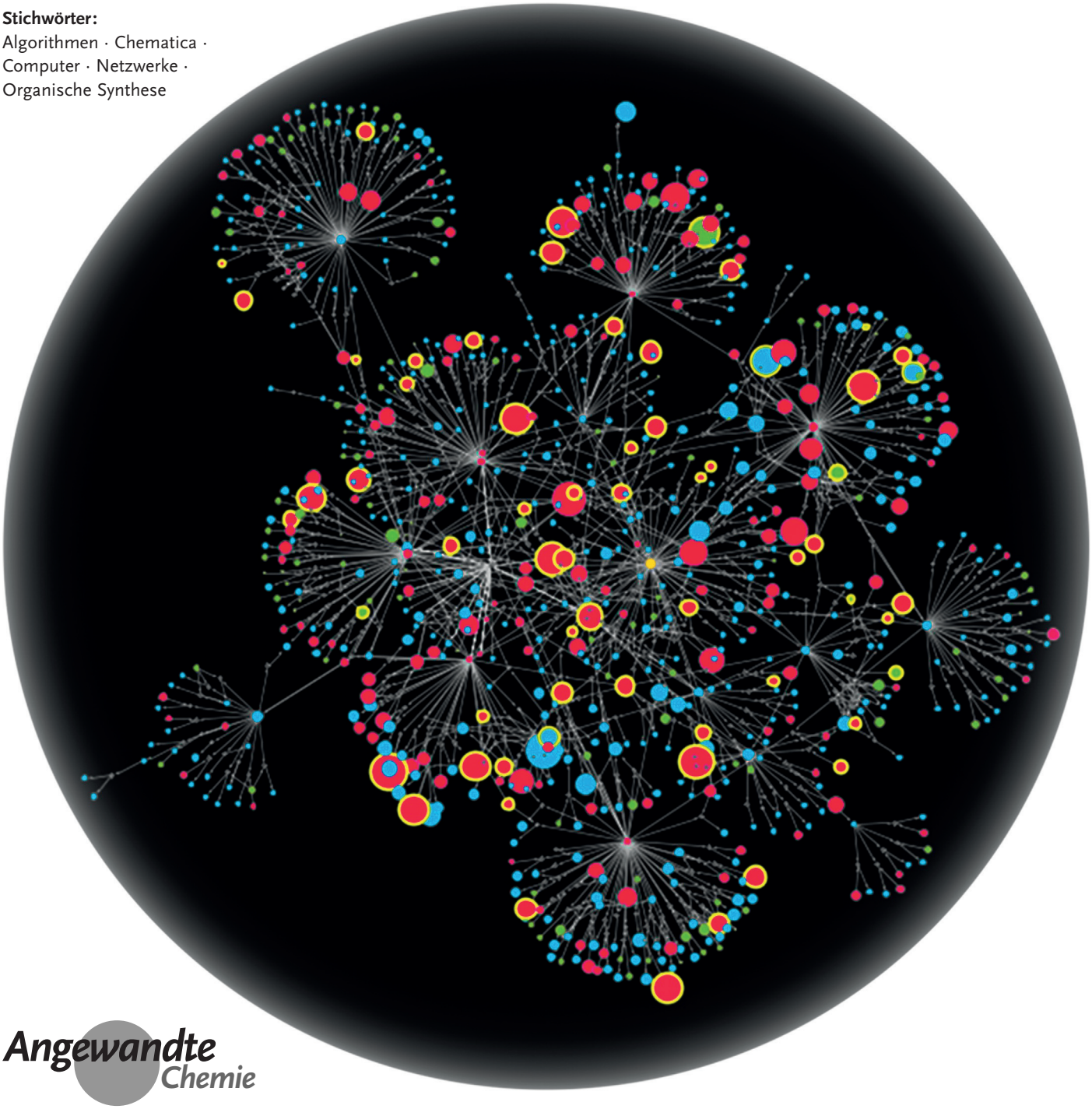
Computergestützte Syntheseplanung

Deutsche Ausgabe: DOI: 10.1002/ange.201506101
Internationale Ausgabe: DOI: 10.1002/anie.201506101

Computergestützte Syntheseplanung: Das Ende vom Anfang

*Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk und Bartosz A. Grzybowski**

Stichwörter:
Algorithmen · Chematica · Computer · Netzwerke · Organische Synthese



Seit der Einführung des ersten dokumentierten Forschungsprojekts (Dendral, 1965) zur rechnergestützten organischen Synthese ist ein halbes Jahrhundert vergangen. In den 1970er und 1980er Jahren wurden viele weitere Programme entwickelt, doch bis zu den 2000er Jahren war der Enthusiasmus dieser Pionierzeit weitgehend verschwunden, und wegen der schwierigen Aufgabe, die Planung organischer Synthesen in Computer einzugeben, erhielt sie den Ruf einer „Mission impossible“. Das ist recht merkwürdig angesichts der Tatsache, dass Computer in der Zwischenzeit viele andere Fähigkeiten „gelernt“ haben, die als alleinige Domänen von Intellekt und Kreativität des Menschen betrachtet wurden – so können Rechner heutzutage besser Schach spielen als menschliche Weltmeister, und sie können klassische Musik komponieren, die für das menschliche Ohr angenehm ist. In der organischen Synthese hat es zwar keine vergleichbaren Leistungen gegeben, aber dieser Aufsatz behauptet, dass es verfrüht wäre, eine Niederlage einzugestehen. Tatsächlich kann dem Computer schließlich „beigebracht“ werden, die Synthesen von komplizierten organischen Verbindungen in Sekunden- bis Minutenschnelle zu planen, indem die Kombination aus moderner Rechenleistung und Algorithmen aus der Graphen-/Netzwerktheorie zusammengeführt wird mit in geeigneten Formaten codierten chemischen Regeln (mit vollständiger Stereo- und Regiochemie) und Elementen der Quantenmechanik. Der Aufsatz beginnt mit einem Überblick über theoretische Grundkonzepte, die essenziell sind für die Analyse der großen Datenmengen chemischer Synthesen. Im Anschluss daran wird auf die Optimierung von Synthesewegen unter Einbeziehung bekannter Reaktionen eingegangen. Einen Schwerpunkt bildet die anschließende Besprechung der Algorithmen, die eine vollständig neue und automatisierte Planung der Synthesen für komplizierte Zielverbindungen ermöglichen, darunter auch solche, die noch nicht hergestellt wurden. Es gibt natürlich noch Verbesserungsmöglichkeiten, letztendlich aber werden Computer für die praktische Planung der organischen Synthese wichtig und hilfreich sein. Churchills berühmte Worte nach dem ersten wichtigen Sieg der Alliierten über die Achsenmächte paraphrasierend, ist es für die computergestützte Syntheseplanung nicht das Ende, nicht einmal der Anfang vom Ende, sondern das Ende vom Anfang. Der Computer ist da und wird bleiben.

Aus dem Inhalt

1. Einleitung

Nur wenige Gebiete der chemischen Forschung haben eine so drastische Schicksalswende – von anfänglicher Begeisterung bis zum heutigen Pessimismus – erlebt wie die computergestützte Syntheseplanung. Dieser Zustand überrascht etwas angesichts der Tatsache, dass in der Organik tätige Chemiker zu den ersten gehörten, die bereits in den 1960er Jahren die Möglichkeiten der modernen Datenverarbeitung in den Naturwissenschaften erkannten. Leider erwies sich die Aufgabe, die diese Pioniere in Angriff nahmen, als zu kompliziert für die Geräte und Algorithmen der damaligen Zeit, und die Methoden, die sie entwickelten, ließen sich nur auf relativ einfache Zielverbindungen anwenden, für die Fachleute eigentlich keine Unterstützung durch einen Rech-

[*] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, Dr. P. Dittwald, M. Bajczyk, Prof. Dr. B. A. Grzybowski
Institute of Organic Chemistry, Polish Academy of Sciences
Kasprzaka 44/52, Warsaw 02-224 (Polen)
E-Mail: nanogrzybowski@gmail.com
Prof. Dr. B. A. Grzybowski
Center for Soft and Living Matter of Korea's Institute for Basic Science (IBS) and Department of Chemistry, Ulsan National Institute of Science and Technology
50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan (Südkorea)
E-Mail: grzybor72@unist.ac.kr
Dr. M. Startek
Faculty of Mathematics, Informatics, and Mechanics
University of Warsaw
Banacha 2, 02-097 Warszawa (Poland)

Hintergrundinformationen zu diesem Beitrag sind im WWW unter http://dx.doi.org/10.1002/anie.201506101 zu finden.

ner benötigten. Zudem war die Verwendung der Programme mühsam, wie Professor P. Judson in seinem Buch über Expertensysteme in der Chemie eloquent berichtet,[1] sodass sie bei der Allgemeinheit keine breite Akzeptanz fanden und langsam, eins nach dem anderen, vom Schauplatz der Geschichte verschwanden. Diese ersten Programme für die chemische Synthese waren weder richtig noch falsch, sie waren schlicht ein wenig irrelevant für die Alltagspraxis in der organischen Synthese. Welche Gründe es auch gewesen sein mögen, die computergestützte Syntheseplanung scheint die große Revolution der Computertechnologie in den 1990er und 2000er Jahren verpasst zu haben: In einer Zeit, als Garry Kasparov im Schach von Deep Blue geschlagen wurde und Computer Gebiete revolutionierten, die von der Ökonomie bis zur Biologie reichten, hielten in der Chemie nur wenige Arbeitsgruppen die aktive Forschung zur computergestützten Syntheseplanung aufrecht. Eine bemerkenswerte Ausnahme, in der die dynamische und fruchtbare Entwicklung anhielt, waren chemische Datenbanken und die automatisierte Verwaltung von Synthesewissen (z. B. elektronische Notebooks). Heute sind Datenspeicher wie Reaxys,[2a] SciFinder,[2b] ChemSpider[2c] und SPRESI[2d] für Chemiker eine unschätzbare Hilfe bei der Suche nach Literaturquellen und/oder Beispielen für „analoge“ Reaktionen. Aber auch auf diesem Gebiet wird die Leistung moderner Computer nur unzureichend genutzt, und die Suchen erfolgen weitgehend manuell und schrittweise mit nur elementaren Leistungen zur Evaluierung von Reaktionssequenzen.

Die Hauptthese dieses Aufsatzes ist, dass sich die genannten Einschränkungen letztlich überwinden lassen und moderne Rechner 1) erheblich verbesserte Suchen in chemischen Datenbanken ermöglichen können, die den gesamten chemischen Raum bekannter Verbindungen umfassen, und 2) „lernen“ können, optimierte Synthesewege zu noch nicht untersuchten Zielverbindungen, darunter auch relativ komplizierte, zu entwerfen. Wie wir sehen werden, haben diese beiden Probleme einen gemeinsamen Nenner: Sowohl die Reaktionsdatenbanken als auch die von Computern während der retrosynthetischen Planung erzeugten Optionen können als Netzwerke beschrieben werden. Diese mathematische Erkenntnis ist ganz entscheidend, denn sie ermöglicht es, die in der Telekommunikation oder in Schachprogrammen allgemein verwendeten Algorithmen zu adaptieren und auf das Durchsuchen und Analysieren der riesigen Netzwerke von Synthesemöglichkeiten anzuwenden.

Der Aufsatz gliedert sich in drei Teile. Der erste Teil (Abschnitt 2) beschreibt die Anwendungen von Netzwerkalgorithmen beim Durchsuchen der gesamten bekannten (d. h. in der Literatur beschriebenen) Reaktionen. Wir besprechen die wesentlichen Unterschiede zwischen dem Durchsuchen von Datenbanken und der Suche in Netzwerken sowie die Vorteile der so genannten bipartiten Darstellung des Network of Organic Chemistry (NOC).[3a,b] Danach stellen wir die Konzepte verschiedener Bewertungsfunktionen (Scoring-Funktionen) vor, die – mit passend entwickelten Suchalgorithmen – hunderte Millionen Synthesemöglichkeiten pro Sekunde genau prüfen und evaluieren können und Synthesewege identifizieren, die für verschiedene nutzerdefinierte Kriterien optimiert sind (z. B. monetäre Gesamtkosten, Popularität der betreffenden Substrate, Vermeidung toxischer/gefährlicher Zwischenverbindungen usw.).

Auch wenn das Durchsuchen aller bekannten chemischen Reaktionen rechnerisch anspruchsvoll sein kann, sind die zugrundeliegenden Algorithmen auf anderen Wissenschaftsgebieten gut entwickelt und können daher verhältnismäßig einfach auf das Problem der chemischen Synthese erweitert

Sara A. Szymkuć beendete ihr Studium an der Faculty of Chemistry der Warsaw University of Technology mit einer Arbeit über Mehrkomponentenreaktionen zur Synthese von Peptidmimetika. Derzeit ist sie Doktorandin am Institute of Organic Chemistry der Polish Academy of Sciences in Warsaw. Zu ihren wissenschaftlichen Interessen gehören die computergestützte organische Chemie, chemische Netzwerke und Mehrkomponentenreaktionen.

Tomasz Klucznik studierte Chemie und Biotechnologie an der Gdańsk University of Technology. Er ist derzeit Doktorand am Institute of Organic Chemistry der Polish Academy of Sciences in Warsaw. Er erhielt bei verschiedenen Tagungen mehrere Preisen für die beste Präsentation und beteiligte sich aktiv an der Verbreitung der Chemie bei Gymnasiasten und Fachoberschülern. Er interessiert sich für chemische Netzwerke, die Theorie der organischen Chemie, Totalsynthesen, Methoden der organischen Synthese, die Chemie von Schwefel und Philosophie.

Ewa P. Gajewska schloss ihr Studium in Chemie und Biotechnologie an der Gdańsk University of Technology mit Auszeichnung ab. 2013 erhielt sie den Outstanding Achievements Award vom polnischen Ministry of Science and Higher Education. Sie ist zurzeit Doktorandin am Institute of Organic Chemistry der Polish Academy of Sciences in Warsaw. Ihre wissenschaftlichen Interessen umfassen chemische Netzwerke und die Grundlagen organischer Reaktionsmechanismen.

Karol Molga beendet sein Studium mit Auszeichnung an der Faculty of Chemistry der Warsaw University of Technology mit einer Arbeit zur Chemie von Organonickelverbindungen. Er ist zurzeit Doktorand am Institute of Organic Chemistry der Polish Academy of Sciences in Warsaw, wo er an chemischen Netzwerken und der Logik der computergestützten retrosynthetischen Analyse arbeitet.

werden. Die Planung einer De-novo-Synthese, die in Abschnitt 3 besprochen wird, verlangte jedoch die Entwicklung einer neuen Methode. Wir beginnen mit einer Schilderung historisch wichtiger Entwicklungsschritte (Abschnitt 3.1). Obwohl heute nur noch wenige der ersten Syntheseplanungsprogramme zur Verfügung stehen, sollte ihre Bedeutung für die Entwicklung auf diesem Gebiet nicht unterschätzt werden, denn sie veranlassten die Schaffung verschiedener maschinenlesbarer Molekülnotationen (z. B. SMILES)[4] und Strukturediteren (z. B. ChemDraw von Stewart Rubenstein, das aus dem von Corey entwickelten LHASA hervorging).[5,6] Programme wie das von P. Y. Johnson entwickelte SYNLMA[7] führten Elemente der formalen Logik in die Syntheseplanung ein, SYNCHEM von Gelernter[8] war eine Pionierleistung beim Durchsuchen expandierender Bäume von Synthesemöglichkeiten, und bei dem von Hanessian entwickelten CHIRON[9] liegt der Schwerpunkt auf dem Erkennen ähnlicher Strukturmuster in Zielverbindungen und Substraten. All dies sind wesentliche Aspekte der Syntheseplanung – es bleibt jedoch die Frage, warum all diese Innovationen nicht zu allgemein anerkannten rechnerischen Methoden/Mitteln geführt haben, wie sie in der Quantenchemie oder der Moleküldynamik genutzt werden.

Unter Berücksichtigung dieser Frage greifen wir (in Abschnitt 3.2) die Grundlagen des Problems auf und diskutieren den gedanklichen Zusammenhang zwischen der Syntheseplanung und anderen Gebieten, darunter Schach, der Zauberwürfel Rubik's Cube (Tabelle 1 und Lit. [10–15]) und sogar Linguistik. Aus diesen Vergleichen ergeben sich drei charakteristische Merkmale der Syntheseplanung: 1) Die Zahl der Regeln/„Züge" ist sehr viel größer als bei anderen Spielen (zehntausende Reaktionsarten gegenüber ungefähr zehn beim Schach und nur wenige bei Rubik's Cube); 2) die Anwendbarkeit jedes gegebenen „Synthesezugs" hängt vom Kontext ab, d. h. vom Vorhandensein anderer chemischer Gruppen im selben Molekül, genau wie die Bedeutung eines Worts vom Kontext des ganzen Satzes abhängen kann; und 3) im Unterschied zur Situation beim Schachspiel gibt es kein wohldefiniertes Kriterium für eine aktuelle „Syntheseposition", die zur Planung künftiger „Züge" systematisch evaluiert werden könnte. Das hat mehrere wichtige Konsequenzen. Erstens müssen die Regeln, die in den Computer eingegeben werden müssen, nicht nur „lokal" („trenne diese spezifische Bindung"), sondern auch kontextbezogen sein („prüfe andere Gruppen in dem Molekül") – diese Kontextabhängigkeit ist der Hauptgrund dafür, dass frühere Methoden (die entweder auf Bindungsbrüchen basierten oder „analoge" Reaktionen aus Literaturbeispielen nutzten) keinen Erfolg hatten. Kurz gesagt: Die chemischen Regeln für Bedingungen und Kontext müssen von menschlichen Experten codiert werden, und es sind sehr viele zu codieren (>10000), bevor der Rechner in Konkurrenz zu einem kompetenten Menschen treten kann. Zweitens muss und kann das Positionskonzept algorithmisch definiert werden, indem die strukturelle und chemische Komplexität der Substratsätze berücksichtigt wird, die in jedem „Reaktionszug" generiert werden. Das bedeutet wiederum, dass hypothetische Pfade sowohl hinsichtlich der durchgeführten Reaktionsschritte als auch der Komplexität der Substrate bewertet werden müssen. Diese Art der dualen Bewertung ist eine deutliche Abkehr von der alleinigen Untersuchung von Bindungsbrüchen und veranlasst uns, die Konzepte der Scoring-Funktionen für Reaktion und Verbindungen einzuführen. Drittens muss es Algorithmen geben, die wissensbasierte und Scoring-Funktionen für eine intelligente Navigation durch den Syntheseraum nutzen, die nicht nur „vorwärts" gerichtet ist, sondern auch von aussichtslosen

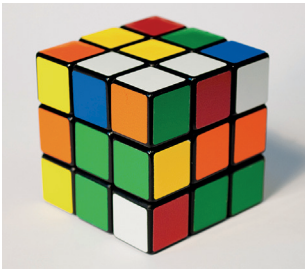
Piotr Dittwald studierte Mathematik und Informatik an der University of Warsaw, von der er auch seinen Ph.D. in Informatik erhielt. Seine Doktorarbeit umfasste die Untersuchung wiederkehrender Umordnungen im menschlichen Genom und die Anwendung rechnerischer Methoden in der Massenspektrometrie. Er erhielt zweimal die START Fellowship von der Foundation for Polish Science. Derzeit ist Postdoktorand am Institute of Organic Chemistry der Polish Academy of Sciences in Warsaw, wo er Algorithmen für die computergestützte chemische Synthese entwickelt.

Michał P. Startek studierte Mathematik und Informatik an der University of Warsaw, wo er 2015 mit höchster Auszeichnung promovierte. Derzeit ist er Postdoktorand an der gleichen Universität. Zu seinen Forschungsinteressen gehören Evolutionsmodelle für das Verhalten transponierbarer Elemente, Methoden der rechnerischen Analyse von chemischen Datenmassen und die computergestützte Planung chemischer Synthesen.

Michał D. Bajczyk erhielt seine M.Sc.-Abschlüsse in Chemie und in Biochemie von der Jagiellonian University in Cracow. Er wurde 2012 und 2013 zweimal mit dem Stipendium TEAM von der Foundation for Polish Science ausgezeichnet. Zurzeit ist er Doktorand am Institute of Organic Chemistry der Polish Academy of Sciences in Warsaw, wo sein Interesse der De-novo-Planung von Mehrkomponentenreaktionen und der physikalischen Biochemie gilt.

Bartosz A. Grzybowski beendete sein Studium an der Yale University 1995 und promovierte 2000 in Harvard. Nach mehr als 10 Jahren an der Northwestern University wechselte er nach Südkorea, wo er Distinguished Professor of Chemistry am Ulsan Institute of Science and Technology und Group Leader am Institute for Basic Science ist. Er ist zudem Professor am Institute of Organic Chemistry der Polish Academy of Sciences in Warschau. Er erhielt zahlreiche Auszeichnungen, darunter 2006 den ACS Unilever Award und 2013 den Nanoscience Prize.

Tabelle 1: Vergleich zwischen Schach, dem Zauberwürfel Rubik's Cube und der chemischen Synthese.[a]

	Schach	**Rubik's Cube**	**Chemische Synthese**
			
Zahl der Spieler	Zwei	Einer	Einer
Züge	Für jede Figur ein kleiner Satz definierter Züge; einige Züge können für manche Positionen nicht erlaubt sein	Drehung der einzelnen Würfelschichten; immer gleich viele Züge erlaubt	Sehr viele (>10000) mögliche Züge (d.h. Reaktionsregeln); anwendbare Züge hängen von der Struktur der Verbindung ab; Datenbanken der Züge können mit Fortschritten in der Chemie wachsen
Ausgangsposition	Immer die gleiche Anordnung der Figuren auf dem Brett; der Spieler mit Weiß beginnt	(Zufällige) Anordnung des Würfels	Zielverbindung, die synthetisiert werden soll
Position	Momentane Anordnung der Figuren auf dem Brett	Anordnung des Würfels	Satz von Substraten/Synthonen in jedem Schritt
Endposition	Schachmatt oder Überschreiten der erlaubten Zeit; Remis ist ebenfalls möglich	Jede der sechs Würfelseiten zeigt nur eine Farbe	Alle Substrate für die Synthese der Zielverbindung sind als „verfügbar" bewertet
Bewertung des Spiels	Gewonnen/verloren/Remis/nicht beendet	Gelöst/nicht gelöst; außerdem könnte die Zeit oder die Zahl der Züge bewertet werden (weniger Züge = bessere Bewertung)	Realisierbare Synthese gefunden/nicht gefunden; Realisierbarkeit schließlich durch experimentelle Ausführung bestätigt; neben „harten" Kriterien (Zahl der Stufen, Ausbeute) können auch „weiche" Kriterien wie „Eleganz" in die Bewertung einfließen
Komplexität	Obergrenze für Positionen ohne Bauernumwandlungen etwa 2×10^{40};[10] nach gängiger Schätzung sind durchschnittlich 35 Züge aus einer bestimmten Position möglich, damit errechnen sich etwa 10^{123} mögliche Spiele mit 80 Zügen[11]	Mehr als 4×10^{19} mögliche Anordnungen; mehr als 2×10^{20} mögliche Sequenzen aus 18 Zügen[12]	Durchschnittlich 80.2 verschiedene Reaktionen lassen sich auf ein nicht triviales Retron anwenden;[13] daraus ergeben sich $\approx3.5\times10^{28}$ mögliche 15-stufige Syntheseswege und $\approx1.2\times10^{57}$ mögliche 30-stufige Pfade
Maximale Zahl der Züge	Theoretisch unbegrenzt (wenn keiner der Spieler die Dreifachwiederholung oder die 50-Züge-Regel anwendet), das längste belegte Turnier ging aber über 269 Züge	Intensiven Rechnungen zufolge lässt sich jeder Zauberwürfel in nicht mehr als 20 Zügen lösen[14]	Kommerziell erhältliches Halaven wird in 62 Syntheseschritten hergestellt;[15] das scheint eine Obergrenze für industriell relevante Synthesen zu sein
Optimale Lösung	Existiert im Allgemeinen nicht; in einer bestimmten Position kann es eine Gewinn-/Remis-Strategie geben	Existiert, ist aber normalerweise schwer zu bestimmen	Generell kann keine Einzellösung objektiv als „optimal" angesehen werden, da sie von verfügbaren Substraten und/oder den angewendeten Kriterien (z.B. minimale Stufenzahl, umweltfreundliche Bedingungen, keine Schutzgruppen usw.) abhängt

[a] Abbildungsnachweis: Links/Mitte: Wikimedia Commons, Genehmigung CC-BY-SA-3.0; rechts: modifiziert nach https://www.flickr.com/photos/usdagov/16714715557/, U.S. Department of Agriculture, Genehmigung CC BY 2.0.

Positionen umkehrt, um zu vermeiden, was Corey als „kombinatorische Explosion" der Möglichkeiten bezeichnet hat. In den Abschnitten 3.3 und 3.4 gehen wir auf Methoden ein, die diese Vorgaben erfüllen, wobei auch die ganze Stereochemie, Regiochemie, Schutzgruppeninformationen und ansatzweise auch die Quantenmechanik berücksichtigt wird.

Obwohl die im Folgenden beschriebenen Beispiele für tatsächliche computergeplante Synthesen schon nicht trivial sind und die Fähigkeit des Computers widerspiegeln, mit fast so gutem Erfolg zu planen wie ein hoch qualifizierter menschlicher Chemiker, bleiben für die künftige Forschung auch weiterhin vielfache Herausforderungen und interessante

Möglichkeiten, auf die wir in Abschnitt 4 eingehen. So gab es interessante Entwicklungen bei der Vorhersage von Ergebnissen stereoselektiver Reaktionen, von Reaktionsausbeuten und sogar Reaktionsbedingungen. Zudem kamen neue Maße für die Synthesekomplexität auf, mit denen sich die „Synthetisierbarkeit“ in großen Bibliotheken[16] aus „virtuellen Verbindungen“, die heutzutage in der pharmazeutischen Industrie und der Materialwissenschaft routinemäßig erstellt werden, rasch abschätzen lässt. Nicht zuletzt gibt es neue Methoden zur Vorhersage neuer Reaktionsarten/Mechanismen, die von quantenmechanischen Rechnungen bis zu Methoden der Graphentheorie und des maschinellen Lernens reichen.

Alles in allem glauben wir, dass moderne Computer für Praktiker in der organischen Chemie eine wertvolle Hilfe sein können. Auch wenn die Rechner wahrscheinlich noch nicht an die Kreativität von Spitzenforschern in der Totalsynthese heranreichen, können sie eine unglaubliche Menge an chemischem Wissen kombinieren und die Daten intelligent und mit einer Schnelligkeit verarbeiten, die von Menschen niemals erreicht wird. In der Retrosyntheseplanung können sogar preiswerte Desktop-Rechner Tausende von passenden Reaktionsmotiven je Sekunde prüfen und diejenigen identifizieren, die selbst erfahrene Chemiker nur schwer erkennen würden; tatsächlich können auch Desktop-Computer mit ihrer Fähigkeit, komplizierte Anordnungsmuster und Mehrkomponentenreaktionen zu erkennen, Menschen deutlich überlegen sein. Man könnte natürlich behaupten, dass diese Strukturmotive mithilfe der menschlichen Intuition zu erkennen wären. Das ist aber etwa so, als würden wir argumentieren, wir könnten mit Papier und Stift „letztendlich“ zwei zehnstellige Zahlen bis zur Genauigkeit von zehn Dezimalstellen dividieren – warum sollten wir das tun, wenn wir über einen Taschenrechner verfügen? Unser Ansicht nach sollten alle Synthesehilfsprogramme gerade als „chemische Rechner“ betrachtet werden, die eine Syntheseplanung beschleunigen und erleichtern und schnell mehrere Synthesemöglichkeiten bieten, die ein menschlicher Experte anschließend beurteilen und vielleicht kreativ verbessern kann.

2. Navigieren in einem bekannten chemischen Raum: Syntheseplanung auf der Basis literaturbekannter Reaktionen

2.1. Einfache und nicht so einfache Suchen in Datenbanken

Es könnte zwar den Anschein haben, dass die Praxis der organischen Synthesechemie nicht unbedingt von Themen der Datenverarbeitung abhängt oder überhaupt mit ihnen zusammenhängt, tatsächlich gehörte die Synthesechemie aber zu den ersten Naturwissenschaften, in denen die moderne Informationstechnologie umfangreiche Verwendung fand. Bereits 1957 – ein Jahr, bevor Jack Kilby den ersten praktischen integrierten Schaltkreis vorführte, stellten sich die beiden sowjetischen Wissenschaftler G.E. Vléduts und V. K. Finn eine „*information machine for chemistry*“ vor,[17] die eine „*practically unlimited amount of chemical information*“ speichern konnte und diese Information anschließend zur Lösung von verschiedenen anwenderspezifischen Aufgaben weiterverarbeitete, unter anderem „*(…in ascending order of complexity): (i) search for information about an individual chemical compound, (ii) search for chemical compounds possessing a certain given combination of characteristics (including structural indices), (iii) search for the classes of reactions into which a definite individual compound can enter, (iv) search for the class of reactions producing a particular chemical compound, (v) search for the class of reactions which are of the same type chemically and are characterized by a transfer of given structural elements … from the initial molecules into other definite structural elements of the final molecules, (vi) search for the reaction that will take place between given compounds under given conditions, (vii) search for ways of synthesizing a given compound from a definite number of permissible initial compounds, and so on.*“ Im Rückblick nach mehr als 50 Jahren kann man die Vision der Verfasser nur bewundern, denn sie legten einen recht genauen Entwurf moderner chemischer Datenbanken und ihrer Leistungsfähigkeit vor. Tatsächlich können Millionen Chemiker weltweit heute Speicher mit veröffentlichten Reaktionen wie die zuvor genannten Reaxys und SciFinder nutzen, um mühelos nach bestimmten Verbindungen oder Substrukturen, spezifischen Umwandlungsarten, nach Verbindungen/Reaktionen mit strukturellen Ähnlichkeiten zu einer gewünschten Verbindung und vielem mehr zu suchen. Dennoch wurden bis vor kurzem selbst mit moderner Rechenleistung nicht alle Möglichkeiten auf der Liste von Vléduts und Finn realisiert. Der betreffende Punkt ist hier die effiziente Navigation durch den bekannten chemischen Raum (d.h. durch die Millionen publizierten und patentierten Reaktionen), sodass sich Einzelreaktionen zu optimalen Synthesewegen kombinieren lassen, die von einer gewünschten Verbindung zu kommerziell erhältlichen Substraten zurückverfolgt werden können (Punkt [vii] der Liste). Der Leser könnte Einwände erheben und auf Hilfsprogramme verweisen, z.B. Auto Plan von Reaxys,[18] das scheinbar optimale Reaktionswege (bis zu zehn Stufen) liefert, indem bei jeder Stufe eine (oder wenige) Bestmöglichkeiten gewählt werden, oder SciPlanner von SciFinder,[19] das dem Anwender erlaubt, bei jeder Stufe eine optimale Wahl (abhängig von den Nutzerkriterien) zu treffen und sie letztlich in Syntheseplänen anzuordnen. Wir stellen jedoch fest, dass man, wenn man bei jeder Stufe die beste verfügbare Option wählt, nicht unbedingt zum Aufbau der besten Gesamtsequenz gelangt. Angenommen, wir führen eine retrosynthetische Suche auf der Basis der veröffentlichten Literatur durch und bewegen uns schrittweise von der Zielverbindung „rückwärts“, bis wir schließlich kommerziell erhältliche Substrate finden. Zur Veranschaulichung nehmen wir an, dass wir eine Stufe von der Zielverbindung entfernt („Syntheseabstand“ von der Zielverbindung $d=1$) zwei geeignete Reaktionen finden – eine mit der experimentell bestimmten Ausbeute 80%, die andere mit nur 60%. Wir wählen die Option mit 80%, aber alle folgenden Möglichkeiten erweisen sich als schlecht (z.B. ergibt die beste Option bei $d=2$ nur 40% Ausbeute). Die maximale Ausbeute, die wir daher durch Untersuchen dieses „Zweigs“ der Synthesemöglichkeiten erreichen können, beträgt $80\% \times 40\% = 32\%$. Hätten wir bei $d=1$ die 60%-Option gewählt, könnten wir

vielleicht zu einem zweiten Schritt gelangen, der eine Ausbeute von 90% bietet, sodass die Gesamtausbeute über den Reaktionsweg $60\% \times 90\% = 54\%$ betragen würde. Das hier besprochene Beispiel ist zwar trivial, aber die generelle Schlussfolgerung gilt für beliebige Reaktionswege: Die Optimalität (sei es in Form der Ausbeute oder der Atomökonomie oder einer anderen auf jeden Schritt angewendeten Messgröße) des Gesamtwegs ist erst bekannt, wenn der Reaktionsweg vollständig ist. Das bedeutet, dass zuerst komplette Suchen durchgeführt und vollständige Reaktionswege identifiziert werden müssen, und erst wenn alle vorliegen, können der oder die besten bestimmt werden.

Dies erschwert das Problem allerdings erheblich. Bei der Strategie „Das beste Ergebnis jeder Stufe“ (in der Informatik als „gierige“ Suche bezeichnet) wird in jedem Schritt nur eine Option ausgewählt, und die Gesamtzahl der Möglichkeiten, die zur Bestimmung des „besten“ Pfads aus L Stufen zu untersuchen und zu bewerten sind, ist lediglich die Summe der Zahl von Möglichkeiten, n_i, die in jedem Schritt i bewertet werden, $\Sigma_i^L n_i$; normalerweise liegt die Zahl der Optionen für jeden Schritt zwischen Dutzenden und ein paar tausend, und diese Suchen sind auch mit einem Desktop-Rechner machbar. Wenn wir jedoch unsere Suchen erweitern und m „beste“ Möglichkeiten bei jeder Stufe prüfen wollen, steigt die Zahl der durchzuführenden Synthesebewertungen exponentiell mit m^L. Solange wir in jeder Stufe nur wenige „beste“ Möglichkeiten berücksichtigen und die Pfade nicht zu lang sind (bei Auto Plan von Reaxys: $L < 11$), sind die Zahlen noch zu bewältigen, zumindest mit einem leistungsfähigen Computer (z.B. $3^{10} \approx 60\,000$ für $m = 3$ und $L = 10$). Wenn wir aber nach längeren optimalen Synthesen für strukturell komplexere Zielverbindungen suchen wollen, würde die Zahl der zu prüfenden Möglichkeiten rasch astronomisch werden. So wären für Synthesen mit $L = 30$ Stufen (z.B. die Synthese von (+)-Manzamin A durch Fukuyama oder die Synthesen von Strychnin durch Woodward, Shibasaki oder Magnus) und $m \leq 5$ $5^{30} \approx 10^{20}$ einzelne Syntheseschritte zu bewerten. Das Durchsuchen eines so gewaltigen Raums an Möglichkeiten erfordert ziemlich komplizierte Algorithmen sowie Datenstrukturen, die zur Speicherung chemischer Informationen am besten geeignet sind.

2.2. Vergleich von Reaktionsdatenbanken mit dem Network of Organic Chemistry (NOC)[3]

Generell können Sätze von Daten/Einträgen, zwischen denen paarweise Zusammenhänge/„Verknüpfungen“ bestehen, entweder als Liste dieser Verknüpfungen oder als Netzwerke dargestellt werden. Als Beispiel betrachten wir die kürzlich in Betrieb genommenen Schnellzugverbindungen in Polen. Der linke Teil der Abbildung 1a listet die einzelnen Züge zwischen den größeren Städten Polens auf, während der rechte Teil die gleiche Information als Netzwerk darstellt. Offenbar ist es viel leichter, anhand der Netzwerkdarstellung sofort zu bestimmen, dass eine Verbindung zwischen Stettin und Warschau existiert und dass Warschau das Zentrum dieses Eisenbahnnetzes ist. Diese Leichtigkeit der „visuellen“ Betrachtung ist ein eindrucksvoller Grund, Netzwerkdarstellungen in einem breiten Anwendungsbereich zu verwenden – von Karten über GPS-Navigatoren bis zur Analyse großer Datenmengen in der Soziologie[20] und Biologie.[21] Der zweite, noch wichtigere Grund ist, dass die Suche nach Verknüpfungen über Netzwerke – mit den in Abschnitt 2.3 zu besprechenden Algorithmen – rechnerisch effizienter ist als die Suche nach Verknüpfungen über Listen, wobei der Unterschied mit wachsender Größe des Datensatzes signifikan-

a)

Städte

Gdańsk	–	Warszawa
Gdańsk	–	Poznań
Szczecin	–	Poznań
Warszawa	–	Wrocław
Warszawa	–	Katowice
Warszawa	–	Kraków
Poznań	–	Wrocław
Poznań	–	Warszawa

b)

Reaktionen

1) A → D + E
2) B + C → F
3) C + G → F
4) E + F → H
5) F + I → G

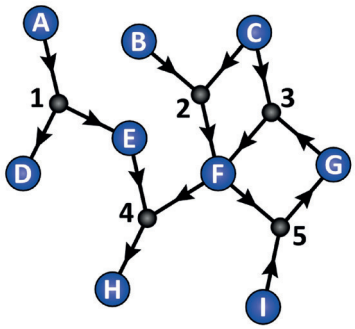


Abbildung 1. Listen und Netzwerke. a) Intercity-Verbindungen der polnischen Eisenbahn, dargestellt als Liste (links) und als Netzwerk/Karte (rechts). Das Netzwerk lässt unmittelbar erkennen, dass man von Stettin nach Warschau (über Posen) reisen kann. b) Liste chemischer Reaktionen (links) und ihre äquivalente Netzwerkdarstellung (rechts). Mit zwei Arten von Knoten (blau für Verbindungen, dunkel für Reaktionen) werden alle Beziehungen zwischen Substraten und Produkten erfasst. Würde man lediglich Pfeile von allen Substraten zu allen Produkten ziehen, gäbe es irreführende Verknüpfungen wie B→F und C→F; in Wirklichkeit muss B mit C zu F umgesetzt werden. Diese Art der Zwei-Knoten-Darstellung wird als Petri-Netz oder bipartiter Graph bezeichnet. Genehmigte Wiedergabe von Teil (b) aus Lit. [26a].

ter wird. In der Syntheseplanung werden große Datensätze für mehrstufige Verknüpfungen zwischen der Zielverbindung und den verfügbaren Substraten abgefragt, daher wollen wir Chemie als Netzwerk und nicht als Liste von Einträgen in den üblichen chemischen Datenbanken darstellen.

Zunächst ist es wichtig, die für chemische Reaktionen am besten geeignete Netzwerkdarstellung zu wählen. Unter Bezug auf Abbildung 1b betrachten wir eine Reaktion des Typs B + C→F. Wenn wir alle Zusammenhänge zwischen Substraten und Produkten (B→F, C→F) zeichnen, führen wir möglicherweise chemisch unsinnige Verknüpfungen in das Netzwerk ein. Handelt es sich bei der Reaktion, beispielsweise, um die Acylierung einer etwas komplizierteren Verbindung B, z. B. mit Acetylchlorid (C), dann würde die Verknüpfung C→F implizieren, dass die strukturell komplexe acylierte Verbindung F aus Acetylchlorid herstellbar ist. Zur Vermeidung derartiger Probleme und Erfassung aller relevanten chemischen Informationen wird die so genannte bipartite oder Petri-Netzdarstellung verwendet.[22] Sie enthält zwei Arten von Knoten, wovon eine für Substrat-/Produktmoleküle (blaue Kreise in Abbildung 1b) und die andere für Reaktionsabläufe (schwarze Kreise in Abbildung 1b; im betrachteten Beispiel der mit „2“ bezeichnete Knoten) steht. Man kann sich dann vorstellen, dass die Substanzen B und C in ein Reaktionsgefäß (Knoten „2“) gegeben werden und daraus als Produkt F hervorgeht.

Mit diesen Überlegungen lässt sich jeder Speicher mit chemischen Reaktionen in ein Netzwerk übertragen. Anfang der 2000er Jahre haben wir mit der Arbeit an dieser Übertragung begonnen, und heute enthält das Network of Organic Chemistry (NOC) etwa zehn Millionen Verbindungen und ähnlich viele Reaktionen, die sie verknüpfen. Das ist in jeder Hinsicht ein sehr großes Netzwerk (Abbildung 2a,b), etwa 1000-mal größer als ein menschliches Metabolom.[23] Obwohl dieses riesige „Universum“ der bekannten organischen Chemie von so vielen unabhängigen Chemikern geschaffen wurde, entwickelte es sich von Anfang an überraschend vorhersehbar und unveränderlich. Beispielsweise sind die Zahlen der Verbindungen mit einer bestimmten Anzahl „eingehender“ Verknüpfungen/Reaktionen, k_{in} (d. h. die Häufigkeit, mit der jede dieser Verbindungen als Reaktionsprodukt erhalten wurde), wie auch die Zahlen der Verbindungen mit einer bestimmten Anzahl „ausgehender“ Verknüpfungen/Reaktionen, k_{out} (d. h. die Häufigkeit, mit der jede dieser Verbindungen als Substrat einer Reaktion verwendet wurde), auf einer doppelt-logarithmischen Skala linear (Abbildung 2c). Diese mathematische Gesetzmäßigkeit verrät uns, dass das NOC die so genannte skalenfreie Architektur[24a] hat – ähnlich wie das WWW,[24b,c] das Internet,[24d] metabolische Netzwerke[24e] und sogar Gesellschaften.[24f] Diese Architektur ist charakterisiert durch das Vorliegen hoch verknüpfter Zentralmoleküle, über die der Großteil des Syntheseverkehrs stattfindet (zu einer eingehenden Diskussion und der Liste dieser Verbindungen siehe Lit. [3b]).

Entscheidend ist, dass das NOC *rasche* Suchen nach Synthesewegen ermöglicht, indem es eine spezielle Datenstruktur nutzt, die auf große Graphen/Netzwerke zugeschnitten ist, und eine, in der Verbindungen und auch Reaktionen mit gewünschten Eigenschaften (Molekülmassen, Löslichkeiten, Ausbeuten usw.) gekennzeichnet werden können, sodass der Nutzer während der Netzwerksuche Kriterien und/oder Bedingungen festlegen kann. Damit kommen wir zum Thema der Suchalgorithmen.

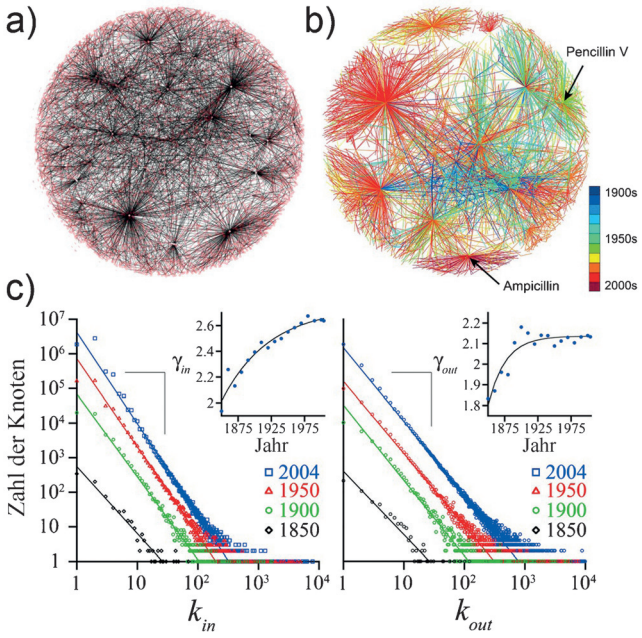


Abbildung 2. Struktur und Dynamik des Network of Chemistry (NOC). a) Kleines Fragment (≈5500 Knoten) des NOC; Einzelknoten stellen Verbindungen und Pfeile Reaktionen dar. Das ganze „Universum“ der bekannten organischen Reaktionen ist mehr als 1200-mal größer als das hier gezeigte Teilnetzwerk und etwa 1000-mal größer als das metabolische Netzwerk des Menschen.[24] b) Die Farbcodierung der Reaktionspfeile entspricht den Zeiten, in denen diese Reaktionen erstmals beschrieben wurden; sie verdeutlicht das explosionsartig steigende Interesse an bestimmten Gebieten der Chemie, z. B. die in den 1960er Jahren nach der ersten Totalsynthese einsetzende Syntheseaktivität um den Penicillin-V-Knoten. c) Die Graphiken zeigen, wie viele Verbindungen („Knoten“, vertikale Achse) im NOC eine bestimmte Zahl „ankommender“ (k_{in}, horizontale Achse der linken Graphik) oder „abgehender“ (k_{out}, horizontale Achse der rechten Graphik) Syntheseverknüpfungen haben. Die Linearität der Kurven auf der doppelt-logarithmischen Skala spricht für eine exponentielle Verteilung, $p(k) \propto k^{-\gamma}$, wie sie für skalenfreie Netzwerke charakteristisch ist (siehe Haupttext). Die Einschübe zeigen, dass sich die Exponenten γ der Verteilungen (d. h. die Steigung der doppelt-logarithmischen Kurven) mit der Zeit asymptotisch den Werten $\gamma_{in} = 2.67$ und $\gamma_{out} = 2.14$ und damit den für das gerichtete Netzwerk des WWW charakteristischen Werten (2.71 bzw. 2.1) nähern. Das bedeutet, das NOC und das WWW sind topologisch ähnlich. Genehmigte Wiedergabe der Abbildung aus Lit. [3a,25b].

2.3. Bewertungsfunktionen und Suche nach „optimalen“ Reaktionswegen

Zunächst wollen wir die Komplexität der zu lösenden Suchprobleme genauer abschätzen (vgl. Abschnitt 2.1). Abbildung 3a zeigt eine realistische Ansicht aus dem „Inneren“ des NOC – ein hoch verknüpftes Netzwerk mit vielen „Zweigen“, die aus jedem Knoten hervorgehen. Auch intuitiv lässt sich erfassen, dass die Zahl möglicher Synthesewege mit der Zahl der gemachten Schritte explosionsartig steigt. In

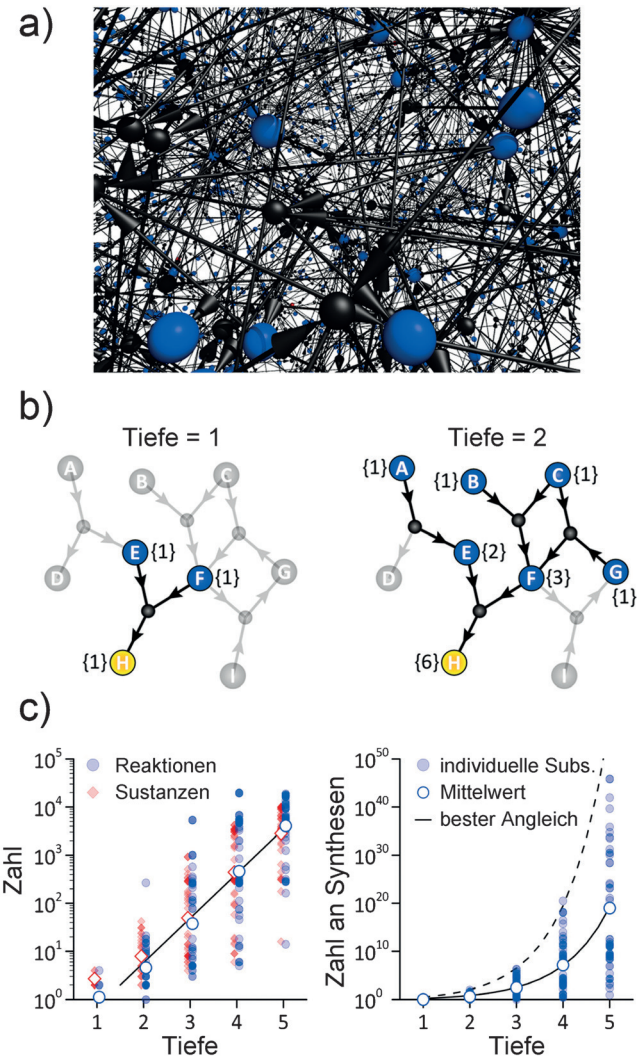


Abbildung 3. Komplexität des Network of Chemistry. a) Ein realistischer Blick vom Innern des NOC verdeutlicht seine hohe Konnektivität. b) Anzahl möglicher Synthesen bei zwei verschiedenen Tiefen *d* ausgehend von der Zielverbindung (gelber Knoten H). Bei der Tiefe = 1 gibt es nur eine mögliche Synthese mit den Substraten E und F. Bei der Tiefe = 2 gibt es sechs mögliche Synthesen der Zielverbindung. Dabei ist zu beachten, dass die Verbindung F auf drei Arten erhalten werden kann: 1) Sie kann gekauft und als Ausgangssubstanz eingesetzt werden; 2) sie kann aus den Substraten B und C oder 3) aus den Substraten C und G synthetisiert werden. Analog lässt sich Verbindung E auf zwei Arten erhalten. Abhängig davon, wie die Substanzen E und F erhalten werden, gibt es sechs mögliche Synthesen für das Produkt H. c) Ausgehend von Netzwerksuchen in der Umgebung von 51 verschiedenen Zielverbindungen (zu Einzelheiten siehe Lit. [25a]) nimmt die Zahl der für die Synthese jeder Zielverbindung relevanten Einzelreaktionen (blau) und Verbindungen (rot) mit zunehmendem Abstand von der Zielverbindung exponentiell mit $\approx (8.5)^d$ (links) zu. Die Zahl möglicher Synthesen (d. h. Plänen, die Einzelreaktionen kombinieren) steigt sogar noch schneller, in diesem Fall mit $\approx (1.4)^{(2.7)^d}$ (rechts, schwarze durchgezogene Kurve). Transparente Markierungen entsprechen den Ergebnissen für jede der 51 Verbindungen, offene Markierungen bezeichnen das geometrische Mittel dieser Daten; die durchgezogene Kurve ist die Anpassung an diese Daten durch kleinste Fehlerquadrate; die gestrichelte Linie ist eine Obergrenze für die Schätzungen. Genehmigte Wiedergabe der Teile (b) und (c) aus Lit. [25a].

früheren Arbeiten haben wir untersucht,[25a] wie die Zahl der Synthesemöglichkeiten vom Abstand zu einer gewünschten Zielverbindung (der so genannten Suchtiefe, Abbildung 3b) abhängt. Den in Abbildung 3c zusammengefassten Ergebnissen zufolge variieren die Zahlen zwar in Abhängigkeit von bestimmten Synthesezielen, aber innerhalb von nur fünf Stufen Entfernung vom Zielmolekül kann die durchschnittliche Zahl der zu berücksichtigenden Synthesewege 10^{16} erreichen.

Angesichts dieser Komplexität ist es unser Ziel, Synthesewege zu identifizieren, die eine gewünschte Verbindung aus kommerziell erhältlichen Substraten zugänglich machen, wobei die Gesamtbewertung (im Unterschied zur Bewertung jeder Stufe, vgl. Abschnitt 2.1) des Reaktionswegs optimiert wird. Die kommerziell erhältlichen Substrate sowie ihre aktuellen Preise je Masseneinheit[25a] stammen aus Lieferantenkatalogen; das NOC ist in seiner Standardform mit dem Sortiment von Sigma-Aldrich verknüpft, es kann aber auch leicht mit anderen Kataloge als Textdateien verbunden werden. Der „Punktwert" für jeden Syntheseweg kann von Reaktionsmerkmalen (z. B. Ausbeuten, Arbeitskosten für das Durchführen der Reaktionen) und/oder von den Eigenschaften der beteiligten Verbindungen (vor allem von den Preisen der Reaktanten, aber auch von Messgrößen, die sich aus dem Netzwerk ableiten, wie der Konnektivität innerhalb des Netzwerks, k_{in} und/oder k_{out}, siehe Abbildung 2c,d) abhängen. Wie in Abschnitt 2.2 abschließend besprochen wurde, gehen diese Eigenschaften einher mit den Knoten/Verbindungen und Reaktionen in dem Graph-Datenbankformat des NOC und ermöglichen es, verschiedene Arten von „Bewertungsfunktionen" aufzustellen, nach denen die Reaktionswege beurteilt werden. Im Folgenden beschreiben wir zwei allgemeine Arten dieser Funktionen.

2.3.1. *Kostenfunktionen*

Das Ziel ist hierbei, Reaktionswege zu finden, deren reale monetäre Kosten minimal sind. Das Bewertungsmaß für diese Art der Suche sind die Gesamtkosten eines Synthesewegs, C_{tot}, die sich als Summe der Kosten der Einzelreaktionen (einschließlich Arbeits- und Festkosten sowie Reinigungsverfahren) und den Kosten der kommerziell erhältlichen Ausgangsverbindungen (die von der mit dem NOC verknüpften Lieferantenliste abhängt) ausdrücken lassen. Mit der sinnvollen Näherung, dass die Arbeitkosten für die Durchführung jeder Reaktion pro Masseneinheit etwa konstant sind, $C°_{rxn}$, ist eine solche Funktion am einfachsten zu schreiben[25a] als $C_{tot} = C°_{rxn}N_{rxn} + \Sigma_i C_{sub}(i)$, wobei N_{rxn} die Zahl der Stufen/Reaktionen im Syntheseweg ist. Natürlich können auch ausführlichere Funktionen verwendet werden, in denen die Kosten der Substanzen anhand der experimentell erhaltenen Reaktionsausbeuten, γ, korrigiert sind, sodass die „Effizienz" der Einzelumwandlungen einfließt. Zudem ist $C°_{rxn}$ als praktisch wichtige Größe zu berücksichtigen, denn sie ermöglicht es zu spezifizieren, wie kostenintensiv Arbeit im Vergleich zu kommerziell erhältlichen Materialien ist. Arbeitet man beispielsweise in einer Ökonomie, in der die Arbeitskosten relativ günstig, Substrate aber recht teuer sind, sollte für $C°_{rxn}$ ein niedriger Wert angesetzt werden; in diesem

Fall begünstigt die Bewertungsfunktion längere Reaktionswege, die zu billigeren Substraten führen. Ist dagegen die Arbeit teuer (wie in Labors in den USA, Deutschland oder der Schweiz), sollte C°_{rxn} hoch angesetzt werden, sodass die Funktion kürzere Pfade begünstigt, dabei aber teurere Substrate nutzt.

2.3.2. Popularitätsfunktion

Eine praktische Suchanfrage könnte Synthesen gelten, die sehr populäre Substanzen nutzen, da diese Chemikalien im Allgemeinen leicht zu handhaben und bewährt sind. Im Netzwerkformalismus lässt sich die Popularität in der Synthese anhand der Konnektivität der Verbindung messen, die durch die Indizes k_{in} und/oder k_{out} quantifiziert ist (siehe Abschnitt 2.2 und Abbildung 2c,d). Auf der Basis dieser Indizes könnte die Bewertungsfunktion für die Popularität, P_{tot}, die Summe der inversen Konnektivitätsindizes, $\Sigma_i 1/k(i)$, minimieren.

2.3.3. Suchalgorithmen

Die genannten Beispiele für Bewertungsschemata lassen sich natürlich modifizieren, und es ist einfach, Kombinationen der obigen Funktionen zu definieren oder weitere Variable hinzuzufügen. Auf einige dieser Möglichkeiten kommen wir in Abschnitt 2.5 zurück, in dem die Syntheseoptimierung mit Bedingungen (Synthesis Optimization with Constraints, SOCS) besprochen wird. Unabhängig von der verwendeten Funktion benötigt man jedoch einen Algorithmus, der den Graph effizient durchläuft, sodass die zu bewertenden Reaktionswege aufgebaut werden können. Eine gute Möglichkeit für eine solche Traversierung ist eine Variante des so genannten Breitensuchalgorithmus (breadth-first search, BFS),[26] ein Algorithmus, der vom Syntheseziel ausgehend rekursiv auf dem NOC propagiert (vgl. Pseudocode in Abbildung 4a). Soll beispielsweise der Reaktionsweg mit den niedrigsten monetären Kosten gefunden werden (siehe Abschnitt 2.3.1), prüft der erste „Rückwärtsschritt“ des Algorithmus alle Reaktionen, die zur Zielverbindung führen, und berechnet für jede von ihnen die Minimalkosten. Diese Berechnung hängt wiederum von den Minimalkosten der zugehörigen Reaktionspartner ab, die gekauft oder synthetisiert werden können. Auf diese Weise setzt sich die Kostenrechnung rekursiv fort und bewegt sich von der Zielverbindung rückwärts, bis eine kritische, nutzerdefinierte Suchtiefe (d.h. die maximal erlaubte Länge des Synthesewegs) erreicht ist.

Die auf diese Weise erstellten Reaktionspfade müssen auf ihre synthetische Durchführbarkeit geprüft werden. Das ist ein wichtiger Punkt, denn das Netzwerk ist kein einfacher „Baum“, und es gibt die Möglichkeit von Schleifen und Verbindungszweigen. Die Abbildungen 4b und 4c zeigen ein anschauliches Beispiel: Die Reaktionswege sind zwar ähnlich, und beide haben kommerziell erhältliche Substrate als Endpunkte, aber nur der Syntheseweg in Abbildung 4c ist realisierbar. Der Reaktionsweg in Abbildung 4b ist nicht durchführbar, da die Zwischenstufen 2, 3 und 4 nicht aus Ausgangsverbindungen synthetisiert werden können (z.B., 2 erfordert 3, und 3 erfordert 2, beide sind nicht kommerziell erhältlich). Der Algorithmus zur Bewertung der Durchführbarkeit einer Synthese ist in den Hintergrundinformationen zu Lit. [25a] enthalten.

a)
```
MinCost(substance s, depth d)
· if s.cost(d) < 0  // substance not yet visited
· · if s.type == substrate
· · · s. cost(d) = s.purchase_price
· · else
· · · s. cost(d) = INF  // infinite cost
· · if d < d_max
· · · for each reaction r ∈ {incoming reactions of s}
· · · · if r.mrk(d) == 0 // reaction not currently being explored
· · · · · if r.cost(d) < 0 // reaction not yet visited
· · · · · · r.cost(d) = c°_rxn
· · · · · · r.mrk(d) = 1
· · · · · · for each substance u ∈ {reactants of r}
· · · · · · · r.cost(d) = r. cost(d) + MinCost (u, d + 1)
· · · · · · r.mrk(d) = 0
· · · · · if r.cost(d) < s.cost(d)
· · · · · · s.cost(d) = r.cost(d)
· return s. cost (d)
```

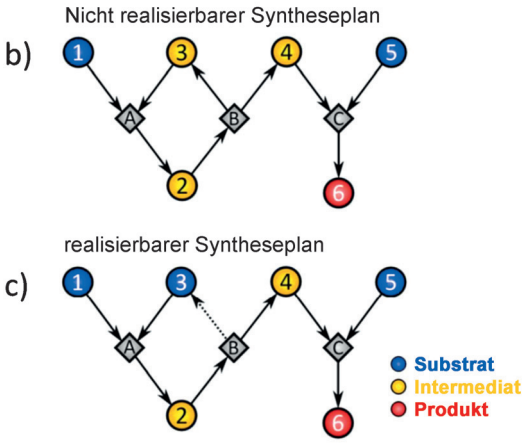


Abbildung 4. a) Pseudocode eines rekursiven Algorithmus für die Suche nach den kostengünstigsten Synthesewegen. b,c) Nicht durchführbare und durchführbare Synthesewege. Reaktionen, die sich mit kommerziell erhältlichen Substraten oder daraus zugänglichen Verbindungen durchführen lassen, werden als ***realisierbar*** bezeichnet; analog werden Verbindungen, die aus kommerziell erhältlichen Substraten synthetisiert werden können, als ***herstellbar*** bezeichnet. Das Beispiel in (b) ist ein nicht durchführbarer Syntheseplan. Keine der drei Reaktionen ist realisierbar, da die Zwischenstufen 2, 3 und 4 nicht herstellbar sind. Das Beispiel in (c) ist ein realisierbarer Syntheseplan. Reaktion A ist realisierbar, weil alle Reaktionspartner kommerziell erhältlich sind; demzufolge ist Substanz 2 herstellbar. Reaktion B ist realisierbar, weil ihr Reaktant herstellbar ist; demzufolge ist Substanz 4 herstellbar. Schließlich ist Reaktion C realisierbar, weil jeder ihrer Reaktanten entweder herstellbar (4) oder kommerziell erhältlich (5) ist. Demnach ist die Zielverbindung (6) herstellbar. Genehmigte Wiedergabe der adaptierten Abbildung aus Lit. [25a].

Die Suche liefert interessante und syntheserelevante Ergebnisse,[25a–c] wie das Beispiel der kostenoptimalen Synthese von Zolpidem zeigt (Abbildung 5). Bei hohen Arbeitskosten, $C_{rxn} = 7.5$, nutzt der Kostenalgorithmus relativ teure Substrate (speziell Propiolsäure), vollendet die Synthese aber in nur zwei Schritten über eine Dreikomponentenreaktion. Sind die Arbeitskosten hingegen niedrig ($C_{rxn} = 0.075$), umfasst die optimale Synthese sieben Stufen und geht von den einfachen und billigen Substraten 2-Aminopicolin und *p*-Methylacetophenon aus.

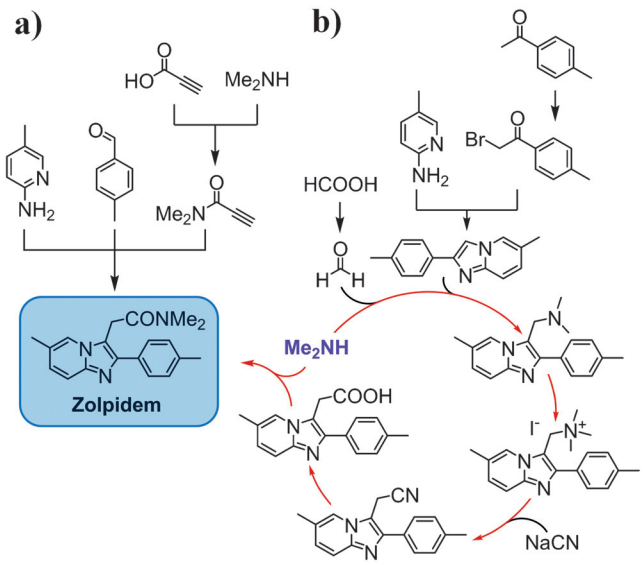


Abbildung 5. Zwei optimale Synthesen von Zolpidem bei relativ niedrigen (rechts, $C_{rxn} = 0.075$) und hohen Arbeitskosten (links, $C_{rxn} = 7.5$).

Bemerkenswerterweise wird einer der Bausteine (Dimethylamin, violett) nicht nur für die abschließende Amidierung genutzt, sondern auch zur Einführung einer passenden Abgangsgruppe, die für die Homologisierung des im vorausgehenden Schritt benötigen Mannich-Addukts benötigt wird. Die Identifizierung dieses Reaktionswegs durch herkömmliche Suchen, die immer nur einen Schritt zur Zeit berücksichtigen, wäre äußerst unwahrscheinlich. Der Grund dafür ist, dass bei einer rückwärts gerichteten Suche von der Zielverbindung aus der „linke“ und der „rechte“ Unterbaum im Syntheseplan divergieren, und die Wahrscheinlichkeit einen Weg zu finden, der die beiden „Zweige“ verknüpft, ist sehr gering.

Trotz dieser interessanten Beispiele sind die grundlegenden BFS- und DFS-Algorithmen in zwei wesentlichen Punkten begrenzt. Zum einen sind ihre Geschwindigkeiten für relativ einfache Verbindungen vielleicht ausreichend (Sekunden bis Minuten), aber bei größeren Synthesezielen und längeren Synthesen (einige Dutzend Stufen) kann die Zahl der zu betrachtenden Möglichkeiten auch mit verhältnismäßig großen Computerclustern nicht bearbeitet werden. Zum anderen fehlt den Datenstrukturen und Algorithmen eine gemeinsame, optimale Systemarchitektur, was die Durchführung der Aufgabenstellung häufig verlangsamt. Diese Anforderungen werden mit einer vereinheitlichten Software-Umgebung erfüllt, die wir Chematica genannt haben.

2.4. NOC-Suchen in Chematica

Chematica unterstützt verschiedene Arten von NOC-Suchen. Die einfachste, als Network Travel bezeichnete Suche (siehe Filme S1 in den Hintergrundinformationen) zeigt im NOC Reaktionen, die, abhängig von der Nutzerpräferenz, entweder zur interessierenden Verbindung führen (Abbildung 6a; hier Methylindol-3-carboxylat) oder von ihr

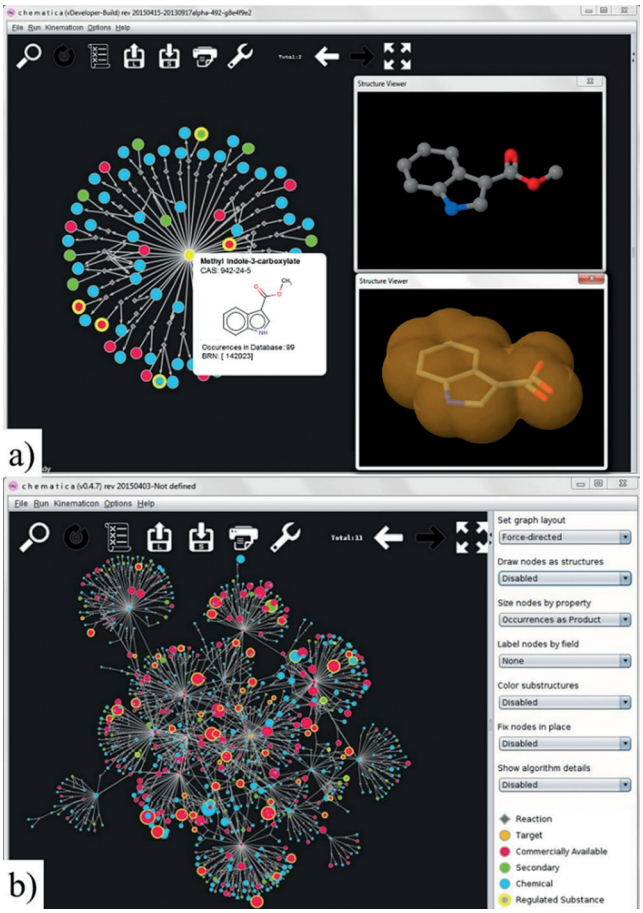


Abbildung 6. Basistraversierung im Netzwerk von Chematica. Der Screenshot in (a) zeigt als Beispiel Verbindungen, die direkt aus Methylindol-3-carboxylat hergestellt werden können. Die Knoten lassen sich als Molekülstrukturen und auch als 3D-Modelle darstellen, an denen eine grundlegende Modellierung möglich ist (die Unterfester zeigen hier z. B. die geometrische Optimierung und Connolly-Oberflächen). Blaue Knoten stehen für Produkte, grüne Knoten für Nebenprodukte, rote Knoten für kommerziell erhältliche Verbindungen, gelbe Ringe kennzeichnen reglementierte Substanzen (z. B. auf den DHS- oder EPA-Listen, die mit Chematica gekoppelt sind). Zum Netzwerk hinführende (statt „wegführender“) Reaktionen sowie weiteren Anzeigeoptionen siehe Abbildung S2. b) Jeder „Tochterknoten“ kann weiter expandiert werden (siehe Film S1). Innerhalb nur weniger Schritte wird das Netzwerk ziemlich komplex. Im hier gezeigten Anzeigemodus sind die Größen der Knoten proportional zu ihrer „Popularität in der Synthese“, d. h. der Konnektivität der Verbindungen im NOC. Einige der größten Knoten im Netzwerk bezeichnen Verbindungen, die in Zehntausenden von Synthesen verwendet werden!

weg (Abbildung S2a). Die kleinen rautenförmigen Reaktionsknoten enthalten Basisinformationen über die jeweilige Reaktion (z. B. Literaturquellen), grüne und blaue Knoten bezeichnen bekannte Verbindungen (wobei Nebenprodukte der Reaktion grün dargestellt sind), rote Knoten bezeichnen kommerziell erhältliche Chemikalien, und gelbe Ringe kennzeichnen regulierte und/oder toxische Substanzen. Jeder oder alle Molekülknoten können als zweidimensionale Molekülstrukturen (Abbildung 6 sowie Abbildung S2b) oder als 3D-Modelle dargestellt werden, an denen sich verschiedene Arten von Analysen (Geometrieoptimierungen, Connolly-

Oberflächen usw.) durchführen lassen (Unterfenster in Abbildung 6a). Außerdem war jeder Molekülknoten mit Informationen über seine „Synthesepopularität“ (d.h. die Konnektivität der Verbindung im NOC, quantifiziert durch die in Abschnitt 2.2 besprochenen Indizes k_{in} und k_{out}) verknüpft. Die Größen der Knoten lassen sich dann proportional zu verschiedenen Moleküleigenschaften darstellen, unter anderem die genannte „Synthesepopularität“ (siehe Abbildung 6b). Die zeitliche Entwicklung dieser Popularität ist ebenfalls abrufbar und spiegelt oft aktuell gefragte Gebiete der Chemie wider (vgl. Abbildung S3a). Jeder „Tochterknoten“ lässt sich zudem erweitern (siehe Film S1), was innerhalb weniger Stufen zu einem komplizierten Netz von Verknüpfungen führen kann. Darstellungen wie in Abbildung 6b verdeutlichen, warum manuelle Suchen ziemlich hoffnungslos sind und warum eingebaute automatische Suchvorgänge unter Verwendung der zuvor besprochenen Kosten- und/oder Popularitätsfunktionen wichtig sind, um optimale Synthesewege zu identifizieren.

Abbildung 7a veranschaulicht die Suche nach der kostengünstigsten Synthese des Krebsmedikaments Camptothecin mit bis zu $L=10$ Stufen und gleich hoch angesetzten Kosten für Arbeit und für die Reaktanten. Der so ermittelte Reaktionsweg ist mit Molekülstrukturen dargestellt, die sich zu Knotendarstellungen umschalten lassen (Abbildung 7b; Knotenfarben wie in Abbildung 6). Die bei den Knoten stehenden Zahlen sind Dollarpreise je Gramm Substrat (hier aus dem Sigma-Aldrich-Katalog). Durch Ändern des Kostenver-

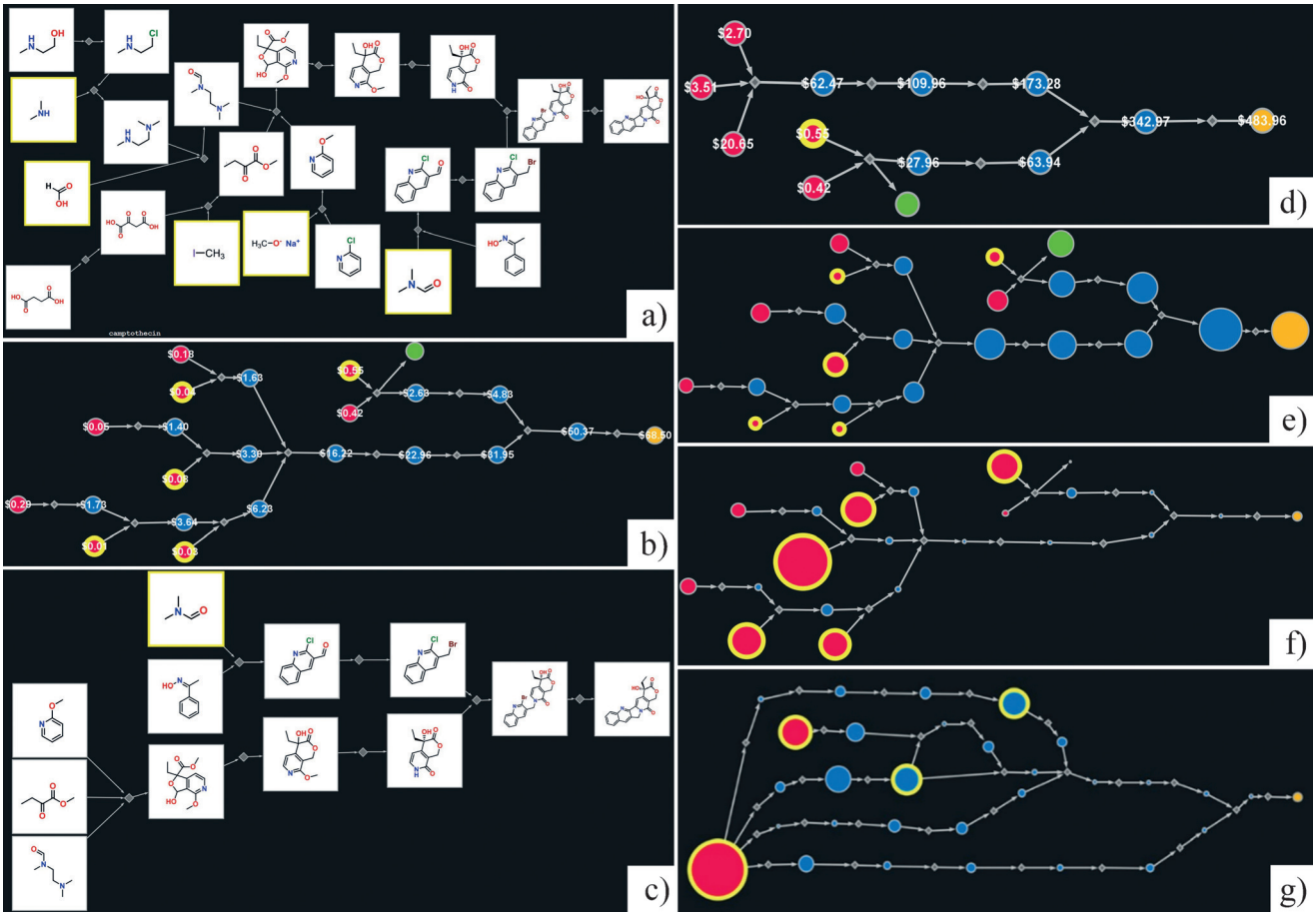


Abbildung 7. Chematica-Speicherauszüge der optimalen Synthesewege zu Camptothecin auf der Basis literaturbekannter Reaktionen. Bildschirmanzeigen der a) Molekülstrukturen und b) Knoten für den kostengünstigsten Syntheseweg mit bis zu zehn Stufen. In (b) bezeichnen die Zahlen in den Knoten die Dollarpreise je Gramm, berechnet aus den realistischen Preisen der Substrate von Sigma-Aldrich (rote Knoten). c,d) Kostenoptimierte Synthesewege, wobei die Arbeitskosten auf das 20fache der Chemikalienkosten festgesetzt wurden. Die Synthesewege sind zwar kürzer, gehen aber von teureren Substraten aus (um Arbeitskosten zu sparen); der Gesamtpreis der Synthese in (d) ist ca. sechsmal höher als in (b). e,f) Gleiche Synthesewege wie in (a,b), aber die Knotengrößen sind proportional zu den Molmassen der Verbindungen (e) bzw. ihrer Popularität (f). Diese Größendarstellung der Knoten ermöglicht die rasche Beurteilung, ob die Synthese von kleineren zu größeren und von gefragteren zu weniger gefragten Verbindungen verläuft, was tatsächlich erwünscht ist (anderenfalls würde man kleinere Syntheseziele aus größeren Substraten und verbreitete Chemikalien aus speziellen/unpopulären Substraten herstellen). Die gezeigten Beispiele erfüllen beide Bedingungen. Schließlich sind neben minimalen Kosten mehrere andere Suchoptionen verfügbar. So kann wie in (g) die Optimalität des Syntheseplans definiert werden durch die Popularität (d.h. die Konnektivität im Netzwerk) aller beteiligten Verbindungen, die „global“ über den gesamten Syntheseweg maximiert wird. Mit dieser Art der Suche werden Synthesewege identifiziert, die auf robusten (d.h. populären) Reaktionen beruhen, auch wenn minimale Dollarkosten nicht unbedingt typisch für sie sind. Farbcode der Knoten: blau = Zwischenverbindungen, grün = Nebenprodukte, rot = kommerziell erhältlich, gelbe Ringe = reglementierte Substanzen. Detaillierte Informationen zur Reaktion sind durch Klicken auf die rautenförmigen Reaktionsknoten erhältlich.

hältnisses Arbeit:Chemikalien werden im Allgemeinen völlig andere optimale Reaktionswege erhalten. Setzt man in diesem Beispiel die Arbeitskosten mit dem 20fachen der Substratkosten an, dann sind kürzere Synthesewege, die aber von teureren Substraten ausgehen, erheblich begünstigt; Abbildung 7c zeigt die zugehörigen Verbindungen in der Netzwerkdarstellung und Abbildung 7d die Kostenkalkulation.

Für die bisher betrachteten, relativ kurzen Reaktionswege waren die herkömmlichen BFS-Suchen, die an die Syntheseplanung adaptiert (vgl. Abbildung 4) und mit der effizienten Graph-Datenbankstruktur des zugrundeliegenden NOC kombiniert wurden, schnell genug und lieferten innerhalb von Sekunden Antworten. Die gleichen Programme reichen jedoch nicht aus, wenn nach optimalen Synthesen von komplizierten Zielverbindungen gesucht wird, für die es eine riesige Zahl von Möglichkeiten gibt. Da die gleiche Einschränkung auch bei anderen häufig verwendeten Suchalgorithmen (Tiefensuche,[27a] Dijkstra[27b] usw.) auftreten würde, könnte das Problem unüberwindbar scheinen, lässt sich aber in Anlehnung an Schachspielprogramme lösen. Eine algorithmisch effiziente Strategie dieser Programme besteht darin, eine zuvor berechnete „Bibliothek aus Endspielen" vorzuhalten, sodass der Möglichkeitenbaum nicht vollständig expandiert werden muss, wenn das Programm potenzielle Züge berechnet, sondern bereits eine optimale Lösung zur Verfügung hat, wenn es auf eins der bekannten Endspiele trifft (d.h., das Programm muss die Analyse nur durchführen, bis es auf dieses Endspiel trifft). In unserem Fall berechnet Chematica zuerst die optimalen, relativ kurzen Reaktionswege mit M Stufen zu allen Verbindungen im NOC. Wird anschließend nach einem langen, N-stufigen Syntheseweg ($N > M$) für eine bestimmte Verbindung gefragt, muss das Programm nur Suchen bis zu einer Tiefe von $N-M$ durchführen, während die Endwege der Länge M bereits verfügbar sind, was die Suche faktisch beschleunigt. Die algorithmischen Daten sind zwar recht kompliziert, aber durch eine überlegte Wahl von M kann die Gesamtsuche um mehrere Größenordnungen beschleunigt werden.

Mit diesem Algorithmus wurden auch sehr lange Reaktionswege zu sehr komplizierten Verbindungen innerhalb weniger Sekunden identifiziert. Ein Beispiel dafür zeigt Abbildung 8a, wonach das Programm die kostengünstigste, bis zu 50 Stufen lange Synthese von Paclitaxel (Taxol) findet.[28a,b] Die gesamte Suche, die 345 Millionen Sequenzen möglicher Schritte und über 400 Millionen Kombinationen der beteiligten Substanzen berücksichtigte, wurde an einem Desktop-Computer mit Vierkernprozessor durchgeführt und dauerte nur sieben Sekunden (siehe Echtzeitfilm S2). Wir möchten nochmals betonen, dass bei dieser und allen anderen Suchen des NOC die Durchführbarkeit des Synthesewegs außer Frage steht, da alle einzelnen Reaktionsschritte experimentell durchgeführt und wurden und veröffentlicht sind. Der Algorithmus setzt diese Einzelreaktionen lediglich zu einem optimalen Gesamtweg zusammen. In einigen Fällen spiegeln die optimalen Reaktionswege die Totalsynthesen durch einzelne Arbeitsgruppen wider. So folgt die Synthese von Taxol in Abbildung 8a in weiten Teilen der von Danishefsky entwickelten Strategie (rote Pfeile), aber der Algorithmus findet kostengünstigere Alternativen für die Synthesen von Ausgangsverbindungen wie dem Wieland-Miescher-Keton,[28c] 2-Methylcyclohexandion[28d] sowie den Silylierungs- und Mesylierungsreagentien, die zwar einfach, aber recht teuer sind. Allgemeiner ist jedoch das Ergebnis, dass die optimierten Synthesen aus den Methoden verschiedener Forschungsgruppen „zusammengesetzt" sind, die manchmal an der gleichen Zielverbindung arbeiten, manchmal aber ein völlig anderes Thema haben. Das lässt die kostenoptimierte Route zu Vardenafil (Wirkstoff zur Behandlung der erektilen Dysfunktion) erkennen, die zwischen 1952 (Alkylierung von Benzamid mit Ethylbromid) und 2008 (Sulfonylierung der Vardenafil-Vorstufe) veröffentlichte Reaktionen umfasst (Abbildung 8d,e). Den Zeitraum und die „Struktur" des Synthesewegs verdeutlichen die farbigen Reaktionspfeile und/oder die Angabe der Jahreszahlen (Abbildung 8b,d).

2.5. Syntheseoptimierung mit Bedingungen (SOCS)

Monetäre Kosten oder die Popularität der beteiligten Verbindungen sind nur zwei von mehreren Faktoren, die im Verlauf einer Syntheseplanung berücksichtigt werden können. Beispielsweise könnte es auch erwünscht sein, eine Syntheseroute zu planen, die nicht nur die wirtschaftlichste ist, sondern auch über eine bestimmte gewünschte Zwischenverbindung verläuft (z.B. eine leicht zugängliche, gefragte Verbindung) oder alle Substrate und Zwischenstufen vermeidet, die reglementiert und/oder toxisch sind oder nur wasserlösliche Substanzen verwendet (für umweltfreundliche Chemie).[29] Solche Vorgaben werden auf die Auswertung von außerordentlich vielen (wieder Milliarden) möglichen Synthesewegen übertragen, wobei mehrere Optimierungskriterien/Bedingungen *gleichzeitig* angewendet werden. Eine solche Syntheseoptimierung mit Bedingungen (Synthetic Optimization with Constraints, SOCS) liegt ohne Zweifel außerhalb des menschlichen Erkenntnisvermögens, sowohl was die Zahl der zu berücksichtigenden Möglichkeiten angeht als auch hinsichtlich der komplexen logischen Verknüpfungen, die, wenn sie manuell durchgeführt würden, eine Querreferenz von Synthesemachbarkeit, Kostenkatalogen, Toxizitätsdaten, Listen reglementierter Substanzen oder Löslichkeitswerten bedingen würden. Computer können dagegen mit Leichtigkeit mehrere logische Operationen (ORs, ANDs, IFs) an verschiedenen chemischen Kriterien/Bedingungen unter Aufwand von nur wenigen zusätzlichen Codezeilen durchführen.

Neben dem bereits besprochenen Kosten/Arbeit-Parameter unterstützt das SOCS-Schema von Chematica verschiedene Arten von Bedingungen, darunter die maximale Zahl der Reaktionsprodukte, die Zeitspanne der berücksichtigten Reaktionen, die Löslichkeit der beteiligten Substanzen, die Umgehung nutzerdefinierter Zwischenstufen oder die Vermeidung jeglicher reglementierter (d.h. giftiger oder gefährlicher) Chemikalien (Einzelheiten siehe Hintergrundinformationen, Abschnitt S4).

Abbildung 9 zeigt einen Vergleich der kostenoptimierten Synthese des Antikonvulsivums Gabapentin ohne weitere Vorgaben mit der Synthese des gleichen Wirkstoffs unter der

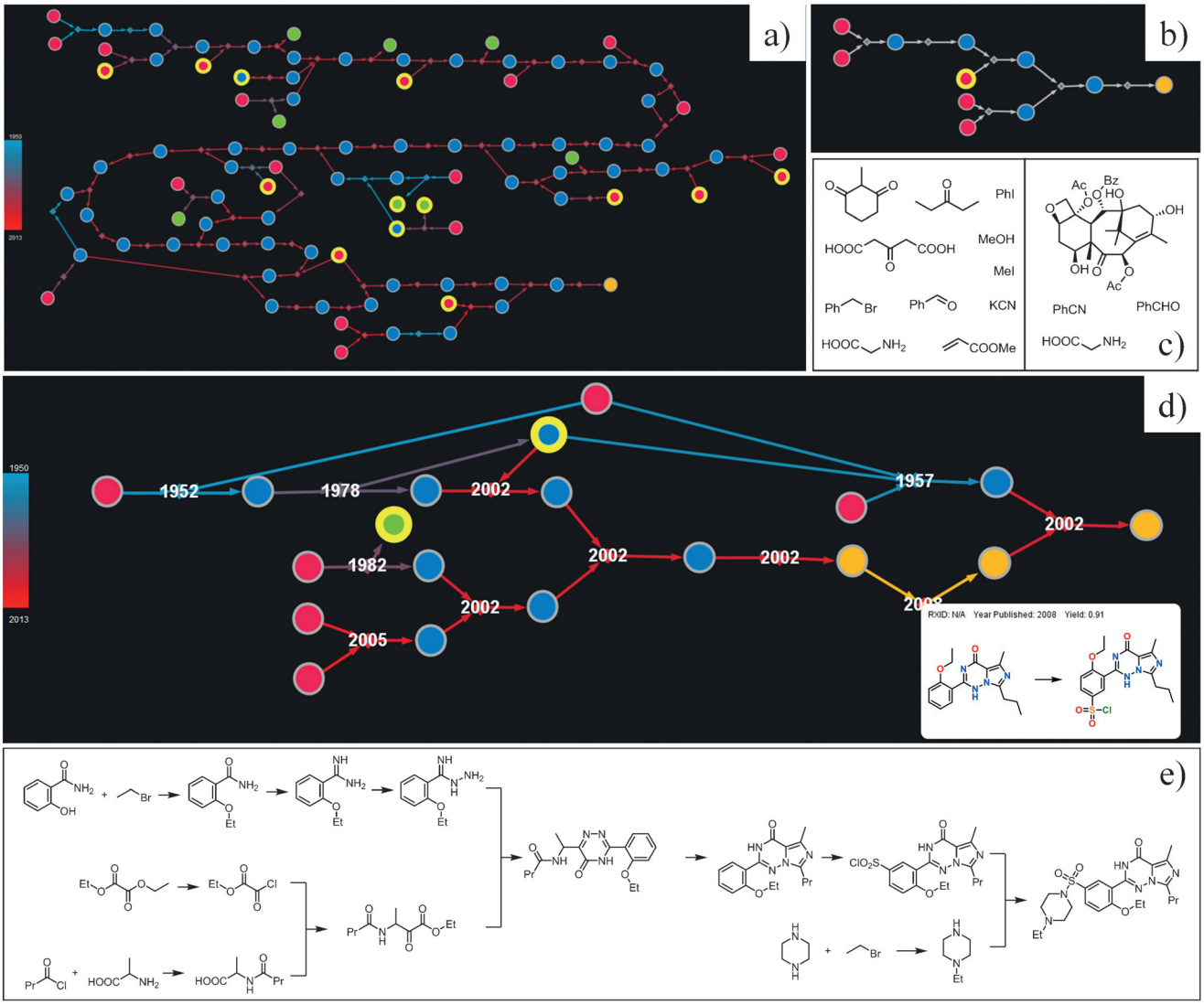


Abbildung 8. Kostenoptmierte Synthesen von Paclitaxel (Taxol) und Vardenafil. a) Knotendarstellung der kostengünstigsten, auf $L = 50$ Stufen begrenzten Synthese von Paclitaxel (vgl. Film S2). Die Farben der Reaktionspfeile entsprechen den Jahren, in denen eine bestimmte Reaktion beschrieben wurde. Der gezeigte Reaktionsweg (rote Pfeile) basiert größtenteils auf der 1995 von Danishefsky beschriebenen Synthese.[28a,b] b,c) Eine Erhöhung der Arbeitskosten um den Faktor zehn verteuert lange Synthesen sehr stark, daher geht die kostengünstigste Lösung, die der Algorithmus findet, von natürlich vorkommendem und kommerziell erhältlichem Taxan aus (Baccatin III; Kosten bei Sigma-Aldrich: mehrere tausend Dollar je Gramm). Diese Route erinnert an die erste industrielle Herstellungsmethode für Paclitaxel[28e] und wurde später weiterentwickelt.[28f] c) Sätze kommerziell erhältlicher Ausgangsverbindungen für die Synthesen von Paclitaxel entsprechend (a) und aus 10-Desacetylbaccatin III wie in (b). Für Schutzgruppen verwendete Reagentien wurden der Übersicht halber weggelassen. d,e) Kostenoptimierter Syntheseweg zu Vardenafil. Der als optimal identifizierte Reaktionsweg umfasst Reaktionen aus dem Zeitraum von 1952 bis 2008. Zu beachten ist die kostensparende Verwendung des gleichen Reagens in zwei verschiedenen Schritten (Ethylbromid, roter Knoten oben Mitte). Knotenfarben wie in Abbildung 7.

Bedingung, dass keine toxischen/reglementierten Substanzen (Knoten in gelben Ringen) verwendet werden. Ohne Einschränkung durch reglementierte Chemikalien geht der kosteneffizienteste Reaktionsweg (Abbildung 9a) von Cyclohexanon aus, das in großen Mengen zur Verfügung steht, aber reglementiert ist. Es reagiert mit Dimethylmalonat unter Knoevenagel-Kondensation zum ungesättigten Diester (Schritt a), der dann mit dem billigen, aber äußerst giftigen Kaliumcyanid umgesetzt wird. Nach einer Michael-Addition (Schritt b) folgen die nickelkatalysierte Reduktion des Nitrils zum primären Amin mit gasförmigem Wasserstoff (Explosionsrisiko) und die Aminolyse des Esters zum Spiroamid (Schritt c). Die Synthese der Zielverbindung endet mit der Decarboxylierung und Hydrolyse des so erhaltenen Amids (Schritt d). Wenn der Algorithmus den wirtschaftlichsten Reaktionsweg finden und dabei jegliche reglementierten Chemikalien vermeiden soll, schlägt er den Syntheseplan in Abbildung 9b vor, der keine reglementierten (Cyclohexanon), giftigen (Cyanide) oder explosiven Chemikalien (Wasserstoff) nutzt. Die Synthese mit einer Zusatzbedingung ist natürlich teurer (um etwa 40%) als die Synthese ohne Vorgaben – tatsächlich zieht jede Art von Bedingung stets einen höheren Preis nach sich.

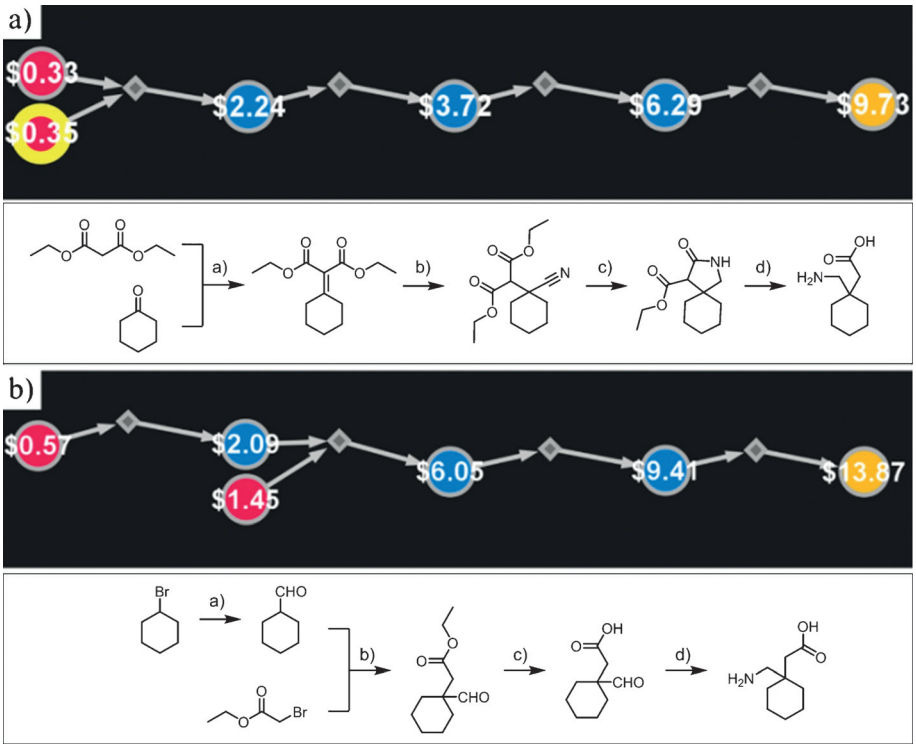


Abbildung 9. Kostenoptimierte Synthese des Antikonvulsivums Gabapentin a) ohne Zusatzbedingungen und b) mit der Bedingung, dass keine giftigen/reglementierten Substanzen (in gelben Ringen) verwendet werden. Zu den verwendeten Reagentien siehe Haupttext. Zur Vermeidung reglementierter Verbindungen im Syntheseweg wurden drei behördliche Datenbanken verwendet (Australia Group, Department of Homeland Security, Environmental Protection Agency List of Regulated Chemicals). Die Preise je Gramm wurden anhand der aktuellen Katalogpreise von Sigma-Aldrich angesetzt.

Eine weitere Reihe von Beispielen für die SOCS sind die unter verschiedenen Bedingungen optimierten Synthesen von Ketoprofen (Abbildung 10). Die durch rote Reaktionsknoten gekennzeichnete, kostenoptimierte Route beschränkt sich auf Reaktionen, die nur ein Produkt ergeben; da Nebenprodukte in der Industrie unerwünscht sind, hofft man, mit dieser Vorgabe einen industriell relevanten Syntheseplan zu erhalten. Der vorgeschlagene Plan umgeht einige, in anderen Synthesewegen häufig auftretende Zwischenstufen (z. B. Benzaldehyd, Ethylbenzophenonpropionat und Benzoylchlorid) und gleicht einer industriellen, von Rhône-Poulenc patentierten[30b] Synthesemethode für Ketoprofen.[30a] Die grün markierte kostenoptimierte Synthese erlaubt mehrere Produkte, verbietet aber reglementierte Substanzen. Der vorgeschlagene konvergente Syntheseweg beginnt mit der Herstellung von α-Bromacrylat durch DMSO-vermittelte Dehydrohalogenierung von α,β-Dibrompropanoat[30c] und von 3-Iodbenzophenon durch lösungsmittelfreie Iodierung von

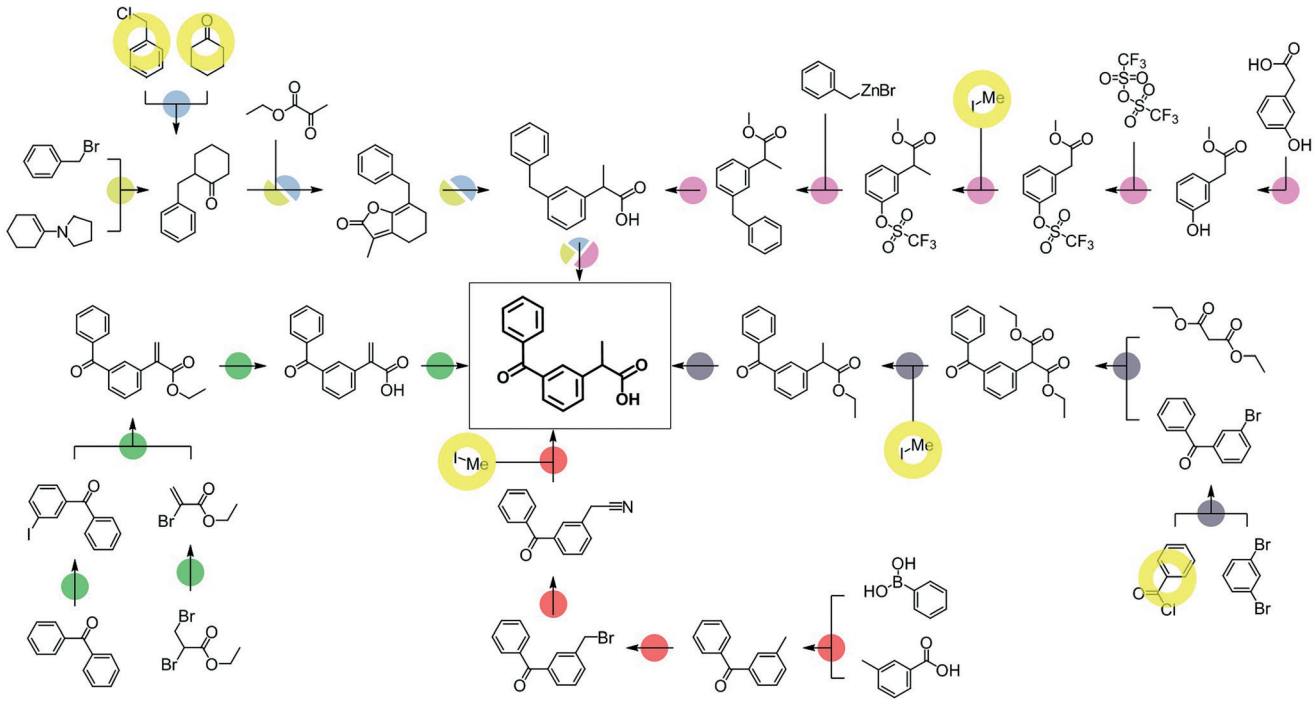


Abbildung 10. Syntheseoptimierung mit Bedingungen. Kostenoptimierte Synthesen von Ketoprofen mit verschiedenen Zusatzbedingungen; jede wurde von Chematica innerhalb weniger Sekunden geplant (zu Einzelheiten siehe Haupttext). Gelbe Ringe bezeichnen reglementierte Verbindungen.

Benzophenon[30d] (beide Reaktionen sind mild und selektiv). Die nachfolgende Kupplung einer aus dem α-Bromacrylat erhaltenen Organozinkverbindung mit dem Aryliodid[30e] liefert den ungesättigten Ester, der hydrolysiert und danach zu Ketoprofen hydriert wird. Bei dem rosa markierten Reaktionsweg handelt es sich um eine „moderne" Synthese, die auf neuere, nach 1998 publizierte Reaktionen beschränkt ist; mit dieser Bedingung identifiziert das Programm als kosteneffizienteste Synthese eine Methode,[30f] die mit Ausnahme der letzten Stufe[30g] in einer Arbeit beschrieben wurde. Die blau markierte Route[30h,g] ist kostenoptimiert für ein hohes Verhältnis Arbeit:Chemikalien und soll die Atomökonomie verbessern, sie geht aber von den reglementierten Verbindungen Benzylchlorid und Cyclohexanon aus. Mit der Zusatzbedingung, reglementierte Substanzen zu vermeiden, werden diese beiden unerwünschten Chemikalien durch nicht reglementiertes Benzylbromid und Cyclohexenylpyrrolidin ersetzt (grünlich-gelb markierte Route). Der grau-bräunlich markierte Weg ist schließlich ein Beispiel dafür, wie der Ausschluss nur weniger Verbindungen (z. B. Benzaldehyd, 3-Chlorbenzophenon und 3-Methylbenzophenon) eine gänzlich andere kostenoptimierte Route ergeben kann, die sich eng an die in Lit. [30i] anlehnt.

Die Kernaussage dieses Abschnitts ist, dass die eigentliche Bedeutung einer „optimalen Synthese" von den auferlegten Bedingungen abhängt. Bei der Suche nach solchen Synthesen im NOC haben Computer gegenüber Menschen den Vorteil, dass sie sehr viel schneller mehrere Informationsquellen kombinieren und gewünschte Suchkriterien einführen können. Zwar sind auch andere Arten von Suchen und Analysen über einen bekannten Reaktionsraum möglich und nützlich (siehe z. B. Hintergrundinformationen, Abschnitt S5 zur Simultanoptimierung bei mehreren Zielverbindungen[25a] und Abschnitt S6 zur Neuverdrahtung von Synthesenetzwerken durch den Austausch von Stufensequenzen durch Eintopf-„Abkürzungen"),[25b] aber definitionsgemäß können keine neuen Synthesestrategien und/oder Reaktionswege vorgeschlagen werden, die zu bisher noch nicht synthetisierten Zielverbindungen führen. Den Computern dieses De-novo-Design beizubringen, war die große Herausforderung der computergestützten Syntheseplanung, auf die wir im Folgenden eingehen werden. Netzwerksuchen, Bewertungsfunktionen und Bedingungen werden weiterhin genutzt, aber diesmal sind die Netzwerke nicht a priori bekannt und statisch, sondern expandieren kontinuierlich, wenn der Computer gesteuert durch die ihm eingegebenen chemischen Regeln Synthesemöglichkeiten wählt.

3. Computergestützte Planung von De-novo-Synthesen

3.1. Die großen Erwartungen

Vor genau einem halben Jahrhundert, 1965, initiierte ein Team aus Informatikern und Chemikern (Edward Feigenbaum, Bruce Buchanan, Joshua Lederberg und Carl Djerassi) an der Stanford University das so genannte Dendral-Projekt.[31] Das Ziel dieses Projekts war, 1) die Struktur organischer Verbindungen aus spektroskopischen Daten zu bestimmen und anschließend 2) Methoden der künstlichen Intelligenz für die Planung von Synthesewegen durch Computer zu nutzen. Auch wenn die zweite Aufgabe nie vollendet wurde, markierte diese Pionierleistung den Beginn der computergestützten Synthese. Bereits zwei Jahre vorher hatte Vléduts 1963 die Entwicklung einer Software vorgeschlagen, die es Chemikern ermöglichen würde, in Datenbanken gespeicherte Informationen und bestimmte „chemische Analogien" zu nutzen, um von einer gewünschten Zielverbindung „rückwärts" zu ihren Vorstufen zu navigieren.[17] Dieser Vorschlag lag zeitlich nur wenige Jahre vor Coreys wegweisender Arbeit von 1967, in der er die Philosophie der „retrosynthetischen Analyse" darlegte und ihre Hauptregeln festschrieb.[32] Da es zu diesem Thema ausgezeichnete Bücher[33a,b] und Übersichten[33c,d] gibt, erwähnen wir nur kurz, dass Coreys Beitrag wesentlich über den Vorschlag, sich auf einem wachsenden Baum der Synthesemöglichkeiten iterativ „rückwärts" zu bewegen, hinausging, und er in der Lage war, die Heuristik für die Wahl bestimmter synthetischer Bindungsbrüche zu identifizieren (d. h. Regeln, die vorschlagen, welche Bindungen „getrennt" werden sollten). Coreys Problemlösungsmethode revolutionierte das Gebiet, indem sie die Planung neuer Synthesewege systematischer machte, und sie lieferte außerdem Regeln, die – zumindest dachte man das damals – in die Rechner eingegeben werden konnten. In der Tat stellten Corey und Wipke bereits 1969 die erste Software für computergestützte Syntheseplanung mit der Bezeichnung OCSS (Organic Chemical Simulation of Synthesis) vor.[34] Das Projekt war von kurzer Dauer und zerfiel in zwei Richtungen: LHASA (Logic and Heuristics Applied to Synthetic Analysis)[5] stand unter der Leitung von Corey, und Wipke entwickelte SECS (Simulation and Evaluation of Chemical Synthesis).[35] LHASA hatte unter anderem Bedeutung als eins der ersten Retrosyntheseprogramme, das eine graphische Benutzeroberfläche zur Eingabe und Darstellung chemischer Strukturen verwendete. Technisch lässt es sich als halbempirische Software für die Retrosyntheseplanung klassifizieren, die auf verschiedenen Arten von heuristischen Transformationen in der dem Englischen ähnlichen Sprache CMTRN (Chemistry TRaNslator) sowie mehreren Planungsstrategien und einigen Daten zu Schutzgruppen beruht.[5b] Zu den Hauptnachteilen des Programms gehörten die begrenzte Fähigkeit, die Stereochemie zu berücksichtigen[36,33d] und die interaktive (schrittweise) anstelle der automatisierten Vorgehensweise beim Auffinden ganzer Synthesewege. Unseres Wissens wurde LHASA seit mehreren Jahren nicht aktiv weiterentwickelt,[5b] und die Internetseite lhasa.harvard.edu scheint inaktiv zu sein, während dieser Aufsatz verfasst wurde.

Nach der Entwicklung von LHASA wurden zahlreiche Anstrengungen zur Schaffung anderer Synthesplanungsprogramme unternommen. Da die Geschichte ihrer Entwicklung – und leider auch des endgültigen Verschwindens der meisten – sehr eloquent in dem ausgezeichneten Buch von Professor Philip Judson mit dem provokativen Titel „*Knowledge-based Expert Systems in Chemistry: Not Counting on Computers*" beschrieben ist,[1] fassen wir hier nur die wichtigsten Methoden kurz zusammen.

Das oben erwähnte, von Todd Wipke entwickelte Programm SECS[35] baute weitgehend auf LHASA auf, erweiterte aber dessen Wissensbasis. Obwohl es beträchtliche Unterstützung durch ein Konsortium pharmazeutischer Unternehmen aus der Schweiz und Deutschland erhielt, wurde es aus nicht ganz geklärten Gründen schließlich eingestellt.

SYNLMA[7] wurde von P. Y. Johnson et al. am Illinois Institute of Technology entwickelt und war wichtig, weil es die Wissensbasis von ihrer „Schlussfolgerungskomponente" trennte, die auf logischen, während der Retrosynthese anzuwendenden Operationen basiert. Leider lief das Programm in die „kombinatorische Explosion", wobei es übermäßig große Retrosynthesebäume generiert, die es nicht sinnvoll stutzen kann. Es verschwand bereits 1989.

SYNCHEM[8] und seine Folgeprogramme wurden an den Universitäten Stanford und Stony Brook bereits zur Zeit der ersten Veröffentlichung von LHASA entwickelt, aber erst 1977 bekannt. Das eigentlich Innovative dieser Methode – vor allem in Zeiten, in denen die moderne Datenverarbeitung in den Anfängen steckte – war der Versuch, (mit BFS-ähnlichen Suchen) ganze Retrosynthesebäume aufzubauen und zu untersuchen, die zu einigen Tausend eingespeicherten kommerziellen Produkten führten und einige Hundert von Fachleuten codierte (aber allgemeingültige) Umwandlungen nutzten. Leider wurden die Umwandlungen häufig erst codiert, nachdem ein Chemiker die Zielverbindung geprüft hatte, und es gab zusätzliche Schwierigkeiten mit der Anwendbarkeit der Umwandlungen auf bestimmte Verbindungen. Die entwickelten Strategien erwiesen sich als zu „kurzlebig" ohne eine ausreichend signifikante Berücksichtigung der Pfadhistorien.[33d] Wie bei allen früheren Programmen war die Stereochemie ein Hauptproblem, und die Regiochemie wurde nicht betrachtet. Die letzte Publikation zu SYNCHEM erschien 1998[8b] und beschrieb Arbeiten zur Parallelisierung des Codes. Danach schien sich SYNCHEM anderen Retrosyntheseprogrammen im Walhall der computergestützten Chemie angeschlossen zu haben.

Das Programm SYNGEN wurde von Jim Hendrickson und seiner Gruppe bei Brandeis entwickelt. Jim hat in diesem Bericht einen besonderen Platz, denn er war vielleicht der einzige erfahrene Kollege, der 2001 den damals zurückgekehrten Postdoktoranden B.G. ermutigte, an chemischen Netzwerken und der Retrosynthese zu arbeiten. Hendrickson entwickelte das Programm SYNGEN in den 1970er und 1980er Jahren[36] und konzentrierte sich darauf, mithilfe verschiedener Arten der Heuristik Retrosynthesebäume überschaubarer Größe und mit den besten, hoch konvergenten Synthesewegen zu identifizieren. Vereinfacht wurde die Synthese dadurch, dass der Schwerpunkt auf den Aufbau des Grundgerüsts gelegt und die Refunktionalisierung ignoriert wurde, da dies die kürzesten Synthesewege ergeben würde.[36d] Die empirische Beobachtung, wonach drei von vier Bindungen im Syntheseziel aus den Ausgangsverbindungen stammen, diente zur Ermittlung der Ausgangssubstanzen. Dadurch konnte der Computer mögliche Bindungssätze im Syntheseplan errechnen, der sich verfeinern ließ, indem der Syntheseraum nach Pfaden untersucht wurde, die zu Synthonen mit den unveränderten Bindungen führten. Da mit dem Programm Probleme auftraten, wenn die Funktionalisierung schon fehlte, bevor der ganze Reaktionsweg gefunden war, wurde das Zusatzprogramm FORWARD zur Wiedereinführung der Funktionalisierung in Syntheserichtung entwickelt. Diese Arbeiten wurden aber nie beendet.

Im Zusammenhang mit der nun folgenden Besprechung von konzeptionell anderen Ansätzen[37] wäre kein Artikel zur computergestützten Synthese vollständig, ohne die Beiträge von Ivar Ugi zu nennen, der die Konzepte der logikorientierten, auf chemischem Grundwissen beruhenden Syntheseplanung einführte, um die Durchführbarkeit nicht nur bekannter, sondern auch potenziell neuer Reaktionen zu evaluieren. In den 1980er und 1990er Jahren entwickelten Ugi et al. Programme wie IGOR und IGOR2,[38,31b] in denen die Moleküle als Bindungselektronen(BE)-Matrizen dargestellt wurden und die Reaktionen als R-Matrizen, die durch Subtraktion von Substrat- und Produkt-Matrizen erhalten wurden. Die interaktive (d. h. eine Reaktion nach der anderen) Analyse potenzieller Reaktionen beruhte auf den Anordnungen der in den Matrizen gespeicherten Valenzelektronen, und außerdem wurde die Wahl von möglichen gegenüber unsinnigen Reaktionen durch Berechnungen von Größen wie den Reaktionsenthalpien gesteuert. Mit IGOR konnten mehrere neue pericyclische Reaktionen und eine neue Umlagerung von α-Aminoalkylboranen in die entsprechenden β-Dialkylaminomonoalkylborane identifiziert werden, die später experimentell verifiziert wurden.[31b] Zur Planung mehrstufiger Synthesen wurde das Programm dagegen kaum genutzt, was vielleicht darin begründet ist, dass Rechenprozesse an BE- und R-Matrizen aufwändig sind und nur eine begrenzte Zahl von Reaktionen untersucht werden konnte. Das für die Entwicklung auf dem Gebiet recht bedauerliche Endergebnis ist, dass mit dem Ruhestand und danach dem Tod von Prof. Ugi 2005 die Arbeiten naturgemäß eingestellt wurden; heute wird die Abkürzung IGOR2 sogar in der Algorithmenentwicklung für Software auf Cabell-Basis für die „induktive Synthese von funktionellen Programmen", nicht von Verbindungen genutzt.[39]

Eine weitere bemerkenswerte Leistung aus den 1990er Jahren ist das von Johann Gasteiger et al. entwickelte Programm WODCA.[40] Ähnlich wie IGOR trennt sich diese Vorgehensweise vom Dogma der auf Synthonen basierenden retrosynthetischen Planung und den Methoden für funktionelle Gruppen. Stattdessen sind die Grundeigenschaften der Bindungen (z. B. Polarität, induktive Effekte, Resonanz, Polarisierbarkeit) die Basis für Vorschläge, welche Bindungen für retrosynthetische Brüche geeignet sind. Außerdem unterscheidet sich das Programm dadurch, dass es dem Nutzer eine bidirektionale Analyse ermöglicht, bei der verbreitete, im Computer gespeicherte Substrate mit der Zielverbindung abgeglichen und Routen vorgeschlagen werden, die den Chemiker zu diesen Zielverbindungen leiten. Da das Programm auf der Matrixschreibweise beruht, sind die Analysen von Verbindungen zwangsläufig langsamer als mit einer alphanumerischen Darstellung wie SMILES. Das hat jedoch kein Auswirkungen auf die Zielsetzung von WODCA, das per se kein Mittel zur automatischen Syntheseplanung ist, sondern den Chemiker bei der Syntheseplanung unterstützen soll.

Das Programm CHIRON,[9] das von Stephen Hanessian an der University of Montreal entwickelt wurde, nutzt ebenfalls den Gedanken, verfügbare Substrate mit den vom Nutzer spezifizierten Synthesezielen abzugleichen und ihn so zu Synthesen zu leiten, die die Überlappung maximieren. Charakteristisch für diese Methode ist, dass sie die Stereochemie während des Abgleichs berücksichtigt. Das Programm sucht hingegen nicht nach kompletten Retrosynthesebäumen und kann daher als interaktives Instrument klassifiziert werden, das ähnlich wie WODCA einen Chemiker bei der Syntheseplanung unterstützen soll. Die letzte Veröffentlichung zu CHIRON erschien 2005.[9b]

Schließlich beruht der von SymBioSys entwickelte ARChem Route Designer auf der Idee, Ähnlichkeiten zu nutzen.[41] Diese Methode weicht drastisch vom Konzept der expertencodierten Reaktionen ab, stattdessen stützt sie sich auf Reaktionsumwandlungen/„Reaktionsregeln“, die der Computer ähnlichen Literaturbeispielen entnommen hat (etwa 100000; es ist aber auch ein Satz aus etwa 50 manuell generierten Regeln eingegeben). Das Programm untersucht relativ kurze Reaktionsbäume vollständig, berücksichtigt aber keine Stereo- und/oder Regiochemie. Auf ähnliche Weise beruht IC_{SYNTH} von InfoChem auf Reaktionszentren, die verschiedenen Datenbanken entnommen wurden[42a] und dann anwenderkontrolliert genutzt werden, um Vorschläge für Synthesebäume aufzubauen. Diese Vorschläge basieren zwar auf „analogen“, an anderen Verbindungen durchgeführten Reaktionen (vgl. Abschnitt 3.2.2), sie können aber die Intuition eines Chemikers in der Praxis ergänzen und so als „Ideengenerator“ für die Synthese dienen.[42b] Anders als viele andere in diesem Abschnitt beschriebenen Programme sind ARChem und IC_{SYNTH} kommerziell erhältlich.

3.2. Fehleranalyse und wichtige Aufgaben

Trotz zahlreicher Versuche scheint keiner der oben besprochenen Ansätze zu einer Software geführt zu haben, die für die tägliche Arbeit erfahrener Chemiker in der organischen Synthese relevant gewesen wäre. Vielleicht wurde die Aufgabe zu früh in Angriff genommen, als Computer noch in den Anfängen steckten und viele der notwendigen Algorithmen einfach noch nicht ausreichend entwickelt waren. Ungeachtet der Gründe für diese Entwicklung ist es sehr zu bedauern, dass Computer seit Beginn der 2000er Jahre zwar viele Forschungsgebiete revolutioniert haben, in der Chemie die Fragestellungen der Syntheseplanung aber weitgehend aufgegeben wurde – und wenn Corey keinen Erfolg hatte, wer sollte es dann noch versuchen? Ein Teil des Problems könnte die starke Vereinfachung der Aufgabenstellung bei den ersten Methoden gewesen sein. Die unzureichende Rechenleistung in den 1970er und 1980er Jahren verlangte von den Forschern, „Abkürzungen“ zu nehmen, indem sie verschiedene Arten der Syntheseheuristik und Vereinfachungen einführten; rückblickend und mit den heutigen Kenntnissen wissen wir, dass sich bestimmte komplexe rechnerische Aufgaben nicht zu stark vereinfachen lassen. Ein gutes Beispiel dafür sind Schachprogramme wie Deep Blue, das menschliche Großmeister nicht dadurch schlägt, dass es wenige heuristische Regeln nutzt, sondern weil es alle praktikablen Möglichkeiten vollständig durchsucht. In gleichem Sinne kann auch eine Syntheseplanung nicht dadurch erfolgen, dass in den Computer einige hundert allgemeine Regeln eingegeben werden oder er in Analogie zu literaturbekannten Reaktionen arbeitet. Den Computer müssen eine enorme Zahl präziser chemischer Regeln und deren Anwendung beigebracht werden, und sie müssen in der Lage sein, Milliarden von Synthesemöglichkeiten zu untersuchen, bevor ihre wahre Leistung offenkundig wird. Zunächst wollen wir diese und einige andere Faktoren, die die Syntheseplanung zu einer so anspruchsvollen Aufgabe machen, erneut prüfen.

3.2.1. Die Bedeutung „seltener Ereignisse“

Ähnlich wie die polnische Grammatik ist die organische Synthese voll von Ausnahmen. Diese Aussage wird durch die Graphik im linken Teil von Abbildung 11 quantifiziert, für die wir aus mehr als 1.2 Millionen literaturbekannten Reaktionen einzelne Reaktionsarten und Reaktionszentren (d.h. Substrukturen, die sich während der Reaktionen ändern) ausge-

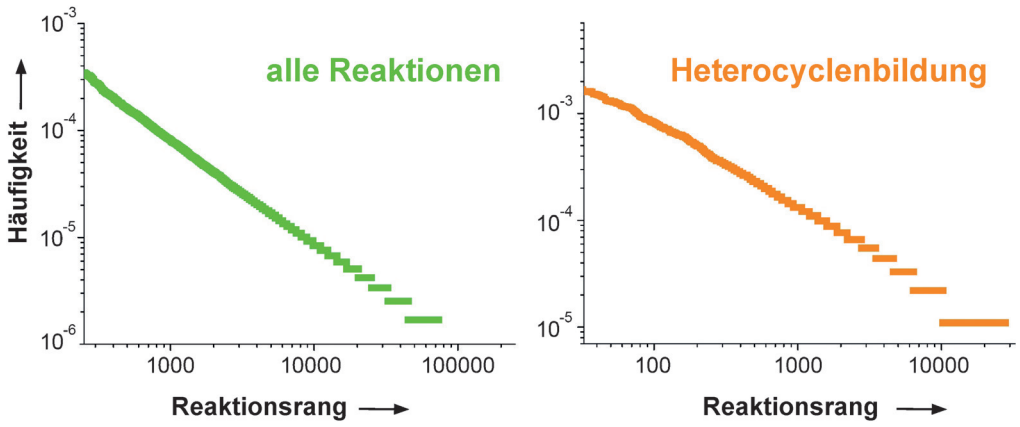


Abbildung 11. Rang-Häufigkeits-Diagramme verschiedener Reaktionsarten. Das linke Diagramm basiert auf der Analyse von 1.2 Millionen in der Literatur beschriebenen und zufällig aus dem NOC ausgewählten Reaktionen. Das rechte Diagramm bezieht sich auf Reaktionen, die aromatische Heterocyclen bilden. In beiden Fällen sind die Verteilungen Exponentialfunktionen (d.h. linear auf der doppelt-logarithmischen Skala), was die relative Bedeutung von seltener vorkommenden Reaktionen erkennen lässt. Reaktionsrang = 1 steht für die häufigste Reaktion, 2 für die zweithäufigste usw.

wählt, sie nach der Häufigkeit ihres Auftretens (d.h., die häufigste Reaktionsart erhielt Rang 1, die zweithäufigste Rang 2 usw.) geordnet und daraus ein Rang/Häufigkeits-Diagramm erstellt haben. Hierbei ist zu beachten, dass die Kurve im doppelt-logarithmischen Maßstab linear ist. Ähnliche Exponentialfunktionen traten auch „im Kleinen“ für spezielle chemische Teilgebiete wie die Bildung aromatischer Heterocyclen auf (90000 Reaktionen, rechts in Abbildung 11). Der entscheidende Punkt dieser Analysen ist, dass das Vorliegen einer Exponentialfunktion die Bedeutung der geringen Eintrittswahrscheinlichkeit anzeigt – so genannte „Schwarze Schwäne“, d.h., unvorhergesehene Ereignisse in dem der Verteilung zugrundeliegenden Prozess. Das bedeutet für uns, dass in der Chemie einige relativ seltene oder spezielle Reaktionen in einer bestimmten Synthese entscheidende Bedeutung haben können (siehe Beispiele in den Hintergrundinformationen, Abschnitt S7). Auch wenn das Problem der „Schwarzen Schwäne“ in der Synthese meist strukturell komplexe oder „exotische“ Zielverbindungen betrifft, müssen wir in der Lage sein, mit diesen Fällen umzugehen, wenn wir ein echtes Expertensystem aufbauen wollen (und kein „Spielzeug“, das nur einfache Moleküle bearbeitet). Das bedeutet leider auch, dass wir in den Computer nicht hunderte allgemeine Umwandlungen eingeben müssen, sondern zehntausende, einschließlich der speziellen Reaktionen.

3.2.2. Keine einfache Automation

Es wäre sehr verlockend, diese zehntausende Reaktionsarten automatisch aus den Datenbanken bekannter/veröffentlichter Reaktionen auszulesen. Tatsächlich sind wir das Problem zunächst auf diese Weise angegangen, und einige andere Softwarepakete nutzen das als ihre Wissensdatenbank.[41,42] Bei diesen Methoden extrahiert der Rechner zum einen die Gruppe von Atomen/Bindungen, die sich in jeder der in der Datenbank gespeicherten Reaktionen ändern, und fügt dieser Kernstruktur unter Umständen eine zuvor festgelegte Zahl von Nachbaratomen hinzu, um eine einzelne Syntheseumwandlung zu präzisieren. Leider treten bei einer solchen automatischen Datenauslese größere Probleme auf. Angenommen, wir haben das Reaktionszentrum der Friedel-Crafts-Reaktion ausgelesen: ein aromatisches Kohlenstoffatom mit daran gebundener Alkyl- oder Acyleinheit. Das Hauptproblem bei der Verwendung einer solchen Umwandlung in der Syntheseplanung ist, dass die Effekte anderer Substituenten (die der aromatische Ring möglicherweise trägt) nicht berücksichtigt werden, da sie per se nicht an dieser aromatischen Substitutionsreaktion beteiligt sind, bekanntlich aber ihr Resultat bestimmen. Man kann versuchen, das Reaktionszentrum um einige Atome „nach links“ und/oder „nach rechts“ zu erweitern, aber das ist immer willkürlich und in Wahrheit aussichtslos angesichts der vielen möglichen Substitutionsmuster, Substituenten und aromatischen Systeme, an denen die Friedel-Crafts-Reaktion durchführbar ist. Dies ist nur ein relativ einfaches Beispiel, es gibt viele weitere (siehe Hintergrundinformationen, Tabelle S1 in Abschnitt S8), bei denen die automatische Datenauslese mit Schwierigkeiten verbunden ist, die einfache Fehler in den zugrundeliegenden Datenbankeinträgen ebenso betreffen wie das Unvermögen, sterische und/oder elektronische Effekte, Reaktivitätskonflikte und die Stereo- und/oder Regiochemie der Reaktionen zu berücksichtigen (vgl. Abschnitt 3.2.4). Um die Sache richtig zu machen, müssen die Reaktionen von menschlichen Experten codiert werden, die genau festlegen, welche Substituenten erlaubt sind und welche nicht, und die sterische sowie elektronische Faktoren und anderes berücksichtigen. Dieser auf Experten basierende Zugang ist eigentlich keine Ausnahme, wenn Computer lernen sollen, komplizierte Aufgaben zu lösen: Deep Blue konnte Schachpositionen bewerten, weil das Programm die unglaubliche Zahl von 700000 Großmeisterpartien „gelernt“ hatte; Mathematica begann erst, seine Wunder in der formalen Mathematik zu vollbringen, nachdem es von Menschen eine bestimmte Zahl von Regeln, Heuristik und Algorithmen „gelernt“ hatte, deren Entwicklung zum Teil Jahre und deren Beschreibung Bücher beanspruchte (z.B. umfasst der Risch-Algorithmus für unbestimmte Integration in Lit. [43] allein mehr als 100 Seiten).

3.2.3. Die Bedeutung des molekularen Kontexts

Der vielleicht wichtigste Grund, warum die Retrosyntheseplanung so schwierig ist, ist, dass es nicht genügt festzustellen, ob eine bestimmte Bindung getrennt werden kann (alle *einzelnen* Bindungsarten können irgendwie getrennt werden), sondern vielmehr, ob sie in einer bestimmten Verbindung getrennt werden kann. Nach welchen Regeln die Bindungsbrüche auch erfolgen – Coreys Heuristik der strategischen Bindungsbrüche, Gasteigers auf Bindungseigenschaften basierende Wahl, die Präferenz von Bindungen mit maximalem Informationsinhalt[44] usw. –, sie sollten zusätzliche Informationen zum *Kontext* bieten, vor allem darüber, welche anderen Gruppen in der/den reagierenden Verbindung(en) während einer potenziellen Reaktion geschützt werden müssen und welche Gruppen unüberwindbare Reaktivitätskonflikte darstellen. Ein einfaches Beispiel ist die Synthese von Ketonen aus Weinreb-Nahm-Amiden, die mit dem Substrat in Abbildung 12a ablaufen kann, mit dem Substrat in Abbildung 12b dagegen nicht, weil es eine kreuzreaktivere Aldehydgruppe enthält, sodass ein Alkohol und kein Keton gebildet wird. Es gibt unzählige weitere, ähnliche Beispiele, daher werden die Retrosyntheseregeln allgemein in Form von Konditionalen ausgedrückt: „*Reaktion*

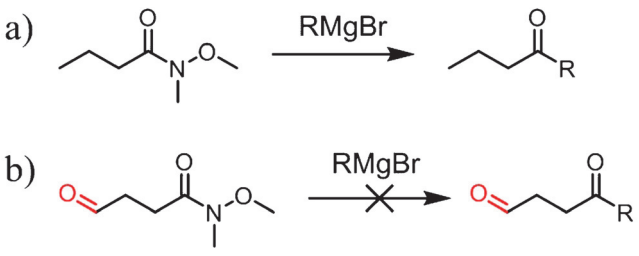


Abbildung 12. Bedeutung des molekularen Kontexts am Beispiel der Synthese von Ketonen aus Weinreb-Nahm-Amiden. a) Die Grignard-Reaktion führt direkt zu einem Keton. Durch Verwendung der Weinreb-Nahm-Amide wird eine weitere Addition verhindert. b) Da eine reaktivere (Aldehyd-)Gruppe vorliegt, würde die gleiche Grignard-Reaktion einen Alkohol statt eines Ketons ergeben.

X kann ablaufen, wenn die Gruppen Y,Z nicht vorhanden sind“ oder „*Reaktion X kann ablaufen, wenn die Gruppen Y,Z entsprechend geschützt sind*“. Die Bedeutung des Kontexts ist ähnlich wie bei einer Sprache, in der die gleichen Wörter abhängig vom übrigen Satz unterschiedliche Bedeutungen haben; wir konnten vor kurzem tatsächlich nachweisen, dass zwischen der organischen Synthese und der Linguistik einige formale Ähnlichkeiten bestehen (zu Einzelheiten siehe Lit. [44]).

3.2.4. Berücksichtigung von Stereochemie und Regiochemie

Der Überblick über die Retrosynthesprogramme in Abschnitt 3.1 hat gezeigt, dass der weitaus größte Teil dieser Programme die Stereo- oder Regiochemie nicht berücksichtigt. Das ist kein unabsichtliches Versäumnis, sondern ein inhärentes Problem der Datenstrukturen. In der Matrixschreibweise ist beispielsweise die Konnektivität der Atome einfach zu programmieren, die absolute Konfiguration dagegen nicht. Das Problem besteht auch bei den modernen SMILES[4]/SMARTS[45]-Notationen, bei denen sich die Konfigurationen einzelner Verbindungen leicht zuweisen lassen, ihre Änderungen während der Reaktion aber oft nur schwer zu verfolgen sind.

3.2.5. Das Fehlen genau definierter „Synthesepositionen“

Schachprogramme waren unter anderem deshalb so erfolgreich, weil sie auf der Basis der aktuellen Position, die durch die Anordnung der Figuren auf dem Schachbrett bestimmt wird, Vorhersagen zum Ausgang des Spiels machen konnten. In der Synthese ist die „Position“ ungenau definiert, auch wenn wir intuitiv fühlen, dass sie etwas mit der Komplexität der in einem bestimmten Schritt gebildeten Substrate zu tun haben muss. Anders als im Schach, wo die Geschichte vorheriger Züge belanglos ist, bringen Synthesepositionen aber auch „Kosten“ im Zusammenhang mit der Zahl bereits durchgeführter Schritte mit sich. Ein bevorzugtes Szenario ist, mit möglichst wenigen Synthesezügen an die „Position“ von einfachen Substraten zu gelangen.

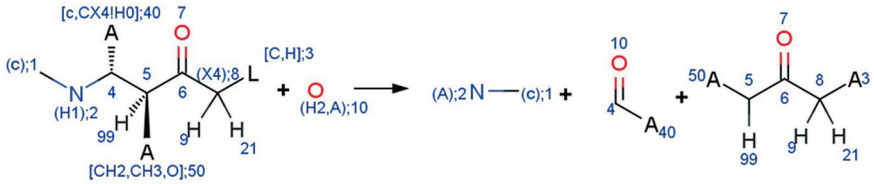
3.2.6. Größe des Suchraums und das Fehlen intelligenter Algorithmen

Die ermutigende Nachricht ist, dass die Zahl der in der retrosynthetischen Planung zu berücksichtigenden Möglichkeiten zwar nach wie vor sehr groß ist (bei langen Sequenzen in der Größenordnung 10^{30}–10^{50}), aber viel kleiner als die Zahl möglicher Schachspiele ($\approx 10^{230}$ für Partien mit 80 Zügen, Tabelle 1). Da Computer irgendwie Schach spielen, ist demnach die Annahme vernünftig, dass wir ihnen beizubringen könnten, wie sie die vom Maßstab her kleineren Syntheseaufgaben lösen. Zu den größten Hürden gehörte bisher jedoch das Fehlen geeigneter Algorithmen, mit denen der riesige Raum von Synthesemöglichkeiten untersucht werden kann. Die meisten Programme in Abschnitt 3.1, die überhaupt eine Expansion des Retrosynthesebaums versuchten, taten dies auf erschöpfende Weise oder über einfache BFS-Suchen, beides ist bei einem so riesigen Suchraum nicht durchführbar. Um die menschliche Syntheselogik nachzuempfinden, sollte sich der Algorithmus nicht einfach „vorwärts“ bewegen, sondern in der Lage sein, von aussichtlosen Pfaden umzukehren, lokale Alternativen zu untersuchen und, wenn diese versagen, zu völlig neuen Strategien zu wechseln (wie wir in Abschnitt 3.4.3 tatsächlich sehen werden).

3.3. Syntaurus

In den vergangenen zehn Jahren hat unsere Arbeitsgruppe Software entwickelt, die die meisten der oben geschilderten Schwierigkeiten berücksichtigt. Das aus diesen Arbeiten hervorgegangene Programm heißt Syntaurus, eine Verknüpfung von *Syn*thesis und *taurus*, dem lateinischen Wort für Bulle.

Das Herzstück von Syntaurus ist natürlich eine Kollektion von Reaktionsumwandlungen, die bei der Retrosynthese von gewünschten Verbindungen angewendet werden. Zunächst stützten wir uns auf maschinell aus Literaturbeispielen ausgelesene und in etwa 115000 einzelne Reaktionsklassen eingestufte Umwandlungen. Diese wissensbasierte Methode erwies sich zwar als unkompliziert (die gesamte Auslese aus dem NOC dauerte nur wenige Wochen), führte aber leider zu völlig bedeutungslosen Synthesevorhersagen. Wie aus Abschnitt 3.2.2 und Tabelle S1 (siehe Hintergrundinformationen, Abschnitt S8) hervorgeht, ist der wesentliche Grund für diesen Misserfolg, dass die maschinelle Auslese voraussetzt, dass die Syntheseplanung „durch Analogie“ erfolgen kann, selbst wenn das ursprüngliche, in der Literatur beschriebene Beispiel und ein bestimmtes, aktuell interessierendes Syntheseziel ganz verschiedene „Kontexte“ haben können (d. h. Schutzgruppenbedingungen, Inkompatibilitäten usw.; vgl. Abschnitt 3.2.3). Dieses Problem könnte bei manchen einfacheren Verbindungen vernachlässigbar sein, aber im Allgemeinen ist es ein verhängnisvoller Fehler der automatisierten Auslese, die chemisches Fachwissen nicht ersetzen kann. Das bedeutet aber, dass Abertausende von Reaktionsarten durch Experten manuell codiert und sorgfältig auf Fehler geprüft werden müssen. Aus diesem Grund dauerte die Erstellung der Wissensbasis von Syntaurus so viele Jahre; sie resultierte schließlich in rund 20000 von Experten codierten und geprüften Umwandlungen, die von einfachen S_N2-Reaktionen bis zu komplizierten Umlagerungen und Kaskadenreaktionen reichen. Ein typischer Eintrag in unserer Datenbank enthält das in der SMILES/SMARTS-Notation codierte Reaktionsmotiv (hier die Prolin-katalysierte Mannich-Reaktion) sowie eine Liste der zu schützenden Gruppen, eine Liste der mit der gegebenen Reaktion inkompatiblen Gruppen, die typischen/vorgeschlagenen Reaktionsbedingungen und einige Literaturzitate zur Reaktionsart (Abbildung 13). Damit diese und andere chemische „Regeln“ aussagefähige Ergebnisse während der Syntheseplanung liefern, müssen alle Arten von Atomen und Substituenten sorgfältig bedacht und definiert werden, sodass Stereochemie, Regiochemie ebenso wie sterische und elektronische Effekte angemessen berücksichtigt werden (siehe auch die Beispiele in den Hintergrundinformationen, Abschnitt 9).



rxn_id: 8382,
name: "Proline-catalyzed Mannich Reaction",
reaction_SMARTS:[c:1][NH:2][C@H:4]([c,CX4!H0:40])[C@:5]([#1:99])([CH2,CH3,O:50])[C:6](=[O:7])[CX4:8]([#1:9])([#1:21])[#6,#1:3].[OH2:10]>>[c:1][N:2].[*:40][C:4]=[O:10].[*:50][C:5]([#1:99])[C:6](=[O:7])[C:8]([#1:9])([#1:21])[*:3]"
products:["[c][NH][C@H]([c,CX4!H0])[C@]([#1])([CH2,CH3,O])[C](=[O])[CX4]([#1])([#1])[#6,#1]", "[OH2]"]
groups to protect: ["[#6][CH]=O", "[CX4,c][NH2]", "[CX4,c][NH][CX4,c]", "[#6]C([#6])=O"]
protection_conditions_code: ["NNB1", "EA12"]
incompatible_groups: ["[#6]O[OH]", "c[N+]#[N]", "[NX2]=[NX2]", "[#6]OO[#6]", "[#6]C(=[O])OC(=[O])[#6]", "[#6]N=C=[O,S]", "[#6][N+]#[C-]", "[#6]C(=O)[Cl,Br,I]", "[CX3]=[NX2][*!O]", "[#6]C(=[SX1])[#6]", "[#6][CH]=[SX1]", "[#6][SX3](=O)[OH]", "[CX4]1[O,N][CX4]1", "[#6]=[N+]=[N-]","[CX3]=[NX2][O]"]
typical reaction conditions: "(S)-proline. Solvent, e.g., DMSO",
general references: "DOI: 10.1021/ja001923x or DOI: 10.1021/cr0684016 or DOI: 10.1021/ja0174231 or DOI: 10.1016/S0040-4020(02)00516-1"

Abbildung 13. Prolin-katalysierte Mannich-Reaktion, wie sie in Syntaurus codiert ist. Die Informationsdatenbank des Programms umfasst rund 20000 von Fachleuten codierte Reaktionseinträge wie diesen. In diesem speziellen Beispiel spezifiziert der Untereintrag SMARTS beispielsweise, dass das Atom mit der Nummer 40 ein aromatisches Kohlenstoffatom („c“) oder eine Alkylgruppe mit mindestens einem Wasserstoffatom (codiert als „CX4!H0“) sein kann. Diese Bedingung schließt sperrige/verzweigte Substituenten aus,[53a] die bekanntlich zu niedrigeren Ausbeuten und Enantioselektivitäten führen. Außerdem enthält die Liste der erlaubten Substituenten an der Position 50 ein aliphatisches Kohlenstoffatom mit zwei oder drei Wasserstoffatomen oder einem Sauerstoffatom (codiert als „[CH2,CH3,O]“). Substituenten an dem Atom mit der Nummer 8 (mit zwei explizit codierten Wasserstoffatomen) beschränken mögliche Ergebnisse auf primäre oder Methylkohlenstoffatome. @-Zeichen an den Kohlenstoffatomen 4 und 5 ermöglichen es (mit den in Abschnitt 3.4.3 beschriebenen Operationen), die aktuelle Stereochemie der Reaktion beizubehalten und die Reaktionsprodukte auf eins der möglichen *syn*-Diastereomere zu beschränken. Durch die Art der Codierung wird auch die Regiochemie dieser Umwandlung bewahrt, wenn das Kohlenstoffatom 5 einen Sauerstoffsubstituenten trägt (z. B. in Hydroxyaceton), was mit experimentell erhaltenen Ergebnissen in Einklang ist.[53b] Schließlich wird der Substituent an der Amineinheit als aromatisches Kohlenstoffatom codiert („c“), weil in allen bekannten Prolin-katalysierten direkten Mannich-Reaktionen nur aromatische Amine verwendet werden. Weiterhin enthält der Reaktionseintrag die Liste von Gruppen, die beim Vorliegen in einem der Substrate geschützt werden müssen, den Code, der optimale Schutzgruppen spezifiziert (wenn das Schützen erforderlich ist), die Liste von Gruppen, die mit der Reaktion inkompatibel/kreuzreaktiv sind, die typischen Reaktionsbedingungen (die in bestimmten Synthesen genau angepasst sein können) sowie einige erläuternde Literaturzitate zu dieser Art von Reaktion.

Die Reaktionsregeln sind im SMILES/SMARTS-Format codiert, das seit einigen Jahren zu den chemischen Standardnotationen gehört und Moleküle und Reaktionen als alphanumerische Zeichenketten (Strings) darstellt. Das ist bei umfangreichen rechnerischen Aufgaben von entscheidender Bedeutung, da Rechenoperationen an Strings sehr viel schneller ablaufen als an Matrizen, denen beispielsweise .mol-Dateien zugrunde liegen (vgl. Hintergrundinformationen, Abschnitt S10). Allerdings gibt es bei der Verwendung von SMILES/SMARTS zwei größere Probleme. Das erste betrifft die Konfiguration, die in Einzelverbindungen durch die Symbole @ und @@ angegeben wird. Bei einfachen Reaktionen und mit einer Software wie RDKit[45b] kann die Stereochemie von Reaktionen, die mit allen Atomzuordnungen codiert sind, normalerweise richtig zugewiesen werden (z. B. steht die Änderung von @ zu @@ oder von @@ zu @ für eine Inversion der Konfiguration). Für kompliziertere Reaktionen mit mehreren Chiralitätszentren (vor allem proximalen) gab es aber keine ausreichenden Algorithmen. Das zweite Problem ist ähnlich und betrifft die Symbole // und /\ für die Bezeichnung der Regiochemie von Doppelbindungen in einzelnen Molekülen. Leider gab es keine Methoden, diese Symbole über Reaktionen in SMARTS zu verfolgen und der Reaktion die richtige Regiochemie zuzuordnen. Zur Lösung dieser Probleme entwickelten wir ein Software-Modul, das zwischen Retronen und Synthonen nicht nur die Informationen @, @@ und/oder //, /\ weitergibt, sondern auch nach den Massen der Substituenten geordnete Listen von Bindungen in Nachbarschaft zu jedem in der Umwandlung abgebildeten Atom; ist das spezielle Atom an der Bindungsbildung oder dem Bruch der Bindung beteiligt, dann haben seine Nachbarbindungen im Retron und im Synthon normalerweise unterschiedliche Reihenfolgen. Bei der Erstellung der Listen ist es wichtig, fehlende Wasserstoffatome zu ergänzen, was Mehrdeutigkeiten vermeidet und zu einer korrekten Reihenfolge beiträgt (in Grenzfällen kann es nötig sein, die nächstgelegenen Bindungen zu berücksichtigen). Schließlich wird bei der Ausführung einer Umwandlung ihre Stereo-/Regiochemie anhand der Übereinstimmung von Bindungslistenreihenfolge und Stereo-/Regiochemiesymbolen in der SMARTS-Notation bestimmt.

Es sollte aber erwähnt werden, dass bei einigen Reaktionsklassen auch die detaillierte Angabe des Reaktionszentrums/-motivs nicht ausreicht, um vorherzusagen, wo die Reaktion in einer bestimmten Verbindung abläuft. Ein gutes Beispiel hierfür sind aromatische Substitutionen, bei denen „entfernte“ Substituenten am aromatischen System entscheidende Auswirkungen auf die Reaktivität an anderen Positionen haben können. Bei einfachen Benzolsystemen könnte man zwar eventuell alle möglichen *ortho/meta/para*-Substitutionsmuster aufzählen, aber bei anderen aromatischen Systemen (kondensierte Ringe, heterocyclische Aromaten) ist es unmöglich, die Zahl der Möglichkeiten zu berücksichtigen. Weiß man hingegen, ob die Substitution elektrophil oder nukleophil abläuft, dann lässt sich die Reaktivität jedes Atoms auf der Basis seiner hohen/niedrigen Elektronendichte oder der Elektronendelo-

kalisierungsenergie ermitteln. Syntaurus berechnet ohne zeitliche Verzögerung die Delokalisierungsenergien je Atom (siehe Lit. [46] und Hintergrundinformationen, Abschnitt S11) von aromatischen Systemen und bestimmt mit diesen Werten, wo welche Substitutionen erlaubt sind (elektrophile und nukleophile Substitutionen sind erlaubt, wenn die Delokalisierungsenergie unter- bzw. oberhalb bestimmter Schwellenwerte liegt).

Schließlich wird eine Datenbank aus einigen Tausend „unmöglichen" Molekülfragmenten (z. B. solche, die die Bredt-Regeln verletzen) verwendet, um strukturell unsinnige Ergebnisse zu vermeiden, deren Ausmaß größer ist als die Reaktionszentren der Einzelumwandlungen (zu einer kleinen Auswahl vgl. Hintergrundinformationen, Tabelle S2 in Abschnitt S12).

3.3.1. Schrittweise Syntheseplanung im Allgemeinen und mit Syntaurus

Die einfachste Art der Syntheseplanung ist die, bei der der Nutzer in jedem Syntheseschritt eine Wahl trifft. Um diese Wahlen zu leiten, nutzten mehrere ältere Syntheseplanungsprogramme verschiedene Arten der Heuristik: strategische Bindungsbrüche in LHASA, die Maximierung der strukturellen Überlappung zwischen Substraten und dem Syntheseziel (CHIRON) oder die Maximierung der Ausbeuten von „ähnlichen", in der Literatur beschriebenen Reaktionen (ARChem). Allerdings ist äußerst unwahrscheinlich, dass eine einzige Heuristik universell anwendbar ist; bei polycyclischen Synthesezielen würde man vermutlich strategische Bindungsbrüche zur Bildung neuer Ringe favorisieren, und bei Zielverbindungen mit mehreren Chiralitätszentren sind stereoselektive Synthesen wahrscheinlich die erste Wahl. In Syntaurus haben wir eine skriptähnliche Sprache geschaffen, die vordefinierte Variablen zur Bewertung von Syntheseschritten nutzt. So definiert die Variable RINGS, wie viele Ringe in einer bestimmten Reaktion entstehen, STEREO spezifiziert die Zahl der gebildeten Chiralitätszentren, MREL bewertet Reaktionen positiv, die zu Substraten mit ähnlichen Massen führen („Schnitt in gleiche Fragmente"), BUY unterstützt kommerziell erhältlich Substrate, CONFLICT und PROTECT bewerten Reaktionen negativ, bei denen Inkompatibilitäten von Gruppen bzw. die Notwendigkeit von Schutzgruppenreaktionen erkannt wird. Aus diesen und einigen anderen Variablen (siehe Hintergrundinformationen, Abschnitt S13) kann der Nutzer frei wählbare Ausdrücke („Bewertungsfunktionen") definieren, mit denen die Syntheseoptionen eingeordnet werden können.

So wurde die Retrosynthese von Aripiprazol, einem Antipsychotikum der zweiten Generation, mit der einfachen Funktion MREL gelenkt, die in jeder Synthesestufe Schnitte in gleiche Größen unterstützt (Abbildung 14 und Film S3). Dabei wurde rasch ein kurzer und hoch konvergenter Syntheseweg gefunden, der von kommerziell erhältlichem 5-Hydroxyindanon, 1-Brom-4-chlorbutan, Piperazin und 2,3-Dichloranilin ausgeht. Die wichtigen Retrosyntheseschritte sind die Beckmann-Umlagerung zur Herstellung des ersten Synthesebausteins, die Aminierung des Arylbromids und Alkylierungen von O- und N-Nukleophilen. Der Synthesebaustein mit einer Lactameinheit wird aus kommerziell erhältlichem 5-Hydroxyindanon hergestellt. Dieses reagiert mit Hydroxylaminchlorid (Schritt a) zu einem Oxim, dessen Beckmann-Umlagerung (Schritt b) zu 3,4-Dihydrochinolinon führt (alternativ lässt sich diese Umwandlung unter den Bedingungen der Schmidt-Reaktion durchführen, dann aber muss das phenolische Sauerstoffatom geschützt werden). In Schritt c folgt die chemoselektive Alkylierung mit 1,4-Bromchlorbutan. Der zweite Synthesebaustein, das *N*-Arylpiperazin wird aus kommerziell erhältlichem Piperazin und 2,3-Dichloranilin hergestellt. Dieses wird durch sequenzielle Diazotierung/Bromierung (Sandmeyer-Reaktion) in das zugehörige Arylbromid überführt (Schritt d), das anschließend unter Buchwald-Hartwig-Bedingungen (Pd-NHC-Katalysator, t-BuOK) mit Piperazin verknüpft wird (Schritt e). Für diesen Schritt erkennt das Programm richtig, dass eins der Stickstoffatome im Piperazinsubstrat (blau markiert) ein Syntheseproblem darstellen kann, das sich (außer durch Verschieben des Molverhältnisses der beiden Reaktionspartner und/oder sorgfältige Optimierung der Reaktionsbedingungen) durch Schützen einer der Amingruppen umgehen lässt. Hierfür schlägt das Programm Schutzgruppen vor, die für diese spezielle Reaktionsart/Bedingungen am besten geeignet sind (Unterfester rechts unten in Abbildung 14). Angenommen, das Schutzgruppenproblem ist gelöst, dann ist der letzte Syntheseschritt f die Alkylierung des *N*-Arylpiperazins. Die Planung dieser Synthesestrategie wurde fortlaufend durch die Ranking-Fenster wie dem im rechten Teil von Abbildung 14 unterstützt, wobei alle Reaktionen gemäß der oben erwähnten nutzerdefinierten MREL-Funktion bewertet wurden.

Im Fall einfacher Moleküle wie Aripiprazol ist ohne Weiteres vorstellbar, dass ein erfahrener Chemiker auch ohne Planungssoftware ähnliche Synthesewege wie Syntaurus vorschlagen würde – wahrscheinlich aber nicht innerhalb von Minuten und vielleicht mit einigen Schwierigkeiten bei der Identifizierung von Schritten wie der Beckmann-Umlagerung. Dennoch wäre ein geübter Synthesechemiker dazu in der Lage. Eine interessantere Aufgabe für Computer sind Synthesen, die auch für menschliche Meister eine Herausforderung darstellen. Ein solches Beispiel, das uns von Prof. Dirk Trauner vorgeschlagen wurde, ist die Syntheseplanung für Epicolacton, ein erst 2012 isolierter[47a] Metabolit des Schimmelpilzes *Epicoccum nigrum*, für den nur ein plausibler Biosyntheseweg, aber keine eigentliche Totalsynthese beschrieben ist (Abbildung 15).[47b] Die Untersuchung der Synthese von Epicolacton mit Syntaurus erfolgte mit ähnlichen Bewertungsfunktionen wie die Syntheseplanung von Aripiprazol und lieferte einen beeindruckenden Reaktionsweg, dessen Screenshot Abbildung 15 a zeigt. Wie zuvor waren alle verwendeten Umwandlungen der Gesamtsynthese (Abbildung 15 b) allgemein anwendbar, d. h. in keiner Weise auf diese spezielle Zielverbindung zugeschnitten. Wir stellten fest, dass die Vorschläge des Computers die „Symmetrie" der Zielverbindung nutzten, sodass die komplexe polycyclische Struktur effizient aus einer(!) Ausgangsverbindung aufgebaut wurde. So entstand zu Beginn der Hauptteil des Kohlenstoffgerüsts durch Wasserstoffperoxid-initiierte oxidative Kupplung von polyhydroxylierten Benzolen (Schritt a).[47c]

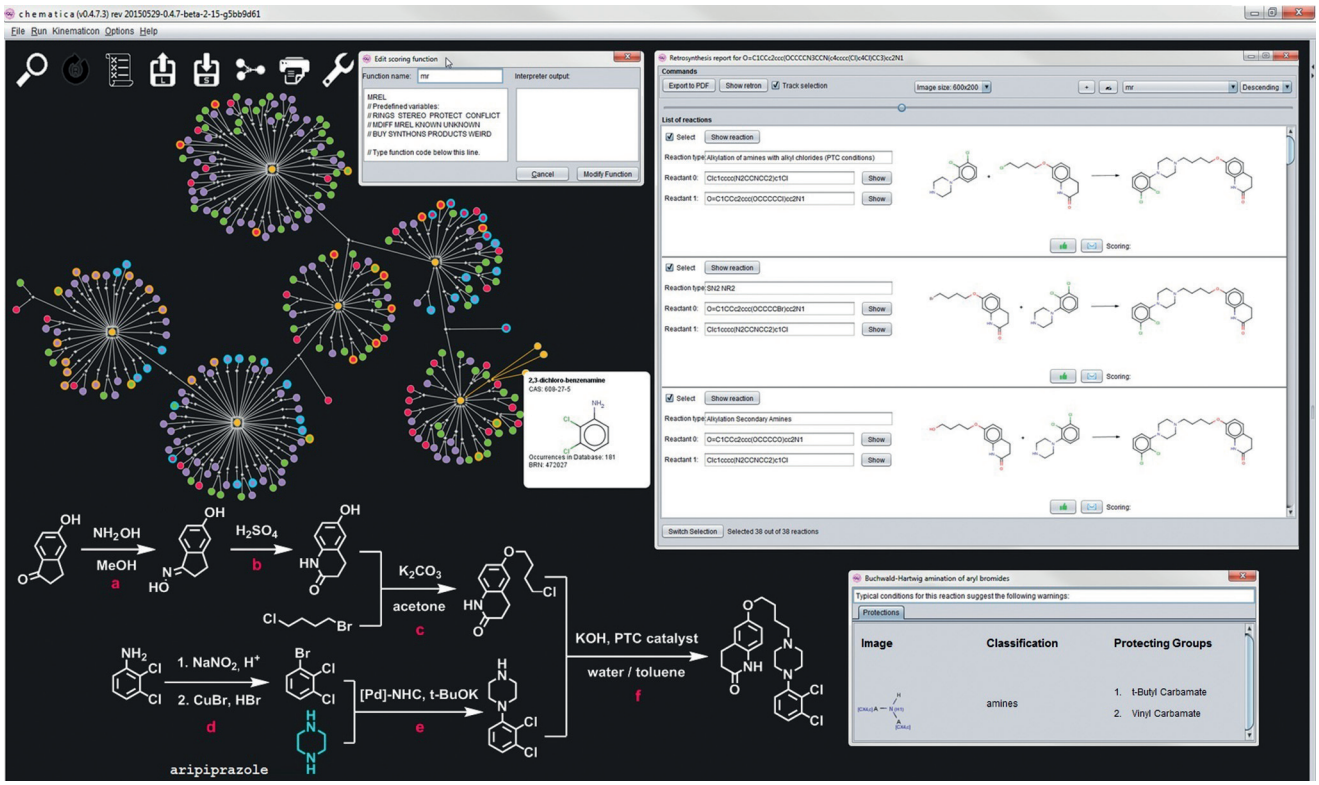


Abbildung 14. Syntheseplanung für Aripiprazole (Handelsname Abilify, ein Antipsychotikum) in Syntaurus mit der Bewertungsfunktion MREL. Oben links ist die unbearbeitete Syntaurus-Ausgabe der „Reaktionsspinnen" für die Synthesemöglichkeiten bei jeder Stufe gezeigt (siehe auch Film S3). Violette Knoten bezeichnen unbekannte und grüne Knoten bekannte Verbindungen (deren Synthese beschrieben und im NOC hinterlegt ist), rote Knoten stehen für kommerziell erhältliche Chemikalien, blaue Ringe signalisieren Schutzgruppenbedarf und orangefarbene Ringe kennzeichnen erhebliche Reaktivitätskonflikte zwischen Gruppen. Unterhalb der „Spinnen" sind die gewählten Schritte zu einem Schema zusammengefasst. Alle Reaktionen lassen sich bei einem gegebenen Schritt in Form einer Liste wie der im oberen rechten Teil der Abbildung anzeigen. Im hier gezeigten Fenster sind die Reaktionen Optionen in einer Stufe Abstand zur Zielverbindung, deren Rangfolge der MREL-Funktion entspricht. Das kleine Nebenfenster unten rechts enthält schließlich Informationen zu Gruppen, die das Programm als N-Schutzgruppe für Piperazin vorschlägt. Alle Bedingungen im Reaktionsschema unten links wurden vom Programm vorgeschlagen.

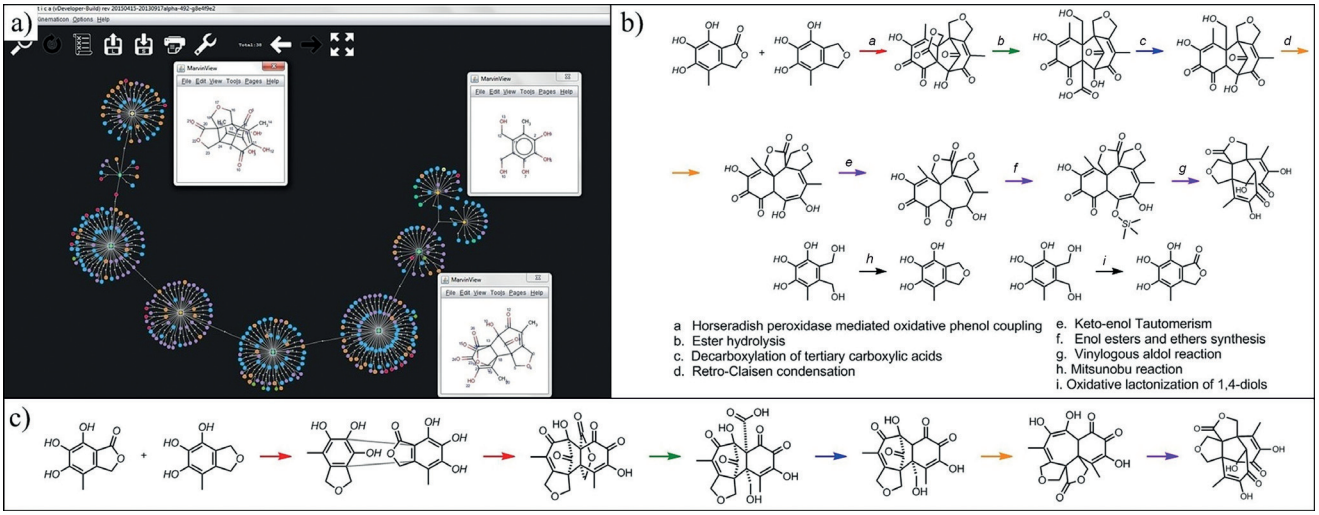


Abbildung 15. Planung eines Synthesewegs zu Epicolacton. a) Syntaurus-Screenshot des Suchbaums. Nebenfenster der Rangfolgen, die die Suche steuern, sind in den Hintergrundinformationen, Abschnitt S14 enthalten. b) Details zum eigentlichen geplanten Reaktionsweg mit einer Zusammenfassung der bei jedem Schritt genutzten Reaktionsarten (alle Reaktionsarten wurden in generalisierter Form und nicht spezifisch für das Syntheseziel oder auf der Basis von Literaturbeispielen codiert, siehe Abbildung 14). c) In Lit. [47b] vorgeschlagener Biosyntheseweg zu Epicolacton.

Danach führten einfache Umwandlungen wie eine Esterhydrolyse und eine Decarboxylierung (Schritte b und c) zu einer wichtigen Zwischenverbindung. Die Gerüstumlagerung über eine Retro-Claisen-Kondensation[47d] (der vielleicht eleganteste und schwer zu erkennende Schritt d) lieferte das tautomere Gemisch (e) aus Enol und Keton einer Zwischenstufe, die als Silylenolether aktiviert wurde (f). Die Synthese endete mit einer wohlbekannten vinylogen Aldolreaktion[47e] (Schritt g). Für die gesamte computergestützte Analyse, die zur Ermittlung dieses Synthesewegs führte, benötigte einer der Autoren nur wenige Stunden (zu Einzelheiten des Planungsvorgangs siehe Hintergrundinformationen, Abschnitt S14). Der gefundene Syntheseweg ist der vermuteten Biosynthese von Epicolacton sehr ähnlich (Abbildung 15c). Die Farben der Pfeile in der Abbildung geben die entsprechenden Reaktionsschritte in den beiden Synthesewegen wieder – der einzige Unterschied in der Methode von Syntaurus ist, dass sie separate mechanistische Schritte auswählt, während die vorgeschlagene Biosynthese diese Schritte zu „oxidativen Kaskaden“ kombiniert.

3.4. Automatisierte Suchen nach vollständigen Synthesewegen

In den Beispielen des vorherigen Abschnitts wurde der Chemiker durch den Computer unterstützt, indem dieser bei jedem Schritt schnell Synthesemöglichkeiten erstellte und in eine Reihenfolge brachte. Diese Fähigkeiten sind zwar nützlich, gehen aber nicht auf die große Herausforderung der computergestützten organischen Synthese ein – die vollkommen automatisierte Planung ganzer Synthesewege. Wegen der sehr großen Zahl von Möglichkeiten bei den „expandierenden“ Netzwerken der Retrosyntheseoptionen kann dieses Problem nicht mit erschöpfenden Suchen gelöst werden (vgl. das ähnliche Problem bei den NOC-Suchen, Abschnitt 2), sondern erfordert stattdessen intelligentere Algorithmen für Netzwerksuchen, die wiederum auf entsprechenden Maßsystemen zur Beurteilung und Lenkung der Suchen beruhen. Es ist daher von entscheidender Bedeutung, die auf die Syntheseplanung anwendbaren Konzepte von „Position“ und „Bewertungsfunktionen“ zu überdenken und genau zu definieren.

3.4.1. Definieren von „Synthesepositionen“

Der Grundgedanke einer „Position“, die es Schachprogrammen ermöglicht, aktuelle und künftige Anordnungen der Figuren auf dem Schachbrett zu bewerten, fehlte bisher in der Syntheseplanung durch Computer. Bei den früheren Methoden standen überwiegend Strategien für Bindungsbrüche (d.h., „Züge“) im Mittelpunkt, und die Realisierbarkeit der Synthone wurde nur intuitiv beurteilt. Es ist unumgänglich, von diesem Konzept abzuweichen und formal *beides*, die Reaktionen und die in jedem Retrosyntheseschritt entstehenden Substratsätze (d.h. vollständige Moleküle und keine „virtuellen“ Synthone) zu bewerten; diese Sätze sind die „Synthesepositionen“ nach jedem „Reaktionszug“. Die Syntheseplanung unterscheidet sich von Schach auch dadurch, dass jeder vollzogene Schritt die Kosten für den Gesamtweg erhöht, während es beim Schach nicht darauf ankommt, mit wie vielen Züge eine bestimmte Position erreicht wird, denn für das Endergebnis der Partie ist nur die aktuelle Stellung ausschlaggebend. In dieser Hinsicht gleicht die Syntheseplanung eher Rubiks Zauberwürfel, wenn man bestrebt ist, das Rätsel in möglichst wenigen Schritten zu lösen (vgl. Tabelle 1). Daraus folgt, dass Computer während der Syntheseplanung sowohl die Substratsätze, die bei jedem Schritt gebildet werden, als auch die Reaktionen, die zu diesen Sätzen führen, bewerten sollten.

3.4.2. Bewertungsfunktionen für Verbindungen und Reaktionen

Der oben genannten Logik zufolge müssen die vom Computer ausgewählten Synthesen durch zwei Bewertungsfunktionen evaluiert werden: die Funktion für Verbindungen (Chemicals' Scoring Function, CSF) bewertet die „Synthesepositionen“ (d.h. die Substratsätze), und die Bewertungsfunktion für Reaktionen (Reaction Scoring Function, RSF) beurteilt die „Synthesezüge“. Die Summe dieser Funktionen, CSF + RSF, kann als Maß für die Gesamtschwierigkeit (oder die „Kosten“) der Synthese betrachtet werden, daher sind Synthesen gefragt, die CSF + RSF minimieren. In der Skriptsprache von Syntaurus können CSF und RSF mit vordefinierten Variablen, die Strukturmerkmale der Verbindungen sowie Eigenschaften der Reaktionen wiedergeben, angelegt werden (z.B. RINGS, STEREO, KNOWN, BUY, PROTECT, CONFLICT; siehe Abschnitt 3.3.1 und Hintergrundinformationen, Abschnitt S13).

1) CSF. Die wichtigste Voraussetzung der CSF ist, dass sie möglichst einfache Substrate favorisiert (sodass die Reaktionen bevorzugt sind, die zu größter „Komplexität“ führen). Die Funktion summiert die für jede Verbindung im Substratsatz charakteristischen Variablen (oder Konstanten). Angenommen wir definieren die CSF nach der Variablen RINGS (Zahl der Ringe, die durch eine bestimmte Reaktion gebildet werden). Wir nehmen weiter an, dass die Zielverbindung einer bestimmten Reaktion zwei Ringe hat und durch zwei Reaktionen aus zwei verschiedenen Substratsätzen herstellbar ist. Im ersten Substratsatz gibt es nur ein Substrat mit zwei Ringen (CSF = RINGS = 2), der zweite Satz enthält dagegen zwei Substrate, wovon eins einen Ring hat, sodass CSF = 1 + 0 = 1. Dem CSF-Kriterium zufolge, so viele Ringe wie möglich zu bilden, ist der zweite Substratsatz eindeutig besser (d.h. sein CSF-Wert ist niedriger). Analog arbeitet der Parameter STEREO, der die Chiralitätszentren zählt. Die Variable MASS entspricht der Masse einer Verbindung und ist hinsichtlich Anwendungsbereich und Möglichkeiten etwas weiter entwickelt. Nehmen wir an, eine Reaktion schneidet die Zielverbindung mit der Molmasse 400 in zwei kleinere Substrate. Bei CSF = MASS ist die Bewertung unabhängig davon, wo der Schnitt erfolgt (z.B. $200 + 200 = 300 + 100$). Definiert man aber $CSF = MASS^2$, wird der Wert für CSF minimal (d.h. am besten) bei einem Schnitt in Hälften sein (z.B. $200^2 + 200^2 < 300^2 + 100^2$). Bei $CSF = MASS^x$ begünstigen steigende Werte von x im Allgemeinen die Trennung in Vorstufen ähnlicher Größe (was oft gut ist, um strategische Bindungsbrüche im Grundgerüst zu finden), sinkende Werte für x ermöglichen dagegen eher „periphere

Schnitte“ (wie bei der Funktionalisierung vorhandener Grundgerüste; die kleineren Verbindungen sind dann oft bekannt oder kommerziell erhältlich) (Abbildung 16). Von den übrigen Variablen bezieht sich SMILES_LEN auf die Länge der Verbindung in der SMILES-Notation und hängt

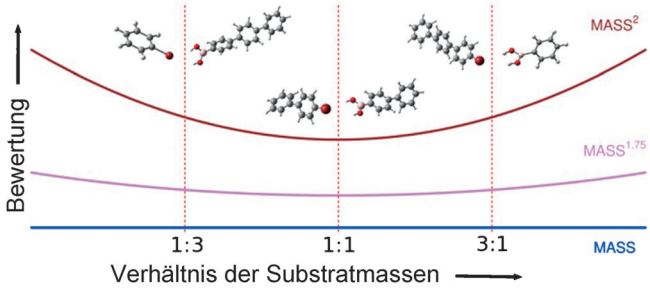


Abbildung 16. Favorisierung von Bindungsbrüche in Substrate mit bestimmten relativen Größen durch Verwendung der Bewertungsfunktion CSF mit nur einer Variablen. Die Funktion CSF ist hier definiert als die Summe (über den Substratsatz) der Substratmassen, wovon jede als Potenz x genommen wird. Wenn $x=1$ und $CSF=\Sigma_{substrates}MASS$ ist, ergibt die Trennung der Zielverbindung (hier *p*-Quaterphenyl, $MASS\approx300$) in zwei gleiche Fragmente ($150+150$) für CSF den gleichen Wert wie die Trennung in ungleiche Fragmente (z. B. $75+225$ oder $225+75$). Ist aber $x>1$, begünstigt die Funktion, die $MASS^x$ über die Substrate summiert, die Trennung in gleich große Fragmente – tatsächlich lässt sich leicht beweisen, dass für $MASS_1+MASS_2=MASS_{target}$ die Funktion $MASS_{substrate1}{}^x+MASS_{substrate2}{}^x$ minimiert ist, wenn $MASS_{substrate1}=MASS_{substrate2}$ ist. Da die besten Synthesepositionen CSF minimieren, sind solche Trennungen in gleich große Fragmente während der Synthesesuchen bevorzugt. Zu beachten ist auch, dass höhere Werte für x zu höheren CSF-Werten führen, sodass Schnitte in ungleiche Fragmente nachteiliger sind als kleinere Exponenten x (hier die Kurven für $x=2$ und $x=1.75$).

mit ihrer Masse sowie ihrer Gesamtkomplexität zusammen (da Klammern, Zahlen und die Symbole @/@@ für Verzweigungen, Ringe bzw. Chiralitätszentren stehen, nimmt SMILES_LEN zu). Die Variable WEIRD wurde dagegen definiert, um die Suche abzubrechen, wenn sie auf sehr kleine, aber „seltsam“ erscheinende und unbekannte Verbindungen trifft ($MW<100$, Verhältnis von Hetero- zu Kohlenstoffatomen >1.5; wenn eine solche niedermolekulare Verbindung in der Literatur noch unbekannt ist, gibt es sie sehr wahrscheinlich einfach nicht). Zwei weitere nützliche Variable sind BUY ($+1$ für eine kommerziell erhältliche Verbindung, anderenfalls 0) und KNOWN ($+1$, wenn die Verbindung nicht kommerziell erhältlich, aber im NOC bekannt ist, anderenfalls 0).

2) RSF. Diese Funktion soll Synthesen favorisieren, die möglichst kurz sind, keine gravierenden Reaktivitätskonflikte beinhalten oder vielleicht Schutzgruppen erfordern. In der RSF könnte man gewisse konstante Kosten für die Durchführung eines Reaktionsschritts sowie eine Kombination der Variablen PROTECT (ein bestimmter Nachteil für jede zu schützende Gruppe), CONFLICT (Nachteil für jede Gruppeninkompatibilität), und YIELD (theoretisch geschätzte Ausbeuten)[48] festlegen. So ist $RSF=30+1000\cdot CONFLICT$ besonders nachteilig für alle Reaktionen, in der Reaktivitätskonflikte auftreten, und $RSF=30+1000\cdot PROTECT$ benachteiligt Reaktionen, bei denen Schutzgruppen erforderlich sind.

3) Bewertungsfunktionen für verschiedene Sucharten. Zu den entscheidenden Vorteilen einer Bewertung von Reaktionen wie auch Substratsätzen gehört, dass man die allgemeine Suchstrategie flexibel und genau einstellen oder modifizieren kann, indem für CSF und RSF verschiedene Werte festgelegt werden. Beispielsweise bewertet die einfache Kombination $CSF=0$ und $RSF=0$ alle Reaktionen, Substrate und Synthesewege gleich (mit Null), sodass faktisch erschöpfende Suchen im Syntheseraum durchgeführt werden. Diese Art der Suche ist offensichtlich nicht „intelligent“ und mit dem Risiko astronomisch hoher Suchzeiten für jedes noch so einfache Syntheseziel verbunden. Die Kombination $CSF=0$ und $RSF=1$ ist schon sehr viel gezielter. Hierbei haben alle Verbindungen den Wert Null, aber jeder einzelne Reaktionsschritt hat die „Kosten“ $+1$. Diese Suchen minimieren die Summen von Einsen, sie berücksichtigen daher keine Molekülдetails und ermitteln tatsächlich die kürzesten Synthesewege zu bekannten oder kommerziell erhältlichen Substraten. Die diesem $CSF+RSF$ entsprechende Suchstrategie gleicht klassischen BFS-Suchen und dürfte angesichts der Zahl möglicher Synthesestrategien (vgl. Tabelle 1), auf die zugegriffen wird, sehr lange Rechenzeiten benötigen. Die Suchzeiten lassen sich drastisch verkürzen, wenn Reaktionen umgangen werden, die mit gravierenden Inkompatibilitäten/Konflikten von Gruppen verbunden sind (z. B. mit Funktionen wie $RSF=10+1000\cdot CONFLICT$). Analog lassen sich Reaktionen mit erforderlichen Schutzgruppen umgehen (und schutzgruppenfreie, Baran-ähnliche Synthesewege ermitteln), indem die Variable PROTECT eine hohe Gewichtung erhält, z. B. $RSF=20+10\,000\cdot PROTECT$.

In der CSF sollten die Variablen RINGS und STEREO bei Verbindungen mit vielen Ringen bzw. Chiralitätszentren höhere Gewichtungen haben; umgekehrt sollten diese Variablen offensichtlich die Gewichtung Null erhalten (d. h. in der CSF fehlen), wenn die Verbindung keine Ringe/Chiralitätszentren enthält. Hinsichtlich bevorzugter Bindungsbrüche entweder in Hälften oder peripher ist $SMILES\_LEN^{3/2}$ (oder $MASS^{3/2}$) der beste Kompromiss, im Allgemeinen sollte der Exponent zwischen 1.2 und 2 liegen. Spezielle Synthesebeispiele zeigen, dass die vielseitigsten, auf die überwiegende Mehrzahl der Syntheseziele anwendbaren Funktionen die Form $CSF=SMILES\_LEN^{3/2}+\alpha RINGS+\beta STEREO$ haben, wobei α und β für Zielverbindungen mit mehreren Ringen bzw. Chiralitätszentren merkliche Gewichtungen haben.

3.4.3. „Intelligente“ Suchen

Mit RSF und CSF werden die Suchen über den Bereich verfügbarer Syntheseoptionen bewertet. Da bei diesem Prozess in jedem Schritt Substratsätze und keine Einzelverbindungen evaluiert werden, ist die mathematische Formulierung des Problems stärker beteiligt und erfordert eine Darstellung als dualer Graph, in der der Algorithmus ein Netzwerk aus Substratsätzen durchläuft, das das Netzwerk spezi-

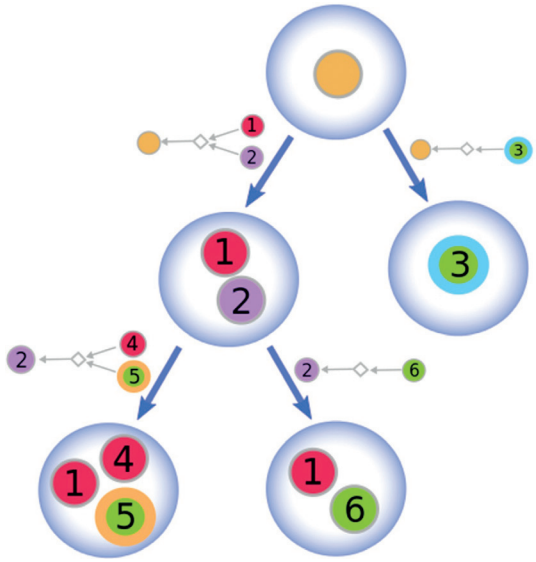


Abbildung 17. Darstellung eines dualen Graphen, der den automatisierten Suchen von Syntaurus zugrunde liegt. Der Algorithmus durchsucht und bewertet keine Einzelsubstrate, sondern Substratsätze („Synthesepositionen", hier durch große Kreise wiedergegeben). Die Beziehungen zwischen der Positionsknoten bilden einen Suchgraph. Da spezifische Reaktionen aber von Einzelverbindungen und nicht von „Positionen" eingegangen werden, muss der Algorithmus auch die Reaktionsverknüpfungen zwischen allen Verbindungen verfolgen (hier als Reaktionsminiaturen neben den blauen Pfeilen gezeigt). Dies erzeugt einen zweiten, „überlagernden" Graph.

fischer Reaktionen von Molekül zu Molekül „überlagert" (Abbildung 17).

Mit dieser Darstellung fordern wir, dass der Suchalgorithmus **1) nicht lokal** ist – das heißt, nicht nur jeweils einen „Synthesezweig" der Syntheselösungen untersuchen kann, sondern zahlreiche verschiedene Möglichkeiten gleichzeitig berücksichtigt; **2) strategisch** vorgeht – das heißt, einige einzelne „Reaktionszüge" durchführen kann, die lokal suboptimal erscheinen mögen, aber schließlich zu einer erfolgreichen Syntheselösung führen können; **3) selbstkorrigierend** ist – das heißt, in der Lage ist, von aussichtslosen Zweigen umzukehren und auf völlig andere Synthesemethoden zu wechseln. Darüber hinaus verlangen wir, dass die Suchen stets bei bekannten oder kommerziell erhältlichen Verbindungen enden (wobei die Grenzen für die Molmassen vom Nutzer spezifiziert werden).

Das Schema in Abbildung 18 veranschaulicht die Arbeitsweise eines Algorithmus, der diese Anforderungen erfüllt; seine Grundfunktionen und die Verwendung der Prioritätsreihe gehen über die spezielle Software hinaus und sind auf die Syntheseplanung allgemein anwendbar. Die Knoten in diesem Schema stehen für Substratkollektionen, die in jeden „Reaktionszug" gebildet werden (vg. Abbildung 17), hellgraue Knoten bilden den Gesamtraum an Synthesemöglichkeiten (d. h. „potenzielle" Substratsätze, die nicht unbedingt aufgesucht werden). Die Suche beginnt an dem zentralen grünen Knoten, der das Syntheseziel darstellt (Abbildung 18a). Zunächst werden alle möglichen „Züge" im Abstand von einer Stufe zum Syntheseziel untersucht und jede wird mit dem Betrag von CSF + RSF bewertet (schwarz umrandete Knoten, die Zahlen entsprechen den Bewertungen). Unter den verfügbaren Möglichkeiten bewegt sich der Algorithmus zum Knoten mit der niedrigsten Bewertung/den geringsten „Kosten" (hier Punktzahl = 4) und erweitert ihn (Abbildung 18b). Die anderen bereits bewerteten Optionen (Knoten den Punktzahlen 5, 35, 63, 71, 85, 93, 97) werden im Computer als so genannte Prioritätsreihe (priority queue, PQ) gespeichert und die „aktuellen Endpunkte" möglicher Synthesewege weiterverfolgt. Der neue Zug führt zu dem Knoten mit der Punktzahl = 35. Wir weisen darauf hin, dass diese Punktzahl die RSF-Kosten für nun *zwei* Schritte von der Zielverbindung entfernt berücksichtigt (und, wie zuvor, das CSF-Maß für die Komplexität der Synthone im aktuellen Knoten 35 enthält). An diesem Punkt werden die Kosten für die Durchführung der Reaktion entlang des aktuellen Wegs (d. h. 35) aber höher als der Punktwert der zweitbesten, in der PQ gespeicherten Syntheseoption – der am Knoten mit der Punktzahl = 5 (Abbildung 18c). Dementsprechend kehrt die Suche zum Knoten mit der Punktzahl 4 (blau) zurück, findet dort aber keine Möglichkeiten für eine „vorteilhafte" lokale Expansion. Da die lokale Verbesserung um Knoten 4 nicht möglich ist, behält der Algorithmus 1) Knoten 35 in der PQ (die nun Knoten mit den Bewertungen 35, 63,71, 85, 93, 97 umfasst), 2) behält Knoten 4 als bereits aufgesucht im Gedächtnis (aber nicht in der PQ, da er kein Endpunkt eines bereits untersuchten Synthesewegs ist) und wechselt 3) zu dem ganz neuen Reaktionsweg mit Knoten 5 als der besten (d. h. am niedrigsten bewerteten) aktuellen Option, die nun nicht mehr in der PQ enthalten ist. Ausgehend von Knoten 5 untersucht der Algorithmus drei verfügbare Möglichkeiten mit den Punktzahlen 8, 10 und 91. Hiervon wählt er den Knoten mit der Punktzahl 8, die niedriger ist als die aller anderen Knoten in der PQ, und untersucht zwei Tochterknoten mit den Bewertungen 61 und 76 (Abbildung 18d). Diese haben jedoch höhere Bewertungen als die beste PQ-Option (35), daher kehrt der Algorithmus „lokal" zu Knoten 8 zurück, wo er nichts zu untersuchen findet, kehrt dann zu Knoten 5 zurück (besser als 35 in der PQ) und expandiert in den Knoten mit der Punktzahl 10 (so weit, so gut), wo er dann aber auf drei Knoten mit hohen Bewertungen (60, 92, 98) trifft (Abbildung 18e). Da der Algorithmus keine aussichtsreichen, von Knoten 5 ausgehenden Synthesen gefunden hat, speichert er die Knoten 8 und 10 als bereits aufgesucht, fügt der PQ 60, 61, 76, 91, 92 und 98 hinzu (die nun aus den Knoten 35, 60, 61, 63, 71, 76, 85, 91, 92, 93, 97, 98 besteht) und zieht weiter zur besten verfügbaren PQ-Option. In diesem speziellen Beispiel ist die beste Option der bereits untersuchte Endpunkt 35. Dieser wird nun zu den Knoten 56, 68, 73 und 88 expandiert, von denen der erste besser ist als alle Einträge in der PQ (nun 60, 61, 63, 68, 71, 73, 76, 85, 88, 91, 92, 93, 97, 98) (Abbildung 18f). Demnach zieht der Algorithmus zu Knoten 56 und expandiert ihn in die Optionen mit den Bewertungen 57, 64 und 77 (Abbildung 18g). Von diesen ist der Knoten 57 (orange markiert) nicht nur besser als die anderen PQ-Optionen, sondern erfüllt letztlich das STOP-Kriterium, wonach alle zugehörigen Substrate kommerziell erhältlich oder bekannt sind (Abbildung 18h). Damit ist der erste realisierbare Syntheseweg (57→56→35→4→Zielverbindung)

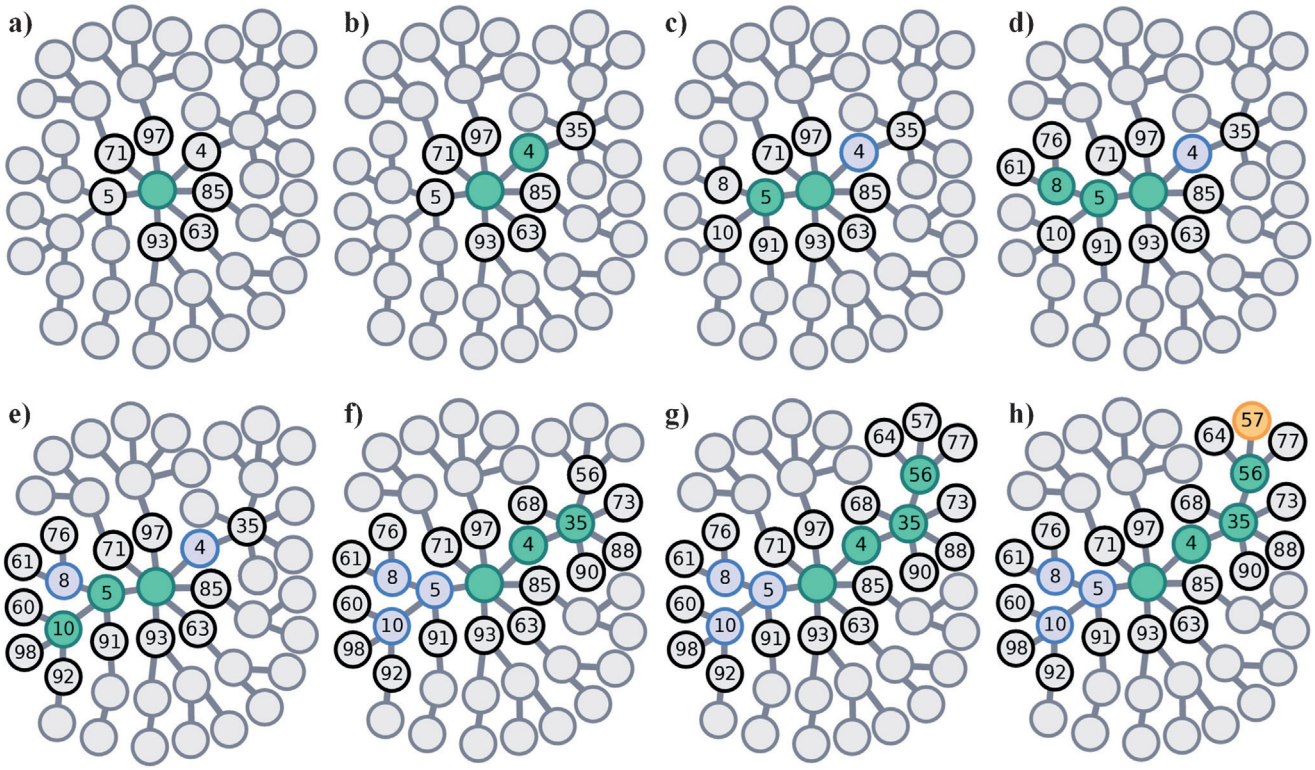


Abbildung 18. Vereinfachtes Schema des retrosynthetischen Graphsuchalgorithmus von Syntaurus. Jeder Knoten steht für einen mit jedem „Reaktionszug" gebildeten Substratsatz; die aktuelle Beschreibung gibt nicht explizit die „Expansion" von Einzelverbindungen innerhalb jedes Knotens wieder – eine solche vollständige Darstellung (siehe Abbildung 17) ist zu kompliziert, um die Grundlagen der Arbeitsweise des Algorithmus zu erklären. Hellgraue Knoten bilden den hypothetische Raum von Synthesemöglichkeiten (d. h. „potenzielle" Substratsätze, nicht unbedingt aufgesuchte). Die Suche setzt sich fort, bis alle Verbindungen im Synthonsatz, deren Massen unterhalb einer vom Nutzer festgelegten Grenze liegen, kommerziell erhältlich oder in der Literatur beschrieben sind.

gefunden, seine Bewertung umfasst die Kosten von vier Synthesestufen plus die Punktzahl der terminalen Substrate an Knoten 57. Die Suche beginnt erneut, um andere Reaktionswege zu finden – wobei der Algorithmus keine bereits identifizierten Synthesewege durchsuchen darf.

Die obige Beschreibung ist natürlich ziemlich abstrakt. So haben wir stillschweigend angenommen, dass die hellgrauen (noch nicht aufgesuchten) Knoten von Anfang an nach Art eines statischen Netzwerks dort sind und darauf warten, untersucht zu werden (wie das NOC). In Wirklichkeit bilden diese verfügbaren Synthesemöglichkeiten ein dynamisches Netzwerk, das nicht a priori bekannt ist und mit jeder untersuchten Syntheseoption expandiert. Anhand eines typischen Beispiels zeigen die Abbildungen 19a und b das expandierende Netzwerk (hier entsprechen die Knoten Verbindungen und keinen ganzen Substratsätzen) der Möglichkeiten, das Syntaurus bei der Suche nach Synthesen von Donepezil, einem Wirkstoff gegen die Alzheimer-Erkrankung, berücksichtigte. Dabei wurde eine Bewertungsfunktion verwendet, die Substrate abnehmender Komplexität favorisiert und jegliche Konflikte durch Kreuzreaktivität strikt ausschließt ($CSF = SMILES\_LEN^{3/2} + SMILES\_LEN$ und $RSF = 10 + 1000 \cdot CONFLICT^2$). Das kleinere Netzwerk in Abbildung 19a entspricht acht Expansionen der einzelnen „Spinnen", das größere in Abbildung 19b 35 Expansionen (ca. 2 min Suchzeit). Diese Netzwerke enthalten alle vom Algorithmus berücksichtigten Knoten (wovon manche weiter expandiert wurden und manche nicht, siehe Abbildungslegende). Dagegen enthalten die Unternetzwerke in den Abbildungen 19c und d nur Knoten, die im Zusammenhang mit gefundenen realisierbaren Synthesen stehen, die von käuflichen und/oder bekannten Substraten ausgehen. Erwartungsgemäß wächst das Spektrum dieser durchführbaren Synthesen, da der Algorithmus weiterläuft und nicht nur mehr, sondern auch besser bewertete Synthesewege findet. Zu den am besten bewerteten, kurzen und chemisch sinnvollen identifizierten Synthesewegen gehört der in Abbildung 19e: Nach der Umwandlung des Alkohols in das Alkylbromid und anschließender reduktiver Aminierung von Benzaldehyd endet die Synthese der Zielverbindung mit der direkten Alkylierung eines Ketons.

3.4.4. Validierung und typische Synthesebeispiele

Der letzte Test für jede Syntheseplanungs-Software ist die Frage, ob sie realisierbare Synthesepläne erstellen kann. Eine Methode der Validierung besteht darin, die Ausgabe des Programms mit Synthesewegen zu vergleichen, in denen alle Schritte bereits experimentell umgesetzt wurden. Eine wichtige Bedingung bei dieser Art der Validierung ist, dass dem Computer keine der Reaktionsregeln im Hinblick auf das spezielle Syntheseziel beigebracht wurde – oder anders

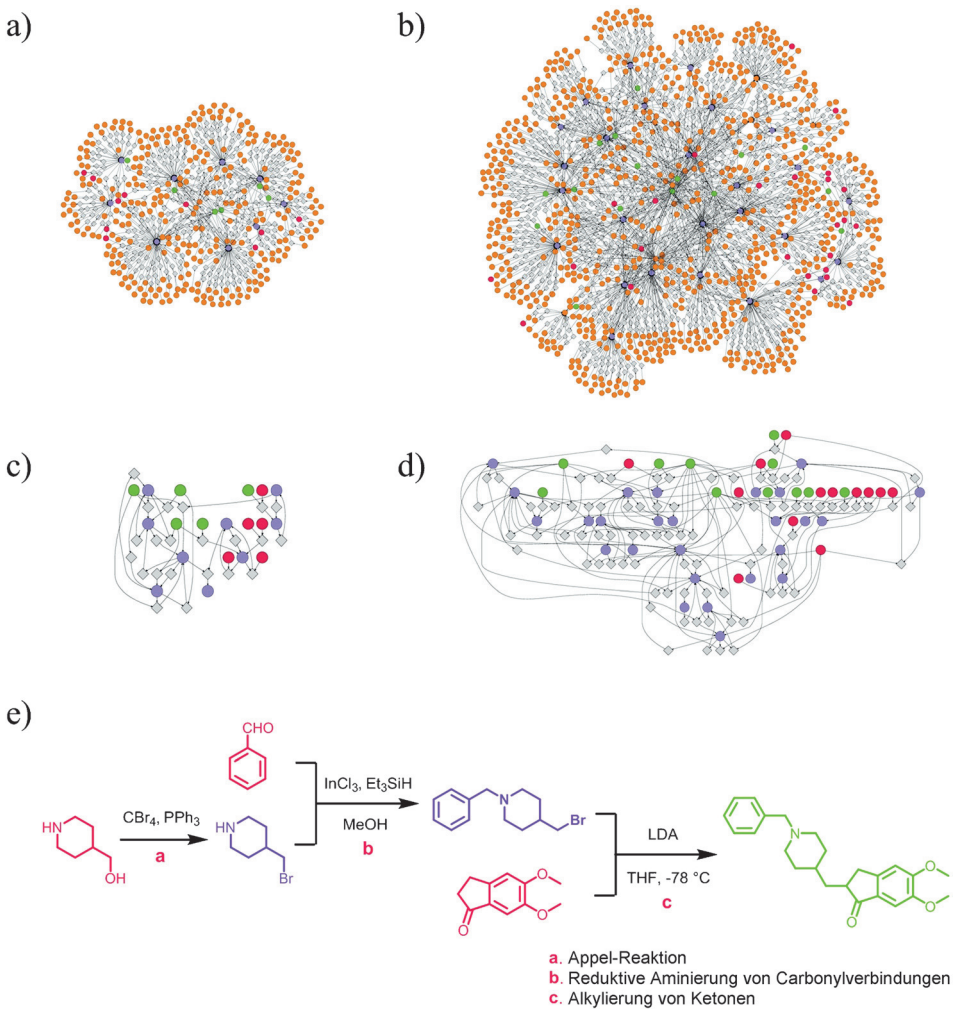


Abbildung 19. Das expandierende Netzwerk der Synthesemöglichkeiten, das Syntaurus bei der Suche nach Synthesen von Donepezil berücksichtigt, nach a) 8 und b) 35 Expansionen. Kreise stehen für Einzelverbindungen; rot: kommerziell erhältlich, grün: im NOC bekannt, orange: durch die Algorithmen berücksichtigt, aber noch nicht aus kommerziell erhältlichen oder bekannten Vorstufen herstellbar, violett: Verbindungen, für die bereits realisierbare Synthesewege gefunden wurden. Ist die erste Synthese im Zusammenhang mit einer bestimmten Verbindung gefunden, wechselt die Farbe des betrachteten Knotens von orange zu violett. c) und d) Unternetzwerke [von (a) bzw. (b)] der gefundenen realisierbaren Synthesen. e) Einer der identifizierten Reaktionswege mit Bestbewertung. Die Farbgebung der Verbindungen ist die gleiche wie bei den Knoten der Netzwerke (insbesondere rot = kommerziell erhältlich). Die Suchen erfolgten mit RSF = 10 + 1000·CONFLICT^2 und CSF = SMILES_LEN$^{3/2}$ + SMILES_LEN.

gesagt, das Programm kann nicht an den Beispielen geprüft werden, an denen es trainiert wurde. Diese Bedingung ist für Syntaurus erfüllt, dem allgemeingültige und nicht für die Zielverbindung spezifische Reaktionsregeln zugrundeliegen (siehe Abschnitt 3.3). Der Ergänzungsabschnitt S15 und der Film S4 bieten mehrere Beispiele für vollkommen automatisierte Suchen mit Syntaurus, die innerhalb von Minuten Synthesewege zu relativ komplizierten Zielverbindungen lieferten, wobei alle Einzelschritte dieser Synthesen in der Literatur beschrieben sind.

Die Planung realisierbarer Synthesewege für Naturstoffe, die erst vor kurzem isoliert wurden, sodass es nur wenige oder keine vorherigen Synthesen gibt, erforderte ähnlich kurze Zeiten (Abbildung 20). Für Fälle wie diese ist die computergestützte Planung von De-novo-Synthesen besonders nützlich. Ein Beispiel ist der vorgeschlagene Syntheseweg für Tacamonidin, ein natürliches, aus der Baumart *Tabernaemontana corymbosa* isoliertes pentacyclisches Alkaloid,[49a] das bisher noch nicht im Labor synthetisiert wurde (Abbildung 20a). Zunächst wird Tryptophol [3-(2-Hydroxyethyl)indol] mit Bernsteinsäureanhydrid acyliert (Schritt a). Das Programm schlägt vor, die freie Hydroxygruppe zu schützen (blauer Ring um den Knoten), wofür die Methoxymethylgruppe unter den gegebenen Reaktionsbedingungen am besten geeignet ist. Das so erhaltene *N*-Acylindol wird dann mit Oxalylchlorid zum Acylchlorid umgesetzt und anschließend unter Friedel-Crafts-Bedingungen[49b] zur tricyclischen Zwischenverbindung acyliert. Danach führen die enantioselektive Alkylierung nach der Methode von Enders[49c,d] (Schritt c) und die reduktive Aminierung[49e] (Schritt d) zum 1,2-*syn*-Amin. Nach dem iodvermittelten[49f] Ringschluss (Schritt e) folgte eine asymmetrische Sharpless-Dihydroxylierung zum gewünschten Diol, das anschließend zur Zielverbindung cyclisiert wird.

Ein weiteres Beispiel ist die Synthese von Goniothalesdiol A, das aus der Pflanze *Goniothalamus amuyon* isoliert wurde (Abbildung 20b).[50a] Verbindungen aus dieser Klasse sind beliebte Syntheseziele, aber Synthesen von Goniothalesdiol wurden erst vor kurzem beschrieben.[50b–e] Syntaurus identifiziert den Schlüsselschritt (*syn*-selektive Oxa-Michael-Addition) und schlägt einen kurzen Syntheseweg ausgehend von Methylacrylat und But-3-enal vor. Durch Umsetzung dieser beiden Verbindungen mit einem Grubbs-II-Katalysator der zweiten Generation[50f] (Bedingungen der Alkenmetathese) entsteht Methyl-(*E*)-5-oxopent-2-enoat. Das so erhaltene Produkt kann im folgenden Schritt mit Hydroxyacetophenon nach der Methode von Shibasaki reagieren, bei der ein *anti*-selektives, heterobimetallisches Katalysatorsystem auf der Basis von BINOL[50g] die Kontrolle über die Diastereo- und Enantioselektivität des erhaltenen 1,2-Diol ermöglicht. Die Synthese der gewünschten Verbindung endet dann auf ähnliche Weise wie die von Reddy und Fadnavis beschriebene Methode: mit einer Ein-

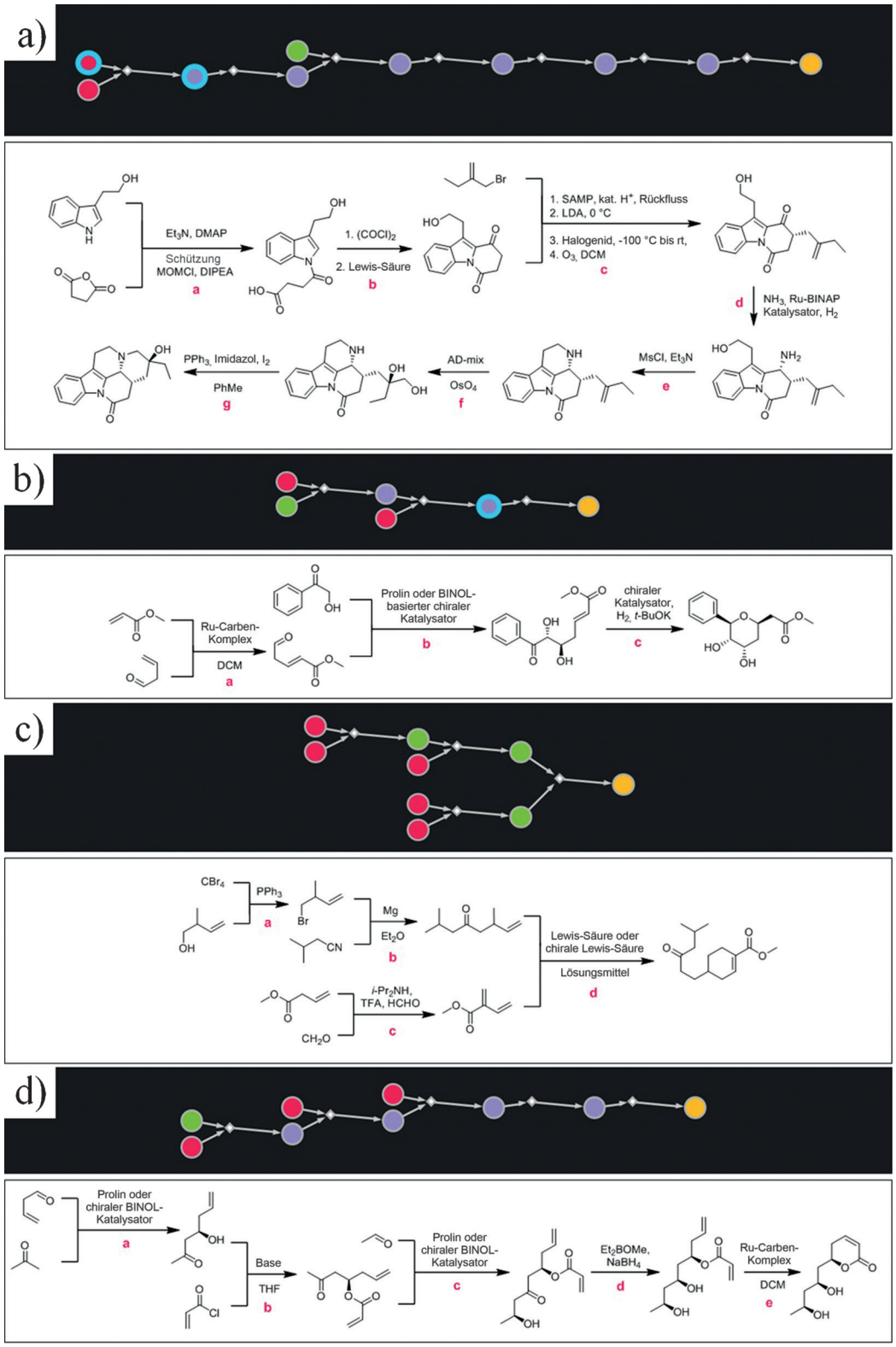


Abbildung 20. Automatisch mit Syntaurus geplante Synthesewege zu kürzlich isolierten Naturstoffen. a) Synthese von Tacamonidin,[49a] ermittelt mit CSF = SMILES_LEN$^{3/2}$ und RSF = 40 + 50·PROTECT + 1000·CONFLICT. Vom Programm vorgeschlagene „generische" Bedingungen stehen über den Reaktionspfeilen. Hinweis: Wegen der endständigen Doppelbindung sollten die Bedingungen für die Abspaltung des Enders-Auxiliars modifiziert werden, um eine Oxidation zum Keton zu vermeiden.[49d] b) Synthese von Goniothalesdiol A,[50a] identifiziert mit CSF = (RINGS + STEREO)·SMILES_LEN^2 + (RINGS + 10·STEREO) und RSF = 20 + 20·CONFLICT; c) Synthese von racemischem Juvabion,[51a] gefunden mit CSF = SMILES_LEN^2 und RSF = 100 + 5·PROTECT + 10·CONFLICT; d) Synthese eines aus *Cryptocarya latifolia* isolierten polyhydroxylierten Naturstoffs.[52a] Der Syntheseweg wurde mit CSF = 10·RINGS + 10·STEREO und RSF = 5 + 20·PROTECT + 50·CONFLICT ermittelt. Farbgebung der Knoten: rot = kommerziell erhältlich; grün = im NOC bekannt; violett = unbekannt, gelb = Zielverbindung; blaue Ringe = Schutzgruppe erforderlich.

topfkaskadenreaktion aus Ru-vermittelter enantioselektiver Reduktion und intramolekularer Oxo-Michael-Reaktion, die in der Literatur für die Synthese von Centrolobin beschrieben ist.[50h]

Das dritte Beispiel ist die Synthese von racemischem Juvabion, ein aus *Abies lasiocarpa* isoliertes[51a] Analogon eines Insektenjuvenilhormons (Abbildung 20c). Der von Syntaurus gefundene Syntheseweg beginnt mit der Umwandlung eines kommerziell erhältlichen, verzweigten ungesättigten Alkohols in das Bromid mit Br_2/PPh_3 (Schritt a), das danach in ein Grignard-Reagens überführt wird.[51b] Dessen Umsetzung mit einem Nitril (Schritt b)[51c] liefert das für die abschließende Diels-Alder-Reaktion benötigte Dienophil (Schritt d; zu weiteren Details siehe Hintergrundinformationen, Abschnitt S17). Das zugehörige Dien wird durch Methylenierung eines ungesättigten Carbonsäureesters hergestellt (Schritt c).[51d]

Abbildung 20d zeigt schließlich die Synthese eines polyhydroxylierten Naturstoffs, der aus *Cryptocarya latifolia* isoliert wurde.[52a] Diese Synthese verdeutlicht einige Details, die einer sorgfältigen Prüfung durch den Nutzer bedürfen. Der vom Computer gefundene Syntheseweg gleicht den in der Literatur beschriebenen Methoden.[52b,c] Das Kohlenstoffgerüst wird durch zwei enantioselektive Aldolreaktionen synthetisiert, wobei in Schritt a ein Prolinkatalysator[52d–f] für die Stereokontrolle und in Schritt c ein chirales Borenolat[52g] oder Lithiumenolat[52h] für die korrekte Diastereoselektivität sorgten. Die zweite Aldolreaktion scheint der schwierigste Schritt im vorgeschlagenen Syntheseweg zu sein; mit chiralen Borauxiliarien wurde zwar gute Diastereokontrolle beschrieben, man könnte aber erwarten (und eingehend prüfen), dass die ungünstige 1,5-*anti*-Selektivität[52i] Auswirkungen auf das Ergebnis der Reaktion hat (siehe Lit. [52j–l]). Die Acylierung des Alkohols mit Acryloylchlorid (Schritt b) und die Alkenmetathese (Schritt e)[52b] ermöglichen die effiziente Bildung der Lactoneinheit, und in Schritt d wird das dritte Chiralitätszentrum mit der Narasaka-Prasad-Methode eingeführt.[52c,j,m]

4. Herausforderungen und Chancen

Die Synthesebeispiele im vorigen Abschnitt lassen erkennen, dass Computer letztlich imstande sind, Synthesen für die tägliche Praxis in der organischen Chemie zu planen. Zugleich ist das – wie der Titel des Aufsatzes betont – erst „das Ende vom Anfang“, und es sind noch viele Schwierigkeiten zu bewältigen. In diesem Teil betrachten wir die Herausforderungen als hochinteressante Möglichkeiten für die künftige Forschung und weisen, wenn möglich, auf die wichtigsten und vielversprechenden neueren Entwicklungen hin.

4.1. Rationalisieren der Suchen, universelle Bewertungsfunktionen und „Maße für die Synthetisierbarkeit“

Algorithmen wie die in Abschnitt 3.4.3 beschriebenen sind ansatzweise zu strategischem Vorgehen fähig, indem sie Synthesezüge wählen, die „lokal“ suboptimal erscheinen, aber insgesamt zu optimalen Synthesewegen führen. In der Synthese von (−)-Curvularin (Abbildung S27d) führt der Algorithmus beispielsweise drei Schritte aus („d–f“), die in Bezug auf den Aufbau molekularer Komplexität scheinbar keinen offensichtlichen unmittelbaren Nutzen bieten, aber wichtig für die gesamte Synthesestrategie sind, die zum entscheidenden Arin/Ketoester-Bindungsbruch (Schritt g) führt.

Bei sehr großen Netzwerken von Synthesemöglichkeiten (z. B. in der Syntheseplanung für große/komplizierte Zielverbindungen), die mehrstufige Strategien ermitteln, könnten allerdings übermäßig viele einzelne Hin- und Zurückzüge für die Sondierung nötig sein. Ein auf LHASA zurückgehender Gedanke[5b] ist, in diesen Fällen die Suchen zu rationalisieren, indem bestimmte häufige Reaktionssequenzen zuvor in verbindungsprogrammierten „Strategien“ zusammengefasst werden (Abbildung 21a). Trifft die Suche während der Syntheseplanung dann auf eine Reaktion, die dem ersten Schritt einer Strategie entspricht, sollte das Programm automatisch den/die folgenden Strategieschritt(e) berücksichtigen. Strategien können die Suchzeiten tatsächlich verkürzen, wie das Beispiel in Abbildung 21b verdeutlicht. Danach verkürzte sich die Suchzeit für die Planung der Synthese von *N*,*N*-Dimethylbispidin (ein Grundgerüst von Verbindungen, die als potenzielle Wirkstoffe zur Behandlung von Herzrhythmusstörungen untersucht werden und strukturell verwandt sind mit dem zur Raucherentwöhnung verwendeten Cytisin) durch Nutzung einer Strategie, die aus einer Mannich-Reaktion mit anschließender Desoxygenierung des Ketons besteht (vgl. fünfte Zeile in Abbildung 21a), um den Faktor drei. Andererseits muss man bedenken, dass die Einführung zu vieler Strategien die Suchen zu sehr in bestimmte Zweige der Synthesemöglichkeiten lenkt und so die Diversität der ermittelten Synthesewege begrenzt. Nach unseren Erfahrungen scheinen etwa zehn bis hundert Strategien optimal zu sein, allerdings muss eine systematische Untersuchung den Nutzen durch höhere Suchgeschwindigkeiten mit den Einschränkungen vergleichen, die Strategien für die Diversität von Syntheselösungen bedeuten.

Eine weitere Frage im Zusammenhang mit der Sucheffizienz ist, welche Arten der Funktionen CSF und RSF in kürzestmöglicher Zeit Lösungen liefern. Analysen wie die in Abbildung S28 (Abschnitt S17) lassen darauf schließen, dass Bindungsbrüche in gleich große Fragmente zu bevorzugen sind (aber mit Spielraum für Schnitte in unterschiedliche Größen), und dass Variablen, die die molekulare Komplexität wiedergeben, wichtig sind (z. B. gibt die Länge eines Moleküls in SMILES bessere Ergebnisse als beispielsweise die Molekülmasse). Dieses Kriterium hängt mit jüngsten Fortschritten bei der Definition aussagekräftiger Maße für die Komplexität von Verbindungen und Synthesen zusammen. Li und Eastgate haben kürzlich einen Komplexitätsindex vorgeschlagen,[53a] der „intrinsische“ Moleküleigenschaften (Randic-Konnektivitätsindex[53b,c] und die Zahl der Heteroatome an und in aromatischen Ringen) kombiniert mit „extrinsischen“ Maßen der Synthesekomplexität (Zahl der in der Synthese gebildeten Chiralitätszentren, Gesamtzahl der Reaktionsschritte, Idealität des Syntheseroute),[53d] die sich zu Fortschritten im Synthesewissen entwickeln. Sie konnten nach-

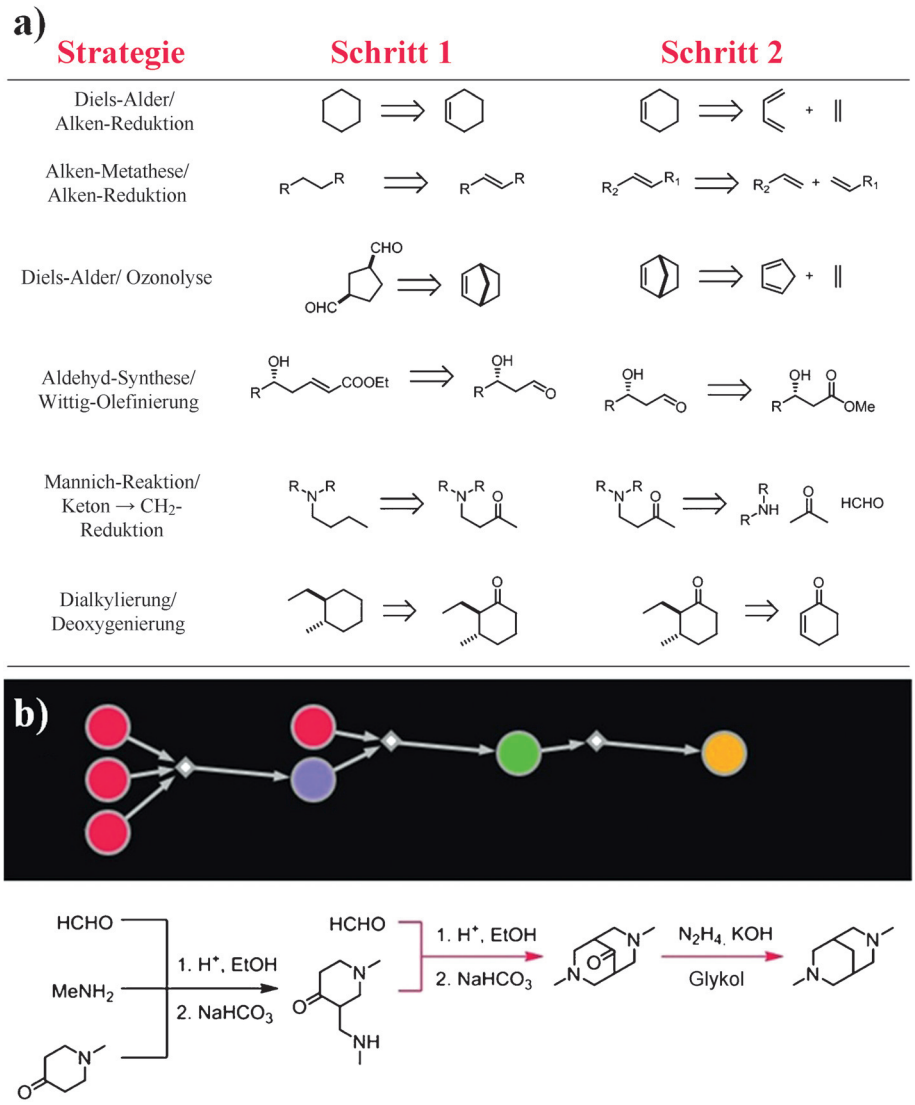


Abbildung 21. a) Beispiele für zweistufige Strategien. Die Schritte sind in Retrosyntheserichtung nummeriert. b) Eine von Syntaurus geplante Synthese von *N*,*N*-Dimethylbispidin. Aus Sicht des Programms scheint die Einführung einer Carbonylgruppe im ersten Retrosyntheseschritt nicht unmittelbar von Vorteil zu sein, da sie zu einem Substrat höherer Komplexität führt. Durch Verwendung einer der eingebauten Strategien [hier in Retrorichtung: „nach Einführung eines Ketons, untersuche Mannich-Reaktion"; Schritte mit rosa Pfeilen markiert, siehe auch fünfte Zeile in a)] verringerte sich die Zahl der Suchiterationen von 151 auf 48. In den Suchen mit und ohne Strategien waren alle Parameter identisch.

weisen, dass dieser zusammengesetzte intrinsisch-extrinsische Index gut mit der Erkenntnis erfahrener Chemiker korreliert, die gemeinsam eine Rangfolge der Schwierigkeiten von Synthesen mehrerer komplizierter Zielverbindungen erarbeitet haben.

Die „intrinsischen" Maße, wie sie von Eastgate und auch von Gasteiger[53e,f] vorgeschlagen wurden, können in die CSF-Bewertungsfunktionen, die die Wahlen der Syntheseplanungsprogramme während der Planung des Reaktionswegs steuern, eingebaut werden. Andererseits kann die Kombination aus „intrinsischen" und „extrinsischen" Maßen (siehe auch Lit. [53g]) nützlich für die Evaluierung und die Rangfolge ganzer Synthesewege sein, die diese Programme schließlich identifizieren. Insbesondere die Rangfolge und Statistiken über sehr viele (vgl. Beispiel in Abschnitt S18 und Film S4) computergenerierte Synthesepläne könnten bezüglich der Syntheseschwierigkeit („Synthesisierbarkeit") aufschlussreicher sein als die Bewertung von nur einem einzigen Reaktionsweg, der im Labor gelingen könnte oder auch nicht. Idealerweise sollten solche Rangfolgen die „Synthesisierbarkeit" nicht nur von Zielverbindungen, die sich in ihren Massen oder der Zahl ihrer Chiralitätszentren oder Ringe deutlich unterscheiden, differenzieren können, sondern auch von strukturell ähnlichen Synthesezielen. Dies verdeutlichen die Beispiele in Abbildung 22, in der ein relativ einfaches Maß für die Synthetisierbarkeit (basierend auf der Zahl der Syntheseschritte und der strukturellen Komplexität der Ausgangsverbindungen) signifikant unterschiedliche Histogramme für die „Synthesisierbarkeit" ergibt (d.h. Anzahl der computergenerierten Synthesen mit unterschiedlichen Bewertungen). Ein Gebiet, auf dem solche Analysen unmittelbare Bedeutung hätten, ist das Screening „virtueller" Leitverbindungen, die heutzutage in den Abteilungen für computergestützte Chemie nahezu aller pharmazeutischen Unternehmen in großen Mengen[16] entworfen werden, oft aber nicht oder zumindest nur schwer zu synthetisieren sind.[54] Die Möglichkeit, solche Verbindungen herauszufiltern, kann beträchtliche Kosteneinsparungen und vor allem die schnellere Entwicklung von der Leitverbindung zum Wirkstoff bedeuten.

4.2. Fehlende versus implausible Reaktionsvorschläge und Nachbearbeitung von Synthesewegen

Bei den ersten Syntheseplanungsprogrammen gehörte die unzureichende Zahl von Reaktionsregeln zu den Hauptproblemen. Wenn ein solches Programm keine plausiblen Reaktionen identifizierte, bedeutete das nicht unbedingt, dass es keine durchführbare Synthese gab, sondern vielmehr, dass einige wichtige Reaktionsregeln/Umwandlungen in der Informationsdatenbank des Computers fehlten (wie bei LHASA[5] oder SYNCHEM).[8] Moderne Programme wie ARChem,[41] IC_{SYNTH},[42] und Syntaurus beruhen dagegen auf

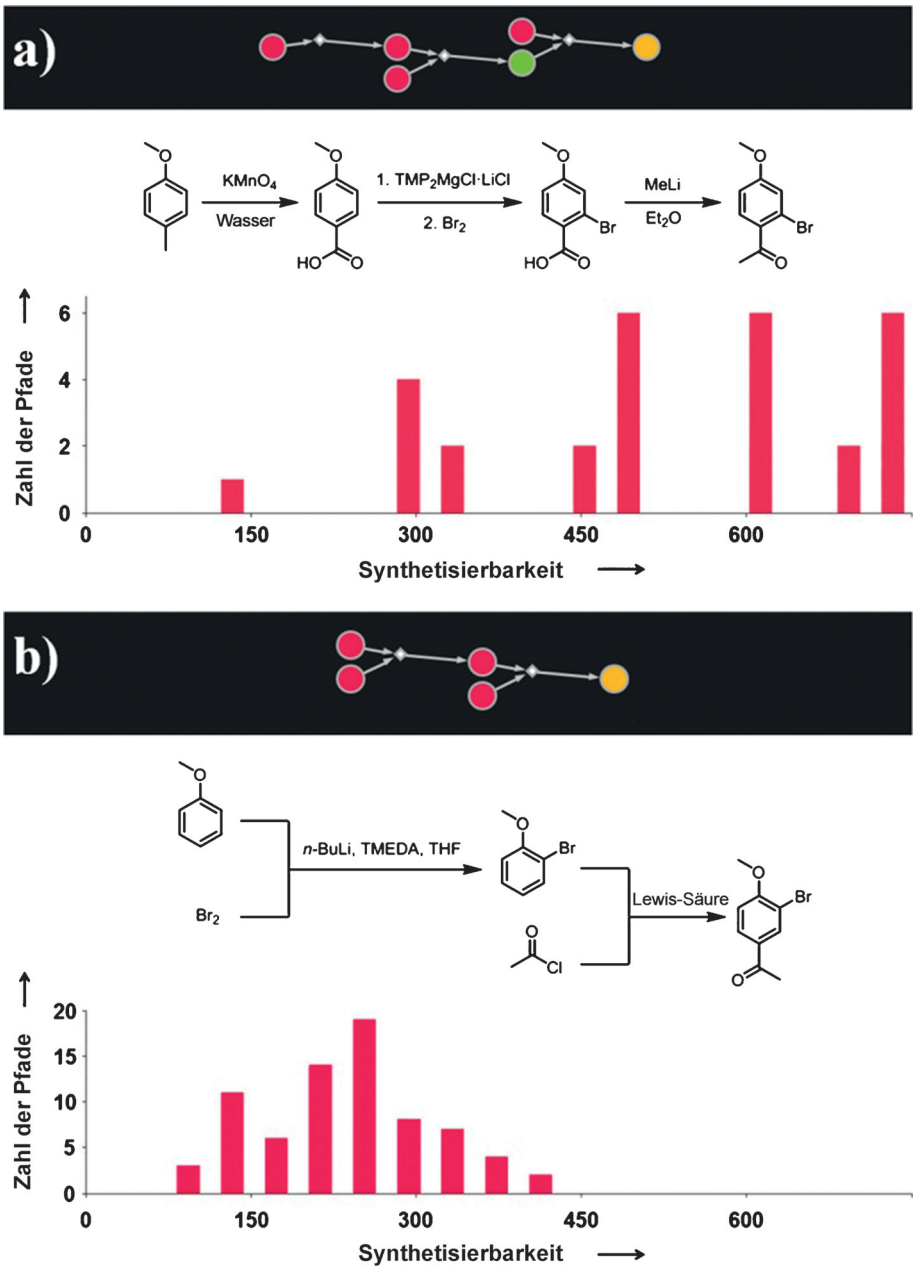


Abbildung 22. „Synthetisierbarkeit", beurteilt nach der Verteilung von Bewertungen computergenerierter Reaktionspfade. Selbst kleine Unterschiede in den Substitutionsmustern können zu ausgeprägten Unterschieden in der Synthetisierbarkeit führen, wie das Beispiel von zwei Brommethoxyacetophenonen zeigt. In (a) hat der beste/am niedrigsten bewertete Pfad den Synthetisierbarkeitswert = 123; der Durchschnittswert für alle Synthesewege beträgt 546.74 mit der Standardabweichung 187.89. Das Isomer in (b) ist deutlich einfacher herzustellen, wie der Punktwert = 83 für den besten Pfad zeigt (Durchschnittswert = 231.15, Standardabweichung = 79.45). Der Computer findet für diese Verbindung auch weitere Synthesepläne. Alle Analysen wurden in Syntaurus mit CSF = SMILES_LEN$^{3/2}$ + SMILES_LEN, RSF = 40 + 60·PROTECT + 50·CONFLICT^2 sowie Nachverfolgen der Suchen zu kommerziell erhältlichen Reagentien mit MW < 125 durchgeführt.

vollständigen oder fast vollständigen Reaktionsdatenbanken; wenn sie falsch liegen, betrifft dies die Vorhersage von Reaktionen, die auf dem Papier akzeptabel aussehen mögen, im Labor aber nicht ablaufen werden. Wie wir bereits ausgeführt haben, lässt sich die Mehrzahl dieser Probleme auf dem Niveau geeignet programmierter Reaktionsregeln vermeiden, die erlaubte Substituenten, Stereo- und Regiochemie, Schutzgruppenchemie und Reaktivitätskonflikte ebenso berücksichtigen wie elektronische und die meisten sterischen Effekte, wobei diese auf der Ebene des Reaktionsmotivs codiert werden. Vielleicht am schwierigsten ist die Berücksichtigung sterischer Effekte in strukturell komplexen Verbindungen, wobei die dreidimensionale Gesamtstruktur die Reaktivität oder ihr Fehlen bestimmen kann (Abbildung 23 und Lit. [55]). Die Untersuchungen zur Auswirkung sterischer Effekte auf das Resultat der Reaktion gehen zurück auf die Arbeit von Taft über die relativen Geschwindigkeiten der säurekatalysierten Esterhydrolyse.[56a] Seither wurden mehrere, auf topologischen Deskriptoren basierende Modelle vorgeschlagen,[53b,56b,c] wovon Cao und Liu[56d] das bekannteste entwickelten, das in der Marvin-Software von ChemAxon enthalten ist.[56e] In dem von Cao und Liu vorgeschlagenen Index TSEI (Topological Steric Effect Index) wird jedem Atom eine Zahl proportional zu den Beiträgen seiner Nachbaratome zur sterischen Hinderung zugeordnet, die mit der reziproken dritten Potenz des topologischen Abstands (d.h. Anzahl der Bindungen) zum betreffenden Atom gewichtet werden. Der wichtigste Vorteil des Modells ist, dass es anders als viele frühere alkanspezifische Maße auch auf Heteroatome anwendbar ist und sich auch für sehr komplizierte Verbindungen im Bruchteil einer Subsekunde berechnen lässt. Dagegen berücksichtigen diese und ähnliche Methoden (z.B. die von Sello vorgeschlagenen sterischen Deskriptoren,[56f] in die die sterische Hinderung aller Reaktionssubstrate und Produkte einfließt) unter Verwendung topologischer statt kartesischer Abstände die dreidimensionalen Formen der Moleküle eher nicht. Dieses Problem könnte umgangen werden, indem man molekülmechanische Rechnungen zur Konformationsanalyse durchführt, die Wahrscheinlichkeiten der verschiedenen Molekülkonformationen aus Energien herleitet (über die Boltzmann-Beziehung $p(E) \propto \exp(-E/kT)$ und anschließend allen Atomen

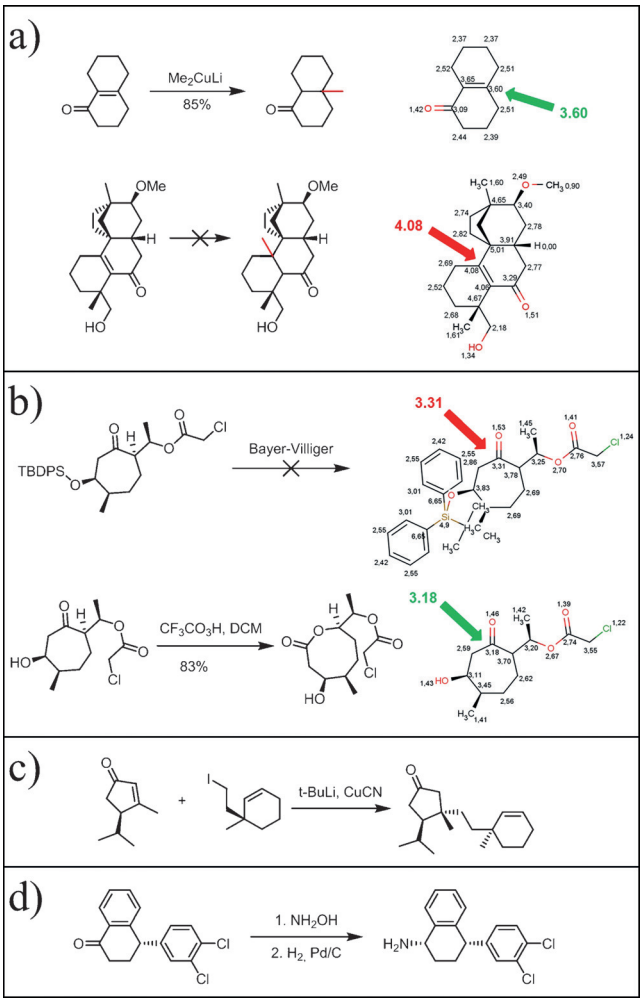


Abbildung 23. Beispiele für Reaktionen, deren Ergebnisse nicht mit „lokalen“ Reaktionsregeln, sondern mit den „globalen“ 3D-Konformationen der Moleküle ermittelt wurden. Die Zahlen in den Strukturen ganz rechts sind die Werte für den Index der sterischen Hinderung TSEI,[56d,e] der in der Marvin-Software enthalten ist. a) Die Einführung einer quartären Methylgruppe durch 1,4-Addition von Cyanocuprat scheint einfach zu sein, aber bei der Synthese von Scopadulcinsäure B durch Overman[55a] entstand das gewünschte Produkt unter keinen der untersuchten Bedingungen. Diese Schwierigkeit beruhte auf dem Vorliegen sterisch anspruchsvoller Substituenten in Nachbarschaft zum β-Kohlenstoffatom des Enons. In diesem Fall gibt der hohe Wert des TSEI-Index (4.08, roter Pfeil) die fehlende Reaktivität an der betreffenden Position korrekt wieder. b) Bei der Synthese der (+)-Octalactine A und B durch Clardy[55b] war die versuchte Bayer-Villiger-Oxidation eines geschützten Ketons wegen einer vorhandenen sterisch anspruchsvollen TBDPS-Gruppe unmöglich. Nach Abspaltung dieser Gruppe wurde das gewünschte Lacton jedoch in guter Ausbeute erhalten. In diesem Fall reagiert die TBDPS-geschützte Verbindung nicht, obwohl der Wert des TSEI-Index am Carbonylkohlenstoffatom sehr viel kleiner ist (3.31) als in Beispiel a). c) Die diastereoselektive Addition eines Organocuprats an ein ungesättigtes Keton (eine in der stereokontrollierten Synthese verbreitete asymmetrische Induktion, die hier in der Totalsynthese von Guanacastepen A verwendet wird)[55c] erfolgt wegen der sterischen Hinderung bevorzugt an die weniger substituierte Seite des Enons. In diesem Fall sind die topologischen Indizes, die auf gemessenen Abständen in Form von Bindungen vom Reaktionszentrum (wie TSEI) basieren, unempfindlich gegenüber der sterischen Induktion und daher nicht anwendbar. d) In der Synthese von Zoloft durch Aggarwal[55d] verlief die reduktive Aminierung im letzten Schritt mit hoher Diastereoslektivität (Isomerenverhältnis 1,4-*syn*:*anti* 96.5:3.5). Dieser Effekt ist auf das Vorliegen eines entfernten Substituenten zurückzuführen, der die Annäherung des Katalysators von der sterisch stärker gehinderten Seite der Verbindung verhindert. Auch hier können topologische Indizes wie TSEI nicht dazu beitragen, das Ergebnis einer Reaktion zu ermitteln.

konformationsgemittelte sterische Hinderungsparameter $\langle S\rangle = \Sigma_i p_i S_i$ zuordnet, die typisch für die tatsächliche Molekülform sind.

Solche Rechnungen sind jedoch zwangsläufig zeitaufwändig; selbst wenn eine Konformationsanalyse etwa 1 s dauert, ist die Auswertung von sehr vielen (Millionen) Verbindungen, die während der Syntheseplanung berücksichtigt werden, offenbar unmöglich. Für eine begrenzte Zahl (etwa einige hundert) ermittelter kompletter Synthesewege mit Bestbewertung sind derartige Analysen dagegen vernünftig. Diese Methode erinnert an die rechnerische Wirkstoffplanung, bei der nur die am besten bewerteten Wirkstoffkandidaten aus großen Computer-Screenings eingehend untersucht werden. Bei der Syntheseplanung könnte eine derartige „Nachbearbeitung“ der besten Reaktionswege über die oben besprochene Analyse der sterischen Hinderung hinausgehen und auch Rechnungen zur Molekülmechanik, Quantenmechanik oder sogar zum Übergangszustand umfassen, sodass auf diese Weise stark gespannte Moleküle markiert, elektronische Effekte eingehend untersucht bzw. die Ergebnisse stereoselektiver Reaktionen prognostiziert werden können (z. B. mit schnellen molekülmechanischen Methoden wie dem von Moitessier et al. entwickelten ACE-Programm).[56g,h]

Als weiterer Punkt in der Nachbearbeitung von Synthesewegen ist ein wichtiges Problem der Handhabung von Schutzgruppen zu erwähnen. Wie vorherige Synthesebeispiele (vgl. Abbildungen 14, 20 a,b) erkennen lassen, kann der Rechner die Blockierungen an Einzelschritten identifizieren. Ein Programm, das „rückwärts“ in Retrorichtung plant, weiß dagegen nicht, welche Reaktionen als „nächste“ gewählt werden und ob die Schutzgruppen, die für einen noch nicht vollzogenen Schritt benötigt werden, den „aktuellen“, bereits betrachteten Schritt überstehen werden. Dieses Problem lässt sich nur dadurch umgehen, dass zuerst der ganze Retrosyntheseweg erstellt und anschließend „vorwärts“, von den Substraten bis zum Produkt verfolgt wird – diesmal aber mit dem ganzen Wissen darüber, welche Schutzgruppen in jedem Schritt vorliegen und ob sie die nachfolgenden Schritte überstehen oder nicht. Auch wenn es derzeit keine rechnerischen Mittel gibt, die optimale Handhabung von Schutzgruppen über ganze Synthesewege zu bestimmen, gehen wir davon aus, dass das Problem gelöst werden kann, indem Algorithmen optimiert werden, die individuelle Regeln für das Blockieren/Entblockieren bei jedem Schritt (wie sie bereits in Syntaurus enthalten sind) mit der Minimierung aller erforderlichen Blockierungs-/Entblockierungsschritte über den gesamten Reaktionsweg kombinieren. Wir arbeiten aktiv an der Entwicklung solcher Algorithmen und hoffen, die Ergebnisse demnächst vorstellen zu können.

4.3. Vorhersage von Reaktionsbedingungen

Neben der Vorhersage, ob eine bestimmte Reaktion ablaufen wird oder nicht, ist es auch wünschenswert, dass der Computer geeignete Reaktionsbedingungen (Lösungsmittel, Katalysator usw.) vorschlägt. Diese Fähigkeit fehlt in der vorhandenen Software weitgehend, und der Nutzer wird entweder auf die Literatur zu „ähnlichen" Reaktionen (ARChem,[41] IC_{SYNTH}[42]) oder auf wichtige Publikationen zu einer bestimmten Reaktionsart (Syntaurus) verwiesen. Varnek et al. berichteten erst kürzlich über Fortschritte auf der Grundlage des von ihnen entwickelten Formalismus des komprimierten Reaktionsgraphen (Condensed Graph of Reaction, CGR).[57a] Im CGR wird der Satz aller Reaktanten und Produkte durch einen einzigen Graph codiert, sodass die Reaktion als Pseudomolekül dargestellt wird, für das Molekül deskriptoren generiert und in verschiedenen Fragestellungen der Chemoinformatik (z. B. Ähnlichkeitssuchen bei Reaktionen[57b] oder Entwicklung von Vorhersagemodellen für kinetische Parameter der S_N2-Reaktion[57c]) weiterverwendet werden können. In ihrer jüngsten Veröffentlichung beschreiben Varnek et al., dass sich CGR-Deskriptoren mit maschinellen Lernmethoden (Support Vector Machines, Naive Bayes und Random Forests) kombinieren und so die Reaktionsbedingungen für Michael-Additionen erfolgreich vorhersagen lassen.

Die Anwendung von Methoden wie der von Varnek beschriebenen auf andere Reaktionsklassen könnte dazu beitragen, Syntheseplanungs-Software mit den in jüngster Zeit immer gefragteren automatisierten Synthesesystemen[58a–h] zu koppeln. Diese Systeme basieren oft auf einer Durchflussapparatur[58b–h] (z. B. ist in dem kürzlich bekannt gewordenen Programm Make-It von DARPA[58i] die Durchflusschemie zwingend notwendig) – in diesen Fällen ist entscheidend vorhersagen zu können, ob Reaktionen in einem bestimmten, mit den automatisierten Durchflussmethoden kompatiblen Lösungsmittel ablaufen können oder nicht.

4.4. Vorhersage neuer Reaktionsarten und Mechanismen

Wenn die Syntheseplanung mit bekannten Methoden schließlich ausgereift ist, stellen wir uns als nächste große Herausforderung die computergestützte Entdeckung neuer Reaktionsarten und Mechanismen vor. Wir wissen, dass dies prinzipiell möglich ist, seit Ivar Ugi bereits in den 1990er Jahren die rechnerische Entdeckung einiger neuer Reaktionen unter Verwendung der von ihm entwickelten Bindungselektronen-Matrizen nachgewiesen hat.[31b] Heute kombinieren Softwarepakete wie RMG[59a,b] oder EXGAS[59c] die Vorhersage von Reaktionsgeschwindigkeiten, indem sie die Übergangszustandstheorie mit quantenchemischen Berechnungen der Aktivierungsschwellen nutzen, um plausible mehrstufige Reaktionsmechanismen für Verbrennungsprozesse zu erstellen. In einer interessanten neueren Arbeit[60a] beschreiben Aspuru-Guzik et al., wie eine Kombination aus heuristischen chemischen Regeln (die durch „Verschieben von Pfeilen" erhalten wurden)[60b,c] mit präzisen quantenmechanischen Rechnungen und Netzwerktheorie die meisten Zwischenverbindungen und Produkte wiedergeben kann, die an dem komplizierten Mechanismus der wohlbekannten, aber immer noch unvollständig verstandenen Formose-Reaktion beteiligt sind.[61] Das sind vielversprechende erste Beispiele, und das Zusammenspiel von ausgereifter Theorie und mechanistischer organischer Chemie kann in der Tat richtungweisend werden, vor allem wenn die Software für quantenmechanische Berechnungen nutzerfreundlich und damit für Chemiker in der Praxis zugänglich wird.

5. Schlussbemerkungen

Zusammenfassend glauben wir, dass mit modernen Computern eine Fülle hochinteressanter chemischer Entdeckungen gemacht werden können. Wir sind nicht mehr in den 1960er Jahren, als die Rechner für die Probleme der organischen Chemie einfach unzureichend waren – heute können Computer bereits realisierbare Synthesen für recht komplizierte Syntheseziele vorschlagen, und mit der Weiterentwicklung rechnerischer Methoden können sie nur besser werden. Die Zeit ist reif, rechnerische Methoden in die tägliche Praxis der organischen Synthese zu integrieren und den Dialog mit unseren Kollegen in der Informatik zu initiieren (oder vielleicht wieder zu beginnen). Nach unseren eigenen Erfahrungen ist das ein fruchtbarer Diskurs.

Danksagung

Wir danken für Förderung durch den Symfonia Award (UMO-2014/12/W/ST5/00592) des Polish National Science Center, NCN. B.A.G. dankt dem Institute for Basic Science Korea, Projekt-Nr. IBS-R020-D1 für Förderung. Die Entwicklung des automatischen Syntaurus wurde aus Privatmitteln von B.A.G. gefördert. P.D. und M.S. danken GSI, Inc. für Unterstützung. Wir danken Prof. Kyle J. M. Bishop (Penn State), Dr. Mikołaj Kowalik, Dr. Jarosław Granda und Piotr Janiuk für ihre Unterstützung bei der Entwicklung von Algorithmen und hilfreiche Diskussionen sowie Prof. Bartłomiej Furman (Warsaw), Dr. Nicolas Armanino (München) und Prof. Dariusz Witt (Gdańsk) für ihre eingehende Prüfung der von Syntaurus geplanten Synthesen. B.A.G. dankt allen früheren Studenten und Postdoktoranden, die zur Entwicklung von Chematica beigetragen haben.

Zitierweise: *Angew. Chem. Int. Ed.* **2016**, *55*, 5904 – 5937
Angew. Chem. **2016**, *128*, 6004 – 6040

[1] P. Judson, *Knowledge-based Expert Systems in Chemistry: Not Counting on Computers*, RSC, Cambridge, **2009**.
[2] a) http://www.elsevier.com/online-tools/reaxys; b) https://scifinder.cas.org; c) http://www.chemspider.com; d) http://www.infochem.de/products/databases/index.shtml.
[3] a) M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2005**, *44*, 7263 – 7269; *Angew. Chem.* **2005**, *117*, 7429 – 7435; b) K. J. M. Bishop, R. Klajn, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2006**, *45*, 5348 – 5354; *Angew. Chem.* **2006**, *118*, 5474 – 5480; c) B. A.

Grzybowski, K. J. M. Bishop, B. Kowalczyk, C. E. Wilmer, *Nat. Chem.* **2009**, *1*, 31 – 36.
[4] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31 – 36.
[5] a) E. J. Corey, W. J. Howe, R. D. Cramer, *J. Am. Chem. Soc.* **1972**, *94*, 421 – 430; b) http://cheminf.cmbi.ru.nl/cheminf/olp/history.shtml.
[6] D. A. Evans, *Angew. Chem. Int. Ed.* **2014**, *53*, 11140 – 11146; *Angew. Chem.* **2014**, *126*, 11320 – 11325.
[7] P. Y. Johnson, I. Bernstein, J. Crary, M. Evans, T. Wang, *Designing an Expert System for Organic Synthesis in Expert Systems Application in Chemistry* (Hrsg.: B. A. Holme, H. Pierce), ACS Symposium Series, Am. Chem. Soc. Washington, **1989**.
[8] a) H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer, J. E. Searleman, *Science* **1977**, *197*, 1041 – 1049; b) D. Krebsbach, H. Gelernter, S. M. Sieburth, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 595 – 604.
[9] a) S. Hanessian, J. Franco, B. Larouche, *Pure Appl. Chem.* **1990**, *62*, 1887 – 1910; b) S. Hanessian, *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 798 – 819.
[10] S. Steinerberger, *Int. J. Game Theory*, **2015**, *44*, 761 – 767.
[11] L. V. Allis, *Searching for Solutions in Games and Artificial Intelligence*, Dissertation, University of Limburg, Maastricht, **1994**, S. 171.
[12] R. E. Korf, *Proc. AAAI'97/IAAI'97*, **1997**, S. 700 – 705.
[13] Schätzung basierend auf der Anwendung der vollständigen Reaktionsdatenbank von Syntaurus auf Verbindungen aus 99 aus http://chemistrybydesign.oia.arizona.edu/ ausgewählten Synthesewegen; diese umfassten 2 bis 19 Stufen mit durchschnittlich 9 Stufen.
[14] http://www.cube20.org/.
[15] http://www.eisai.com/news/news201133.html.
[16] a) J. L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722 – 730; b) Q. Hu, Z. Peng, J. Kostrowicki, A. Kuki, *Methods Mol. Biol.* **2011**, *685*, 253 – 276.
[17] Der ursprüngliche Bericht von 1957 an die sowjetische Akademie der Wissenschaften wurde später revidiert, erweitert und in Englisch veröffentlicht: a) G. E. Vléduts, V. K. Finn, *Inf. Storage Retr.* **1963**, *1*, 101 – 116. Eine weitere Publikation aus dem selben Jahr: b) G. E. Vléduts, *Inf. Storage Retr.* **1963**, *1*, 117 – 146.
[18] https://vimeo.com/63343963.
[19] http://www.cas.org/etrain/scifinder/sciplanner.html.
[20] a) D. J. Watts, S. H. Strogatz, *Nature* **1998**, *393*, 440 – 442; b) M. Girvan, M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7821 – 7826.
[21] a) A. L. Barabási, Z. N. Oltvai, *Nat. Rev. Genet.* **2004**, *5*, 101 – U15; b) E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A. L. Barabasi, *Science* **2002**, *297*, 1551 – 1555.
[22] C. Chaouiya, *Briefings Bioinf.* **2007**, *8*, 210 – 219.
[23] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, B. O. Palsson, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 1777 – 1782.
[24] a) R. Albert, A. L. Barabasi, *Rev. Mod. Phys.* **2002**, *74*, 47 – 97; b) R. Albert, H. Jeong, A. L. Barabasi, *Nature* **1999**, *401*, 130 – 131; c) A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, *Comput. Netw.* **2000**, *33*, 309 – 320; d) M. Faloutsos, P. Faloutsos, C. Faloutsos, *Comput. Commun. Rev.* **1999**, *29*, 251 – 262; e) H. Jeong, B. Tombor, R. Albert, Z. N. Oltval, A. L. Barabasi, *Nature* **2000**, *407*, 651 – 654; f) F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Aberg, *Nature* **2001**, *411*, 907 – 908.
[25] a) M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski, K. J. M. Bishop, *Angew. Chem. Int. Ed.* **2012**, *51*, 7928 – 7932; *Angew. Chem.* **2012**, *124*, 8052 – 8056; b) C. M. Gothard, S. Soh, N. A. Gothard, B. Kowalczyk, Y. H. Wei, B. Baytekin, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2012**, *51*, 7922 – 7927; *Angew. Chem.* **2012**, *124*, 8046 – 8051; c) P. E. Fuller, C. M. Gothard, N. A. Gothard, A. Weckiewicz, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2012**, *51*, 7933 – 7937; *Angew. Chem.* **2012**, *124*, 8057 – 8061.
[26] a) L. C. Lee, *IRE Trans. Electron. Comput.* **1961**, *EC10*, 346 – 365; b) S. Skiena, *The Algorithm Design Manual*, Springer, London, **2008**, S. 480.
[27] a) S. Even, *Graph Algorithms*, 2. Aufl., Cambridge Univ. Press, Cambridge, **2011**, S. 46 – 48; b) K. Mehlhorn, P. Sanders, *Algorithms and Data Structures: The Basic Toolbox*, Springer, Berlin, Heidelberg, **2008**.
[28] a) J. J. Masters, S. J. Danishefsky, J. T. Link, L. B. Snyder, W. B. Young, *Angew. Chem. Int. Ed.* **1995**, *34*, 1723 – 1726; *Angew. Chem.* **1995**, *107*, 1886 – 1888; b) S. J. Danishefsky, J. J. Masters, W. B. Young, J. T. Link, L. B. Snyder, T. V. Magee, D. K. Jung, R. C. A. Isaacs, W. G. Bornmann, C. A. Alaimo, C. A. Coburn, M. J. Di Grandi, *J. Am. Chem. Soc.* **1996**, *118*, 2843 – 2859; c) P. Wieland, K. Miescher, *Helv. Chim. Acta* **1950**, *33*, 2215 – 2228; d) M. L. Miller, P. S. Ray, *Synth. Commun.* **1997**, *27*, 3991 – 3996; e) M. Colin, D. Guenard, F. Gueritte-Voegelein, P. Potier (Rhone-Poulenc Sante), US4924012, **1990**; f) B. Ganem, R. R. Franke, *J. Org. Chem.* **2007**, *72*, 3981 – 3987.
[29] P. T. Anastas, M. M. Kirchoff, *Acc. Chem. Res.* **2002**, *35*, 686 – 694.
[30] a) R. S. Vardanyan, V. J. Hruby, *Synthesis of Essential Drugs*, Elsevier, Amsterdam, **2006**, S. 46; b) D. Farge, V. Marne, M. N. Messer, E. Moutonnier, C. Moutonnier (Rhône-Poulenc S. A.), U.S. Pat. 3641127, **1972**; c) W. Li, J. Li, Z.-K. Wan, J. Wu, W. Massefski, *Org. Lett.* **2007**, *9*, 4607 – 4610; d) A. R. Hajipour, A. E. Ruoho, *Org. Prep. Proced. Int.* **2002**, *34*, 647 – 651; e) A. A. Jalil, N. Kurono, M. Tokuda, *Synthesis* **2002**, *18*, 2681 – 2686; f) M. Allegretti, R. Bertini, M. C. Cesta, C. Bizzarri, R. D. Bitondo, V. D. Ciocco, E. Galliera, V. Berdini, A. Topai, G. Zampella, V. Russo, N. D. Bello, G. Nano, L. Nicolini, M. Locati, P. Fantucci, S. Florio, F. Colotta, *J. Med. Chem.* **2005**, *48*, 4312 – 4331; g) K. Mahesh, WO2013168001, **2013**; h) L. Baiocchi, M. Giannangeli, M. Bonanomi, G. Picconi, P. Ridolfi, *Gazz. Chim. Ital.* **1985**, *115*, 199 – 216; i) R. Ugo, P. Nardi, R. Psaro, D. Roberto, *Gazz. Chim. Ital.* **1992**, *122*, 511 – 514.
[31] a) R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, *Artif. Intell.* **1993**, *61*, 209 – 261; b) I. Ugi, J. Bauer, K. Bley, A. Dengler, A. Dietz, E. Fontain, B. Gruber, R. Herges, M. Knauer, K. Reitsam, N. Stein, *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 201 – 227; *Angew. Chem.* **1993**, *105*, 210 – 239.
[32] E. J. Corey, *Pure Appl. Chem.* **1967**, *14*, 19 – 37.
[33] a) E. J. Corey, X. M. Cheng, *The Logic of Organic Synthesis*, Wiley, New York, **1989**; b) S. Warren, P. Wyatt, *Organic Synthesis: The Disconnection Approach*, Wiley, Chichester, **2008**; c) K. C. Nicolaou, D. Vourloumis, N. Winssinger, P. S. Baran, *Angew. Chem. Int. Ed.* **2000**, *39*, 44 – 122; *Angew. Chem.* **2000**, *112*, 46 – 126; d) M. H. Todd, *Chem. Soc. Rev.* **2005**, *34*, 247 – 266.
[34] E. J. Corey, W. T. Wipke, *Science* **1969**, *166*, 178 – 192.
[35] a) W. T. Wipke, W. J. Howe, Computer-Assisted Organic Synthesis ACS Centennial Meeting, New York, **1976**; b) W. T. Wipke, G. I. Ouchi, S. Krishnan, *Artif. Intell.* **1978**, *11*, 173 – 193.
[36] a) J. B. Hendrickson, A. G. Toczko, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 137 – 145; b) J. B. Hendrickson, *J. Am. Chem. Soc.* **1977**, *99*, 5439 – 5450; c) J. B. Hendrickson, P. Huang, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 145 – 151; d) J. B. Hendrickson, *Angew. Chem. Int. Ed. Engl.* **1990**, *29*, 1286 – 1295; *Angew. Chem.* **1990**, *102*, 1328 – 1338.
[37] Andere Arten von Retrosyntheseprogrammen, die wir identifizieren konnten, im Haupttext aber nicht erklärt haben: Synthesezugänglichkeit von organischen Verbindungen: a) SILVIA: http://www.molecular-networks.com/products/sylvia; b) F. Pennerath, G. Niel, P. Vismara, P. Jauffret, C. Laurenço, A. Napoli, *J. Chem. Inf. Model.* **2010**, *50*, 221 – 239; Systeme, die auf der Vorhersage von Reaktionsprodukten basieren: c) J. H. Chen, P. Baldi, *J. Chem. Inf. Model.* **2009**, *49*, 2034 – 2043; d) BEPPE: G.

Sello, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 713 – 717; e) ROBIA: I. M. Socorro, J. M. Goodman, *J. Chem. Inf. Model.* **2006**, *46*, 606 – 614; f) I. M. Socorro, K. Taylor, J. M. Goodman, *Org. Lett.* **2005**, *7*, 3541 – 3544; g) LILITH, über die Bindungspolarität gesteuert: L. Baumer, G. Sala, G. Sello, *J. Am. Chem. Soc.* **1991**, *113*, 2494 – 2500; h) SYNSUP, berücksichtigt Wechselwirkungen zwischen Reagens und funktionellen Gruppen: A. Tanaka, H. Okamoto, M. Bersohn, *J. Chem. Inf. Model.* **2010**, *50*, 327 – 338; i) CROSS, berücksichtig stabile Funktionalisierung von Seitenketten und Gerüstumbau: A. Evers, G. Hessler, L. H. Wang, S. Werrel, P. Monecke, H. Matter, *J. Med. Chem.* **2013**, *56*, 4656 – 4670; j) ELN-Mining-Software: C. D. Christ, M. Zentgraf, J. M. Kriegl, *J. Chem. Inf. Model.* **2012**, *52*, 1745 – 1756.

[38] a) J. Bauer, R. Herges, E. Fontain, I. Ugi, *Chimia* **1985**, *39*, 43 – 53; b) *Cheminformatics Developments*: *History*, *Reviews and Current Research* (Hrsg.: J. H. Noordik), IOS Press, Amsterdam, **2004**.

[39] http://www.cogsys.wiai.uni-bamberg.de/effalip/installguide.html.

[40] a) „The Prediction of Chemical Reactions“: J. Gasteiger in *Chemoinformatics. A Textbook* (Hrsg.: J. Gasteiger, T. Engel), Springer, Heidelberg, **1990**, S. 542 – 567; b) R. Höllering, J. Gasteiger, L. Steinhauer, K. Schultz, A. Herwig, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 482 – 494.

[41] a) O. Ravitz, *Drug Discovery Today Technol.* **2013**, *10*, e443 – e449; b) J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, H. Y. Ando, *J. Chem. Inf. Model.* **2009**, *49*, 593 – 602.

[42] a) H. Kraut, J. Eiblmaier, G. Grethe, P. Loew, H. Matuszczyk, H. Saller, *J. Chem. Inf. Model.* **2013**, *53*, 2884 – 2895; b) A. Bøgevig, H. J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Loew, C. Oppawsky, T. Rein, H. Saller, *Org. Process Res. Dev.* **2015**, *19*, 357 – 368.

[43] K. O. Geddes, S. R. Czapor, G. Labahn, *Algorithms for Computer Algebra*, Springer US, **1992**.

[44] A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2014**, *53*, 8108 – 8112; *Angew. Chem.* **2014**, *126*, 8246 – 8250.

[45] a) SMARTS theory manual, Daylight Chemical Information Systems Inc., Aliso Viejo, CA 92656, USA. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, Zugriff am 23. Juni 2015; b) http://www.rdkit.org.

[46] F. A. Van-Catledge, *J. Org. Chem.* **1980**, *45*, 4801 – 4802.

[47] a) F. D. Da Silva Araújo, L. C. De Lima Fávaro, W. L. Araújo, F. L. De Oliveira, R. Aparicio, A. J. Marsaioli, *Eur. J. Org. Chem.* **2012**, *27*, 5225 – 5230; b) P. Ellerbrock, N. Armanino, D. Trauner, *Angew. Chem. Int. Ed.* **2014**, *53*, 13414 – 13418; *Angew. Chem.* **2014**, *126*, 13632 – 13636; c) S. Sang, J. D. Lambert, S. Tian, J. Hong, Z. Hou, J. H. Ryu, R. E. Stark, R. T. Rosen, M. T. Huang, C. S. Yang, et al., *Bioorg. Med. Chem.* **2004**, *12*, 459 – 467; d) C. B. Rao, D. C. Rao, D. C. Babu, Y. Venkateswarlu, *Eur. J. Org. Chem.* **2010**, *2010*, 2855 – 2859; e) G. Casiraghi, L. Battistini, C. Curti, G. Rassu, F. Zanardi, *Chem. Rev.* **2011**, *111*, 3076 – 3154.

[48] Ausbeuteschätzungen beruhen auf Thermodynamikrechnungen zusammen mit mehrdimensionaler Optimierung. Dieses Modell ist beschrieben in: F. S. Emami, A. Vahid, W. K. Wylie, S. Szymkuc, P. Dittwald, K. Molga, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2015**, *54*, 10797 – 10801; *Angew. Chem.* **2015**, *127*, 10947 – 10951.

[49] a) K. H. Lim, V. J. Raja, T. D. Bradshaw, S.-H. Lim, Y.-Y. Low, T.-S. Kam, *J. Nat. Prod.* **2015**, *78*, 1129 – 1138; b) X. Zhong, Y. Li, F. S. Han, *Chem. Eur. J.* **2012**, *18*, 9784 – 9788; c) H. Iio, M. Monden, K. Okada, T. Tokoroyama, *J. Chem. Soc. Chem. Commun.* **1987**, 358 – 359; d) D. Enders, T. Hundertmark, R. Lazny, *Synlett* **1998**, 721 – 722; e) A. L. Gutman, M. Etinger, G. Nisnevich, F. Polyak, *Tetrahedron: Asymmetry* **1998**, *9*, 4369 – 4379; f) M. A. Potopnyk, B. Lewandowski, S. Jarosz, *Tetrahedron: Asymmetry* **2012**, *23*, 1474 – 1479.

[50] a) Y. H. Lan, F.-R. Chang, Y. L. Yang, Y. C. Wu, *Chem. Pharm. Bull.* **2006**, *54*, 1040 – 1043; b) J. S. Yadav, N. Rami Reddy, V. Harikrishna, B. V. Subba Reddy, *Tetrahedron Lett.* **2009**, *50*, 1318 – 1320; c) M. Venkataiah, P. Somaiah, G. Reddipalli, N. W. Fadnavis, *Tetrahedron: Asymmetry* **2009**, *20*, 2230 – 2233; d) J. Li, H. Zheng, Y. Su, X. Xie, X. She, *Synlett* **2010**, *15*, 2283 – 2284; e) J. S. Yadav, R. Nageshwar Rao, R. Somaiah, V. Harikrishna, B. V. Subba Reddy, *Helv. Chim. Acta* **2010**, *93*, 1362 – 1368; f) G. S. Forman, R. P. Tooze, *J. Organomet. Chem.* **2005**, *690*, 5863 – 5866; g) N. Yoshikawa, N. Kumagai, S. Matsunaga, G. Moll, T. Ohshima, T. Suzuki, M. Shibasaki, *J. Am. Chem. Soc.* **2001**, *123*, 2466 – 2467; h) J. H. Xie, L. C. Guo, X. H. Yang, L. X. Wang, Q. L. Zhou, *Org. Lett.* **2012**, *14*, 4758 – 4761.

[51] a) J. Manville, C. Kriz, *Can. J. Chem.* **1977**, *55*, 2547 – 2553; b) R. G. Almquist, J. Crase, C. Jennings-White, R. F. Meyer, M. L. Hoefle, R. D. Smith, A. D. Essenburg, H. R. Kaplan, *J. Med. Chem.* **1982**, *25*, 1292 – 1299; c) F. Dehmel, H. G. Schmalz, *Org. Lett.* **2001**, *3*, 3579 – 3582; d) D. J. Chang, S. Lee, J. Jang, S. O. Kim, W. J. Kim, Y. G. Suh, *Bioorg. Med. Chem. Lett.* **2012**, *22*, 6750 – 6755.

[52] a) S. E. Drewes, B. M. Sehlapelo, M. M. Horn, R. Scott-Shaw, P. Sandor, *Phytochemistry* **1995**, *38*, 1427 – 1430; b) M. B. Boxer, H. Yamamoto, *J. Am. Chem. Soc.* **2007**, *129*, 2762 – 2763; c) P. R. Krishna, V. V. R. Reddy, *Tetrahedron Lett.* **2005**, *46*, 3905 – 3907; d) A. Martinez, K. Zumbansen, A. Dohring, M. van Gemmeren, B. List, *Synlett* **2014**, *25*, 932 – 934; e) K. Liu, G. Zhang, *Tetrahedron Lett.* **2015**, *56*, 243 – 246; f) B. List, P. Pojarliev, C. Castello, *Org. Lett.* **2001**, *3*, 573 – 575; g) M. A. Blanchette, M. S. Malamas, M. H. Nantz, J. C. Roberts, P. Somfai, D. C. Whritenour, S. Masamune, M. Kageyama, T. Tamura, *J. Org. Chem.* **1989**, *54*, 2817 – 2825; h) Y. Yamaoka, H. Yamamoto, *J. Am. Chem. Soc.* **2010**, *132*, 5354 – 5356; i) R. Mahrwald, *Modern Methods in Stereoselective Aldol Reactions*, Wiley-VCH, Weinheim, **2013**; für die im Text genannten Schwierigkeiten schlagen wir zwei Lösungsmöglichkeiten vor. Zum einen die Trennung der erhaltenen 1,5-*syn*- und 1,5-*anti*-Diastereomere und die Umwandlung des unerwünschten *anti*- in das *syn*-Derivat durch Mitsunobu-Inversion entsprechend Lit. [52j – l]; j) S. F. Martin, J. A. Dodge, *Tetrahedron Lett.* **1991**, *32*, 3017 – 3020; k) T. Sammakia, J. S. Jacobs, *Tetrahedron Lett.* **1999**, *40*, 2685 – 2688; l) J-M. Vatèle, *Tetrahedron* **2007**, *63*, 10921 – 10929; zum anderen könnte die Reihenfolge der beiden Schritte vertauscht (d. h., die Acylierung mit Acryoylchlorid erfolgt nach der Aldolreaktion mit anschließendem Schützen/Entschützen der Hydroxygruppe) und die 1,5-*syn*-Selektivität über die „Supersilyl“-Schutzgruppe erzwungen werden, siehe Lit. [52b,h]); m) K. M. Chen, G. E. Hardtmann, K. Prasad, O. Repič, M. J. Shapiro, *Tetrahedron Lett.* **1987**, *28*, 155 – 158.

[53] a) J. Li, M. D. Eastgate, *Org. Biomol. Chem.* **2015**, *13*, 7164 – 7176; b) M. Randić, *J. Am. Chem. Soc.* **1975**, *97*, 6609 – 6615; c) M. Randić, *J. Am. Chem. Soc.* **1977**, *99*, 444 – 450; d) T. Gaich, P. S. Baran, *J. Org. Chem.* **2010**, *75*, 4657 – 4673; e) K. Boda, T. Seidel, J. Gasteiger, *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311 – 325; f) J. Gasteiger, *Nat. Chem.* **2015**, *7*, 619 – 620; g) P. Ertl, A. Schuffenhauer, *J. Cheminf.* **2009**, *1*, 8.

[54] B. A. Grzybowski, A. V. Ishchenko, J. Shimada, E. I. Shakhnovich, *Acc. Chem. Res.* **2002**, *35*, 261 – 269.

[55] a) L. E. Overman, D. J. Ricca, V. D. Tran, *J. Am. Chem. Soc.* **1997**, *119*, 12031 – 12040; b) J. C. McWilliams, J. Clardy, *J. Am. Chem. Soc.* **1994**, *116*, 8378 – 8379; c) A. K. Miller, C. C. Hughes, J. J. Kennedy-Smith, S. N. Gradl, D. Trauner, *J. Am. Chem. Soc.* **2006**, *128*, 17057 – 17062; d) S. Roesner, J. M. Casatejada, T. G. Elford, R. P. Sonawane, V. K. Aggarwal, *Org. Lett.* **2011**, *13*, 5740 – 5743.

[56] a) R. W. Taft, *J. Am. Chem. Soc.* **1952**, *74*, 3120 – 3128; b) E. Estrada, E. Molina, L. I. Perdomo, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1015 – 1021; c) O. Ivanciuc, A. T. Balaban, *Croat. Chem. Acta* **1996**, *69*, 75 – 83; d) C. Cao, L. Liu, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 678 – 687; e) https://www.chemaxon.com/products/marvin/; f) G. Sello, *Tetrahedron* **1998**, *54*, 5731 – 5744; g) N. Weill, C. R. Corbell, J. W. De Schutter, N. Moitessier, *J. Comput. Chem.* **2011**, *32*, 2878 – 2889; h) C. R. Corbeil, S. Thielges, J. A. Schwartzentruber, N. Moitissier, *Angew. Chem. Int. Ed.* **2008**, *47*, 2635 – 2638; *Angew. Chem.* **2008**, *120*, 2675 – 2678.

[57] a) A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov′ev, *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693 – 703; b) A. de Luca, D. Horvath, G. Marcou, V. P. Solov′ev, A. Varnek, *J. Chem. Inf. Model.* **2012**, *52*, 2325 – 2338; c) T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, A. V. Bodrov, A. I. Lin, I. I. Baskin, I. S. Antipin, A. Varnek, *Russ. J. Org. Chem.* **2014**, *50*, 459 – 463; d) G. Marcou, J. Aires de Sousa, D. Latino, A. Deluca, D. Horvath, V. Rietsch, A. Varnek, *J. Chem. Inf. Model.* **2015**, *55*, 239 – 250.

[58] a) J. Li, S. G. Balmer, E. P. Gillis, S. Fuji, M. J. Schmidt, A. M. E. Palazzolo, J. W. Lehmann, G. F. Morehouse, M. D. Burke, *Science* **2015**, *347*, 1221 – 1226; b) D. Ghislieri, K. Gilmore, P. H. Seeberger, *Angew. Chem. Int. Ed.* **2015**, *54*, 678 – 682; *Angew. Chem.* **2015**, *127*, 688 – 692; c) K. S. Elvira, X. C. I. Solvas, R. C. R. Wootton, A. J. deMello, *Nat. Chem.* **2013**, *5*, 905 – 915; d) T. Kourti, *Anal. Bioanal. Chem.* **2006**, *384*, 1043 – 1048; e) R. L. Hartman, K. F. Jensen, *Lab Chip* **2009**, *9*, 2495 – 2507; f) S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, R. M. Meyers, *Angew. Chem. Int. Ed.* **2015**, *54*, 3449 – 3464; *Angew. Chem.* **2015**, *127*, 3514 – 3530; g) S. V. Ley, D. E. Fitzpatrick, R. M. Meyers, R. J. Ingham, C. Battilocchio, R. J. Ingham, *Angew. Chem. Int. Ed.* **2015**, *54*, 10122 – 10136; *Angew. Chem.* **2015**, *127*, 10260 – 10275; h) B. Gutmann, D. Cantillo, C. O. Kappe, *Angew. Chem. Int. Ed.* **2015**, *54*, 6688 – 6728; *Angew. Chem.* **2015**, *127*, 6788 – 6832. i) Das Angebot für das Programm Make-It von DARPA kann heruntergeladen werden von http://go.usa.gov/3Pzww.

[59] a) Zur Dokumentierung des Reaction Mechanism Generator siehe http://rmg.mit.edu; b) M. R. Harper, K. M. Van Geem, S. P. Pyl, G. B. Marin, W. H. Green, *Combust. Flame* **2011**, *158*, 16 – 41; c) V. Warth, F. Battin-Leclerc, R. Fournet, P. A. Glaude, G. M. Côme, G. Scacchi, *Comput. Chem.* **2000**, *24*, 541 – 560.

[60] a) D. Rappoport, C. J. Galvin, D. Y. Zubarev, A. Aspuru-Guzik, *J. Chem. Theory Comput.* **2014**, *10*, 897 – 907; b) D. E. Levy, *Arrow-Pushing in Organic Chemistry. An Easy Approach to Understanding Reaction Mechanisms*, Wiley, Hoboken, **2008**; c) G. Knizia, J. E. M. N. Klein, *Angew. Chem. Int. Ed.* **2015**, *54*, 5518 – 5522; *Angew. Chem.* **2015**, *127*, 5609 – 5613.

[61] a) A. Boutlerow, *C. R. Acad. Sci.* **1861**, *53*, 145 – 147; b) T. Zweckmair, S. Böhmdorfer, A. Bogolitsyna, T. Rosenau, A. Potthast, S. Novalin, *J. Chromatogr. Sci.* **2014**, *52*, 169 – 175.

Eingegangen am 3. Juli 2015,
veränderte Fassung am 14. September 2015
Online veröffentlicht am 8. April 2016
Übersetzt von Dr. Kathrin-Maria Roy, Langenfeld

