

INSTYTUT CHEMII ORGANICZNEJ  
POLSKIEJ AKADEMII NAUK



Instytut Chemii Organicznej  
Polskiej Akademii Nauk

**ROZPRAWA DOKTORSKA**

**Rozszerzenie zakresu narzędzi automatycznej retrosyntezy  
do projektowania syntez unikających patentów,  
prowadzących do związków znakowanych izotopowo  
oraz projektowania syntez bibliotek związków.**

mgr inż. Karol Molga

Monotematyczny cykl publikacji z komentarzem przedstawiony  
Radzie Naukowej Instytutu Chemii Organicznej Polskiej Akademii Nauk  
w celu uzyskania stopnia doktora nauk chemicznych

A-21-6  
K-C-130  
K-C-121

Promotor: prof. Bartosz A. Grzybowski



Biblioteka Instytutu Chemii Organicznej PAN

**O-B.455/24**



10000000116425

Warszawa, 2023

<https://rcin.org.pl>



Pragnę serdecznie podziękować

**Profesorowi Bartoszowi Grzybowskiemu**  
za wskazanie drogi naukowej, poświęcony czas  
oraz wszystkie udzielone wskazówki

wszystkim **Koleżankom i Kolegom** z zespołu XI,  
za dyskusje naukowe i lata wspólnej pracy

mojej rodzinie, za wsparcie od najmłodszych lat



Badania do pracy doktorskiej zostały wykonane w ramach projektu:



Defence Advanced Research Projects Agency "Make-It" Award,  
69461-CH-DRP #W911NF1610384

## Spis treści

1. Lista publikacji	- 7 -
Publikacje wchodzące w skład rozprawy doktorskiej	- 7 -
Publikacje nie wchodzące w skład rozprawy doktorskiej	- 8 -
Konferencje	- 9 -
2. Streszczenie w języku polskim	- 10 -
3. Abstract in English	- 12 -
4. Wprowadzenie	- 14 -
5. Założenia i cel pracy	- 16 -
6. Komputerowe planowanie syntez chemicznych – rys historyczny	- 16 -
7. Reguły reaktywności chemicznej	- 19 -
8. Algorytmy	- 22 -
8.1. Wybór rokujących pozycji i rozwijanie grafu	- 22 -
8.2. Analiza ekonomiczna planów syntetycznych	- 24 -
8.3. Planowanie z wykorzystaniem kilkukrokowych sekwencji reakcji	- 28 -
9. Znajdowanie ścieżek syntetycznych omijających patenty	- 30 -
10. Plany syntetyczne prowadzące do bibliotek związków	- 35 -
11. Plany syntetyczne prowadzące do związków znakowanych izotopowo	- 40 -
12. Automatyzacja syntezy bibliotek związków – chemia iteracyjna	- 43 -
13. Podsumowanie	- 48 -
14. Referencje	- 49 -
15. Oświadczenia autorów prac	- 54 -
16. Publikacje oryginalne	- 68 -

## 1. Lista publikacji

### Publikacje wchodzące w skład rozprawy doktorskiej

**P01** Molga, K.; Gajewska, E. P.; Szymkuć, S.; Grzybowski\*, B. A. The Logic of Translating Chemical Knowledge into Machine – Processable Forms: A Modern Playground for Physical-Organic Chemistry. *React. Chem. Eng.* **2019**, *4*, 1506–1521.

**P02** Grzybowski\*, B. A.; Badowski, T.; Molga, K.; Szymkuć, S. Network Search Algorithms and Scoring Functions for Advanced-level Computerized Synthesis Planning. *WIREs Comput. Mol. Sci.* **2022**, e1630.

**P03** Molga, K.; Szymkuć, S.; Grzybowski\*, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094–1106. **Okladka**

**P04** Badowski, T.; Molga, K.; Grzybowski\*, B. A. Selection of Cost-Effective yet Chemically Diverse Pathways from the Networks of Computer-Generated Retrosynthetic Plans. *Chem. Sci.* **2019**, *10*, 4640–4651.

**P05** Molga, K.; Dittwald, P.; Grzybowski\*, B. A. Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways. *Chem* **2019**, *5*, 460–473.

**P06** Szymkuć, S.; Gajewska, E.; Molga, K.; Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated “Synthetic Contingency” Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

**P07** Molga, K.; Dittwald, P.; Grzybowski\*, B. A. Computational Design of Syntheses Leading to Compound Libraries or Isotopically Labelled Targets. *Chem. Sci.* **2019**, *10*, 9219-9232. **Okladka**

**P08** Molga, K.; Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Mlynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58. **Okladka**

## Publikacje nie wchodzące w skład rozprawy doktorskiej

**P9** Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; **Molga\*, K.**; Młynarski\*, J.; Mrksich\*, M.; Grzybowski\*, B. A. Computational Planning of the Synthesis of Complex Natural Products. *Nature* **2020**, *588*, 83–88.

**P10** Badowski, T.; Gajewska, E. P.; **Molga, K.**; Grzybowski, B. A. Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angew. Chem. Int. Ed.* **2020**, *59*, 725–730.

**P11** Roszak, R.; Beker, W.; **Molga, K.**; Grzybowski, B. A. Rapid and Accurate Prediction of pKa Values of C–H Acids Using Graph Convolutional Neural Networks. *J. Am. Chem. Soc.* **2019**, *141*, 17142–17149.

**P12** Grzybowski, B. A.; Szymkuć, S.; Gajewska, E. P.; **Molga, K.**; Dittwald, P.; Wołos, A.; Klucznik, T. Chematica: A Story of Computer Code That Started to Think like a Chemist. *Chem* **2018**, *4*, 390–398.

**P13** Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; **Molga, K.**; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Touthkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *4*, 522–532. **Okladka**

**P14** Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; **Molga, K.**; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937.

**P15** Emami, F. S.; Vahid, A.; Wylie, E. K.; Szymkuc, S.; Dittwald, P.; **Molga, K.**; Grzybowski, B. A. A Priori Estimation of Organic Reaction Yields. *Angew. Chem. Int. Ed.* **2015**, *54*, 10797–10801.

**P16** Buchalski, P.; Pacholski, R.; Gustowski, J.; Buchowicz, W.; **Molga, K.**; Shkurenko, A.; Suwińska, K. Bis-Nickel-Bridged *p*-Terphenyl Dianion - Synthesis and Structures. *J. Organomet. Chem.* **2015**, *789-790*, 40–45.



## Konferencje

### Wystąpienie w formie posteru, autor prezentujący:

*Chess - like algorithms behind Chematica's retrosynthetic planning*

**Molga, K.**; Szymkuć, S. A.; Gajewska, E. P.; Klucznik, T.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A.

IUPAC 2015 45th World Chemistry Congress, 09 – 14.08.2015, Busan, Korea  
Południowa

### Wystąpienie w formie wykładu, autor prezentujący:

*Navigating around patented routes with the help of computer-driven retrosynthetic analysis.* **Molga, K.**; Dittwald, P.; Grzybowski, B. A.

256th American Chemical Society National Meeting, 19 – 23.08.2018, Boston, USA.

## 2. Streszczenie w języku polskim

Planowanie wieloetapowych syntez skomplikowanych związków organicznych jest złożonym zadaniem, wymagającym ogromnej wiedzy i często kreatywnego podejścia. Pierwsze próby wykorzystania komputerów do pomocy chemikom sięgają lat 60-tych XX wieku, jednak żaden z opracowanych algorytmów nie posiadał wystarczająco szerokiej i dokładnej „wiedzy” o regułach reakcji, algorytmów pozwalających na w pełni zautomatyzowane przeszukiwanie ogromnych sieci reakcji ani funkcji pozwalających na ocenę postępu takich poszukiwań. W najlepszym przypadku półautomatyczne programy, takie jak LHASA E. J. Coreya, sugerowały dostępne cięcia retrosyntetyczne krok po kroku, pozostawiając proces faktycznego projektowania syntez użytkownikowi.

Moja praca doktorska była w dużej mierze skoncentrowana na rozwoju programu Chematica - obecnie dystrybuowanego na całym świecie przez Millipore-Sigma/Merck. System ten łączy w sobie reguły reaktywności chemicznej zakodowane na poziomie eksperckim z algorytmami sztucznej inteligencji, za pomocą których można przeszukiwać sieci reakcji. Moim początkowym wkładem było rozszerzenie bazy wiedzy chemicznej programu o około 50 000 zakodowanych reguł reakcji, głównie tych dotyczących stereokontrolowanej syntezy złożonych produktów naturalnych. Następnie, pracowałem nad algorytmami za pomocą których można oszacować realistyczne koszty zaprojektowanych komputerowo planów syntez uwzględniając nie tylko koszty substratów, ale także szacowane wydajności, koszty poszczególnych etapów i ogólną strukturę ścieżki (liniowa vs. zbieżna). Wykazałem, że zastosowanie tych funkcji oceny kosztów może być wykorzystane do projektowania zarówno syntez małoskalowych, prowadzonych np. na początkowych etapach badań klinicznych (gdzie pożądanym jest możliwie krótki plan syntezy), jak i procesów prowadzonych w większej skali (które dążą do minimalizacji cen substratów).

Kolejna część moich badań dotyczyła opracowania algorytmów do projektowania syntez, które omijają opatentowane rozwiązania. W tym celu opracowałem metody oznaczania kluczowych dyskoneksji użytych w opatentowanych syntezach, a następnie priorytetyzowania alternatywnych rozwiązań. Pokazałem, w jaki sposób ta metodologia może być wykorzystywana do znajdowania luk w syntezach leków, które są chronione patentami w najdrobniejszych szczegółach. Praca ta miała dwie uzupełniające się motywacje: z jednej strony umożliwienie producentom chemikaliów lepszej ochrony ich wynalazków, a z drugiej strony umożliwienie producentom leków generycznych (często

z biedniejszych krajów) znalezienia alternatywnych rozwiązań bez naruszania międzynarodowych przepisów handlowych.

Mój kolejny projekt miał na celu umożliwienie Chematicie projektowania syntezy wielu związków jednocześnie, prowadząc do całych bibliotek związków (np. analogów leków). Zaprojektowałem algorytmy pozwalające na planowanie takich syntez z maksymalnym wykorzystaniem wspólnych półproduktów. Pozwoliło to uczynić "globalne" plany syntezy bardziej ekonomicznymi, a czas obliczeń takich planów był znacznie krótszy. Następnie rozszerzyłem te algorytmy na ważną z przemysłowego punktu widzenia sytuację, w której biblioteki izotopowo znakowanych związków mają być syntetyzowane w najbardziej opłacalny sposób. W szczególności, stworzyłem metodę generowania bibliotek izotopomerów, a następnie wybierania z tej biblioteki związku, który jest najłatwiej syntetyzowalny.

Ostatni projekt zrealizowany w ramach mojej pracy doktorskiej dotyczył innego aspektu efektywnego prowadzenia wieloetapowej syntezy organicznej - zdolności do projektowania ścieżek do złożonych związków docelowych przy użyciu tylko kilku iteracyjnie stosowanych typów reakcji. Takie prowadzenie syntezy - uhonorowane nagrodą Nobla za iteracyjną syntezę peptydów i oligonukleotydów - ma ogromne znaczenie dla automatyzacji i robotyzacji syntezy. Liczba znanych iteracyjnych sekwencji reakcji jest niestety bardzo ograniczona, co utrudnia postęp w tej dziedzinie. W związku z tym, postanowiłem opracować algorytm, który wykorzystywałby szeroką wiedzę reguł programu Chematica do automatycznego odkrywania dużej liczby sekwencji iteracyjnych. Ostatecznie, algorytm odkrył tysiące nowych iteracyjnych metodologii syntetycznych, poczynszy od syntezy związków policyklicznych, a skończywszy na syntezie złożonych układów centrów stereogenicznych. Kilka z tych iteracji zostało poddanych zweryfikowanych eksperymentalnie, a sam artykuł - pod wieloma względami wieńczący moje badania - został opublikowany jako artykuł okładkowy w inauguracyjnym numerze *Nature Synthesis*.

### 3. Abstract in English

Planning of multistep syntheses of non-trivial organic molecules is generally a very complex task, requiring immense expertise and often creative insight. First attempts to teach computers to assist in this effort date back to 1960s. However, none of the algorithms offered a broad and accurate enough "knowledge" of reaction rules, algorithms allowing for fully automated searches over the immense networks of retrosynthetic possibilities, or the scoring functions to evaluate progress of such searches. At best, the semi-automatic programs like E.J. Corey's LHASA suggested synthetic options one-step-at-a-time – leaving the process of actual synthesis planning and route design to the human user.

My doctoral work was largely centered on the development of the Chematica program – now distributed worldwide by Millipore-Sigma/Merck. This system combines expert-level rules of chemical reactivity with AI algorithms with which to navigate retrosynthetic searches. My early contribution was extending the chemical knowledge base of Chematica by ca. 50,000 expert-coded reaction rules, mostly those for stereocontrolled synthesis of complex natural products. With this foundational contribution, I then worked on the development of algorithms with which to estimate realistic costs of the computer-designed routes – including not only costs of substrates but also estimated yields, costs of individual steps, and the overall structure of the pathway (linear vs. convergent). I showed that the use of these cost-estimating routines can be used to design and prioritize either small-scale syntheses used in discovery-oriented settings (focusing on synthetic conciseness), or larger-scale processes (which seek to minimize substrate prices).

The next part of my research centered on the development of algorithms to design syntheses that, while chemically correct, navigate around patented solutions. To do so, I designed methods to flag key disconnections used in patented syntheses and then to prioritize alternative solutions. I showed how this methodology can be used to find loopholes in syntheses of blockbuster drugs which, apparently, are patent-protected in exhaustive detail. This work has had two complementary motivations: on one hand, to allow chemical producers to better protect their synthetic inventions but, on the other hand, to allow producers of generic drugs (often from poorer countries) to find alternative solutions without violating international trade laws.

My next project aimed at enabling Chematica to synthesize multiple syntheses at one, leading to entire libraries of targets (e.g., drug analogs). I designed algorithms

to plan such syntheses with the maximal use of common intermediates. This allowed to render the “global” synthetic plans more economical and the calculations times of such plans to be significantly shorter. I then extended these algorithms to an industrially important situation in which libraries of isotopically-labelled compounds are to be synthesized in the most cost-effective manner. In particular, I also created a method to generate libraries of isotopomers, and then of selecting from this library the target that is most readily synthesizable.

The final project described in my thesis aimed at yet another aspect of synthetic efficiency – namely, the ability to design routes to complex targets using only few and iteratively applied types of reactions. Such iterative syntheses have been of immense importance for synthesis automation – two of the transformative and Nobel-winning examples are iterative, couple-deprotect cycles for peptide and oligonucleotide syntheses. Yet, the number of such iterative syntheses is limited, hampering progress of the field. Accordingly, I set out to design an algorithm that would use Chematica’s broad knowledge of synthetic chemistry to automatically discover large numbers of iterative sequences. In the end, this algorithm discovered thousands of new iterative methodologies, ranging from the syntheses of polycyclic compounds to the syntheses of complex stereoarrays. Several of these iterations were subsequently validated by experiment and the paper itself – in many ways crowning my doctoral research – was published as the cover article in the inaugural issue of *Nature Synthesis*.



## 4. Wprowadzenie

Przedstawiony *Komentarz* podzielony jest na dwie części poprzedzone krótkim rysem historycznym opisującym wcześniejsze próby opracowania programu umożliwiającego komputerowe planowanie syntez organicznych (**Rozdział 6**). W następnej kolejności, w **Rozdziałach 7** oraz **8** przedstawione są fundamenty działania naszego własnego oprogramowania Chematica/Synthia, które współtworzyłem w pierwszym etapie moich badań prowadzonych w ramach pracy doktorskiej: reguły reaktywności chemicznej oraz algorytmy, umożliwiające uzyskiwanie kompletnych planów syntetycznych prowadzących do dowolnych związków. W **Rozdziałach 9-12** omówione zostały wyniki mojej pracy badawczej umożliwiające zastosowanie komputerowo wspomaganey analizy retrosyntetycznej do projektowania syntez omijających patenty (**Rozdział 9**), planowania syntez bibliotek związków (**Rozdział 10**), planowania syntez związków znakowanych izotopowo (**Rozdział 11**) oraz planowania syntez bibliotek związków z użyciem narzędzi chemii iteracyjnej (**Rozdział 12**).

Planowanie wielokrokowych syntez prowadzących do oczekiwanych związków jest podstawowym problemem chemii organicznej. Aż do około połowy XX wieku, proces ten bazował na badaniu reaktywności poszczególnych cząsteczek i możliwych do uzyskania z nich związków. W tym czasie, repertuar dostępnych reagentów i możliwych reakcji był mocno ograniczony, a synteza totalna wymagała znalezienia związku o zbliżonej do związku docelowego strukturze. Przełomem, pozwalającym na efektywne planowanie syntez związków o bardzo skomplikowanej strukturze, w tym produktów naturalnych, była analiza retrosyntetyczna wprowadzona przez E. J. Coreya w latach 50-tych XX wieku i uhonorowana nagrodą Nobla w 1991 roku. W tym podejściu, chemik nie szukał podobnego strukturalnie materiału początkowego (mogącego zostać przekształconym w oczekiwany produkt) lecz poddawał związek docelowy szeregowi przekształceń odpowiadającym odwróconym reakcjom chemicznym. Na każdym etapie chemik decydował, które z obecnie dostępnych możliwości wydaje się najbardziej rokująca i może docelowo doprowadzić do dostępnych substratów<sup>1</sup>. Od samego początku, tj. od lat 60-tych XX wieku podejmowane były próby zautomatyzowania tego procesu poprzez stworzenie programów komputerowych, zwracających gotowe rozwiązanie syntetyczne prowadzące do pożądanego przez chemika związku. Mimo tego, że problemem zajmowali się wybitni chemicy – organicy (E. J. Corey<sup>2,3</sup>, C. Djerassi<sup>4</sup>, J. B. Hendrickson<sup>5</sup> czy I. Ugi<sup>6</sup>) przez ponad 50 lat żadne z zaproponowanych rozwiązań nie zostało szeroko zaadoptowane przez społeczność chemików organicznych: przyczyny tego stanu zostaną przeze mnie krótko omówione

w **Rozdziale 6** *Komentarza*, a bardziej szczegółową analizę czytelnik znajdzie w publikacjach *P14*, *P01* oraz *P02*.

Od początku swoich badań prowadzonych w ramach pracy doktorskiej współtworzyłem oprogramowanie Chematica/Synthia, które korzysta nie tylko z bezprecedensowej ilości dokładnych reguł reakcji chemicznych (omówionych w **Rozdziale 7** *Komentarza* i publikacji *P01*), ale i nowatorskich algorytmów (omówionych w **Rozdziale 8.1** i publikacjach *P02*, *P03*, *P14* oraz części **S6** publikacji *P13*) niedostępnych jeszcze kilkanaście lat temu a wykorzystujących wprowadzone przez nasz zespół funkcje oceny cząsteczek i reakcji chemicznych. Zastosowanie tych algorytmów pozwala Chematicie uniknąć tzw. eksplozji kombinatorycznej<sup>2</sup> i wybierać automatycznie najbardziej rokujące cięcia retrosyntetyczne w czasie poszukiwania planu syntetycznego. Każdy z tysięcy wygenerowanych planów syntetycznych jest następnie poddawany przez program szczegółowej analizie (opisanej w publikacji *P04* oraz rozdziale **8.2** *Komentarza*) uwzględniającej strukturę ścieżki (liniowa/zbieżna), koszty materiałów początkowych i szacunkowe wydajności planowanych reakcji co pozwala na efektywne planowanie syntez realizowanych w różnych scenariuszach ekonomicznych. Dodatkowo, jest to jedyne oprogramowanie, które ma możliwość wielokrokowego ‘myślenia’ charakterystycznego dla planowania syntez przez wybitnych chemików-organików dzięki wykorzystaniu wielokrokowych sekwencji „strategicznych” reakcji (opisanych w publikacji *P03* i *P09* oraz **Rozdziale 8.3** *Komentarza*). Oprogramowanie Chematica/Synthia zostało użyte w 2018 r. do zaprojektowania ścieżek syntetycznych do szeregu leków i ich metabolitów. W pełni automatycznie wygenerowane plany zostały następnie poddane walidacji eksperymentalnej<sup>P13</sup>. W każdym przypadku zaproponowane przez program rozwiązanie okazało się być efektywną metodą syntezy pożądanego związku, często przewyższając rozwiązania proponowane przez chemików. W 2020 r. program zaplanował<sup>P09</sup> stereoselektywne ścieżki (również zrealizowane eksperymentalnie) prowadzące do skomplikowanych produktów naturalnych: daurycyny, takamonidyny i lammelodysydyny A oraz został poddany testowi Turinga, w którym jego zdolność planowania syntez skomplikowanych produktów naturalnych okazała się zbliżona do umiejętności czołowych chemików-organików z całego świata.

## 5. Założenia i cel pracy

Mając do dyspozycji sprawdzone eksperymentalnie oprogramowanie pozwalające na przeprowadzanie analizy retrosyntetycznej dowolnych związków, w swojej pracy doktorskiej postanowiłem rozszerzyć możliwości komputera w planowaniu syntez o inne zadania ważne z punktu widzenia przede wszystkim chemików medycznych i firm farmaceutycznych. Pierwszym celem mojej pracy było wykazanie, że oprogramowanie, które do tej pory okazało się skuteczne w analizie retrosyntetycznej pojedynczych związków, może zostać użyte (publikacja **P07**) również do symultanicznego planowania syntez całych bibliotek związków. Dodatkowo, w publikacji **P08** poddałem analizie współtworzony przeze mnie zbiór reguł chemicznych, co pozwoliło na rozszerzenie repertuaru narzędzi chemii iteracyjnej, pozwalającej na automatyzację syntezy bibliotek związków. Kolejnym celem mojej pracy było rozwiązanie problemu efektywnego projektowania syntez związków znakowanych izotopowo (**P07**), wymagającego wyboru najłatwiej dostępnego z izomerów o zadanej masie cząsteczkowej. Ostatnim z celów było rozszerzenie komputerowej analizy retrosyntetycznej do planowania syntez omijających rozwiązania chronione prawem patentowym (**P05**).

## 6. Komputerowe planowanie syntez chemicznych – rys historyczny

Pomysł na wykorzystanie komputerów do wspomagania procesu planowania syntez chemicznych<sup>7</sup> pojawił się wkrótce po pojawieniu się pierwszych mikrokomputerów. Pierwsze praktyczne próby rozwiązania problemu analizy retrosyntetycznej przeprowadzanej przez komputer zostały podjęte wkrótce po opracowaniu metody retrosyntetycznej przez Coreya. Już w 1969 roku zaprezentowany został manuskrypt<sup>8</sup>, opisujący program OCSS (Organic Chemical Simulation of Synthesis), stworzony w zespole E. J. Coreya na Uniwersytecie Harvarda. Program ten był rozwijany przez ponad 30 lat (później jako LHASA, Logic and Heuristics Applied to Synthetic Analysis) i używał zakodowanej przez chemików bazy reakcji, obejmującej około 2270 reakcji w ostatniej wersji<sup>9,10</sup>. Projektowanie syntez z wykorzystaniem programu LHASA było niestety ograniczone do ręcznego wybierania przez chemika-operatora najbardziej rokujących z transformacji wygenerowanych przez program. W związku z tym, efektywność całego procesu projektowania syntezy była zależna od wiedzy chemika, jego umiejętności identyfikacji istotnych podstruktur i wiązań czy jego preferencji, dotyczących określonych klas reakcji chemicznych. Przykłady projektowania syntez z wykorzystaniem oprogramowania LHASA obejmują między innymi projekt syntezy kwasu chinowego<sup>11</sup> z 1980 r. czy projekty<sup>2</sup> syntezy



porantheryny, valeranonu czy erythronolidu B z 1985 r. Jedyna próba eksperymentalnej walidacji predykcji programu LHASA została podjęta na uproszczonym rdzeniu taksolu lecz zakończyła się niepowodzeniem na jednym z początkowych etapów<sup>12</sup>. Mimo niepowodzeń w walidacji eksperymentalnej i wygaszenia projektu (obecnie strona projektu na Uniwersytecie Harvarda jest nieaktywna, a ostatnia dostępna zarchiwizowana wersja<sup>10</sup> pochodzi z 2011 roku), na uwagę zasługują pośrednie rezultaty prac nad tym projektem, obejmujące między innymi kodyfikację reguł analizy chemicznej na potrzeby programu<sup>1</sup>, opracowanie kompendium wiedzy dotyczącej grup zabezpieczających i strategii ich stosowania<sup>13</sup> oraz prace jednego z autorów LHASA – S. Rubesteina, skutkujące w opracowaniu<sup>14</sup> oprogramowania ChemDraw, powszechnie używanego przez chemików do rysowania struktur chemicznych. Wśród innych programów, które zadebiutowały w tym okresie na uwagę zasługuje SYNGEN<sup>15-17</sup>, którego działanie odbiegało od innych zaproponowanych rozwiązań. W przeciwieństwie do pozostałych programów, prowadzących analizę retrosyntetyczną dla całego zadanego retronu, analiza wykonywana z użyciem programu SYNGEN przebiegała w dwóch krokach. W pierwszym z nich program brał pod uwagę tylko szkielet węglowy i systematycznie dokonywał dyskoneksji wiązań C-C, otrzymując najkrótszą i konwergentną ścieżkę aby w drugim kroku dokonać ‘refunkcjonalizacji’ uzyskanych szkieletów węglowych<sup>17</sup>. Innym programem, rozwijanym od 1984 r. i przedstawionym w 1990 r. jest SYNSUP<sup>18</sup>. Program ten bazował na eksperckiej bazie reguł chemicznych (zawierającej 5800 reguł w 2010 roku<sup>19</sup>) oraz dysponował katalogiem komercyjnie dostępnych substratów. SYNSUP był jedynym programem z tego okresu, którego projekt syntezy<sup>20</sup> został zrealizowany - aczkolwiek docelowy azaspiran został uzyskany z wydajnością 3%, znacznie niższą niż uzyskiwaną w znanym podejściu prowadzącym do tego związku. Mimo tego, że w pierwszym okresie rozwoju komputerowo wspomaganą analizę retrosyntetyczną entuzjazm chemików-organików (liczących na pomoc ‘elektronicznego kolegi’) był dość wysoki, żaden z opracowanych programów<sup>21-25</sup> nie radził sobie z bardziej skomplikowanymi związkami i ostatecznie nie został szeroko zaadoptowany w codziennej praktyce. Jako główne przyczyny można wyróżnić między innymi konieczność czasochłonnego manualnego trawersowania przestrzeni syntetycznych, niewielkie bazy reguł reaktywności chemicznej czy wreszcie niewielką moc obliczeniową, ogromny koszt i niską dostępność komputerów (PDP-1<sup>26</sup> używany przez Coreya kosztujący ~120.000 dolarów oferował zaledwie 9kB pamięci operacyjnej i 200kHz zegar, a przez 9 lat wyprodukowano zaledwie 53 egzemplarze).

Zainteresowanie problemem wspomagania planowania syntez chemicznych wróciło na początku XXI wieku wraz ze znacznym wzrostem możliwości oferowanych przez komputery i pojawieniem się zdigitalizowanych baz danych agregujących znaną wiedzę syntetyczną, takich jak komercyjnie dostępne Reaxys<sup>27</sup> wywodzący się z bazy Beilstein/Gmelin, Scifinder<sup>28</sup>, wywodzący się z CAS Chemical Abstracts czy publicznie dostępny zbiór reakcji pochodzących z patentów, USPTO<sup>29</sup>. Pojawienie się tych baz danych dało impuls do rozwoju nowej kategorii programów retrosyntetycznych, których wspólną cechą<sup>30</sup> było odejście od ręcznie kodowanych reguł chemicznych na rzecz zbiorów pozyskanych automatycznie z wymienionych wyżej baz danych. Programy takie jak ARChem Route Designer<sup>31</sup>, ICSynth<sup>32</sup>, ASKCOS<sup>33,34</sup> czy sieć Seglera-Wallera<sup>35</sup> mogły zostać ‘uzbrojone’ w bazy reguł chemicznych zawierające dziesiątki tysięcy reguł wyekstrahowanych z kilkumilionowych baz danych reakcji w ciągu zaledwie kilkudziesięciu godzin. Mimo tak ogromnych baz danych (potencjalnie pokrywających całą dostępną wiedzę chemiczną) przykłady syntez zaprojektowanych przez programy tego nurtu ograniczone są do stosunkowo prostych związków. Dodatkowo, analiza nawet tych prostych ścieżek syntetycznych<sup>POI</sup> pokazuje, że automatyczna ekstrakcja reguł wiąże się z ich znacznie mniejszą dokładnością, objawiającą się między innymi brakiem wymaganych do zachowania reaktywności podstruktur, czy niemożliwością poprawnego przewidywania reakcji stereoselektywnych, zwłaszcza kontrolowanych strukturą substratu. Oczywiście, automatyczne pozyskiwanie reguł z pojedynczych precedensów nie umożliwia też poprawnego określenia, które grupy funkcyjne są niekompatybilne w warunkach reakcji i wymusza stosowanie uproszczeń, eliminujących wszystkie grupy, które nie były obecne w precedensach użytych do pozyskania reguł. Wreszcie, żaden z programów nie został wyposażony w procedury umożliwiające kalkulację kosztu ścieżek uwzględniającego strukturę ścieżki (konwergentna/liniowa) i wydajności poszczególnych etapów<sup>36</sup>, co pozwoliłoby na wybór planów syntetycznych o możliwie niskim koszcie.

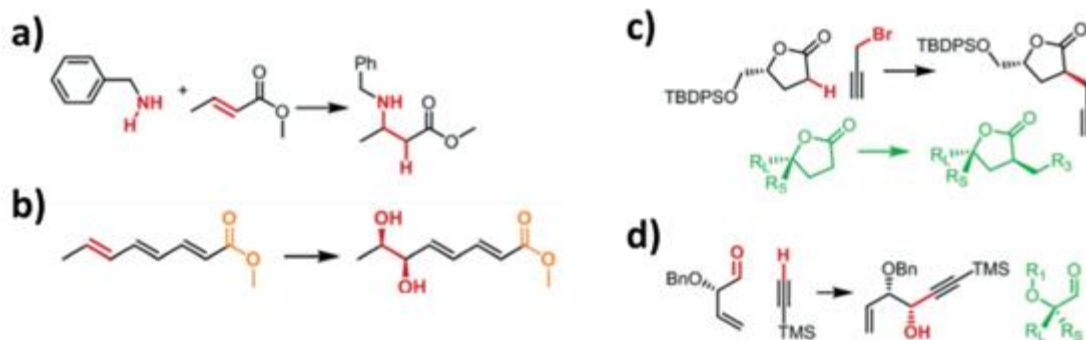
## 7. Reguły reaktywności chemicznej

*Część pracy doktorskiej poświęcona regułom reaktywności chemicznej, ich kodowaniu i porównanie z automatycznym pozyskiwaniem z baz danych została przedstawiona w publikacjach P01 oraz P03.*

Większość programów do automatycznej retrosyntezy (z nielicznymi wyjątkami wykorzystującymi uczone maszynowo oceny par atomów<sup>37,38</sup> lub modele lingwistyczne<sup>39,40</sup> do przewidywania reaktywności) opiera swoje działanie na zbiorach reguł chemicznych, zawierających informacje o możliwych do wykonania cięciach retrosyntetycznych. W literaturze chemicznej można wyróżnić dwa generalne podejścia służące do definiowania tych reguł, opisane szczegółowo i porównane w publikacji P01. Pierwsze z nich obejmuje wspomnianą już wcześniej automatyczną ekstrakcję<sup>31-35,41</sup> z dostępnych baz danych, takich jak Reaxys<sup>27</sup>, Scifinder<sup>28</sup> czy zbiór USPTO.<sup>29</sup> Zaletą tego rozwiązania jest jego szybkość a wadą zaś jego ograniczona dokładność, istotna szczególnie w planowaniu wieloetapowych syntez – nawet jeżeli w zbiorze reguł 9/10 nie będzie zawierać żadnych błędów, uzyskany dziesięciokrokowy plan syntezy będzie poprawny zaledwie w  $0.9^{10} \sim 35\%$  przypadków. Alternatywnym podejściem jest ręczne kodowanie reguł reakcji<sup>42.P01.P14</sup> (uwzględniające ich mechanizm i warunki prowadzenia) przez eksperta-chemika, gwarantujące dokładność reguł zarówno w określaniu rdzenia reakcji jak i grup funkcyjnych, niekompatybilnych w warunkach danej reakcji.

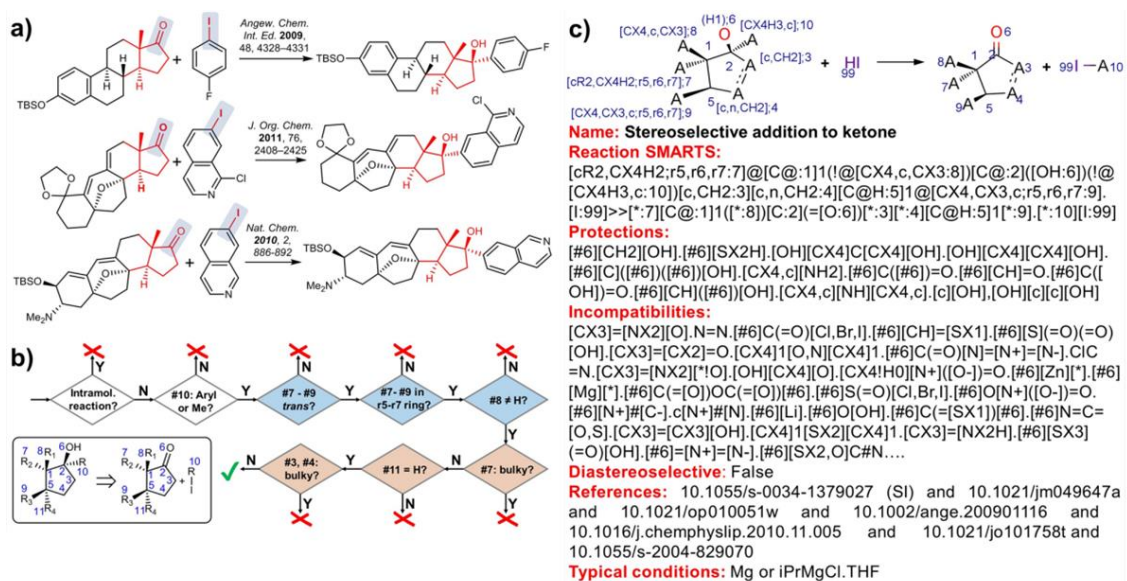
Rdzeniem reakcji wyspecyfikowanym w każdej regule jest fragment cząsteczki, który jest niezbędny do jej przebiegu. Co istotne, rdzeń reakcji jest szerszy niż sam zbiór atomów, który uległ zmianie i musi obejmować zarówno fragmenty niezbędne do zachowania reaktywności (np. obecność grup elektronoakceptorowych w reakcji addycji amin do alkenów, **Schemat 1a**) jak i fragmenty, które odpowiadają za kontrolę regiochemii (**Schemat 1b**) i stereochemii powstającego produktu (**Schemat 1c,d**). Istotnym problemem generowania reguł jest określenie odpowiedniej wielkości rdzenia reakcji – stosowany w automatycznym pozyskiwaniu reguł wybór stałej odległości od atomów, które uległy zmianie prowadzi do otrzymania zbiorów, w którym znaczna część reguł posiada zbyt wąskie lub zbyt szerokie rdzenie. W przeciwieństwie do takiego automatycznego podejścia, ręcznie kodowane reguły opierające się na mechanizmach i modelach stereochemicznych, posiadają wyspecyfikowane przez chemika-organika rdzenie reakcji uwzględniające wymagane czynniki stereoelektronowe i nie obejmują

fragmentów, które są nieistotne dla przebiegu reakcji lecz ograniczałyby zakres stosowalności określonej reguły.



**Schemat 1.** Określenie właściwego rdzenia reakcji. **a)** Addycja amin do alkenów wymaga obecności grup elektroakceptorowych. Poprawnie zdefiniowany rdzeń reakcji jest znacznie szerszy niż atomy, które zmieniły swoje otoczenie (zaznaczone na czerwono). **b)** Regioselektywność stereoselektywnego dihydroksylowania polienów kontrolowana jest przez odległą grupę elektroakceptorową. **c)** Alkilowanie laktonu oraz **d)** addycja związków metaloorganicznych do aldehydów jest kontrolowana przez strukturę substratu. Rdzeń reakcji obejmuje odległy podstawnik stanowiący zawadę steryczną w (**c**) oraz grupę eterową zdolną do wytworzenia cyklicznego kompleksu w (**d**). Schemat i opis zaadaptowane z publikacji **P01**.

W ramach badań prowadzonych w ramach mojej pracy doktorskiej stworzyłem ~50.000 z około 100.000 reguł wchodzących w skład będącej podstawą działania oprogramowania Chematica/Synthia bazy reguł reakcji chemicznych. Zakres mojej pracy obejmował w dużej mierze reguły pokrywające skomplikowane reakcje stereoselektywne, kontrolowane zarówno przez chiralny katalizator, reagent oraz strukturę substratu – reguły te były konieczne do planowania syntez produktów naturalnych (**P09**) oraz posłużyły do przeprowadzenia analiz opisanych w **P08** mających na celu rozszerzenie narzędzi chemii iteracyjnej, pozwalającej na automatyzację syntez bibliotek związków. W publikacji **P01** omówione zostały szczegółowo kryteria które muszą zostać spełnione, aby zapewnić poprawne działanie poszczególnych klas reakcji.



**Schemat 2.** Przykład kodowania reguły reakcji. **a)** Rdzeń reakcji stereoselektywnej addycji do ketonów (szary) jest niewielki, ale uwzględnienie czynników determinujących przebieg stereochemiczny reakcji wymaga jego znacznego rozszerzenia (czerwony). **b)** Drzewko decyzyjne specyfikujące kryteria dla nukleofila (białe pola, pozostałe nukleofile mają inne drzewka i są zakodowane w innych regułach), kryteria determinujące przebieg stereochemiczny reakcji (niebieskie pola) i kryteria wykluczające czynniki strukturalne, które mogłyby prowadzić do zmniejszenia selektywności reakcji (pomarańczowe pola). **c)** Fragment reguły reakcji uwzględniającej wymienione czynniki w notacji SMARTS. Schemat i opis zaadaptowane z publikacji **P03**.

**Schemat 2** przedstawia przykład jednej z takich reguł, opisującej kontrolowaną strukturą substratu addycję związków metaloorganicznych do cyklopentanonów (**Schemat 2a**). W przytoczonym przykładzie, rdzeń reakcji chemicznej jest znacznie szerszy niż atomy, które uległy zmianie (#2, #6, #10) i obejmuje cały *trans*-podstawiony bicykliczny układ, posiadający podstawnik w pozycji angularnej odpowiedzialny za kontrolę stereochemiczną (warunki oznaczone kolorem niebieskim na **Schemacie 2b**). Dodatkowo, zdefiniowany rdzeń reakcji nie zezwala na obecność żadnych podstawników, które mogłyby stanowić zawadę przestrzenną i zmieniać przebieg addycji (warunki oznaczone kolorem pomarańczowym). Reguła reakcji (**Schemat 3c**) zawiera również informację o grupach funkcyjnych (ze zbioru około 400 zdefiniowanych grup, na schemacie przedstawiono tylko część z grup obecnych w tej regule) niekompatybilnych w warunkach reakcji (grupy elektrofilowe: np. chlorki kwasowe i tialdehydy, grupy o właściwościach kwasowych: np. kwasy sulfonowe, kwaśne alifatyczne nitrozwiazki) oraz grupach funkcyjnych, które wymagają stosowania grup zabezpieczających (np. alkohole, aminy, fenole, kwasy karboksylowe). Dodatkowo, każda reguła zawiera informację o typowych warunkach prowadzenia reakcji oraz odnośnik literaturowy prowadzący do precedensu, pokazującego przykład zastosowania

określonego typu reakcji. Wreszcie, w każdej z reakcji zawarta jest informacja (istotna zwłaszcza w czasie planowania syntez enancjoselektywnych), czy dana reakcja prowadzi do otrzymania pojedynczego enancjomeru czy mieszaniny racemicznej.

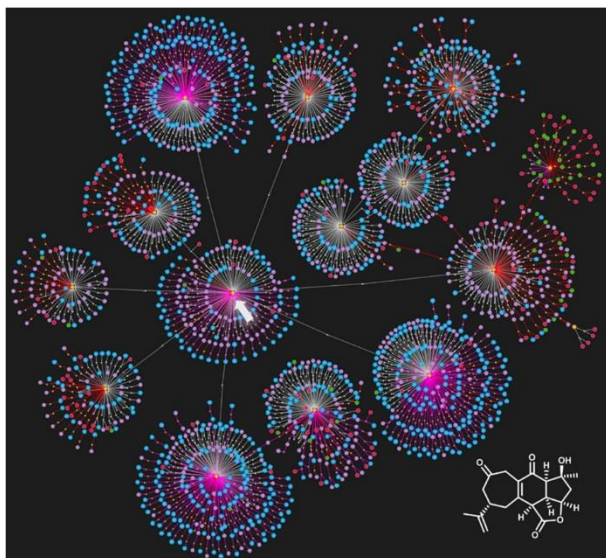
## 8. Algorytmy

Opisane w poprzednim rozdziale reguły reaktywności chemicznej stanowią podstawowe ‘ruchy’ syntetyczne i gwarantują poprawną predykcję pojedynczego kroku podczas automatycznej analizy retrosyntetycznej. Kolejnym elementem, obecnym w każdym oprogramowaniu do planowania wielokrokowych syntez są algorytmy, pozwalające na ‘budowanie’ i ocenę kompletnych planów syntetycznych z przewidzianych kroków syntetycznych. Algorytmy te mogą zostać podzielone na trzy kategorie: i) algorytmy wykorzystywane w ocenie pojawiających się opcji syntetycznych i wyborze rokujących kandydatów oraz służące do iteracyjnego rozwijania grafu reprezentującego przestrzeń syntetyczną; ii) algorytmy umożliwiające wybór najkrótszych/najtańszych oraz różnorodnych planów syntetycznych; oraz iii) algorytmy umożliwiające uwzględnienie w planowaniu kilkukrokowych sekwencji reakcji.

### 8.1. Wybór rokujących pozycji i rozwijanie grafu

*Algorytmy i funkcje oceniające (SF) odpowiedzialne za wybór rokujących pozycji oraz rozwijanie grafów w czasie poszukiwania planu syntetycznego opisano szczegółowo w publikacjach P02, P13 i P14.*

Pierwszą kategorię algorytmów, obecną w każdym oprogramowaniu do planowania wielokrokowych syntez są funkcje oceniające (*scoring functions, SF*), zdefiniowane<sup>P14</sup> po raz pierwszy przez nasz zespół w 2016 i umożliwiające wybór najbardziej rokujących pozycji i ograniczenie przestrzeni wyszukiwania. Są one niezbędne ze względu na eksplozję kombinatoryczną<sup>2</sup> uniemożliwiającą rozwinięcie wszystkich możliwych ścieżek syntezy – przy zaledwie kilkudziesięciu możliwych reakcjach (**Schemat 3**) prowadzących do każdego związku, dla dziesięciokrokowej syntezy liczba możliwości sięga  $n \sim 10^{17}$ .



**Schemat 3.** Początkowe możliwości syntetyczne prowadzące do Scabrolidu A (zaznaczonego białą strzałką). W analizach retrosyntetycznych konieczne jest rozwinięcie i analiza tysięcy zilustrowanych tu ‘pajaków’. Schemat i opis zaadaptowane z publikacji **P03**.

W programach do analizy retrosyntezy można wyróżnić stosowane obecnie dwa generalne trendy (porównane w publikacji **P10**) w sposobie definiowania funkcji oceniających pojawiające się możliwości syntetyczne. Jednym z nich, rozwijanym intensywnie w ostatnich latach<sup>33,35,43</sup> i często połączonym z automatycznym pozyskiwaniem reguł chemicznych jest zastosowanie metod uczenia maszynowego do trenowania funkcji oceniających. Metoda ta sprawdza się w przypadku syntez wykorzystujących proste reakcje chemiczne, reprezentowane przez liczne przykłady w zbiorach

użytych do uczenia<sup>P10</sup> ale jej efektywność znacznie maleje, gdy w czasie planowania syntezy konieczne jest użycie reakcji, która były reprezentowane przez zaledwie kilka przykładów. Alternatywnym podejściem są funkcje oceniające (*SF*) wykorzystujące zestaw zmiennych (oceniających między innymi wielkość powstających substratów, ilość centrów stereogenicznych i pierścieni) do oceny uzyskanych cząsteczek/syntonów (*Chemical Scoring Function, CSF*) i reakcji do nich prowadzących (*Reaction Scoring Function, RSF*) używane w oprogramowaniu Chematica/Synthia. Funkcja oceniająca ma postać sumy funkcji CSF i RSF,  $SF = [RSF] + [CSF]$  oraz

$$CSF = [SMALLER^n + a \cdot RINGS + b \cdot STEREO],$$

$$RSF = p + q \cdot PROTECTIONS + r \cdot (FILTERS + CONFLICTS + NON\_SELECTIVITY)$$

z typowymi wartościami parametrów  $n=1.5 \div 3$ ,  $a, b = 20 \div 50$ ,  $p=20 \div 100$ ,  $r \sim 100000$ ,  $q \sim 2 \cdot p$  lub  $q \sim r$  dla analiz, posiadających zaplanowaną strategię stosowania grup zabezpieczających lub gwarantujących brak stosowania tych grup. Omówienie szczegółów poszczególnych zmiennych oraz sposobu działania poszczególnych funkcji oceny *CSF* i *RSF* Czytelnik znajdzie w **Sekcji 4** publikacji **P02**. Tak zdefiniowana funkcja była wykorzystana w naszym wczesnym algorytmie typu A\* (opisanym szczegółowo w **Części 5** publikacji **P02**) używanym do eksploracji grafu i umożliwiającym projektowanie kilkukrokowych syntez do stosunkowo prostych związków o działaniu biologicznym<sup>P13</sup>. Planowanie kilkunastokrokowych syntez produktów naturalnych przedstawionych w publikacji **P09** było możliwe dopiero

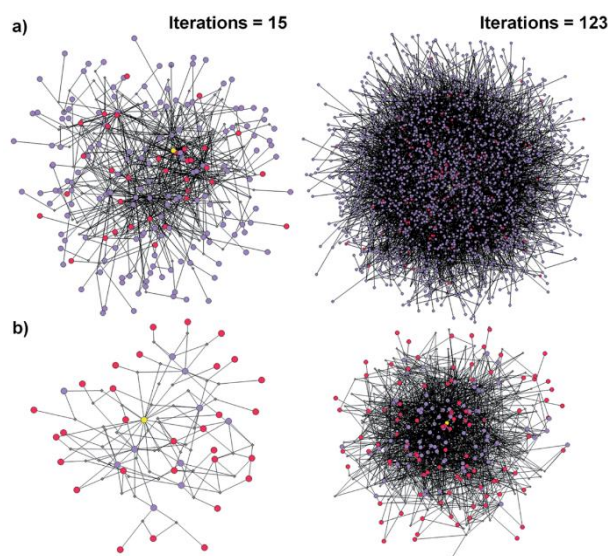
po zastosowaniu zmodyfikowanego algorytmu typu beam-search, wykorzystującego kilka kolejek priorytetowych oraz używającego równocześnie dwóch funkcji oceny cząsteczek. Pierwsza z tych funkcji,  $CSF_1 = \text{SMALLER}^{1.5}$  preferowała modyfikacje peryferyjnych fragmentów retronów i odpowiadała za znajdowanie rokujących początków ścieżek syntetycznych, podczas gdy druga,  $CSF_2 = \text{SMALLER}^3$ , wykazywała preferencję w kierunku dyskoneksji szybko zmierzających do fragmentacji retronu w małe bloki budulcowe i odpowiadała za „dokańczanie” rozpoczętych ścieżek syntetycznych. Szczegółowy opis działania tego algorytmu (opracowanego i zaimplementowanego przez dr Tomasza Badowskiego) umieszczony został w **Sekcji 6** publikacji **P02**.

## **8.2. Analiza ekonomiczna planów syntetycznych**

*Algorytmy odpowiedzialne za wybór planów syntetycznych uwzględniające ich koszt oraz wybór różnorodnych planów syntetycznych opisane zostały w publikacji **P04**.*

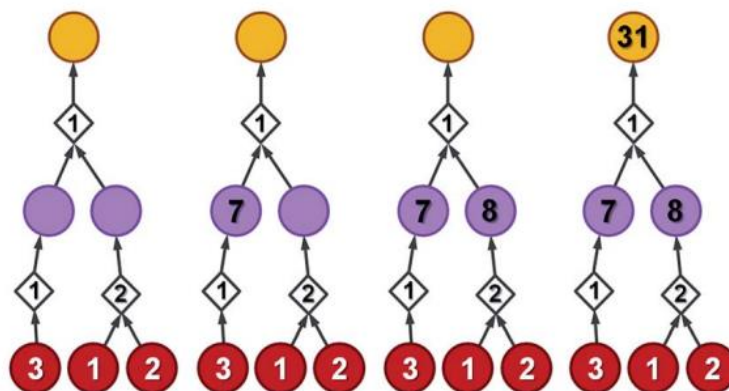
Działanie omówionego w poprzedniej sekcji algorytmu polegające na iteracyjnym rozwijaniu retronów w potomne syntony oraz nawigowaniu po stworzonej przestrzeni syntetycznej prowadzi do utworzenia grafu, reprezentującego całą wyeksplorowaną przestrzeń. Pierwsze rozwiązanie syntetyczne zostaje znalezione, gdy pojawi się co najmniej jedno połączenie prowadzące od retronu do komercyjnie dostępnych substratów. W czasie planowania syntezy, algorytm rozwija setki (w przypadku krótszych analiz wystarczających do zaplanowania kilkukrokowych syntez) do dziesiątek tysięcy pojedynczych ‘pajaków’ (przedstawionych na **Schemacie 3**), a otrzymany graf zawiera setki lub nawet tysiące możliwych planów syntetycznych. Kolejna grupa algorytmów uczestnicząca w procesie automatycznego planowania syntez ma za zadanie (i) uzyskanie z tego grafu ścieżek syntetycznych, które będą charakteryzowały się możliwie niskim kosztem, wyrażonym zarówno jako suma kosztów substratów niezbędnych do uzyskania określonej ilości produktu jak i kosztu wykonania każdej reakcji; oraz (ii) wybranie ścieżek, które będą od siebie znacząco różne. W zaproponowanym rozwiązaniu, opracowanym przez mnie we współpracy z dr Tomaszem Badowskim pierwszym krokiem działania algorytmu jest usunięcie fragmentów grafu reprezentującego odwiedzoną przestrzeń syntetyczną (**Schemat 4a**) i nie prowadzących do komercyjnie dostępnych związków. Ma to na celu uzyskanie grafu zawierającego wyłącznie ‘skończone’ ścieżki syntetyczne, wielokrotnie mniejszego (**Schemat 4b**) od wyjściowego grafu.





**Schemat 4.** Sieci reakcji rozwijane podczas analiz retrosyntetycznych **(a)** i podgrafy wykonalnych rozwiązań **(b)**. **a)** Graf (456 węzłów) po lewej stronie uzyskany na początkowym (15 iteracji, około 20s działania) etapie planowania syntezy prostej triaryloaminy. Sieć po prawej stronie przedstawia graf (5300 węzłów) uzyskany po 123 iteracjach (< 120s działania algorytmu). **b)** Podgrafy sieci z **(a)** zawierające tylko ścieżki prowadzące do komercyjnie dostępnych substratów zawierające odpowiednio 90 i 779 węzłów. Oznaczenia węzłów: żółty - target; fioletowy - półprodukty; czerwony – komercyjnie dostępne substraty; małe szare romby - węzły reakcji. Schemat i opis zaadaptowane z publikacji **P04**.

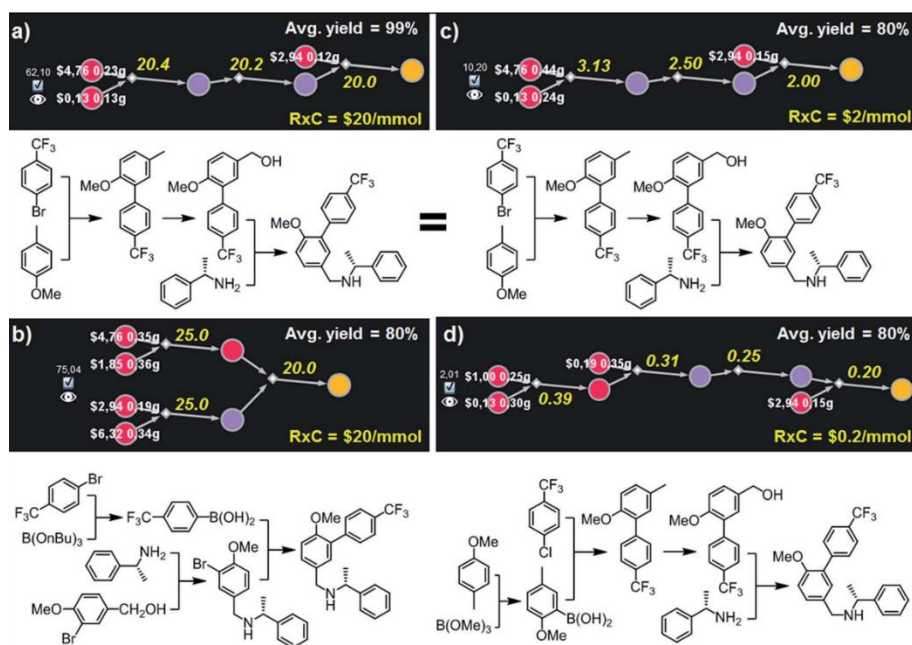
W kolejnych krokach, korzystając z informacji o kosztach poszczególnych substratów (zawartych w katalogu Sigma Aldrich będącym składnikiem oprogramowania Chematica/Synthia) i kosztach reakcji (zdefiniowanych przez użytkownika w USD/mmol produktu i zależnych od planowanej skali syntezy oraz miejsca jej prowadzenia wpływającego na koszt pracy chemika-operatora procesu) algorytm oblicza koszty wszystkich węzłów znajdujących się w grafie zaczynając od substratów i rekursywnie propagując koszty (**Schemat 5**) w stronę retronu. W czasie obliczania kosztów węzłów algorytm uwzględnia wydajność każdej reakcji, używając uśrednionej wartości podanej przez użytkownika. Po przeliczeniu wszystkich kosztów wybrana zostaje pierwsza ścieżka syntetyczna o najniższym koszcie, a koszt wszystkich węzłów będących jej elementami zostaje powiększony o dodatkową karę  $P$ . Operacja ta pozwala na wybór kolejnych planów syntetycznych różniących się od najwyżej ocenionego rozwiązania. Dodatkowo, aby uniknąć zwracania planów syntetycznych zawierających bardzo podobne do siebie reakcje (np. sprzęgania Suzuki pomiędzy kwasem boronowym a bromkiem vs. jodkiem arylowym) algorytm przypisuje wspomnianą karę reakcjom, które prowadzą do wybranego już produktu z co najmniej jednego, takiego samego substratu.



**Schemat 5.** Przykład rekursywnego obliczania kosztów z wykorzystaniem algorytmu propagującego koszty substratów i koszty reakcji. Założona wydajność reakcji  $Y=50\%$ . Na koszt węzła '7' składa się koszt reakcji '1' oraz koszt substratu o cenie 3\$/mmol (2 mmoli niezbędnych do uzyskania 1 mmola produktu przy zadanej wydajności). Analogicznie, na koszt węzła 31 składa się koszt reakcji ('1') oraz koszt 2 mmoli półproduktów o obliczonych kosztach 7 i 8 \$/mmol. Schemat i opis zaadaptowane z publikacji **P04**.

Po znalezieniu kolejnego najlepiej ocenianego rozwiązania, algorytm powtarza operacje aktualizacji kosztu węzłów i wyboru ścieżek syntetycznych aż do osiągnięcia zadanej przez użytkownika ilości ścieżek syntetycznych. Co istotne, zmiana parametrów algorytmu (obejmujących zmianę kosztu prowadzenia reakcji, zmianę przewidywanej uśrednionej wydajności czy nawet zmianę cen materiałów startowych) nie wymaga ponownego wykonywania czasochłonnej analizy retrosyntetycznej – we wszystkich analizach przedstawionych w publikacji **P04** algorytm korzystał z zapisanych grafów obejmujących podgrafy prowadzące do komercyjnie dostępnych substratów a wybór 100 najlepszych planów syntetycznych nie przekraczał 0,5s nawet dla grafów zawierających ponad 12000 węzłów. W jednym z eksperymentów, obejmującym wybór planów syntetycznych prowadzących do AMG641<sup>44</sup> (kalcymimetyka rozwijanego przez Amgen) opracowany algorytm został skonfrontowany z zadaniem (**Schemat 6**) wyboru planów syntezy optymalnych dla różnych etapów badań nad nową substancją czynną, charakteryzujących się zróżnicowanym kosztem wykonywania reakcji (wysokim na początkowych etapach wykonywanych w niewielkiej skali) i malejącym w miarę zbliżania się do produkcji wielkoskalowej. Przy wyjściowych ustawieniach, obejmujących zakładany koszt prowadzenia reakcji na poziomie 20\$/mmol algorytm wybrał jako najtańszy trójfazowy konwergentny plan, używający stosunkowo drogich bromowanych materiałów początkowych. Po zmniejszeniu kosztu pojedynczego etapu do 2\$/mmol, najwyżej oceniony plan syntetyczny obejmował trójfazową liniową syntezę, w której układ biarylowy został otrzymany w eleganckiej sekwencji obejmującej jednoetapowe *o*-litowanie/sprzężanie<sup>45</sup> a niezbędny alkohol został otrzymany w reakcji

utleniania pozycji benzylovej<sup>46</sup>, co pozwoliło zrezygnować z drogiego alkoholu benzylovego używanego jako substrat w pierwszym planie syntetycznym. Wreszcie, zmniejszenie kosztu reakcji do 0,2\$/mmol, symulujące produkcję związku w dużej skali spowodowało dalsze zwiększenie długości ścieżki syntetycznej do czterech kroków, przy jednoczesnym wykorzystaniu relatywnie niedrogich substratów.



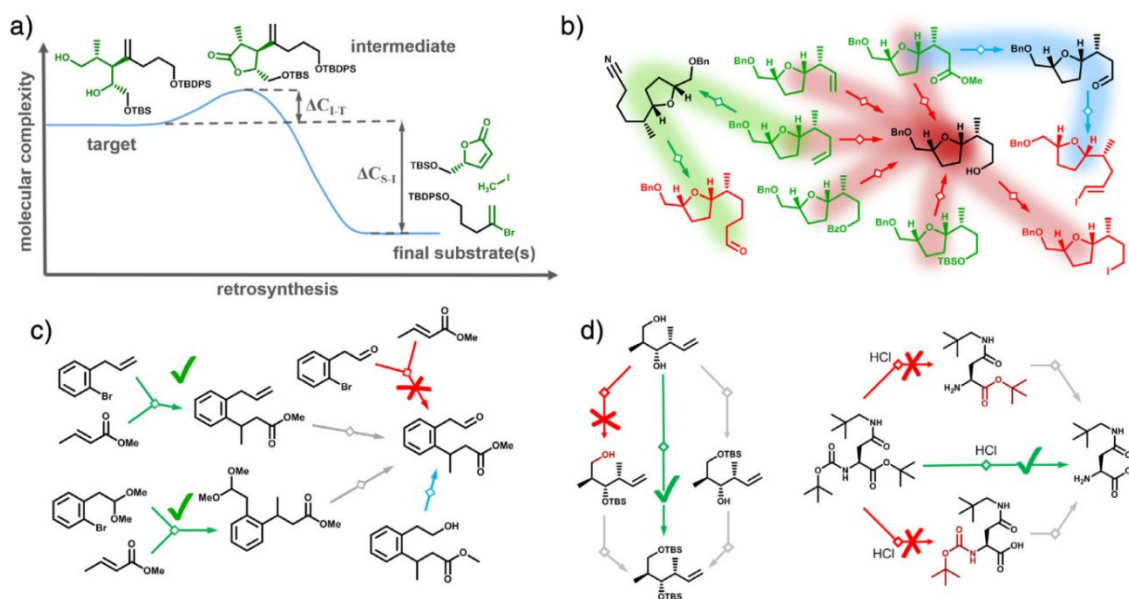
**Schemat 6.** Najwyżej ocenione plany syntetyczne prowadzące do AMG641 opracowane przy różnych ustawieniach wydajności i kosztu reakcji. Wszystkie ścieżki syntetyczne zostały wybrane z zapisanego grafu (zawierającego 5363 węzłów i uzyskanego z analizy retrosyntetycznej przeprowadzonej z użyciem oprogramowania Chematica/Synthia w ciągu 7 minut). RxC – koszt reakcji wyrażony w \$/mmol. Żółte liczby obok węzłów reakcji przedstawiają ich koszty uwzględniające zadaną wydajność reakcji. Schemat i opis zaadaptowane z publikacji **P04**.

Działanie algorytmu pozwalające na generowanie różnorodnych planów syntetycznych zostało zwalidowane między innymi na próbie opracowania planów syntezy prowadzących do prostego produktu naturalnego, *trans*-whisky laktonu.<sup>47</sup> W każdym z najwyżej ocenionych planów syntezy prowadzącym do tego związku uzyskanym bez zastosowania kary (wymuszającej różnorodność), ostatnim etapem syntezy jest kontrolowana strukturą substratu 1,4-addycja związku magnezoorganicznego do cyklicznego laktonu, naśladująca znane metody syntezy<sup>48,49</sup> tego związku. Zbiór najwyżej ocenionych planów syntetycznych uzyskanych z włączoną penalizacją etapów już użytych w zwróconych ścieżkach syntetycznych jest dużo bardziej różnorodny. Związek docelowy otrzymywany jest zarówno poprzez opisaną wyżej 1,4-addycję (najwyżej oceniony plan) jak i oksydacyjną laktonizację odpowiedniego diolu<sup>50</sup> lub laktonizację odpowiedniego hydroksynitrylu.

### 8.3. Planowanie z wykorzystaniem kilkukroowych sekwencji reakcji

*Algorytmy umożliwiające planowanie z wykorzystaniem kilkukroowych sekwencji reakcji zostały opisane w **Rozdziale 6** publikacji **P02**. Wyniki analiz retrosyntetycznych uzyskane z wykorzystaniem tych algorytmów zawarte są w publikacji **P09**.*

Połączenie (i) dokładnych reguł reaktywności chemicznej zakodowanych przez chemików-organików, (ii) algorytmów rozwijających przestrzeń syntetyczną i wykorzystujących funkcje oceny cząsteczek i reakcji oraz (iii) procedur pozwalających na wybór najtańszych i różnorodnych planów syntetycznych okazało się wystarczające do planowania syntez prowadzących do związków o średnim stopniu skomplikowania, opisanych w publikacji<sup>P13</sup>. Mimo znacznego rozszerzenia bazy reguł chemicznych o reakcje stereoselektywne, próby użycia hybrydowego podejścia łączącego eksperckie reguły z metodami uczenia maszynowego<sup>P10</sup> czy wykorzystanie w funkcji CSF dodatkowych zmiennych, odpowiadających Corey'owskiemu regułom retrosyntezy<sup>51</sup> oprogramowanie nadal nie potrafiło poradzić sobie z efektywnym planowaniem stereoselektywnych syntez prowadzących do skomplikowanych związków naturalnych. Rezultaty analiz retrosyntetycznych uzyskiwanych dla tych związków pokazywały, że oprogramowanie nie potrafi planować dalej niż jeden krok naprzód: każdy z kroków syntetycznych oceniany był indywidualnie, a program nie był w stanie korzystać z kilkukroowych sekwencji reakcji, w których początkowe etapy dawałyby niewielki zysk (mierzony jako uproszczenie struktury) lecz przygotowałyby scenę pod kluczowy etap. Dodatkowo, w projektowanych syntezach prowadzących do związków naturalnych program stosował bardzo uproszczone podejście do strategii grup zabezpieczających, zaznaczając etapy, w których wymagane jest ich stosowanie lecz pozostawiając w gestii chemika właściwy dobór i wybór właściwego momentu na ich założenie i usunięcie. Pierwszy z algorytmów, opracowany<sup>52</sup> przez dr Sarę Szymkuć, dr Ewę Gajewską i dr Piotra Dittwalda pozwalał na realizację kilkukroowych sekwencji reakcji, w których pierwsza zwiększała złożoność półproduktu, ale umożliwiała duże uproszczenie struktury na kolejnym etapie (**Schemat 7a**). Analiza wszystkich możliwych kombinacji reguł chemicznych pozwoliła na odkrycie milionów nowych strategii syntetycznych, z których około 100 000 najbardziej użytecznych została umieszczona w oprogramowaniu Chematica/Synthia.



**Schemat 7.** Algoritmy umożliwiające planowanie syntez z wykorzystaniem wielokrokowych sekwencji reakcji. **a)** Kilkukrokowe sekwencje reakcji, w których pierwsza zwiększa złożoność półproduktu, ale umożliwiała duże uproszczenie struktury na kolejnym etapie. **b)** Kilkukrokowe sekwencje pozwalające na przekształcenie wysoce reaktywnych grup funkcyjnych w ich bardziej stabilne odpowiedniki **c)** Algorytm umożliwiający ‘omijanie’ napotkanych problemów związanych z obecnością grupy niekompatybilnej lub nieselektywności. **d)** Wykorzystanie kombinacji reakcji, które mogą zostać zrealizowane w pojedynczym kroku syntetycznym. Schemat i opis zaadaptowane z publikacji **P02**.

Drugi z opracowanych algorytmów (**Schemat 7b**) – w którego tworzeniu uczestniczyłem – umożliwił realizację kilkukrokowych sekwencji (poprzednio ocenianych jako mało produktywnie) pozwalających na przekształcenia wysoce reaktywnych grup funkcyjnych (np. jodków alkilowych czy aldehydów) w ich bardziej stabilne odpowiedniki (etery benzytowe/sililowe, estry czy alkeny). Działanie tego algorytmu wykorzystywało stworzoną na podstawie analizy opublikowanych w literaturze syntez totalnych związków naturalnych bazę obejmującą około 100 dwu- i trzykrokowych sekwencji. Zamaskowanie wysoce reaktywnych grup funkcyjnych zmniejszało w czasie analiz retrosyntetycznych ilość wykrywanych przez program konfliktów wynikających z obecności grupy niekompatybilnej i pozwalało na wykorzystywanie poprzednio niedostępnych opcji syntetycznych. Trzecia z opracowanych metod (**Schemat 7c**) umożliwiała ‘omijanie’ napotkanych problemów związanych z obecnością grupy niekompatybilnej lub nieselektywności, w przeciwieństwie do bazowej wersji algorytmu, który w takiej sytuacji wybierał inne cięcie retrosyntetyczne, często oferujące dużo mniejsze uproszczenie. Ostatnie z udoskonaleń (**Schemat 7d**) pozwoliło na wykorzystanie kombinacji reakcji, które do tej pory stanowiły odrębne reguły (i często były odrzucane podczas rozwijania grafu

przez algorytm jako nieselektywne) lecz w rzeczywistości mogą zostać zrealizowane w jednym kroku. Połączenie tej funkcjonalności z udoskonaleniem pozwalającym na omijanie napotkanych problemów oraz wysoką karą za obecność grup wymagających protekcji umożliwiło planowanie skomplikowanych syntezy posiadających pełną strategię grup zabezpieczających.

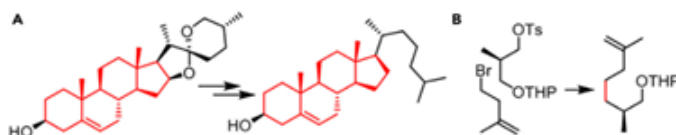
## 9. Znajdowanie ścieżek syntetycznych omijających patenty

*Algorytmy identyfikujące kluczowe wiązania w znanych planach syntezy i umożliwiające analizę retrosyntetyczną z ich pominięciem zostały opisane w publikacji P05.*

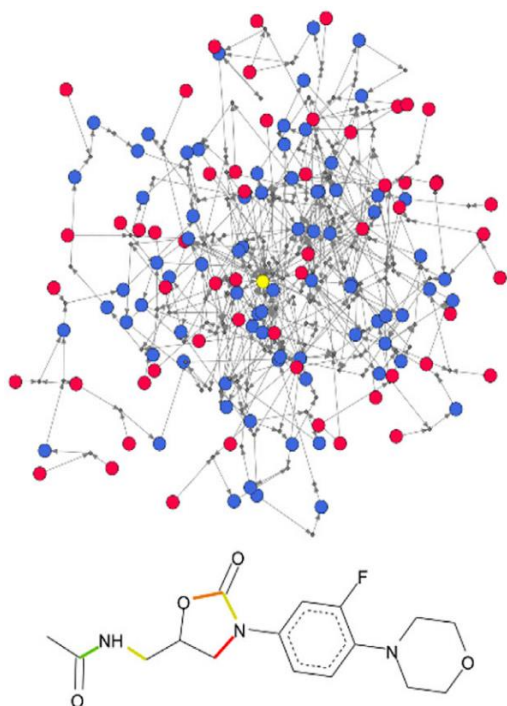
Opracowane do tej pory algorytmy pozwoliły na stworzenie oprogramowania pozwalającego na planowanie syntez prowadzących do dowolnych związków, nawet wysoce skomplikowanych produktów naturalnych. W czasie planowania, oprogramowanie korzystało z dokładnych reguł chemicznych<sup>P01</sup> (gwarantujących poprawność chemiczną pojedynczych etapów), używało algorytmów wykorzystujących funkcje oceny do wyboru rokujących cięć retrosyntetycznych<sup>P02</sup>, wykorzystywało możliwość użycia w czasie analizy wielokrokowych sekwencji reakcji<sup>P03</sup> a wygenerowane plany były poddawane analizie pod kątem ekonomicznym<sup>P04</sup>. Kolejnym etapem mojej pracy badawczej była próba rozszerzenia oprogramowania o planowanie syntez, które gwarantowałyby unikalność rozwiązania, tj. takich, które byłyby znacząco różne od opisanych w literaturze podejść do danego produktu. Zagadnienie to związane jest również z możliwością planowania syntez znacznie różniących się od tych chronionych prawem patentowym. Jednym z rozwiązań tego problemu jest skorzystanie z obecnej w oprogramowaniu możliwości wykluczania określonych substratów, półproduktów czy określonych reakcji. To podejście zostało użyte (szczegółowy opis Czytelnik znajdzie w publikacji P06) do zaplanowania syntez hydroksychlorochiny z wykluczeniem półproduktów wykorzystywanych w znanych rozwiązaniach syntetycznych. Dodatkowo, Cernak i współautorzy<sup>53</sup> wykorzystali możliwość wyłączania określonych klas reakcji oraz wykluczania substratów używanych jako materiały początkowe w planowaniu syntez prowadzących do leków przeciw-wirusowych. Rozwiązanie to ma jednak ograniczenia: nie umożliwia zaplanowania syntezy wychodzącej z tych samych substratów, lecz korzystających z zupełnie innych strategii syntetycznych (lub wykorzystujących te same metodologie, ale wychodząc z innych substratów) oraz jest skuteczne tylko w przypadku związków, do których prowadzi stosunkowo niewiele znanych rozwiązań syntetycznych (ze względu na czasochłonność tworzenia list obejmujących dziesiątki lub setki związków/reakcji

do wykluczenia). W rzeczywistości, synteza związków o udokumentowanym działaniu leczniczym jest często chroniona przez setki patentów, a dany związek można otrzymać na wiele sposobów używając różnych zbiorów substratów (z przykładami umieszczonymi

w publikacji **P05** w sekcjach **S4,S5**). Innym potencjalnie użytecznym podejściem mogłoby być określenie podstruktury, która musi być zachowana na całej ścieżce. Oczywiście, podejście to jest użyteczne wyłącznie w przypadku dużych i unikalnych podstruktur (**Schemat 8a**), podczas gdy często do zaplanowania syntezy odbiegającej od znanych rozwiązań wymagane jest zablokowanie jednego z wielu identycznych wiązań (**Schemat 8b**). W związku z tym, w zaproponowanym przeze mnie rozwiązaniu, planowanie syntez omijających metody chronione prawem patentowym opiera się na ‘zabezpieczeniu’ istotnych wiązań (często rozłącznych), które były miejscem dyskoneksji w znanych rozwiązaniach syntetycznych. Pierwszym krokiem w celu zidentyfikowania wiązań, które są najbardziej istotne było stworzenie sieci obejmujących wszystkie reakcje prowadzące do pożądanego produktu. Jako przykład związku, którego synteza chroniona jest wieloma patentami posłużył mi Linezolid, do którego prowadzi sieć reakcji chemicznych obejmująca odpowiednio 156 reakcji oraz 138



**Schemat 8. Ograniczenia prostych algorytmów zachowujących motywy strukturalne.** Zachowywanie podstruktur jest skuteczne wyłącznie w przypadku zabezpieczenia dużych fragmentów (a), lecz nie może być efektywnie użyte w przypadku konieczności zablokowania jednego z kilku (b) nieunikalnych wiązań. Schemat i opis zaadaptowane z publikacji **P05**.



**Schemat 9 Reprezentacja sieciowa syntez i istotnych wiązań w opatentowanych rozwiązaniach prowadzących do Linezolidu.** W górnej części przedstawiono sieć łączącą wszystkie reakcje w patentach chroniących syntezę Linezolidu. Kolorem żółtym oznaczono docelową cząsteczkę; kolorem niebieskim, półprodukty a kolorem czerwonym, materiały wyjściowe. W dolnej części przedstawiono strukturę Linezolidu z wiązaniami pokolorowanymi według ich istotności w planach syntetycznych w sieci. Najistotniejsze wiązania oznaczono kolorem czerwonym. Schemat i opis zaadaptowane z publikacji **P05**.

cząsteczek. (**Schemat 9a**). W rozwiązaniu, które opracowałem przy współpracy z matematykiem, dr Piotrem Dittwaldem, rozpoczęliśmy od automatycznej analizy tych sieci w celu zidentyfikowania najistotniejszych wiązań. W tym celu, używając bazy reguł reprezentujących reakcje chemiczne<sup>P01</sup> uzyskaliśmy niezbędne sieci zmapowanych reakcji chemicznych, posiadające w pełni uzgodnioną numerację atomów od produktu do początkowych substratów. W kolejnym kroku, algorytm obliczał istotność każdego z wiązań (szczegółowy opis tego procesu Czytelnik znajdzie w publikacji **P05**, w sekcji *Identification of Essential Disconnections in the Networks of Patented Syntheses*) i po normalizacji prezentował rezultat w formie graficznej. Wynik działania tego algorytmu dla Linezolidu przedstawiono na **Schemacie 9b**, dla którego algorytm zidentyfikował jako najbardziej istotne wiązania znajdujące się wewnątrz pierścienia oksazolidynowego. W związku z tym, te wiązania powinny zostać w następnym kroku zablokowane podczas analizy mającej na celu znalezienie rozwiązań odległych od już opublikowanych bądź opatentowanych. W publikacji **P05** Czytelnik znajdzie wyniki podobnych analiz przeprowadzonych dla Sitagliptyny i Panobinostatu, do których sieci reakcji chronionych patentami zawierają odpowiednio 469 i 21 reakcji oraz 410 oraz 28 cząsteczek.

W drugim kroku, naszym zadaniem było opracowanie algorytmu (opisanego szczegółowo w publikacji **P05** w sekcji *Bond Preservation Algorithm to Accompany Retrosynthetic Planning*), umożliwiającego przeprowadzenie w pełni automatycznej analizy retrosyntetycznej wykluczającej dyskoneksje określonych wiązań. W tym celu, oprócz wykorzystania notacji SMILES<sup>54</sup> (używanej w zwykłych analizach) użyty został dodatkowo format Extended Molfile<sup>55</sup> (czarna plansza na **Schemacie 10**), umożliwiający przekazywanie dodatkowych informacji o określonych wiązaniach – w tym przypadku, ‘zablokowaniu’ ich przez użytkownika przy wprowadzaniu struktury retronu. W kolejnym etapie, w reprezentacji retronu zapisanej w notacji SMILES, atomy znajdujące się na końcach zablokowanych wiązań zostają oznaczone numerami (C=CC[C:1]([C:2])[C:3](=[O:4])OC, na przykładzie umieszczonym na **Schemacie 10**) oraz zostaje utworzona lista określająca, które wiązania zostały zablokowane ( $B(T) = \{[1,2], [3,4]\}$ ). Po rozpoczęciu automatycznej analizy retrosyntetycznej, algorytm (opisany uprzednio w **Rozdziale 8**) generuje pierwszą pulę kandydatów zawierających oznaczone atomy. Do kolejnych etapów analizy retrosyntetycznej (wybór rokujących kandydatów poprzez ocenę cząsteczek i reakcji przy użyciu funkcji CSF/RSF, dalsze generowanie zbiorów syntonów, aż do osiągnięcia komercyjnie dostępnych substratów i selekcja ścieżek syntetycznych) algorytm dopuszczał tylko te zbiory syntonów, które zawierają w sobie wszystkie wiązania znajdujące się na liście B(T).





omijających patenty planów syntezy Sitagliptyny. Analiza sieci ponad 450 cząsteczek i 400 reakcji z patentów pozwoliła na zidentyfikowanie istotnych wiązań w otoczeniu centrum stereogenicznego Sitagliptyny oraz wiązania amidowego. Podobnie jak w przypadku Linezolidu, analiza przeprowadzona bez żadnych zablokowanych wiązań skutkowała otrzymywaniem planów syntetycznych zbliżonych do znanych rozwiązań, wykorzystujących kontrolowaną pomocnikiem chiralnym addycję związku metaloorganicznego do sulfinyloiminy bądź reakcję asymetrycznej redukcji odpowiedniego związku dikarbonylowego. Gdy jednak Chematica nie mogła rozłączyć żadnego z zablokowanych wiązań, program obchodził te ograniczenia proponując ścieżki rozpoczynające się od komercyjnie dostępnych chiralnych bloków budulcowych, pochodnych kwasu asparaginowego lub homoseryny. Podobny eksperyment, mający na celu opracowanie planu syntezy w którym obecne centra stereogeniczne pochodzą z materiałów początkowych (*chiral pool*) został przeprowadzony dla Koniceiny (publikacja **P05**, **Figura 7**) – w tym przypadku, uzyskane ścieżki syntetyczne prowadziły do pochodnych proliny lub kwasu pipekolinowego, w zależności od tego, które z wiązań zostało zablokowane. Wreszcie, w ostatnim przykładzie opisującym projektowanie planów syntetycznych prowadzących do Panobinostatu wykazałem, że opracowane algorytmy umożliwiają projektowanie planów syntetycznych omijających patenty przy jednoczesnym zastosowaniu kryteriów chemii procesowej. Zaproponowana przez Chematicę metoda syntezy Panobinostatu omijająca znane rozwiązania wykorzystywała jako jeden z etapów katalizowane palladem reakcje Hecka do wprowadzenia fragmentu cynamyłowego oraz reakcję typu Buchwalda do otrzymania pożądanego indolu. Stosowanie katalizatorów palladowych w produkcji substancji czynnych leków wiąże się oczywiście z koniecznością dokładnego usunięcia ich pozostałości oraz znacznym kosztem samego metalu i towarzyszących mu ligandów. W związku z tym, w ostatniej analizie poza zablokowaniem istotnego wiązania C-N program nie miał możliwości stosowania żadnych reakcji zawierających słowa kluczowe ‘Pd’ oraz ‘palladium’. W tym przypadku, w otrzymanych planach syntetycznych, pożądaný indol otrzymywany był w reakcji Fischera, podczas gdy fragment cynamyłowy został uzyskany w wyniku olefinowania Hornera-Wadswortha-Emonsa lub kondensacji Knoevenagela-Doebnera.

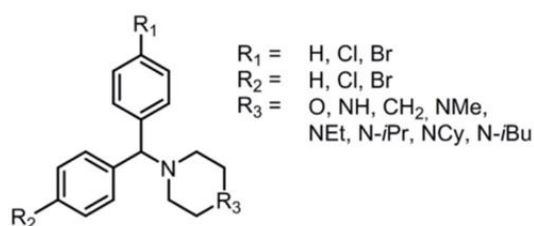
Podsumowując, opracowany został algorytm pozwalający na planowanie syntez omijających znane i/lub opatentowane rozwiązania poprzez zablokowanie określonych wiązań, używanych wcześniej jako miejsca dyskoneksji. Zaproponowane podejście ma szereg zalet w porównaniu z alternatywami przedstawionymi na początku tego rozdziału. W szczególności, algorytm wykorzystujący mapowanie atomów i odpowiadające im listy wiązań nie ma ograniczeń dotyczących wielkości struktury

i ilości chronionych wiązań: jest w stanie zapewnić integralność nawet pojedynczego, nieunikalnego wiązania (np. C-C) czy zabezpieczyć wiele rozłącznych wiązań. Dodatkowo, zaproponowane podejście nie jest ograniczone do oprogramowania retrosyntetycznego Chematica, lecz może poszerzyć możliwości również innych programów retrosyntetycznych, w szczególności trenowanych na bazach opublikowanych reakcji, zwiększając różnorodność proponowanych przez nie rozwiązań.

## 10. Plany syntetyczne prowadzące do bibliotek związków

Algorytmy pozwalające na analizę retrosyntetyczną bibliotek związków zostały opisane szczegółowo w publikacji **P07** oraz w publikacji **P03**.

Opisane dotychczas algorytmy pozwoliły na efektywne planowanie syntez prowadzących do pojedynczych związków. Szerszym i bardziej skomplikowanym problemem – szeroko spotykanym np. w chemii medycznej – jest potrzeba syntezy całych bibliotek związków, posiadających wspólny rdzeń lecz różniących się podstawnikami (**Schemat 11**) wpływającymi na właściwości biologiczne, takie jak zdolność do wiązania

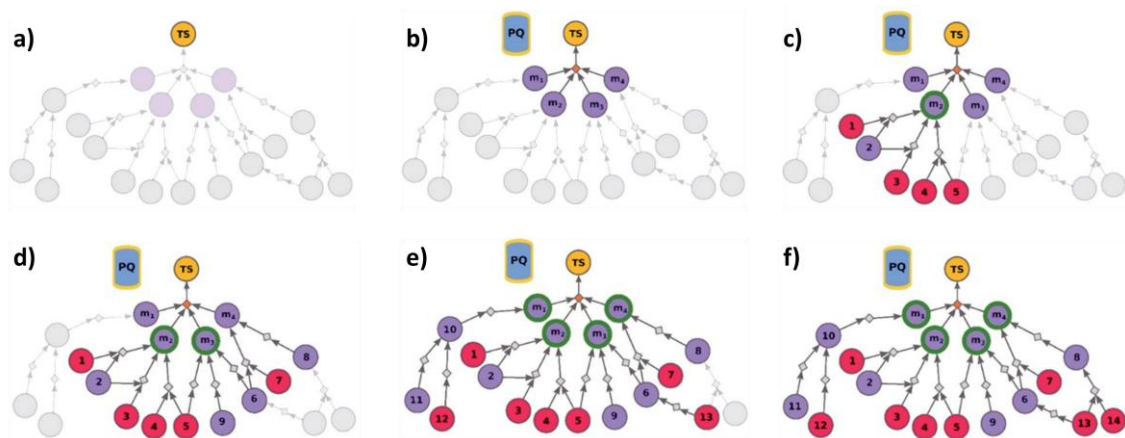


**Schemat 11.** Przykład biblioteki zawierającej wiele związków docelowych. Biblioteka pochodnych chlorocyklizyny badanych<sup>56</sup> pod kątem stosowania w terapii WZW typu C. Schemat i podpis zaadaptowane z publikacji **P07**.

się z określonymi receptorami czy farmakokinetyka i farmakodynamika. Pożądany plan syntetyczny, pozwalający na otrzymanie biblioteki w możliwie krótkim czasie często bywa diametralnie różny od optymalnych strategii prowadzących do pojedynczych składników biblioteki – znalezienie najkrótszego rozwiązania prowadzącego do pojedynczego związku często jest nieoptymalne dla całej biblioteki,

gdy zróżnicowane fragmenty są wprowadzane na samym początku a plan zawiera niewielką ilość wspólnych półproduktów. W opracowanym przy współpracy z matematykiem, dr Piotrem Dittwaldem rozwiązaniu (przedstawionym na **Schemacie 12**) pozwalającym na projektowanie syntezy całej biblioteki zdefiniowanej przez użytkownika (**Schemat 12a**) działanie algorytmu rozpoczyna się od przeprowadzenia fikcyjnej reakcji wielokomponentowej, w której substratami są pojedyncze składniki biblioteki a produktem jest węzeł TS (*target set*) grafu reprezentujący całą bibliotekę (**Schemat 12b**). Ta operacja gwarantuje warunek logiczny ORAZ, tj. że zwrócony globalny plan syntetyczny będzie zawierał ścieżki prowadzące do każdego

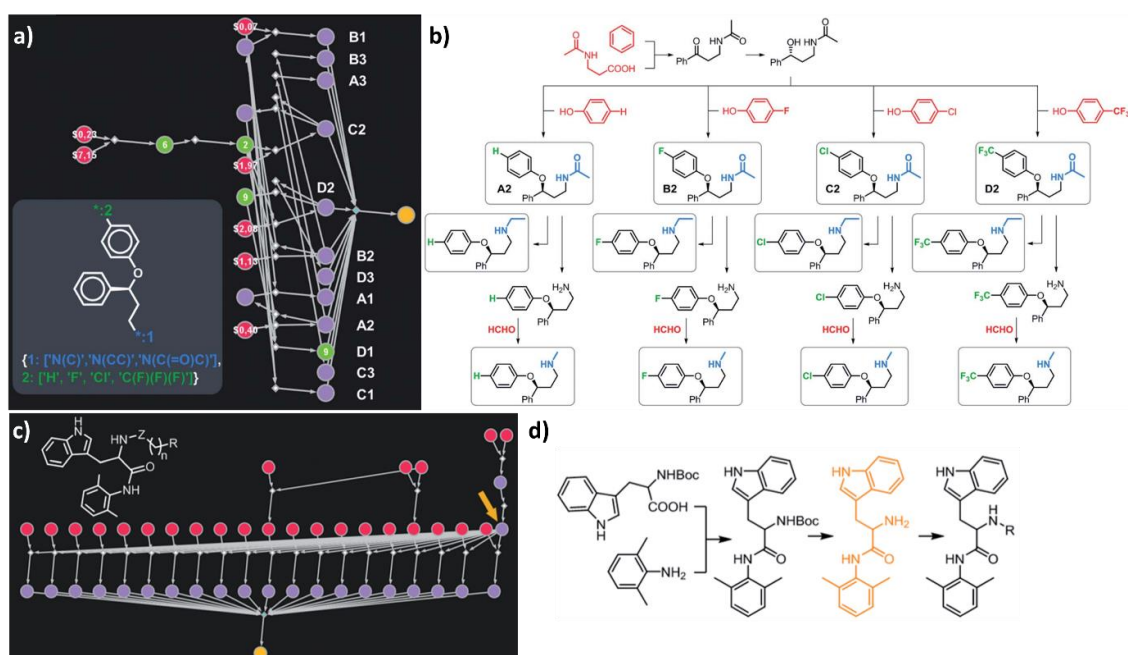
ze składników biblioteki. Oczywiście, jako że jest to tylko przekształcenie matematyczne a nie prawdziwa „reakcja wielokomponentowa”, koszt jej przypisany wynosi zero i w ten sposób nie wpływa na realistyczną wycenę planów syntezy.



**Schemat 12** Schemat działania algorytmu wyszukującego syntezę całej biblioteki. Reakcje chemiczne reprezentowane są przez romby, czerwone węzły reprezentują dozwolone materiały początkowe (wartości na węzłach odpowiadają cenom), fioletowe węzły reprezentują półprodukty, nierozwinięte części grafu wyszarzono. **a)** Użytkownik wprowadza listę cząsteczek docelowych lub definiując lokalizacje i przypisane im listy podstawników (struktury Markusha, zobacz również przykład na **Schemacie 13a**) tworzących zbiór *TS*, *target set* (żółty). **b)** Graf zostaje rozszerzony na elementy zbioru *TS* ( $m_1, m_2, m_3, m_4$ ). Następnie, algorytm rozpoczyna iteracyjne rozwijanie grafu kierowane funkcjami oceny kroków używając wspólnej kolejki priorytetowej *PQ*. **c,d)** Algorytm, *kontynuuje* rozwijanie grafu *oznaczając* elementy zbioru *TS* *do których zostały znalezione plany syntetyczne* (zielone obwódki). **e)** Wszystkie elementy zbioru *TS* są oznaczone jako *syntetyzowalne*: algorytm zwraca pierwsze rozwiązanie problemu syntezy zbioru *TS*. **f)** Wraz z postępem obliczeń graf rozwiązań jest rozbudowywany o znalezione alternatywne rozwiązania. Schemat i podpis zaadaptowane z publikacji **P07**.

Po przekształceniu *TS* w zbiór poszczególnych członków biblioteki, algorytm korzystając ze wspólnej kolejki priorytetowej *PQ* iteracyjnie rozwija graf (**Schemat 12c-f**) kierując się funkcjami oceny opisanymi w **Rozdziale 6**. W miarę rozwijania grafu, algorytm znajduje kończące się na komercyjnie dostępnych substratach plany prowadzące do poszczególnych składników biblioteki (i oznacza je jako syntetyzowalne, zielone obwódki na **Schemacie 12d-f**) – pierwsze rozwiązanie planu syntezy biblioteki jest zwracane, gdy każdy ze składników biblioteki uzyska ten status (**Schemat 12e**). W miarę postępu obliczeń, graf zawierający wyłącznie poprawne (tj., kończące się na znanych/komercyjnie dostępnych substratach) jest rozbudowywany (**Schemat 12f**) o kolejne znalezione rozwiązania. W kolejnym etapie obliczeń na otrzymanym grafie przeprowadzana jest analiza kosztów poszczególnych ścieżek, opisana szczegółowo w **Rozdziale 7**. Algorytm oceny kosztu ścieżek prowadzących do biblioteki (w przeciwieństwie do swojej podstawowej wersji) dodatkowo uwzględnia karę za ilość

unikalnych reakcji obecnych w rozwiązaniu, co gwarantuje możliwie dużą ilość wspólnych dla całej biblioteki półproduktów. Ostatnią, wartą podkreślenia, cechą algorytmu operującego na wspólnym grafie w porównaniu z osobnymi analizami przeprowadzonymi dla pojedynczych związków jest szybkość działania. Jak wykazano w publikacji **P7**, już dla niewielkiej biblioteki zawierającej 12 pochodnych fluoksetyny algorytm używający wspólnego grafu zwracał rozwiązanie sześciokrotnie szybciej (wykorzystując jednocześnie dziesięciokrotnie mniejszy graf) niż pojedyncze analizy przeprowadzane dla każdego ze składników biblioteki.

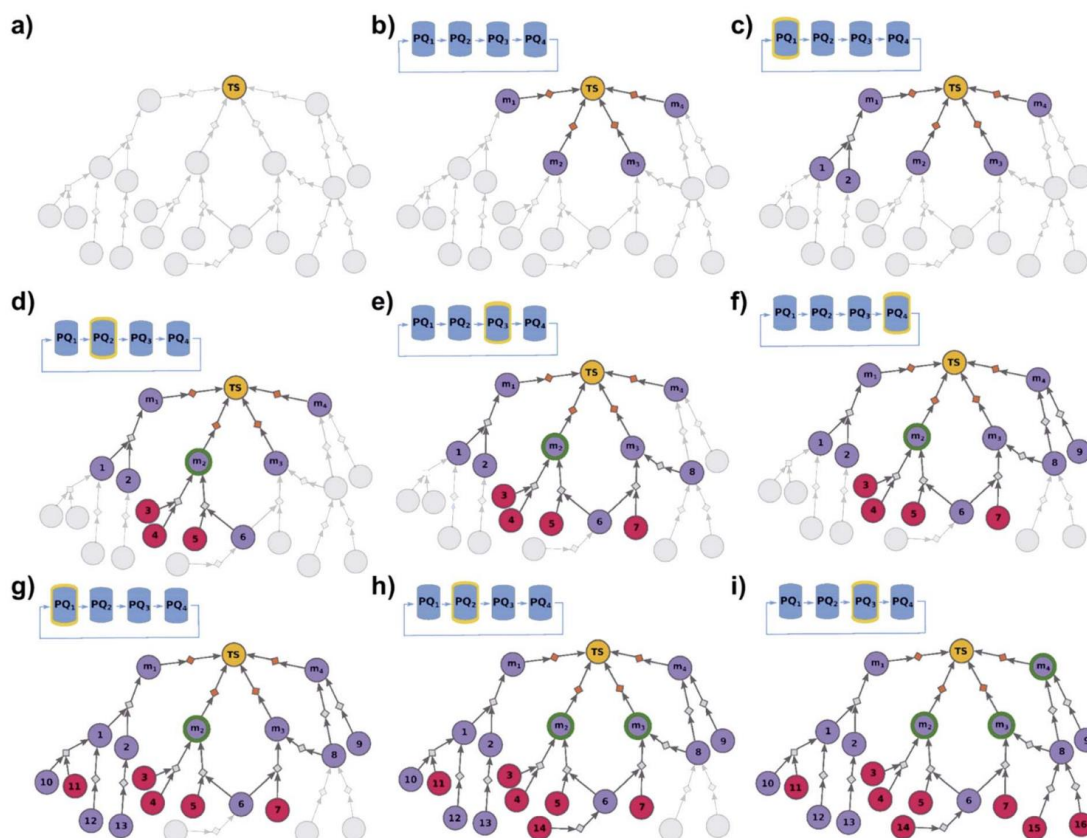


**Schemat 13.** Walidacja algorytmu. **a,b)** Plan syntezy biblioteki zawierającej 12 pochodnych fluoksetyny. **c,d)** Plan syntezy biblioteki zawierającej inhibitory RANK/RANKL. Schemat i podpis zaadaptowane z publikacji **P07**.

Działanie algorytmu zostało przeze mnie zwalidowane (publikacja **P7**, część *Chemical examples implemented in Chematica*) na szeregu obliczeń, obejmujących między innymi bibliotekę pochodnych fluoksetyny (**Schemat 13a,b**) czy porównanie z opisaną w literaturze<sup>57</sup> syntezą inhibitorów RANKL/RANK (**Schemat 13c,d**). Do uzyskania rozwiązania przedstawionego na **Schemacie 13a** i pozwalającego na syntezę całej dwunastoelementowej biblioteki (zdefiniowanej przy użyciu struktury Markusha obejmującej dwie lokalizacje i trzy/cztery przypisane im typy podstawników) przy użyciu zaledwie 18 kroków syntetycznych algorytm potrzebował zaledwie kilku minut. Przedstawiony plan syntetyczny zawiera tak mało kroków syntetycznych ponieważ algorytm zidentyfikował możliwość wykorzystania *N*-acetylowanych pochodnych fluoksetyny jako półproduktów do syntezy pozostałych elementów biblioteki. Dodatkowo, zaproponowany plan różnił się od planów syntezy poszczególnych

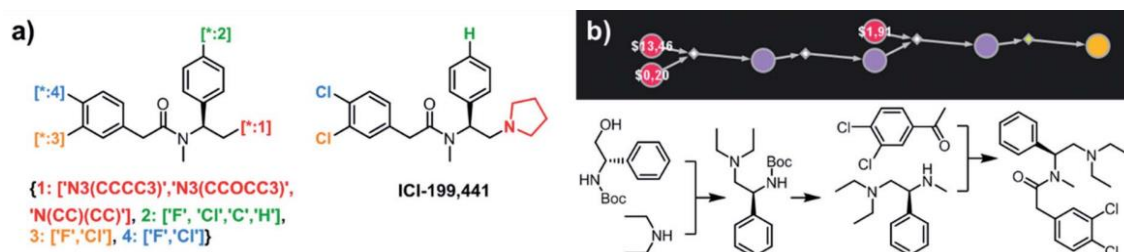
składników biblioteki i pozwalała na jej otrzymanie w mniejszej (ogólnej) ilości kroków. W drugim z przytoczonych przykładów obejmującym bibliotekę inhibitorów RANKL/RANK, algorytm zidentyfikował możliwość wykorzystania jednego, możliwego do otrzymania w dwukrokowej syntezie wspólnego bloku budulcowego (oznaczonego na **Schemacie 13c** kolorem pomarańczowym). Na uwagę zasługuje fakt, że zaproponowane rozwiązanie jest identyczne z zaproponowanym przez autorów publikacji, co potwierdza poprawność działania algorytmu.

Innym zagadnieniem związanym z syntezą bibliotek związków jest ocena syntetyzowalności jej poszczególnych elementów. W niektórych sytuacjach zadaniem chemika jest synteza wyłącznie jednego, najłatwiej osiągalnego związku ze zbioru potencjalnych kandydatów<sup>58</sup>. Oczywiście, jednym z rozwiązań pozwalającym na ocenę syntetyzowalności byłoby przeprowadzenie osobnych obliczeń dla każdego z elementów biblioteki, ale podobnie jak w poprzednio omówionym algorytmie, użycie wspólnego grafu pozwala na znaczne zredukowanie czasu obliczeń. Schemat działania opracowanego algorytmu przedstawiono na **Schemacie 14**. W przeciwieństwie do algorytmu wyszukującego syntezę wszystkich elementów biblioteki, w pierwszym kroku zdefiniowany przez użytkownika zbiór TS (**Schemat 14a**) nie jest przekształcany przy użyciu „reakcji wielokomponentowej”  $m_1 + \dots + m_n \rightarrow TS$  lecz przy wykorzystaniu szeregu „reakcji jednokomponentowych”, przekształcających zbiór TS w jego poszczególne elementy:  $m_1 \rightarrow TS, \dots, m_n \rightarrow TS$ . Ta operacja wprowadza warunek logiczny LUB (tzn. węzeł TS będzie syntetyzowalny jeśli choć jeden ze związków  $m_1, \dots, m_n$  jest syntetyzowalny) i, podobnie jak przy poprzednim algorytmie, ma przypisany zerowy koszt w obliczeniach związanych z wydajnościami czy kosztem syntez. Dodatkowo, algorytm nie używa wspólnej kolejki priorytetowej PQ lecz listy priorytetowej PL składającej się ze zbioru kolejek priorytetowych PQ1, ..., PQn, zainicjalizowanych przez poszczególne elementy zbioru TS. W czasie działania, algorytm cyklicznie (**Schemat 14c-i**) rozwija kolejki priorytetowe PQ1, ..., PQn, co niweluje możliwość „utknięcia” algorytmu w próbie znalezienia rozwiązania/syntezy dla jednego z elementów zbioru TS. Pierwsze rozwiązanie zostaje zwrócone gdy zostanie znalezione rozwiązanie do co najmniej jednego ze związków, będących elementami zbioru TS (zostanie on oznaczony jako syntetyzowalny – zielona obwódka na **Schemacie 14d**). W kolejnym etapie obliczeń, na otrzymanym grafie przeprowadzana jest analiza kosztów poszczególnych ścieżek – w tym przypadku nie jest stosowana kara za ilość unikalnych reakcji, gdyż każdy z planów obejmuje syntezę wyłącznie jednego związku.



**Schemat 14.** Schemat działania algorytmu wyszukiującego syntezę najłatwiej syntetyzowalnego elementu biblioteki. Schemat i podpis zaadaptowane z publikacji *P07*.

Działanie algorytmu zostało zwalidowane na bibliotece pochodnych agonisty receptorów  $\kappa$ -opiodowych ICI-199441, zawierającej trzy lokalizacje podstawników i odpowiadające im trzy/cztery/cztery możliwe typy podstawników zdefiniowane przy użyciu struktury Markush (**Schemat 15a**).



**Schemat 15.** Analiza retrosyntetyczna prowadząca do najłatwiej dostępnego elementu biblioteki pochodnych ICI199441. **a)** Biblioteka pochodnych ICI199441 zdefiniowana przy użyciu struktur Markusha zawierająca 48 elementów. **b)** Plan syntetyczny prowadzący do najłatwiej dostępnego składnika biblioteki z a). Schemat i podpis zaadaptowane z publikacji *P07*.

Analiza całej 48-elementowej biblioteki zakończyła się po około 10 minutach, a najwyższej ocenione rozwiązanie (**Schemat 15b**) prowadziło do analogu, w którym  $R^1 = -NEt_2$ ,  $R^2 = H$  oraz  $R^3, R^4 = -Cl$ . Trój etapowy plan syntetyczny obejmował alkirowanie

odpowiedniej aminy drugorzędowej przy użyciu komercyjnie dostępnego, zabezpieczonego fenyloglicynolu, usunięcie grupy zabezpieczającej i syntezę amidu w reakcji Wilgerodta-Kindlera, co pozwoliło na uniknięcie stosowania kosztownych chlorków kwasowych. Pozostałe analogi podstawione w pozycji **1** grupami morfolinowymi lub cyklopektyloaminowymi zostały ocenione jako trudniej syntezowalne ze względu na wyższy koszt substratów, a analogi podstawione w pozycji **2** nie pojawiły się wśród najprościej syntetyzowalnych elementów biblioteki ze względu na konieczność przygotowania odpowiednich, chiralnych aminoalkoholi.

Podsumowując, w publikacji **P07** wykazałem, że oprogramowanie retrosyntetyczne może zostać rozszerzone o możliwość analizy całych zbiorów cząsteczek pod kątem ich syntetyzowalności. Dodatkowo, opracowanie algorytmu pozwalającego na planowanie syntez najłatwiej dostępnych związków z danej biblioteki pozwoliło również na podjęcie próby rozwiązania problemu efektywnego planowania syntez związków znakowanych izotopowo.

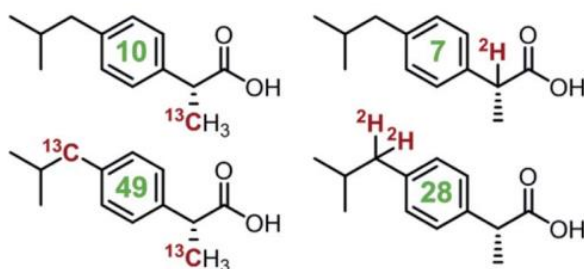
## **11. Plany syntetyczne prowadzące do związków znakowanych izotopowo**

*Algorytmy umożliwiające planowanie z syntez związków znakowanych izotopowo zostały opisane w publikacji **P07**. Analiza biblioteki związków antykoagulacyjnych znakowanych izotopowo opisana została w publikacji **P03**.*

Związki znakowane stabilnymi izotopami są szeroko wykorzystywane w chemii medycznej. Zastąpienie poszczególnych atomów izotopami ( w szczególności  $^2\text{H}$  i  $^{13}\text{C}$ ) nie wpływa znacząco na aktywność biologiczną substancji, a jednocześnie pozwala na przeprowadzanie badań dotyczących wchłaniania, dystrybucji, metabolizmu, wydalania i toksyczności leków<sup>59,60</sup>. Dodatkowo, wprowadzenie cięższych izotopów w połączeniu z technikami spektroskopii mas i technikami rozcieńczenia izotopowego<sup>61</sup> (IDMS) umożliwia szybkie pozyskiwanie danych dotyczących określenia zawartości poszukiwanej substancji w skomplikowanych próbkach środowiskowych (w analizie pozostałości pestycydów<sup>62</sup>), żywnościowych<sup>63,64</sup> czy wreszcie w testach wykrywających niedozwolone substancje i ich metabolity w próbkach pochodzących od ludzi.<sup>65</sup> W każdym z tych zastosowań pożądany związek znakowany izotopowo powinien różnić się od związku wyjściowego o określoną masę (powinien posiadać co najmniej  $n$  znakowanych izotopowo atomów), lecz lokalizacja tych atomów nie jest istotna (oczywiście, znakowane izotopy nie mogą znajdować się w pozycjach, w których ulegały

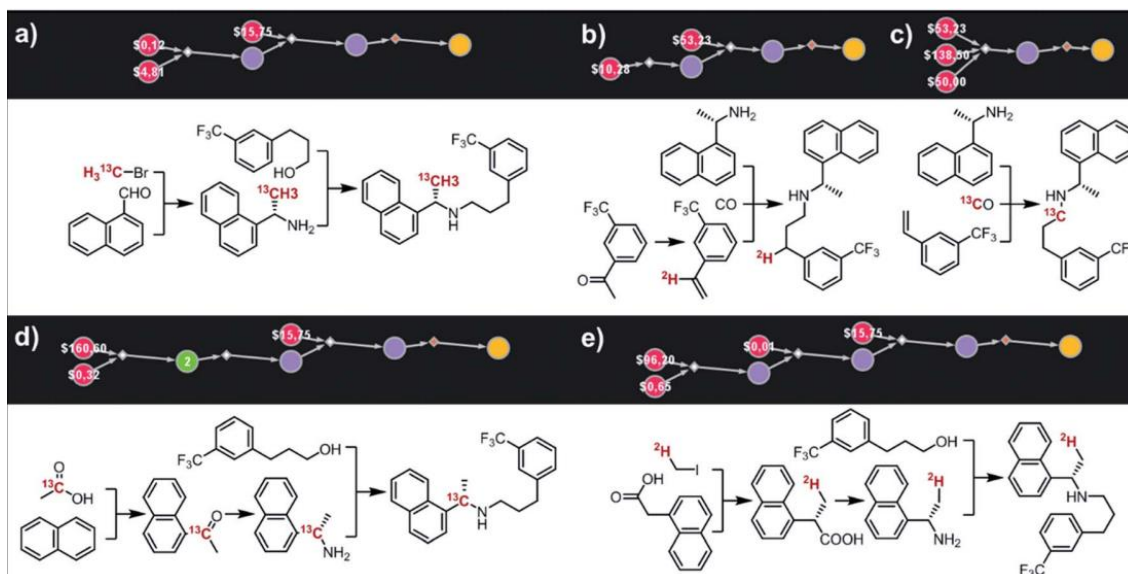


łatwo wymianie, np. w  $^2\text{H}$  znakowanych kwasach karboksylowych czy fenolach). Zbiór takich izotopomerów stanowi więc bibliotekę związków (**Schemat 16**), a zadaniem algorytmu powinno być znalezienie planu syntetycznego prowadzącego do najłatwiej dostępnego elementu tej biblioteki i kończącego się na komercyjnie dostępnych substratach.



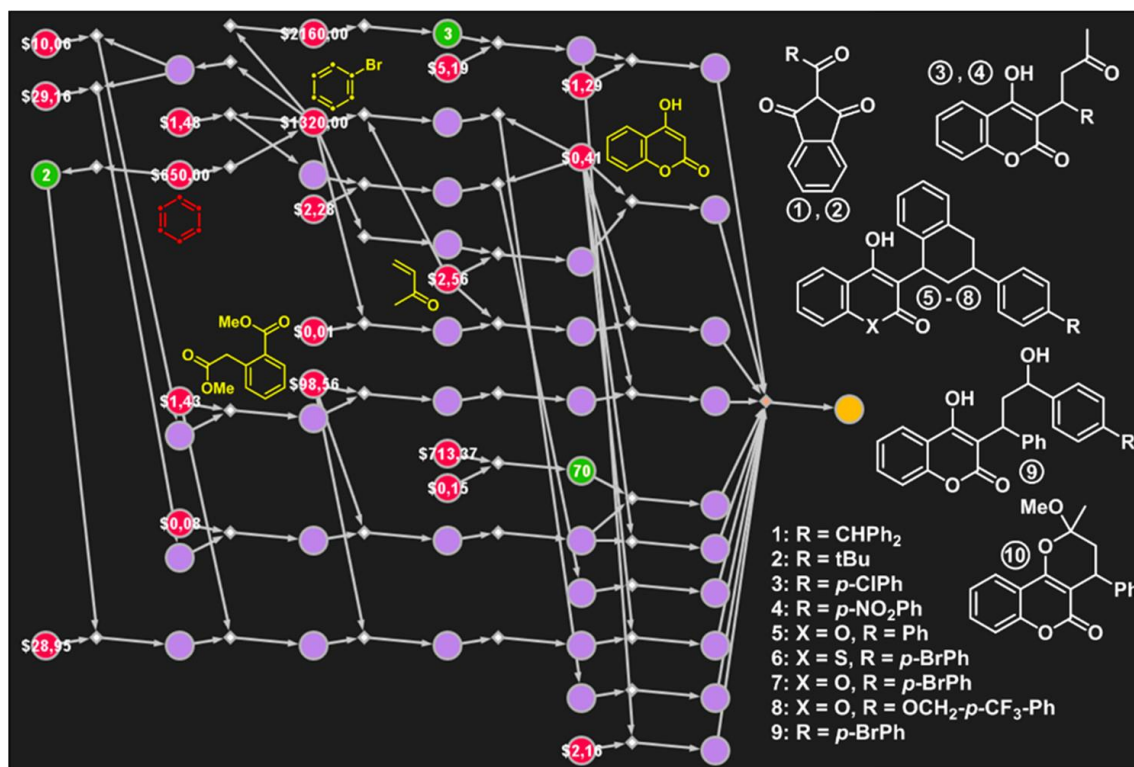
**Schemat 16.** Biblioteka M+1 ( górny rząd) oraz M+2 (dolny rząd) izotopowo znakowanych pochodnych ibuprofenu znakowanych przy użyciu  $^{13}\text{C}$  (lewa strona) i  $^2\text{H}$  (prawa strona). Zielone liczby przedstawiają liczbę możliwych unikalnych izotopomerów z wyłączeniem labilnych  $^2\text{H}$  znakowanych kwasów karboksylowych. Schemat i podpis zaadaptowane z publikacji **P07**.

W zaproponowanym rozwiązaniu wykorzystującym algorytm do znajdowania syntezy najłatwiej dostępnego elementu biblioteki, użytkownik definiuje związek bazowy (nieznakowany), oczekiwaną różnicę masy S oraz listę znakowanych atomów, które mogą być użyte np.  $^{13}\text{C}$  lub  $^{13}\text{C}/^2\text{H}$ . Ze względu na to, że nawet dla stosunkowo prostych cząsteczek ilość dostępnych izotopomerów może dochodzić do dziesiątek (**Schemat 16a**), algorytm automatycznie generuje bibliotekę wszystkich możliwych izotopomerów o określonej różnicy masy S i wytworzonych przy użyciu zdefiniowanej listy znakowanych atomów. Działanie algorytmu zostało przeze mnie zwalidowane na szeregu izotopowo znakowanych leków. W pierwszym przykładzie, opracowany algorytm został użyty do opracowania planu syntezy  $^{13}\text{C}$  lub  $^2\text{H}$  znakowanego Cinacalcetu (dla którego istnieje 39 różnych izotopomerów spełniających warunek M+1). W rozwiązaniu prowadzącym do najłatwiej osiągalnego izotopomeru znakowany atom znajduje się na grupie metylowej i pochodzi z  $^{13}\text{CH}_3\text{Br}$ . W kolejnych przykładach algorytm został użyty do zaprojektowania M+1,  $^{13}\text{C}$  znakowanych pochodnych AMG-319 (21 izotopomerów), Lasmiditanu (14 izotopomerów), Roluperidonu (18 izotopomerów), Pitolisantu (13 izotopomerów) i Almotriptanu (13 izotopomerów).



**Schemat 17.** Synteza izotopowo znakowanych M+1 pochodnych Cinacalcetu znakowanych z użyciem  $^{13}\text{C}$  (a,c,d) lub  $^2\text{H}$  (b,e). Przedstawiono 5 planów syntetycznych o najniższym koszcie otrzymanych w wyniku analizy retrosyntetycznej trwającej poniżej 10 minut. Schemat i podpis zaadaptowane z publikacji *P07*.

Ostatnim zadaniem, w którym sprawdziłem opracowane algorytmy była próba zaplanowania planu syntetycznego pozwalającego na otrzymanie zestawu znakowanych izotopowo pochodnych 10 substancji o działaniu antykoagulacyjnym (**Schemat 18**). Zadaniem algorytmu było opracowanie planu syntezy całej biblioteki, a plan syntetyczny prowadzący do każdego z jej elementów miał prowadzić do najłatwiej dostępnego, M+6 znakowanego izotopomeru. Dla tej biblioteki ilość możliwych izotopomerów przekracza 2 miliony, a większość z nich jest w zasadzie niesyntetyzowalna (ze względu na brak komercyjnie dostępnych M+3 czy M+4 znakowanych pochodnych aromatycznych ich synteza musiałaby wychodzić z alifatycznych substratów oraz obejmować karbocyklizację i aromatyzację, zwiększając znacznie ilość wymaganych kroków syntetycznych). Aby zmniejszyć rozmiar zadania obliczeniowego, uzyskane izotopomery zostały poddane filtrowaniu względem katalogu komercyjnie dostępnych  $^{13}\text{C}$  znakowanych substratów. Po wykluczeniu izotopomerów wymagających użycia więcej niż dwóch znakowanych substratów lub wymagających ‘konstrukcji’ pierścienia benzenowego pozostałe 33 izotopomery zostały użyte jako biblioteka poddawana analizie.



**Schemat 18.** Analiza retrosyntetyczna prowadząca do biblioteki najłatwiej syntetyzowalnych izotopomerów M+6 znakowanych pochodnych 10 związków o działaniu antykoagulacyjnym (przedstawionych po prawej stronie). Żółty węzeł po prawej stronie reprezentuje całą bibliotekę (33 izotomery). Schemat i podpis zaadaptowane z publikacji **P03**.

Rozwiązanie syntetyczne uzyskane po około 18h obliczeń przedstawiono na **Schemacie 18**. W zaplanowanym planie algorytm zidentyfikował możliwość użycia wielu wspólnych substratów i półproduktów, co pozwoliło zmniejszyć ilość kroków syntetycznych potrzebnych do otrzymania dziesięciu związków do 40: komercyjnie dostępny <sup>13</sup>C<sub>6</sub> benzen jest użyty jako jedyne źródło znakowanych atomów, a uzyskany z niego <sup>13</sup>C bromobenzen jest wykorzystywany w syntezie 9 pochodnych.

## 12. Automatyzacja syntezy bibliotek związków – chemia iteracyjna

*Algorytmy umożliwiające odkrywanie nowych iteracyjnych sekwencji reakcji zostały opisane w publikacji **P08**.*

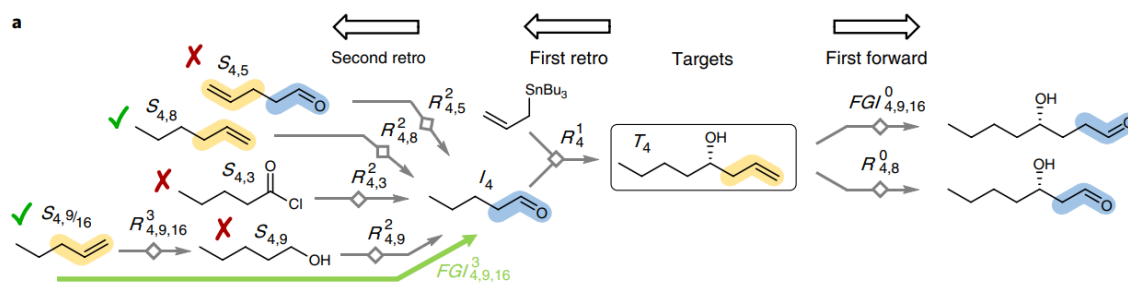
Zautomatyzowana synteza iteracyjna przebiegająca na nośnikach stałych jest jedną z najbardziej efektywnych metod syntezy pozwalającą na uzyskiwanie wysokich wydajności nawet w przypadku kilkunastokrotowych syntez. W tym podejściu, plan syntetyczny opiera się na powtarzającym się użyciu zaledwie kilku reakcji chemicznych (odpowiadających etapom sprzęgania/odbezpieczenia) przebiegających pomiędzy odpowiednio sfunkcjonalizowanymi blokami budulcowymi. W każdym z etapów

sprzęgania co najmniej jeden z użytych substratów zawiera grupę funkcyjną w formie zamaskowanej (**Schemat 19**).



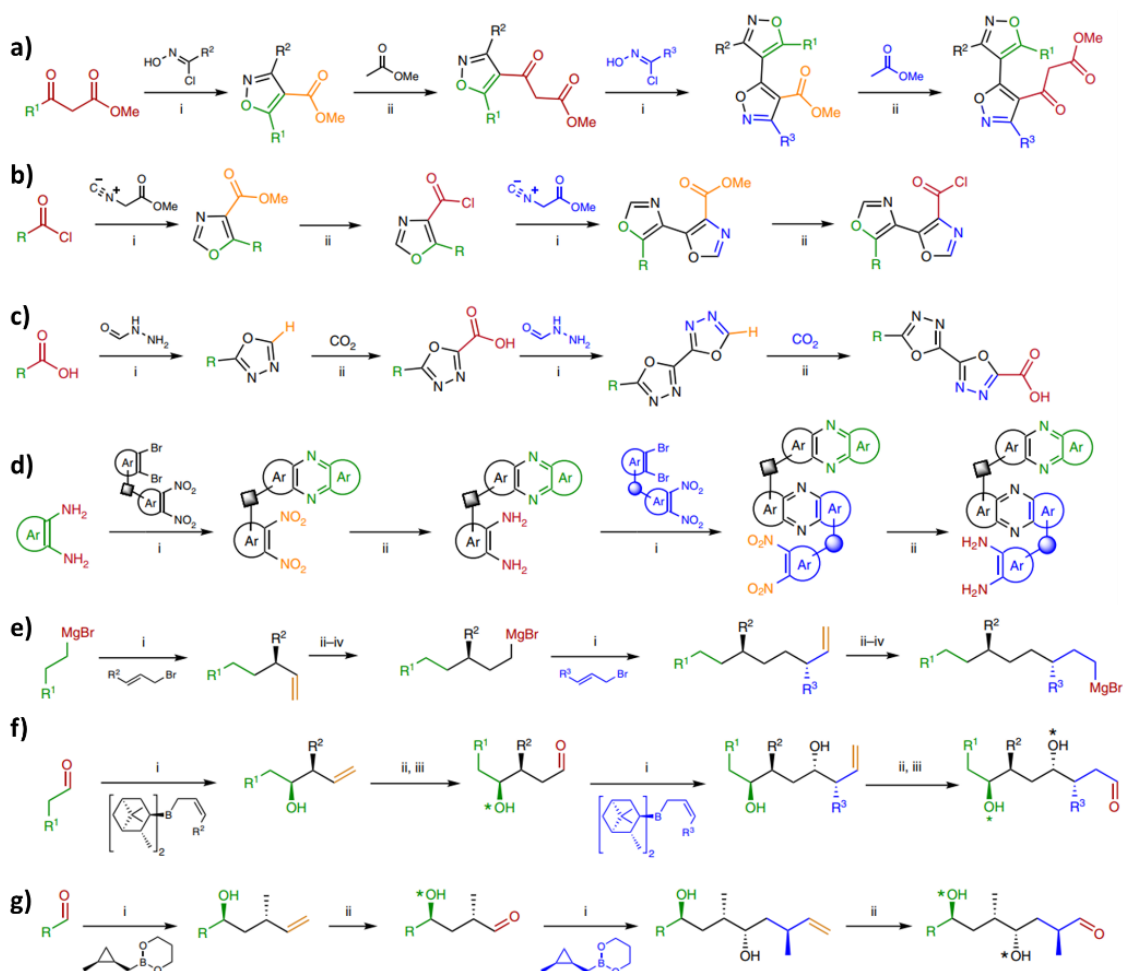
**Schemat 19.** Synteza iteracyjna. Podczas etapu sprzęgania jedna z reagujących grup funkcyjnych (FG) jest obecna w formie zamaskowanej (oznaczone łukiem). X oznacza grupę funkcyjną poddawaną reakcji z FG. W rzeczywistości pojedyncza iteracja może być bardziej skomplikowana a FG może być odtwarzane w kilkukrokowej sekwencji. Schemat i podpis zaadaptowane z publikacji **P08**.

Jak dotąd, opracowane rozwiązania pozwoliły na prowadzenie syntezy bibliotek oligosacharydów<sup>66,67</sup>, oligopeptydów<sup>68</sup> czy oligonukleotydów<sup>69</sup>. W ostatnich latach opracowane zostały również metodologie pozwalające na przeprowadzenie iteracyjnej katalizowanej palladem reakcji Suzuki z udziałem odpowiednich zabezpieczonych haloboranów<sup>70</sup>, co pozwoliło na automatyczną syntezę sp<sup>2</sup> bogatych produktów naturalnych, leków i materiałów funkcjonalnych<sup>71</sup>. Jak dotąd, zbiór opracowanych sekwencji reakcji pozwalających na prowadzenie syntezy iteracyjnej jest dość wąski, a znane kombinacje często obejmują wyłącznie bardzo proste reakcje sprzęgania/odbezpieczenia. Ostatnim zadaniem, które obejmowała moja praca doktorska była próba rozszerzenia znanego repertuaru sekwencji pozwalających na prowadzenie syntezy iteracyjnej. W tym celu, zbiór reguł chemicznych będących podstawą oprogramowania Chematica/Synthia (opisanych w **Rozdziale 6 Komentarza**) został poddany szczegółowej analizie przy użyciu dwóch algorytmów, stworzonych wspólnie z dr Piotrem Dittwaldem i dr Sarą Szymkuć.



**Schemat 20.** Schemat działania algorytmu umożliwiającego znajdowanie sekwencji iteracyjnych. Algorytm przeprowadza reakcję retrosyntetyczną  $R_n^1$  na przykładowym produkcie  $T_n$  ( $n=4$ ) otrzymując półprodukt  $I_n$ , poddawany przekształceniu przy użyciu wszystkich możliwych reakcji retrosyntetycznych  $R_{n,m}^2$  w substraty  $S_{n,m}$ . Sekwencje iteracyjne muszą posiadać co najmniej jeden motyw obecny w substracie i produkcie (a nieobecny w półprodukcie, zaznaczony na żółto) oraz co najmniej jeden motyw obecny w półprodukcie, a nieobecny w substracie i produkcie (niebieski). Schemat i podpis zaadaptowane z publikacji **P08**.

Działanie pierwszego z algorytmów (przedstawionego na **Schemacie 20** i opisanego w szczegółach w publikacji **P08**) rozpoczyna się od przeprowadzenia reakcji retrosyntetycznej  $R_n^1$  dla „typowego” produktu  $T_n$  ( $n=4$ ) przypisanego dla w każdej z ponad 100 tysięcy reguł chemicznych, którymi operuje *Chematica*. Następnie, dla otrzymanego półproduktu  $I_n$  generowane są przy użyciu reakcji  $R_m^2$  wszystkie możliwe prowadzące do niego substraty  $S_{n,m}$ . Na tym etapie, algorytm używa nie tylko pojedynczych reakcji ale bierze również pod uwagę kilkukrokowe sekwencje, umożliwiające przekształcenia grup funkcyjnych (ang. FGI, functional group interconversions, **Rozdział 8.3**). W kolejnym kroku dla otrzymanych kandydatów algorytm sprawdza szereg kryteriów gwarantujących wybranie kombinacji reakcji pozwalających na syntezę iteracyjną. Po pierwsze, sekwencja nie może być prostą pętlą i prowadzić do wyjściowego związku ( $S_{n,m} = T_n$ ), prowadzić do substratu bardziej skomplikowanego niż produkt a w żadnym z kroków nie może pojawiać się grupa niekompatybilna z warunkami danej reakcji. Po drugie, musi istnieć co najmniej jedna podstruktura nieobecna w półprodukcie  $I_n$  a obecna w produkcie  $T_n$  i substracie  $S_{n,m}$  (zaznaczona na Schemacie 20 kolorem żółtym) oraz co najmniej jedna podstruktura obecna w półprodukcie  $I_n$  i nieobecna w produkcie  $T_n$  i substracie  $S_{n,m}$  (zaznaczona na Schemacie 20 kolorem niebieskim). Wreszcie, algorytm sprawdza czy produkt  $T_n$  może uczestniczyć jako substrat w reakcji odwrotnej do reakcji  $R_m^2$ . Opisane warunki strukturalne gwarantują, że znaleziona sekwencja będzie odtwarzać niezbędną do reakcji grupę funkcyjną. Zdefiniowany w ten sposób algorytm pozwolił na zidentyfikowanie ponad 2600 sekwencji, spośród których znalazły się zarówno już znane sekwencje (z których część umieszczona została w publikacji **P08**, część **S4**) jak i zupełnie nowe kombinacje, pozwalające na iteracyjną syntezę układów poliheteroaromatycznych (**Schemat 21a-d** oraz publikacja **P08**, **Figura 3**) czy stereoselektywną syntezę układów obecnych między innymi w produktach naturalnych (**Schemat 21e-g** oraz publikacja **P08**, **Figura 4** oraz **Figury EF1-EF9**).



**Schemat 21.** Przykłady nowoodkrytych sekwencji reakcji umożliwiających (a-d) iteracyjną syntezę układów poliheteroaromatycznych oraz (e-g) iteracyjną stereoselektywną syntezę fragmentów produktów naturalnych zidentyfikowanych przez algorytm. **a)** Iteracyjna synteza izoksazoli z chlorków imidoilowych i ketoestrów, regenerowanych poprzez kondensację produktu z octanem metyłu; **b)** Iteracyjna synteza oksazoli z izocyjanek i chlorków kwasowych; **c)** Iteracyjna synteza 1,3,4-oksadiazoli przebiegająca przez kondensację formylowodniany z kwasami karboksylowymi, regenerowanymi poprzez reakcje metalacji i karboksylację; **d)** Iteracyjna synteza fenazyń poprzez sprzęganie diamin, regenerowanych poprzez redukcję grup nitrowych; **e)** Stereoselektywna synteza deoksypropionianów wykorzystująca asymetryczną substytucję alilową. Związek magnezooorganiczny jest regenerowany przez hydroborowanie alkenu, reakcję Appela i metalację; **f)** Stereoselektywna synteza polipropionianów zawierających ugrupowania metylenowe wykorzystująca reakcje asymetrycznego krotylowania aldehydów regenerowanych jest poprzez hydroborowanie alkenów i utlenienie uzyskanych alkoholi; **g)** Asymetryczna synteza przebiegająca przez homokrotylowanie aldehydów, regenerowanych przez ozonolizy uzyskanych alkenów. Schemat i podpis zaadaptowane z publikacji *P08*.

Drugi z opracowanych algorytmów pozwolił zidentyfikować sekwencje reakcji, w których produkt reakcji  $T_n$  z bazy reguł chemicznych nie zawierał wszystkich wymaganych grup funkcyjnych. Oczywiście, mimo że 'przykładowy' produkt  $T_n$  (najczęściej mający bardzo prostą strukturę) nie zawiera tych fragmentów, dowolny inny

związek uczestniczący w danej reakcji  $R_n$  może zawierać wymagane grupy funkcyjne zlokalizowane w dowolnych miejscach (Schemat 19), a sama sekwencja będzie spełniała warunki wymagane do odtworzenia grupy FG. W tym przypadku, algorytm identyfikował znaną sekwencję jako iteracyjną jeżeli dodatkowa grupa funkcyjna była kompatybilna z reakcją  $R_m^1$ , była nieobecna w półprodukcie  $I_n$  oraz pojawiała się w związku  $S_{n,m}$ . Działanie tego algorytmu na bazie reguł chemicznych umożliwiło identyfikację ponad 500 tysięcy sekwencji reakcji (należących do kilku tysięcy klas, obejmujących podobne do siebie warianty) umożliwiającą prowadzenie syntezy iteracyjnej. Podobnie jak w przypadku poprzedniego algorytmu, wśród zidentyfikowanych strategii znalazły się już znane rozwiązania (między innymi synteza peptydów w obecności odpowiednio zabezpieczonych grup aminowych lub karboksylowych czy synteza Suzuki w obecności zabezpieczonych kwasów boronowych) jak i tysiące nowych strategii, pozwalających na iteracyjną syntezę skoniugowanych heterocykli (**Schemat 21d**), przeprowadzanie syntezy stereokontrolowanej czy nawet uwzględniające reakcje multikomponentowe. Część ze znalezionych sekwencji umożliwiających stereo kontrolowaną syntezę fragmentów produktów naturalnych została poddana walidacji eksperymentalnej, przeprowadzonej przez chemików z IChO PAN.

### 13. Podsumowanie

Moja praca badawcza w dużej mierze dotyczyła stworzenia i rozwoju oprogramowania Chematica/Synthia umożliwiającego w pełni autonomiczne planowanie syntez związków organicznych. Swoją pracę rozpocząłem od rozszerzenia bazy reguł reaktywności chemicznej programu o reakcje obejmujące metodologie wykorzystywane w stereokontrolowanej syntezie produktów naturalnych. W kolejnym etapie uczestniczyłem w opracowaniu i walidacji algorytmów, umożliwiających zastosowanie oprogramowania Chematica/Synthia do efektywnego planowania syntez z uwzględnieniem skali prowadzenia procesu. Wykazałem, że działanie oprogramowania retrosyntetycznego może zostać ukierunkowane zarówno na projektowanie krótkich syntez wykorzystujących bardziej skomplikowane materiały początkowe (np. w małoskalowych syntezach związków lekopodobnych na początkowych etapach badań klinicznych) jak i wielkoskalowych syntez wychodzących z bardzo tanich substratów kosztem nieco większej ilości kroków syntetycznych.

Kolejna część mojej pracy badawczej dotyczyła rozszerzenia zakresu stosowalności stworzonego oprogramowania o możliwość planowania syntez omijających rozwiązania chronione prawem patentowym. W tym celu, opracowałem algorytmy pozwalające na i) analizę sieci reakcji znanych syntez i identyfikację kluczowych wiązań oraz ii) przeprowadzenie analizy retrosyntetycznej pomijającej dyskoneksje w obrębie zidentyfikowanych wiązań. Wykazałem, że umożliwia to projektowanie nowych strategii syntetycznych prowadzących do obecnych na rynku leków, których metody otrzymywania chronione są dziesiątkami patentów.

Kolejnym ze zrealizowanych zadań było wzbogacenie programu o możliwość przeprowadzenia analizy retrosyntetycznej zbioru cząsteczek. W pierwszej części, skupiłem się na możliwości projektowania globalnych planów syntetycznych prowadzących do bibliotek związków posiadających wspólny rdzeń lecz różniących się podstawnikami wpływającymi na przykład na właściwości biologiczne kandydatów na leki. W drugiej części zająłem się problemem syntezy związków znakowanych izotopowo, szeroko wykorzystywanych w chemii medycznej. Ze względu na fakt, że lokalizacja znakowanych atomów na ogół może być dowolna problem ten mógł zostać rozwiązany poprzez stworzenie biblioteki możliwych izotopomerów i jej analizę retrosyntetyczną.

W ostatniej części swojej pracy badawczej podjąłem się rozszerzenia repertuaru reakcji i możliwych do uzyskania związków w tzw. chemii iteracyjnej, polegającej na prowadzeniu syntezy z wykorzystaniem powtarzających się kilku reakcji przebiegających pomiędzy odpowiednio sfunkcjonalizowanymi blokami budulcowymi. Analiza stworzonej w początkowym etapie pracy badawczej bazy reguł pozwoliła na odkrycie tysięcy nowych sekwencji reakcji pozwalających na otrzymanie skoniugowanych heterocykli, umożliwiających przeprowadzenie stereokontrolowanej syntezy fragmentów produktów naturalnych a nawet iteracyjną syntezę z wykorzystaniem reakcji multikomponentowych.



## 14. Referencje

1. Corey, E. J. The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules(Nobel Lecture). *Angew. Chem. Int. Ed.* **1991**, *30*, 455–465.
2. Corey, E.; Long, A.; Rubenstein, S. Computer-Assisted Analysis in Organic Synthesis. *Science* **1985**, *228*, 408–418.
3. Corey, E. J.; Cramer, R. D.; Howe, W. J. Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates. *J. Am. Chem. Soc.* **1972**, *94*, 440–459.
4. Varkony, T. H.; Smith, D. H.; Djerassi, C. Computer-Assisted Structure Manipulation. *Tetrahedron* **1978**, *34*, 841–852.
5. Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. A Logic-Based Program for Synthesis Design. *J. Am. Chem. Soc.* **1985**, *107*, 5228–5238.
6. Ugi, I.; Bauer, J.; Baumgartner, R.; Fontain, E.; Forstmeyer, D.; Lohberger, S. Computer Assistance in the Design of Syntheses and a New Generation of Computer Programs for the Solution of Chemical Problems by Molecular Logic. *Pure Appl. Chem.* **1988**, *60*, 1573–1586.
7. Vléduts, G. É.; Finn, V. K. Creating a Machine Language for Organic Chemistry. *Inf. Storage Retr.* **1963**, *1*, 101–116.
8. Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178–192.
9. LHASA, Radboud University Nijmegen  
<https://web.archive.org/web/20200203095323/http://www.cheminf.cmbi.ru.nl/cheminf/lhasa>,  
dostęp 16 sierpnia 2022
10. Logic and Heuristics Applied to Synthetic Analysis;  
[https://web.archive.org/web/20050519100203if\\_/http://lhasa.harvard.edu:80/?page=retro.htm](https://web.archive.org/web/20050519100203if_/http://lhasa.harvard.edu:80/?page=retro.htm),  
dostęp 16 sierpnia 2022
11. Corey, E. J.; Long, A. K.; Mulzer, J.; Orf, H. W.; Johnson, A. P.; Hewett, A. P. W. Computer-Assisted Synthetic Analysis. Long-Range Search Procedures for Antithetic Simplification of Complex Targets by Application of the Halolactonization Transform. *J. Chem. Inf. Model.* **1980**, *20*, 221–230.
12. van Rozendaal, E. L. M. Some approaches to the synthesis of taxol and its derivatives: Total-synthesis based on a LHASA analysis and semi-synthesis starting from taxine B; Radboud University Nijmegen: Nijmegen, Netherlands, **1994**.  
[https://repository.ubn.ru.nl/bitstream/handle/2066/30052/mmubn000001\\_181282496.pdf](https://repository.ubn.ru.nl/bitstream/handle/2066/30052/mmubn000001_181282496.pdf)  
(dostęp 16 sierpnia 2022)
13. Wuts, P. G. M.; Greene, T. W. *Greene's Protective Groups in Organic Synthesis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006.
14. Evans, D. A. History of the Harvard ChemDraw Project. *Angew. Chem. Int. Ed.* **2014**, *53*, 11140–11145.

15. Hendrickson, J. B.; Toczko, A. G. SYNGEN Program for Synthesis Design: Basic Computing Techniques. *J. Chem. Inf. Model.* **1989**, *29*, 137–145.
16. Hendrickson, J. B.; Toczko, A. G. Systematic Synthesis Design: The SYNGEN Program. *Pure Appl. Chem.* **1989**, *61*, 589–592.
17. Hendrickson, J. B. The SYNGEN Approach to Synthesis Design. *Anal. Chim. Acta* **1990**, *235*, 103–113.
18. Takahashi, M.; Dogane, I.; Yoshida, M.; Yamachika, H.; Takabatake, T.; Bersohn, M. The Performance of a Noninteractive Synthesis Program. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 436–441.
19. Tanaka, A.; Okamoto, H.; Bersohn, M. Construction of Functional Group Reactivity Database under Various Reaction Conditions Automatically Extracted from Reaction Database in a Synthesis Design System. *J. Chem. Inf. Model.* **2010**, *50*, 327–338.
20. Tanaka, A.; Kawai, T.; Takabatake, T.; Oka, N.; Okamoto, H.; Bersohn, M. Synthesis of an Azaspirane via Birch Reduction Alkylation Prompted by Suggestions from a Computer Program. *Tetrahedron Lett.* **2006**, *47*, 6733–6737.
21. Wipke, W. T.; Dyott, T. M. Simulation and Evaluation of Chemical Synthesis. Computer Representation and Manipulation of Stereochemistry. *J. Am. Chem. Soc.* **1974**, *96*, 4825–4834.
22. Wipke, W. T.; Ouchi, G. I.; Krishnan, S. Simulation and Evaluation of Chemical Synthesis—SECS: An Application of Artificial Intelligence Techniques. *Artif. Intell.* **1978**, *11*, 173–193.
23. Gelernter, H. L.; Sanders, A. F.; Larsen, D. L.; Agarwal, K. K.; Boivie, R. H.; Spritzer, G. A.; Searleman, J. E. Empirical Explorations of SYNCHEM. *Science (80-. )*. **1977**, *197*, 1041–1049.
24. Agarwal, K. K.; Larsen, T. D. L.; Gelernter, H. L. Application of Chemical Transforms in Synchem2, a Computer Program for Organic Synthesis Route Discovery. *Comput. Chem.* **1978**, *2*, 75–84.
25. Hanessian, S.; Franco, J.; Larouche, B. The Psychobiological Basis of Heuristic Synthesis Planning - Man, Machine and the Chiron Approach. *Pure Appl. Chem.* **1990**, *62*, 1887–1910.
26. <https://www.computerhistory.org/pdp-1/the-machine/>, dostęp 16 sierpnia 2022
27. Reaxys, <https://www.reaxys.com>, (dostęp 16 sierpnia 2022).
28. SciFinder, <https://scifinder-n.cas.org>, (dostęp 16 sierpnia 2022).
29. Lowe, D. M. Chemical reactions from US pat., (1976-Sep2016), 2017, [https://figshare.com/articles/Chemical\\_reactions\\_from\\_US\\_patents\\_1976-Sep2016\\_/5104873](https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873), (dostęp 16 sierpnia 2022)
30. Ravitz, O. Data-Driven Computer Aided Synthesis Design. *Drug Discov. Today Technol.* **2013**, *10*, e443–e449.
31. Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49*, 593–602.

32. Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H. Route Design in the 21st Century: The IC SYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **2015**, *19*, 357–368.
33. Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.
34. Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; et al. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365*, eaax1566.
35. Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604–610.
36. Hendrickson, J. B. Teaching Alternative Syntheses: The SYNGEN Program; 1996; ACS Symposium Series Vol. 626, *Green Chemistry*, Chapter 16, pp 214–231.
37. Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. NIPS, 2017, 2607–2616.
38. Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370–377.
39. Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103–1113.
40. Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. “Found in Translation”: Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *Chem. Sci.* **2018**, *9*, 6091–6098.
41. Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434–443.
42. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design. *J. Chem. Inf. Model.* **2011**, *51*, 3093–3098.
43. Baylon, J. L.; Cilfone, N. A.; Gulcher, J. R.; Chittenden, T. W. Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification. *J. Chem. Inf. Model.* **2019**, *59*, 673–688.
44. Harrington, P. E.; St. Jean, D. J.; Clarine, J.; Coulter, T. S.; Croghan, M.; Davenport, A.; Davis, J.; Ghiron, C.; Hutchinson, J.; Kelly, M. G.; et al. The Discovery of an Orally Efficacious Positive Allosteric Modulator of the Calcium Sensing Receptor Containing a Dibenzylamine Core. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 5544–5547.
45. Giannerini, M.; Hornillos, V.; Vila, C.; Fañanás-Mastral, M.; Feringa, B. L. Hindered Aryllithium Reagents as Partners in Palladium-Catalyzed Cross-Coupling: Synthesis of Tri- and Tetra- Ortho -Substituted Biaryls under Ambient Conditions. *Angew. Chem. Int. Ed.* **2013**, *52*, 13329–13333.

46. Cort, A. D.; Mandolini, L.; Panaioli, S. Selective One-Pot Oxidation of Methylarenes to Benzyl Alcohols with the Copper(II)-Peroxydisulfate System. *Synth. Commun.* **1988**, *18*, 613–616.
47. Otsuka, K.; Zenibayashi, Y.; Itoh, M.; Totsuka, A. Presence and Significance of Two Diastereomers of  $\beta$ -Methyl- $\gamma$ -Octalactone in Aged Distilled Liquors. *Agric. Biol. Chem.* **1974**, *38*, 485–490.
48. Koschker, P.; Kähny, M.; Breit, B. Enantioselective Redox-Neutral Rh-Catalyzed Coupling of Terminal Alkynes with Carboxylic Acids Toward Branched Allylic Esters. *J. Am. Chem. Soc.* **2015**, *137*, 3131–3137.
49. Mao, B.; Geurts, K.; Fañanas-Mastral, M.; van Zijl, A. W.; Fletcher, S. P.; Minnaard, A. J.; Feringa, B. L. Catalytic Enantioselective Synthesis of Naturally Occurring Butenolides via Hetero-Allylic Alkylation and Ring Closing Metathesis. *Org. Lett.* **2011**, *13*, 948–951.
50. Ito, M.; Osaku, A.; Shiibashi, A.; Ikariya, T. An Efficient Oxidative Lactonization of 1,4-Diols Catalyzed by Cp\*Ru(PN) Complexes. *Org. Lett.* **2007**, *9*, 1821–1824.]
51. Corey, E. J., Cheng X.-M. The logic of chemical synthesis. New York, NY: John Wiley & Sons; 1995.
52. Gajewska, E. P.; Szymkuć, S.; Dittwald, P.; Startek, M.; Popik, O.; Mlynarski, J.; Grzybowski, B. A. Algorithmic Discovery of Tactical Combinations for Advanced Organic Syntheses. *Chem* **2020**, *6*, 280–293.
53. Lin, Y.; Zhang, Z.; Mahjour, B.; Wang, D.; Zhang, R.; Shim, E.; McGrath, A.; Shen, Y.; Brugger, N.; Turnbull, R.; Trice, S.; Jasty, S.; Cernak, T. Reinforcing the Supply Chain of Umifenovir and Other Antiviral Drugs with Retrosynthetic Software. *Nat. Commun.* **2021**, *12*, 7327.
54. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.
55. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.
56. He, S.; Xiao, J.; Dulcey, A. E.; Lin, B.; Rolt, A.; Hu, Z.; Hu, X.; Wang, A. Q.; Xu, X.; Southall, N.; et al. Discovery, Optimization, and Characterization of Novel Chlorcyclizine Derivatives for the Treatment of Hepatitis C Virus Infection. *J. Med. Chem.* **2016**, *59*, 841–853.
57. Jiang, M.; Peng, L.; Yang, K.; Wang, T.; Yan, X.; Jiang, T.; Xu, J.; Qi, J.; Zhou, H.; Qian, N.; et al. Development of Small-Molecules Targeting Receptor Activator of Nuclear Factor-KB Ligand (RANKL)—Receptor Activator of Nuclear Factor-KB (RANK) Protein–Protein Interaction by Structure-Based Virtual Screening and Hit Optimization. *J. Med. Chem.* **2019**, *62*, 5370–5381.
58. Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1*, 8.
59. Mutlib, A. E. Application of Stable Isotope-Labeled Compounds in Metabolism and in Metabolism-Mediated Toxicity Studies. *Chem. Res. Toxicol.* **2008**, *21*, 1672–1689.

60. Unkefer, C. J.; Martinez, R. A. The Use of Stable Isotope Labelling for the Analytical Chemistry of Drugs. *Drug Test. Anal.* **2012**, *4*, 303–307.
61. M. Berglund, w Handbook of Stable Isotope Analytical Techniques, ed. P. A. de Groot, Elsevier, Introduction to Isotope Dilution Mass Spectrometry (IDMS), 2004, pp. 820–834.
62. Woudneh, M. B.; Sekela, M.; Tuominen, T.; Gledhill, M. Isotope Dilution High-Resolution Gas Chromatography/High-Resolution Mass Spectrometry Method for Analysis of Selected Acidic Herbicides in Surface Water. *J. Chromatogr. A* **2006**, *1133*, 293–299.
63. Al-Taher, F.; Banaszewski, K.; Jackson, L.; Zweigenbaum, J.; Ryu, D.; Cappozzo, J. Rapid Method for the Determination of Multiple Mycotoxins in Wines and Beers by LC-MS/MS Using a Stable Isotope Dilution Assay. *J. Agric. Food Chem.* **2013**, *61*, 2378–2384.
64. Abu-El-Haj, S.; Bogusz, M. J.; Ibrahim, Z.; Hassan, H.; Al Tufail, M. Rapid and Simple Determination of Chloropropanols (3-MCPD and 1,3-DCP) in Food Products Using Isotope Dilution GC-MS. *Food Control* **2007**, *18*, 81–90.
65. Wang, I.-T.; Feng, Y.-T.; Chen, C.-Y. Determination of 17 Illicit Drugs in Oral Fluid Using Isotope Dilution Ultra-High Performance Liquid Chromatography/Tandem Mass Spectrometry with Three Atmospheric Pressure Ionizations. *J. Chromatogr. B* **2010**, *878*, 3095–3105.
66. Plante, O. J.; Palmacci, E. R.; Seeberger, P. H. Automated Solid-Phase Synthesis of Oligosaccharides. *Science* **2001**, *291*, 1523–1527.
67. Seeberger, P. H.; Haase, W.-C. Solid-Phase Oligosaccharide Synthesis and Combinatorial Carbohydrate Libraries. *Chem. Rev.* **2000**, *100*, 4349–4394.
68. Merrifield, R. B. Automated Synthesis of Peptides. *Science* **1965**, *150*, 178–185.
69. Caruthers, M. Gene Synthesis Machines: DNA Chemistry and Its Uses. *Science* **1985**, *230*, 281–285.
70. Gillis, E. P.; Burke, M. D. A Simple and Modular Strategy for Small Molecule Synthesis: Iterative Suzuki-Miyaura Coupling of B-Protected Haloboronic Acid Building Blocks. *J. Am. Chem. Soc.* **2007**, *129*, 6716–6717.
71. Li, J.; Grillo, A. S.; Burke, M. D. From Synthesis to Function via Iterative Assembly of N - Methyliminodiacetic Acid Boronate Building Blocks. *Acc. Chem. Res.* **2015**, *48*, 2297–2307.

## 15. Oświadczenia autorów prac

Warszawa, 17.07.2023

### OŚWIADCZENIE

Oświadczam, że mój wkład w powstanie artykułów wchodzących w skład rozprawy doktorskiej:

**Molga, K.;** Gajewska, E. P.; Szymkuć, S.; Grzybowski\*, B. A. The Logic of Translating Chemical Knowledge into Machine – Processable Forms: A Modern Playground for Physical–Organic Chemistry. *React. Chem. Eng.* **2019**, *4*, 1506–1521.

był następujący: stworzyłem znaczną część bazy reguł reaktywności chemicznej (ponad 50.000 reguł, głównie obejmujących reakcje stereoselektywne) oprogramowania Chematica/Synthia, przygotowałem przykłady oraz analizy reakcji i reguł omówione w publikacji, przeprowadziłem porównanie oprogramowania Chematica/Synthia z innymi programami retrosyntezy. Uczestniczyłem w pisaniu manuskryptu.

Grzybowski\*, B. A.; Badowski, T.; **Molga, K.;** Szymkuć, S. Network Search Algorithms and Scoring Functions for Advanced-level Computerized Synthesis Planning. *WIREs Comput. Mol. Sci.* **2022**, e1630.

był następujący: uczestniczyłem w opracowaniu założeń oraz przeprowadziłem walidację funkcji oceny kosztu ścieżek syntetycznych oraz algorytmów umożliwiających wykorzystanie wielokrokowych sekwencji reakcji w planowaniu syntez, uczestniczyłem w pisaniu manuskryptu.

**Molga, K.;** Szymkuć, S.; Grzybowski\*, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094–1106.

był następujący: stworzyłem znaczną część (ponad 50000) bazy reguł reaktywności chemicznej, uczestniczyłem w opracowaniu założeń oraz przeprowadziłem walidację algorytmów umożliwiających wykorzystanie wielokrokowych sekwencji reakcji w planowaniu syntez, przeprowadziłem analizy dotyczące syntezy biblioteki izotopomerów i zinterpretowałem uzyskane wyniki. Uczestniczyłem w pisaniu manuskryptu.

Badowski, T.; **Molga, K.;** Grzybowski\*, B. A. Selection of Cost-Effective yet Chemically Diverse Pathways from the Networks of Computer-Generated Retrosynthetic Plans. *Chem. Sci.* **2019**, *10*, 4640–4651.

był następujący: uczestniczyłem w opracowaniu koncepcji badań i założeń funkcji oceny kosztu ścieżek syntetycznych, przeprowadziłem walidację funkcji zaimplementowanej w oprogramowaniu Chematica, przeprowadziłem wszystkie analizy retrosyntezy opisane w publikacji i zinterpretowałem uzyskane wyniki. Uczestniczyłem w pisaniu manuskryptu.

Kawol Molga  


**Molga, K.;** Dittwald, P.; Grzybowski\*, B. A. Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways. *Chem* **2019**, *5*, 460–473.

był następujący: uczestniczyłem w opracowaniu koncepcji badań i założeń algorytmów identyfikujących kluczowe wiązania i umożliwiających analizę retrosyntezy z ich pominięciem, przygotowałem zbiory syntez chronionych prawem patentowym prowadzących do Linezolidu, Sitagliptyny i Panobinostatu, przeprowadziłem wszystkie analizy retrosyntezy opisane w publikacji. Zinterpretowałem wyniki działania algorytmu identyfikującego kluczowe wiązania oraz wyniki analiz retrosyntezy wykorzystujących opracowane algorytmy. Uczestniczyłem w pisaniu manuskryptu.

Szymkuć, S.; Gajewska, E.; **Molga, K.;** Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated “Synthetic Contingency” Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

był następujący: uczestniczyłem w opracowaniu koncepcji badań, przeprowadziłem analizy retrosyntezy prowadzące do Remdesiviru oraz część analiz prowadzących do hydroksychlorochiny, uczestniczyłem w pisaniu manuskryptu.

**Molga, K.;** Dittwald, P.; Grzybowski\*, B. A. Computational Design of Syntheses Leading to Compound Libraries or Isotopically Labelled Targets. *Chem. Sci.* **2019**, *10*, 9219–9232.

był następujący: uczestniczyłem w opracowaniu koncepcji badań i założeń algorytmów pozwalających na analizę retrosyntezy bibliotek związków, przeprowadziłem walidację algorytmów zaimplementowanych w oprogramowaniu Chematica, przeprowadziłem wszystkie analizy retrosyntezy opisane w publikacji oraz zinterpretowałem ich wyniki. Uczestniczyłem w pisaniu manuskryptu.

**Molga, K.;** Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Mlynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58.

był następujący: stworzyłem znaczną część bazy reguł reaktywności chemicznej (ponad 50.000 reguł) użytej do identyfikacji sekwencji reakcji opisanych w publikacji, uczestniczyłem w opracowaniu koncepcji badań i założeń algorytmów identyfikujących sekwencje reakcji, analizowałem otrzymane wyniki, przygotowałem zbiór sekwencji reakcji do aplikacji webowej umożliwiającej generowanie bibliotek przy użyciu chemii iteracyjnej. Uczestniczyłem w pisaniu manuskryptu.

Karol Molga

Karol Molga



### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Molga, K.; Gajewska, E. P.; Szymkuć, S.; Grzybowski\*, B. A. The Logic of Translating Chemical Knowledge into Machine – Processable Forms: A Modern Playground for Physical-Organic Chemistry. *React. Chem. Eng.* **2019**, *4*, 1506–1521.

polegał na współpracowaniu koncepcji badań, ich nadzorze oraz pisaniu manuskryptu

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Grzybowski\*, B. A.; Badowski, T.; Molga, K.; Szymkuć, S. Network Search Algorithms and Scoring Functions for Advanced-level Computerized Synthesis Planning. *WIREs Comput. Mol. Sci.* **2022**, e1630.

polegał na współpracowaniu koncepcji badań, ich nadzorze oraz pisaniu manuskryptu

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Molga, K.; Szymkuć, S.; Grzybowski\*, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094–1106.

polegał na współpracowaniu koncepcji badań, ich nadzorze oraz pisaniu manuskryptu

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Badowski, T.; Molga, K.; Grzybowski\*, B. A. Selection of Cost-Effective yet Chemically Diverse Pathways from the Networks of Computer-Generated Retrosynthetic Plans. *Chem. Sci.* **2019**, *10*, 4640–4651.

polegał na współpracowaniu koncepcji badań, ich nadzorze oraz pisaniu manuskryptu

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Molga, K.; Dittwald, P.; Grzybowski\*, B. A. Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways. *Chem* **2019**, *5*, 460–473.

polegał na współpracowaniu koncepcji badań, ich nadzorze oraz pisaniu manuskryptu

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Szymkuć, S.; Gajewska, E.; Molga, K.; Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated “Synthetic Contingency” Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

polegał na współpracowaniu koncepcji badań, ich nadzorze oraz pisaniu manuskryptu

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Molga, K.; Dittwald, P.; Grzybowski\*, B. A. Computational Design of Syntheses Leading to Compound Libraries or Isotopically Labelled Targets. *Chem. Sci.* **2019**, *10*, 9219–9232.

polegał na współpracowaniu koncepcji badań, ich nadzorze oraz pisaniu manuskryptu

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Molga, K.; Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Mlynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58.

polegał na współpracowaniu koncepcji badań, ich nadzorze oraz pisaniu manuskryptu

Prof. Bartosz A. Grzybowski



## OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

**Molga, K.:** Gajewska, E. P.; Szymkuć, S.; Grzybowski\*, B. A. The Logic of Translating Chemical Knowledge into Machine – Processable Forms: A Modern Playground for Physical–Organic Chemistry. *React. Chem. Eng.* **2019**, *4*, 1506–1521.

polegał na: byłam jedną z głównych osób odpowiedzialnych za całokształt rozwoju programu Chematica. Uczestniczyłam w pisaniu manuskryptu.

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Grzybowski\*, B. A.; Badowski, T.; **Molga, K.;** Szymkuć, S. Network Search Algorithms and Scoring Functions for Advanced-level Computerized Synthesis Planning. *WIREs Comput. Mol. Sci.* **2022**, e1630.

polegał na: byłam jedną z głównych osób odpowiedzialnych za całokształt rozwoju programu Chematica. Uczestniczyłam w pisaniu manuskryptu.

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

**Molga, K.;** Szymkuć, S.; Grzybowski\*, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094–1106.

polegał na: byłam jedną z głównych osób odpowiedzialnych za całokształt rozwoju programu Chematica. Uczestniczyłam w pisaniu manuskryptu.

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Szymkuć, S.; Gajewska, E.; **Molga, K.;** Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated “Synthetic Contingency” Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

polegał na: byłam jedną z głównych osób odpowiedzialnych za całokształt rozwoju programu Chematica. Uczestniczyłam w pisaniu manuskryptu.

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

**Molga, K.;** Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Mlynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58.

polegał na: uczestniczyłam we współpracowaniu koncepcji badań i analizie uzyskanych wyników. Uczestniczyłam w pisaniu manuskryptu

Dr Sara Szymkuć



Warszawa, 13.07.2023

### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Molga, K.; Dittwald, P.; Grzybowski\*, B. A. Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways. *Chem* **2019**, *5*, 460–473.

polegał na stworzeniu i implementacji algorytmów opisanych w publikacji. Uczestniczyłem również w pisaniu manuskryptu.

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Szymkuć, S.; Gajewska, E.; Molga, K.; Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated “Synthetic Contingency” Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

polegał na: byłem jedną z głównych osób odpowiedzialnych za rozwój algorytmów wyszukiwania ścieżek programu Chematica.

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

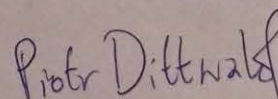
Molga, K.; Dittwald, P.; Grzybowski\*, B. A. Computational Design of Syntheses Leading to Compound Libraries or Isotopically Labelled Targets. *Chem. Sci.* **2019**, *10*, 9219-9232.

polegał na stworzeniu i implementacji algorytmów opisanych w publikacji. Uczestniczyłem również w pisaniu manuskryptu.

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Molga, K.; Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Mlynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58.

polegał na stworzeniu algorytmów pozwalających na wyszukiwanie sekwencji reakcji opisanych w publikacji. Uczestniczyłem również w pisaniu manuskryptu.



Dr Piotr Dittwald

Warszawa, 13.07.2023

### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Grzybowski\*, B. A.; Badowski, T.; Molga, K.; Szymkuć, S. Network Search Algorithms and Scoring Functions for Advanced level Computerized Synthesis Planning. *WIREs Comput. Mol. Sci.* **2022**, e1630.

polegał na: byłem jedną z głównych osób odpowiedzialnych za rozwój algorytmów wyszukiwania, selekcji i oceny ścieżek syntetycznych programu Chematica. Uczestniczyłem w pisaniu manuskryptu.

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Badowski, T.; Molga, K.; Grzybowski\*, B. A. Selection of Cost-Effective yet Chemically Diverse Pathways from the Networks of Computer-Generated Retrosynthetic Plans. *Chem. Sci.* **2019**, *10*, 4640–4651.

polegał na współpracowaniu założeń, stworzeniu oraz implementacji algorytmów selekcji i oceny ścieżek syntetycznych opisanych w publikacji. Uczestniczyłem w pisaniu manuskryptu.

Dr Tomasz Badowski

Tomasz Badowski

### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Molga, K.; Gajewska, E. P.; Szymkuć, S.; Grzybowski\*, B. A. The Logic of Translating Chemical Knowledge into Machine – Processable Forms: A Modern Playground for Physical-Organic Chemistry. *React. Chem. Eng.* **2019**, *4*, 1506–1521.

polegał na: byłam jedną z głównych osób odpowiedzialnych za rozwój bazy reguł chemicznych programu Chematica, uczestniczyłam w pisaniu manuskryptu.

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Szymkuć, S.; Gajewska, E.; Molga, K.; Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated “Synthetic Contingency” Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

polegał na: byłam jedną z głównych osób odpowiedzialnych za rozwój bazy reguł chemicznych programu Chematica, przeprowadziłam część analiz opisanych w publikacji, uczestniczyłam w pisaniu manuskryptu.



Dr Ewa Gajewska

Warszawa, 23.06.2023

### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Szymkuć, S.; Gajewska, E.; **Molga, K.**; Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated “Synthetic Contingency” Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

polegał na pomocy w zebraniu danych i analizie wyników dotyczących kosztów planów syntetycznych

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

**Molga, K.**; Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Mlynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58.

polegał na pomocy w zaprojektowaniu aplikacji webowych związanych z tą publikacją.

Dr Rafał Roszak

Warszawa, 23.06.2023

### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Szymkuć, S.; Gajewska, E.; **Molga, K.**; Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated “Synthetic Contingency” Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

polegał na pomocy w zebraniu danych i analizie wyników dotyczących kosztów planów syntetycznych

  
Dr Wiktor Beker

Warszawa, 23.06.2023

### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Szymkuć, S.; Gajewska, E.; **Molga, K.**; Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated “Synthetic Contingency” Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

polegał na pomocy w zebraniu danych i analizie wyników dotyczących kosztów planów syntetycznych

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

**Molga, K.**; Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Mlynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58.

polegał na zaprojektowaniu aplikacji webowych związanych z tą publikacją.

  
Martyna Moskal

Warszawa, 16.06.2023

### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Szymkuć, S.; Gajewska, E.; Molga, K.; Wołos, A.; Roszak, R.; Beker, W.; Moskal, M.; Dittwald, P.; Grzybowski\*, B. A. Computer-Generated "Synthetic Contingency" Plans at Times of Logistics and Supply Problems: Scenarios for Hydroxychloroquine and Remdesivir. *Chem. Sci.* **2020**, *11*, 6736–6744.

był następujący::

- uczestniczyłam w przygotowaniu danych i przeprowadzeniu analiz dotyczących kosztów planów syntetycznych



Agnieszka Wołos





Instytut Chemii Organicznej  
Polskiej Akademii Nauk

Prof. dr hab. Jacek Młynarski  
Zastępca dyrektora ds. naukowych

+48 22 343 23 22  
jacek.mlynarski@icho.edu.pl

Warszawa, 3 lipca 2023

#### Oświadczenie

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

Molga, K.; Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Młynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58.

polegał na nadzorze syntez przedstawionych w publikacji przeprowadzonych przez dr Patrycję Gołębiowską i dr. Oskara Popika.

Jacek Młynarski

Jacek Jan  
Młynarski

Elektronicznie podpisany  
przez Jacek Jan Młynarski  
Data: 2023.07.03 15:21:00  
+02'00'

Warszawa, 11.07.2023

### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

**Molga, K.**; Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskał, M.; Roszak, R.; Mlynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58.

polegał na przeprowadzeniu części syntez opisanych w publikacji.

  
Dr Oskar Popik

Warszawa, 11.07.2023

### OŚWIADCZENIE

Niniejszym oświadczam, że mój wkład w powstanie artykułu:

**Molga, K.;** Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Młynarski, J.; Grzybowski, B. A. A Computer Algorithm to Discover Iterative Sequences of Organic Reactions. *Nat. Synth.* **2022**, *1*, 49–58.

polegał na przeprowadzeniu części syntez opisanych w publikacji.



Dr Patrycja Gołębiowska

## **16. Publikacje oryginalne**



Cite this: *React. Chem. Eng.*, 2019, 4, 1506

## The logic of translating chemical knowledge into machine-processable forms: a modern playground for physical-organic chemistry†

Karol Molga, <sup>a</sup> Ewa P. Gajewska,<sup>a</sup> Sara Szymkuć<sup>a</sup> and Bartosz A. Grzybowski <sup>\*ab</sup>

Recent years have brought renewed interest – and tremendous progress – in computer-assisted synthetic planning. Although the vast majority of the proposed solutions rely on individual reaction rules that are subsequently combined into full synthetic sequences, surprisingly little attention has been paid in the literature to how these rules should be encoded to ensure chemical correctness and applicability to syntheses which organic-synthetic chemists would find of practical interest. This is a dangerous omission since any AI algorithms for synthetic design will be only as good as the basic synthetic “moves” underlying them. This Perspective aims to fill this gap and outline the logic that should be followed when translating organic-synthetic knowledge into reaction rules understandable to the machine. The process entails numerous considerations ranging from careful study of reaction mechanisms, to molecular and quantum mechanics, to AI routines. In this way, the machine is not only taught the reaction “cores” but is also able to account for various effects that, historically, have been studied and quantified by physical-organic chemists. While physical organic chemistry might no longer be at the forefront of modern chemical research, we suggest that it can find a new and useful embodiment through a conjunction with computerized synthetic planning and related AI methods.

Received 21st February 2019,  
Accepted 7th June 2019

DOI: 10.1039/c9re00076c

rsc.li/reaction-engineering

### Introduction

The idea of computers designing synthetic routes and/or predicting the outcomes of chemical reactions dates back to 1960s.<sup>1</sup> The pioneering efforts of eminent chemists such as E. J. Corey (LHASA program<sup>2,3</sup>), C. Djerassi,<sup>4</sup> I. Ugi,<sup>5</sup> and J. B. Hendrickson<sup>6</sup> were, in many ways, ahead of their time though, but for various reasons (narrated in ref. 7) did not become widely adopted by the synthetic community. Fortunately, recent years have witnessed a revival of interest in this interesting and potentially impactful area of chemical research and several platforms for organic-synthetic analyses have emerged. Our own effort in this area – starting with the 2005 paper on the analysis of large chemical networks<sup>8</sup> – has culminated in the development of Chematica retrosynthesis platform<sup>7,9</sup> that has recently been commercialized by Sigma-Aldrich (under the trade name of Synthia) and validated experimentally *via* execution of several synthetic routes designed by the machine.<sup>9</sup> Other notable ef-

forts have been the ARChem engine<sup>10</sup> (to be incorporated into SciFinder), the ASKCOS tool<sup>11</sup> from MIT, or Segler and Waller's software based on Monte-Carlo searches and described in reference.<sup>12</sup> While the ways in which the algorithms underlying these engines come up with complete synthetic pathways differ in many substantial ways, the common component they all share are the chemical rules (“transforms”) describing individual chemical reactions. In fact, the quality of these rules is absolutely crucial since synthetic pathways are very “unforgiving” to errors in individual steps – if our individual rules are, say, 80% correct, the chance that a *n*-step synthesis is going to be error-free and executable in the laboratory is only 0.8<sup>*n*</sup> (*i.e.*, only 10% for a ten-step synthesis). It is a straightforward but essential conclusion that any synthesis planning software will be only as good as the reaction rules it incorporates. Yet, despite such considerations, the articles on computer-aided synthesis focus mostly on the (often quite advanced) AI routines for concatenating individual steps into pathways<sup>8,9,11–14</sup> while, at the same time, spend little time on how the individual reaction rules are (or should be) coded. We think that given the progress and interest – or even hype<sup>15</sup> – surrounding this re-emerging field of chemical research, the time is ripe to systematize the approaches to and the logic of reaction rule coding. Accordingly, in this Perspective, we will strive to provide an overview of these aspects for reaction rules (i)

<sup>a</sup> Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland. E-mail: nanogrzybowski@gmail.com

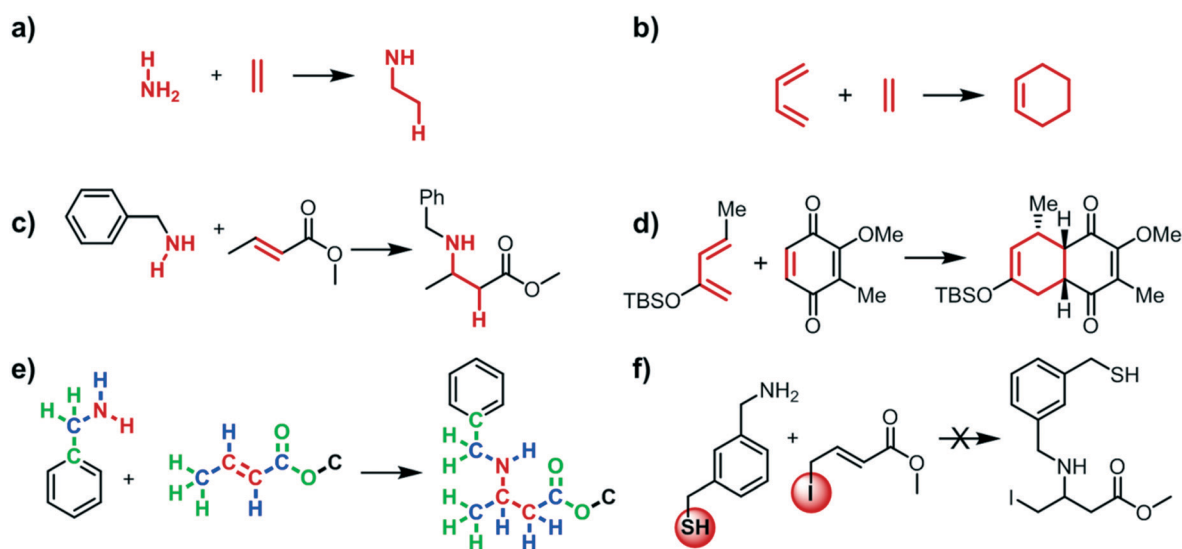
<sup>b</sup> IBS Center for Soft and Living Matter and Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulsan-gun, Ulsan, 689-798, South Korea

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9re00076c

machine-extracted from repositories of published reactions (e.g., Reaxys, SciFinder, InfoChem, USPTO databases<sup>16–19</sup>), or (ii) coded by expert chemists based on the underlying reaction mechanisms. One of the main conclusions of our survey is that while machine extraction approach is very rapid, the chemical correctness of the rules it yields is lower than those coded by experts – this difference becomes all the more pronounced as one becomes interested in the synthesis of more complex targets, for whose synthesis the human experts might actually benefit from computer's help. Another message this article is intended to convey is that translation of chemical knowledge into machine-readable rules is a very nuanced process, in which one has to consider not only the scope of admissible substituents and incompatible groups, but also a range of physical-organic effects including electron densities, steric bulk, molecular strain, and more. In calculating such quantities by quantum mechanical, QM, or molecular mechanics, MM, methods, an added challenge is to make them very fast and compatible with automated synthetic planning, during which the numbers of candidate intermediates considered can be very large (hundreds of thousands<sup>7,9</sup>). Finally, we also aim to illustrate which types of rules can be fine-tuned and improved by AI models, and what precautions should be taken for such models to be robust and meaningful. Altogether, we suggest that rule coding combining mechanistic considerations, elements of QM and MM modelling, and AI can be viewed as a modern embodiment of classic physical-organic chemistry,<sup>20,21</sup> possibly reviving interest in this somewhat forgotten but essential field of organic-chemical research.

## Defining reaction core and its environment

To begin with, a reaction rule must specify atoms that are changing their environments and/or bonding patterns during a chemical reaction. For example, during addition of an amine to an alkene, a bond is formed between an incoming amine nucleophile and a carbon atom, the double C=C bond becomes a single one, and another C–H bond is formed (Fig. 1a). In a Diels–Alder cycloaddition, the double bonds of the diene and dienophile rearrange to form a cyclohexene ring (Fig. 1b). Of course, such a “local” description is chemically incomplete since the flanking atoms are generally also important or even all-important. In the first example, the reaction can proceed only when there is an electron withdrawing group (e.g., an ester or a ketone) activating the alkene for the nucleophilic attack (Fig. 1c); for the Diels–Alder reaction, the electron-withdrawing/accepting ability of the substituents on the diene and dienophile will dictate regio-, site, or diastereoselectivity (Fig. 1d). To account for such effects, the reaction rules need to be extended to include admissible, nearby substituents at various positions. Such extensions can be to different “radii” – to within atoms just flanking the reaction core (radius = 1), to within two atoms from reaction core (radius = 2), *etc.* (Fig. 1e). In addition, chemical reactions might be prohibited by conflicting groups present anywhere in the molecule – in our example of double-bond amination, an iodide and a thiol present in reaction partners introduce an incompatibility because allylic iodide is a better electrophile than  $\alpha,\beta$ -unsaturated alkene and thiol is a more reactive nucleophile than amine (Fig. 1f).



**Fig. 1** Defining reaction core. Atoms changing their environments and bonds modified during a) addition of an amine to alkene and b) Diels–Alder cycloaddition. Literature precedents<sup>22,23</sup> used for extraction of these cores are shown in c) and d), respectively. e) Expansion of the reaction core. The core atoms changing their bonding patterns are coloured in red. Inclusion of the nearest-neighbour atoms (radius = 1, blue) and next-nearest-neighbour atoms (radius = 2; green) increases the accuracy of the reaction rule and begins to cover important substituents – here, the presence of an electron withdrawing group attached to the more distant end of C=C bond. f) Cross-reactive groups (here, allyl iodide and thiol) present anywhere in the substrates must also be considered as they will interfere with a desired reaction outcome.

There are few conceivable ways of capturing such intricacies of chemical reactivity into a machine readable format. Over the past years, two major approaches have emerged (1) extraction of reaction cores from databases such as Reaxys<sup>16</sup> or USPTO<sup>19</sup> and fine-tuning their substituent scope/applicability based on the synthetic latitude of the examples found; or (2) coding the rules manually, by chemists taking into account pertinent mechanistic considerations. In the following, we will focus on these two core-based approaches and will not discuss methods in which reactivity is predicted based on the AI-trained scores for atom pairs<sup>24</sup> or linguistic sequence-to-sequence models.<sup>25,26</sup> The performance of these approaches is described in the ESI† to ref. 27 and also ref. 28 – in addition, examples included in the next section and in the ESI† Fig. S1 evidence that such models produce unacceptably large fraction of chemically problematic predictions.

## Limitations of automatic rule extraction

The advantage of machine-extracting rules from reaction databases is the speed of the method – in fact, with adequate computational power, an entire Reaxys collection can be scripted within few days. In the end, one ends up with tens to hundreds of thousands of reaction rules with the specific number depending on the source database and the radius around the reaction core. For instance, Segler and Waller<sup>12</sup> extracted transformation rules from 12.4 million single-step reactions from the proprietary Reaxys repository. For the transformations having at least 50 literature precedents and encompassing the core atoms plus the radius =1 environments, they extracted ~17 000 rules; for the relaxed requirements of three literature precedents and only the core atoms, the number was ~300 000. In a recent study<sup>29</sup> by Watson *et al.*, the authors extracted from the publicly available United States Patent and Trade Office, USPTO, database (close to 2 million entries) a total of 83 942 unique trans-

forms for radius 0, 180 862 for radius 1 and 325 873 for radius 2.

Of course, not all machine-extracted rules derive from similar numbers of literature precedents – for popular reaction classes, the number of such precedents can be up to hundreds of thousands per one extracted rule (*e.g.*, for generalized cores of Wittig olefination, Suzuki–Miyaura coupling, or formation of an amide from carboxylic acid and amine substrates); for more specialized chemistries, however, there might be just few examples in the literature (*e.g.*, in Reaxys, there are only ~20 examples of stereospecific C–H insertions of carbenes yielding tertiary alcohols; Fig. 2a and ref. 30). This is significant for our discussion since any machine-learning of reaction-rule applicability is possible only for the popular reaction classes. For the ones with just few examples, it is impossible to automatically extract meaningful statistics delineating the scope of the reaction, *i.e.*, which substituents are admissible and which are not, which groups are conflicting with the reaction core, *etc.* This is, in fact, a serious flaw if one wishes to teach the machine some more advanced syntheses – for instance, the number of reported examples related to the stereospecific reduction of tertiary alcohols used in total syntheses of curcumen<sup>31,32</sup> and himachalene<sup>32</sup> is very limited and does not exceed 10. Importantly, all of these reactions were performed on simple substrates (Fig. 2b) complicating prediction of the extracted transformations' applicability to more advanced intermediates, especially bearing potentially cross-reactive functional groups. In another example, anionic 4 + 2 cycloaddition in Fig. 2c is represented by only 30 precedents but was the key step in the synthesis of murrayafoline-A,<sup>33</sup> olivine,<sup>34</sup> clausamine E<sup>35</sup> and claulansine D.<sup>36</sup> This annulation occurs *via* stepwise conjugate addition of a lithiated lactone, subsequent cyclisation forming [2.2.1] bicycloheptane, and elimination–tautomerisation (Fig. 2c) leading to the fused aromatic system. Without mechanistic understanding of this complicated sequence, it is virtually impossible to properly

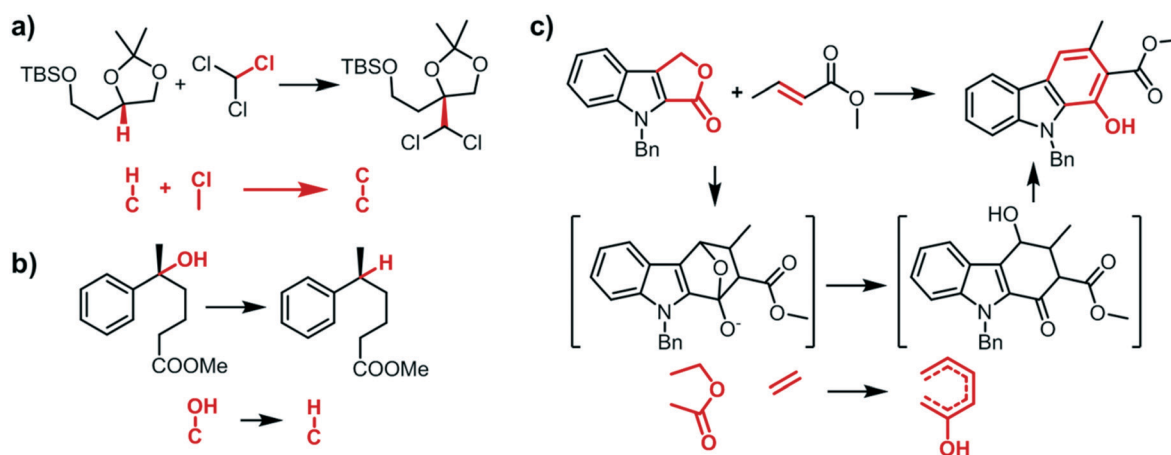
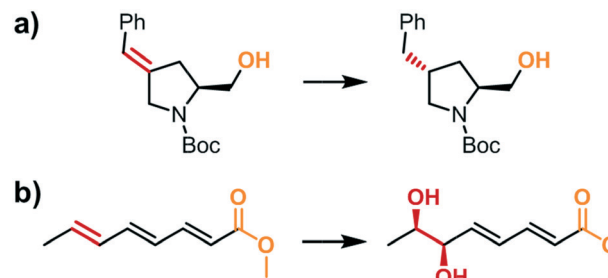


Fig. 2 Examples of rare but useful reactions. a) Base-induced dichloromethylation of derivatives of secondary alcohols (only 20 literature examples). b) Stereospecific deoxygenation of tertiary alcohols (only 10 examples). c) Anionic 4 + 2 cycloaddition forming a fused aromatic system (only 30 examples). Reaction cores without any environments are coloured in red.

define the necessary substituents, *e.g.*, the presence of electron withdrawing group necessary for the conjugate addition to proceed.

Even for the popular reaction classes one has to be careful. Although widening the core to the radius 1 or 2 environments generally increases accuracy of the transforms, it also limits their scope, makes them very much case-specific and non-generalizable (*cf.* previous paragraph), and still does not treat adequately many chemical details. This point is corroborated by examples in Fig. 3 describing popular aldol condensation between an ester and an aldehyde. Limiting the reaction rule to only the core of changing bonds/atoms (b) or supplementing this core (c) with immediate, radius = 1 neighbouring atoms allows for erroneous results such as those shown in the rightmost column of Fig. 3b and c. Even if radius = 2 is applied, the reaction template is incomplete as it allows for the presence of highly acidic H's interfering with expected reaction outcome (Fig. 3d).

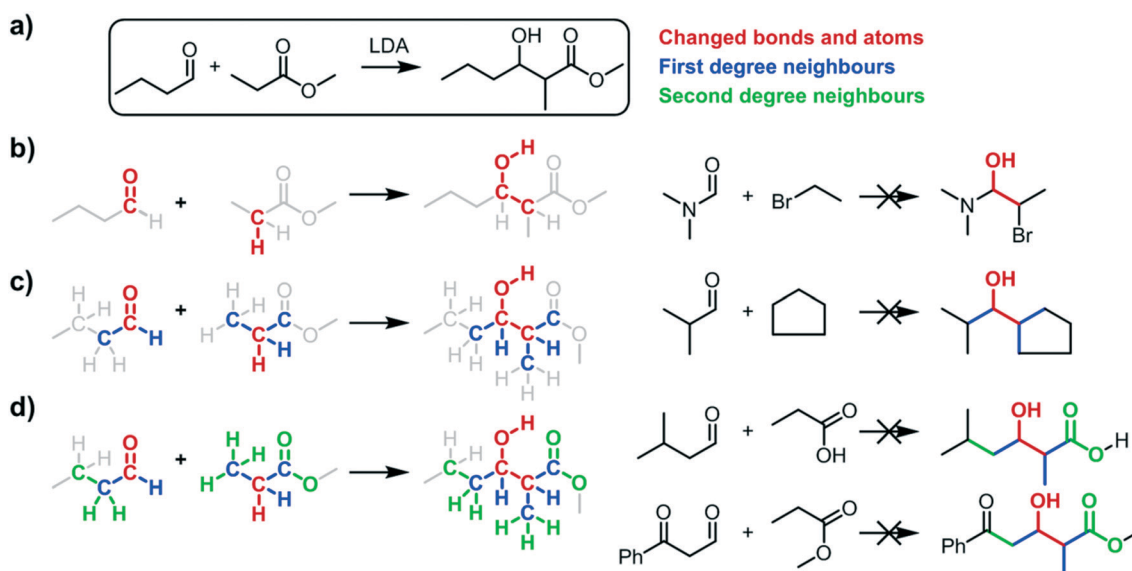
The automated approach is also ill-suited to account for truly long-distance effects that might come from groups many bonds away – for instance, in Fig. 4a, the reduction of a double bond is directed by a hydroxyl group<sup>37</sup> located four atoms away, while the regioselectivity of dihydroxylation of a triene in Fig. 4b is controlled by an electron withdrawing group<sup>38</sup> five atoms away. We observe that extending the environments is not an answer here because in this way one would soon end up with the number of “rules” approaching the number of literature precedents in the database one is learning the rules from – the only way around this problem is to have an expert organic chemist



**Fig. 4** Long-range control of organic reactions. a) Stereoselective reduction of an alkene with Crabtree's catalyst<sup>37</sup> is controlled by a remote hydroxyl group; b) sharpless dihydroxylation of polyene<sup>38</sup> is controlled by a remote electron withdrawing group. The reacting groups are coloured in red and the controlling, distant functionalities, in orange.

determine in which cases narrower cores are adequate and in which they must be extended to capture distant substituent or scaffold effects.

Another relevant issue is the treatment of incompatible groups – the generic automated approach devised so far<sup>11,12,14,39</sup> is to identify which groups are surviving a transformation of interest and deem them as “compatible”; those groups that are not seen in reactions corresponding to a given rule (or are “destroyed” in such transforms), are deemed incompatible. This heuristics is rather crude since the fact that there are no reactions in which a particular functional group is absent does not mean this group is generally prohibited in such a reaction – perhaps no one just tried a particular group or groups' combination, which is all the



**Fig. 3** Defining proper span of reaction rules. a) Base-induced aldol condensation between an ester and an aldehyde. b) Limiting the reaction template to the bare reaction core (red) allows the rule to capture nonsensical results such as the one shown on the right. c) Inclusion of first-degree neighbours (blue) still does not eliminate faulty predictions – here, addition of cyclopentane to aldehyde is still captured though it is chemically nonsensical for this reaction type. d) Extension to next-nearest-neighbour, radius = 2 environments (green) limits the number of nonsensical predictions but even this extended rule allows for the presence of acidic H's from carboxylic acid (top) or benzoyl acetaldehyde (bottom) interfering with expected reaction outcome.



more likely for more specialized transforms having less literature precedents. Second, there are numerous cases in which statistical approach fails because the same group might be incompatible with a given reaction rule in some targets, but compatible in others. As a case in point, consider examples shown in Fig. 5 for which automatic assignment of certain functional groups as compatible is chemically incorrect. Specifically, for methyl ester reduction performed *en route* to jujuboside saponin<sup>40</sup> (Fig. 5a, top), we might “learn” that geminal methyl diester can survive the mono-ester’s reduction; however, for the same reaction conditions applied to an intermediate in plumisclerin A synthesis<sup>41</sup> (Fig. 5a, bottom), we would learn the opposite – namely, that geminal methyl diester reacts while monoester survives. Naturally, these conclusions are chemically faulty: in general, methyl esters are incompatible during reduction of methyl diesters (and *vice versa*) and the examples shown are scaffold-specific exceptions for which one needs to separately code an additional, wider-core reaction rule. Similarly, during the synthesis<sup>42</sup> of milbemycin  $\beta_9$  (Fig. 5b) one lactone is reduced in the presence of another – this, however, is an exceptional example and, in general, lactones should be qualified as incompatible during reduction of another lactone moiety. The same problem of falsely qualifying an alkene as compatible with ozonolytic cleavage of another alkene is illustrated for the specific case in Fig. 5c taken from the synthesis<sup>43</sup> of hippospongiic acid A.

Third, there are important classes of reactions for which the cores are identical yet the mechanisms differ and so the requirements of admissible *vs.* conflicting groups can be markedly different. As an example, consider Buchwald–Hartwig amination *vs.* nucleophilic aromatic substitution, where in both cases amine reacts with an aryl halide (Fig. 6). Although the core of these reactions is identical, the mechanisms, scopes of admissible substituents and incompatible groups are substantially different. In aromatic substitution,

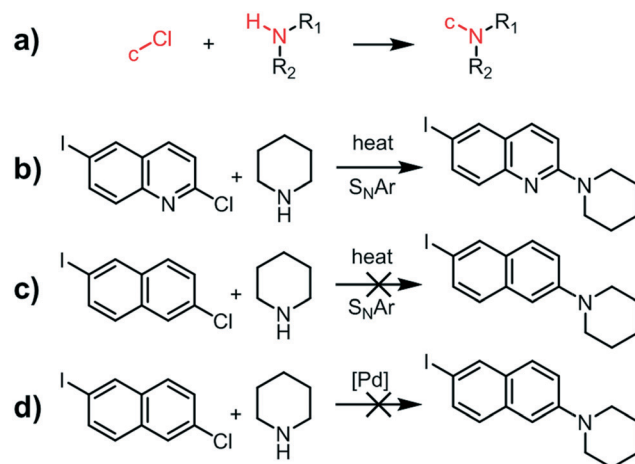


Fig. 6 Reactions with identical cores but different mechanisms, scopes of substituents, and incompatible groups. a) Buchwald–Hartwig amination and nucleophilic aromatic substitution share the same reaction core and convert aryl chloride to arylamine. b) Nucleophilic aromatic substitution can occur in the presence of aryl iodide but requires electron deficient ring to proceed (compare with c). d) In contrast, Buchwald–Hartwig amination of aryl chloride cannot be performed in the presence of a more reactive aryl iodide.

the reacting aryl halide should be attached to an electron deficient ring (*e.g.*, pyridine or nitroarene) while other aryl halides (including iodides) attached to neutral or electron-rich rings – even if present in the same molecule (Fig. 6b) – remain unreactive. In the Buchwald–Hartwig amination, the electronic requirements are less important and reaction is not limited to electron deficient halides, though the reactivity of halogens usually follows the order  $I > Br > Cl$  and thus aryl iodides/bromides cannot be present while the chloride is supposed to react (Fig. 6c).

Finally, machine-extracted rules may not properly handle the stereochemical information. From the very basic level,

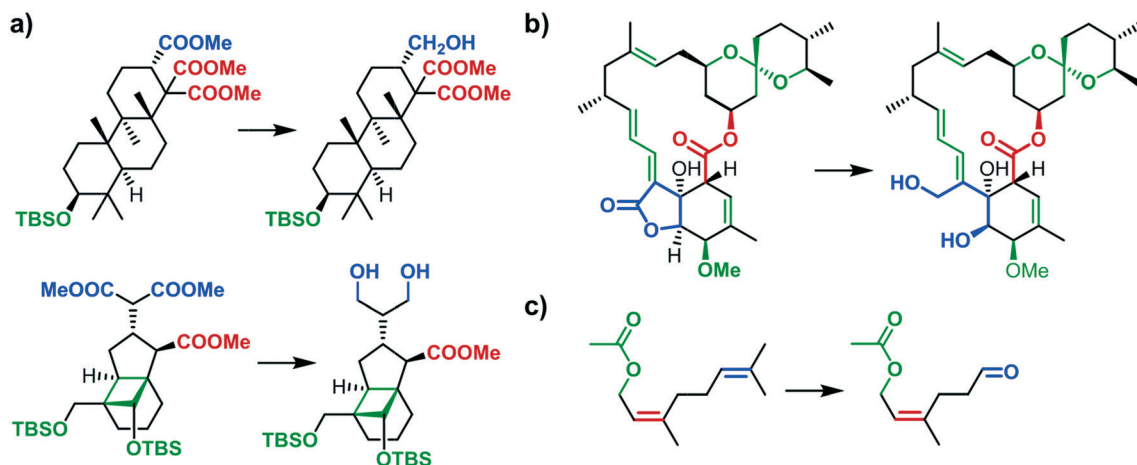


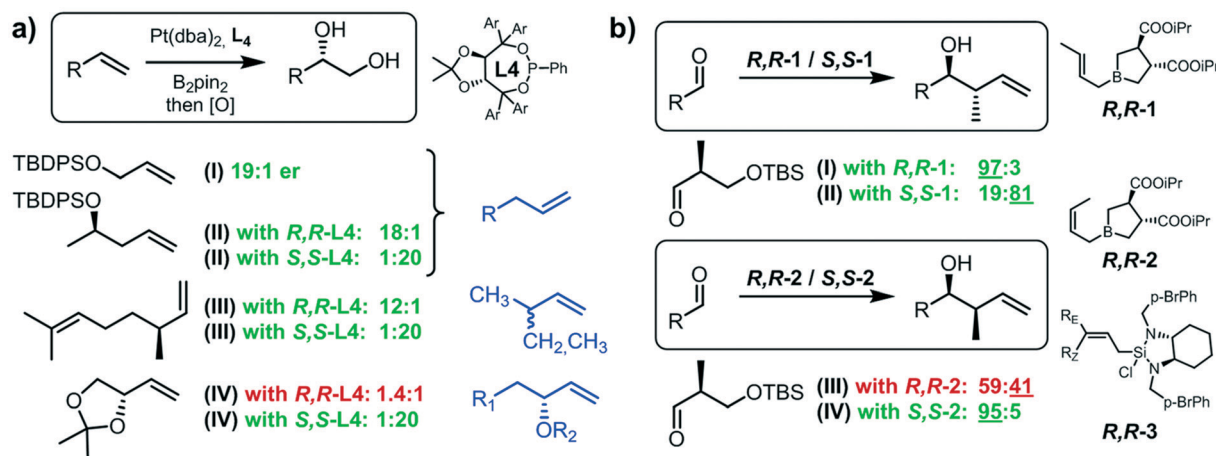
Fig. 5 False-positive assignments of functional groups as compatible or incompatible. Blue – reacting functional group; green – appropriately assigned, stable functional groups; red – false-positive assignments of purportedly “compatible” groups. These groups are compatible only for certain reacting molecules while for others, under the same reaction conditions, they are not stable (*i.e.*, incompatible). For detailed discussion, see main text.

the proper handling of stereochemistry requires incorporation of so called canonical SMILES to ensure identical order of parsing atoms in the product and in the substrates. Even if this requirement is met, however, the tools like RDKit<sup>44</sup> still do not handle stereochemical notation properly and more complex solutions like Stereofix<sup>7</sup> or RDChiral<sup>45</sup> are needed. Moreover, the problem remains how to define the reaction transform to handle relevant stereochemical information not only in the substrate(s) but also chiral catalysts or reagents. In the latter case, properly defined reaction core should not allow for any nearby stereocenters in the substrate causing the so-called “mismatching effects” and lowering stereoselectivity or yield (Fig. 7). For instance, Morken's hydroboration–oxidation<sup>46</sup> catalyzed by phosphonite ligand (denoted L4 in Fig. 7a) should allow for unbiased substrate (I) and “harmless” nearby stereocenters (II, III) but preclude mismatched  $\alpha$ -oxygenated stereocenters (IV, red). Significantly, is it not easy to generalize such catalyst/reagent effects, and the scope of “harmless” vs. “harmful” environments may be quite different for different transformations – and thus not readily amenable to machine rule extraction. For instance, Roush's<sup>47</sup> *anti*-selective crotylation with *E*-boronate (Fig. 7b, denoted *R,R*-1/*S,S*-1) affords products with acceptable diastereoselectivities in both matched (I) and mismatched (II) cases whereas *syn*-selective crotylation with *Z*-boronate denoted *R,R*-2/*S,S*-2 performs well only in the matched (IV) case (Fig. 7b). Additionally, diastereoselectivity of Leighton's<sup>48</sup> *syn*- and *anti*-crotylations with *R,R*-3/*S,S*-3 (Fig. 7b, bottom-right structure) is equally high in each case.

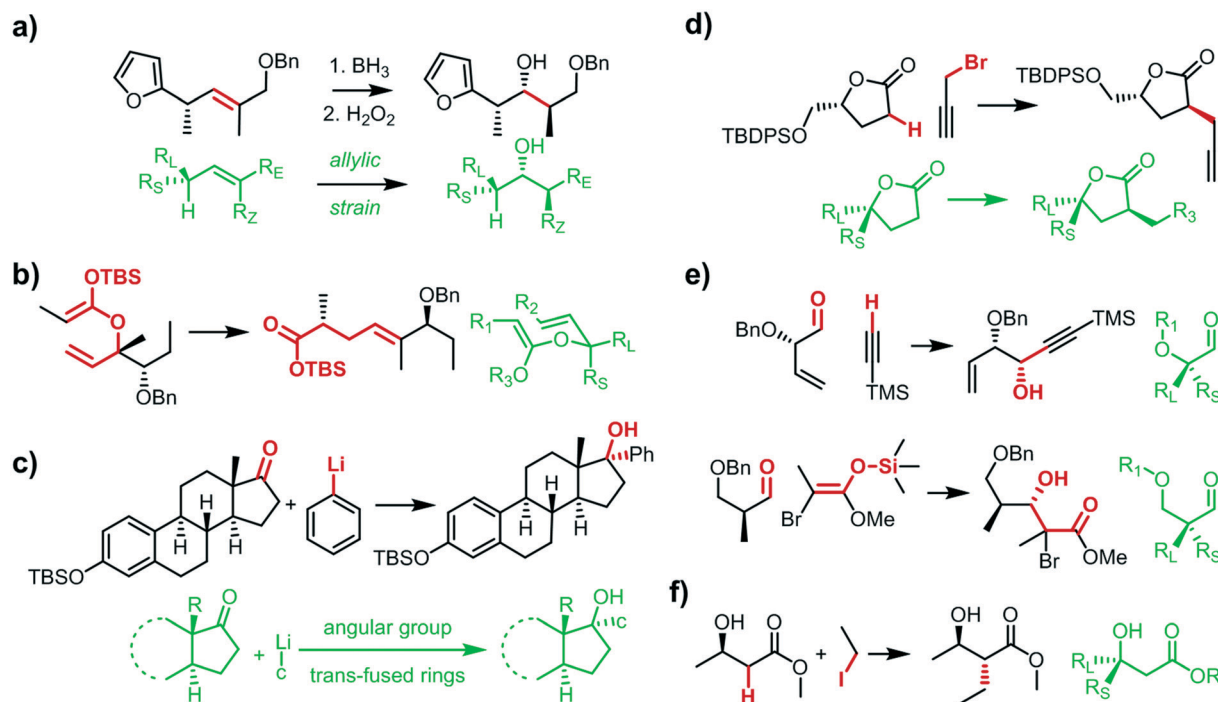
On the other hand, when the observed stereochemical outcome of the reaction is substrate-controlled, one should carefully evaluate the mechanism responsible for the observed stereochemistry and ensure that reaction transform includes all necessary structural features. For example, reactions of

alkenes are often controlled by the so-called allylic strain<sup>49</sup> (Fig. 8a), intramolecular reactions and sigmatropic rearrangements (Fig. 8b) often yield products deriving from the most energetically favourable, pseudoequatorial transition state, face selectivity of additions to cyclic systems is controlled by an entire set of substituents attached to the ring and occurs from the less hindered side (Fig. 8c and d), whereas stereoselective 1,2-additions to carbonyl groups (Fig. 8e) and alkylations of enolates (Fig. 8f) often proceed *via* chelated intermediates. In each case, the environments necessary to capture these stereoelectronic factors (green) are of different sizes and are significantly larger than just the cores of modified bonds or atoms (red). As in some other examples described before, there is – at least currently – no possibility to automate the extraction of such reaction rules.

Without spending much additional time on complications deriving from manual entry errors in reaction databases (Fig. 9a and b) or serious problems with proper atom mapping across the reaction rules<sup>28</sup> (important to ensure that the rules “know” which atom of the substrate becomes which atom of the product; Fig. 9c), our conclusion from this part is that the automatic-extraction approach entails serious chemical problems. In Fig. 10 and in Fig. S1–S25,<sup>†</sup> this conclusion is further corroborated by specific examples of erroneous, automatically extracted rules – in many cases, describing very basic reaction classes – underlying the operation of platforms such as Waller's MCTS<sup>12</sup> or MIT's ASKCOS.<sup>11,14</sup> We, of course, envision a possibility that the quantity and the quality of available literature examples will, one day, increase sufficiently to allow for more meaningful extraction and subsequent machine learning – though one should remember that although the “universe” of chemical reactions grows rapidly,<sup>8</sup> the increase is mostly due to the proliferation of the popular reaction types whereas the statistics on the expert-level transformations, often scaffold-specific and applied to



**Fig. 7** Defining reaction core to handle matched/mismatched cases in catalyst- or reagent-controlled stereoselective reactions. a) Morken's hydroboration–oxidation cannot be performed in the presence of mismatched nearby oxygenated stereocenters. Properly defined reaction cores are shown in blue. b) Lack of general rules for predicting mismatch effects. Scopes of admissible substituents for Roush's *syn*- and *anti*-selective crotylations are significantly different. In contrast, Leighton's crotylation with *R,R*-3/*S,S*-3 performs well in both matched and mismatched reagent/substrate pairs (I–IV).



**Fig. 8** Defining reaction core for substrate-controlled stereoselective transformations. In each case, different model explains the observed stereoselectivity. a) Hydroboration–oxidation of alkenes is controlled by allylic strain and adequate reaction core should address relative size of substituents ( $R_L$ ,  $R_S$  L-large, S-small). b) Stereochemistry of product obtained in Claisen rearrangement is dictated by relative size of substituents and occurs *via* a pseudo-chair transition state; c) addition of a nucleophile to a cyclic ketone is controlled by an angular methyl group. Properly defined reaction core must address the presence of *trans*-fused bicyclic system and angular substituent of any size. d) Alkylation of lactone is controlled by a distant substituent and occurs from the less hindered face. e) Additions of nucleophiles to aldehydes and f) alkylations of enolates are commonly performed and controlled by chelated intermediates. Corresponding reaction rules should capture the presence of chelating groups (here, OBn or OH) and relative sizes of substituents. In all panels, red = reaction cores spanning changing atoms and bonds, green = correct environments capturing substituents responsible for observed stereochemical outcomes. Examples are taken from ref. 50–56.

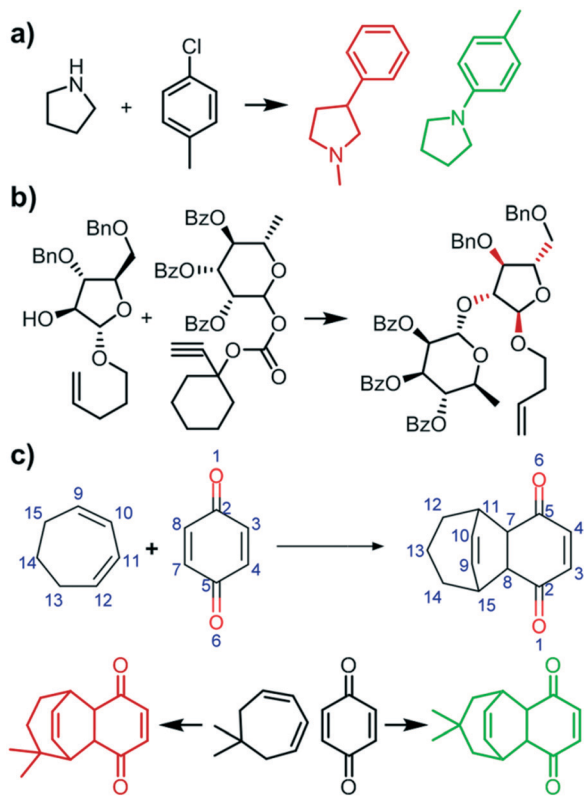
natural products, is growing much slower. How to ensure that more such expert-level data becomes available to the community is currently unclear to us given that total synthesis is, unfortunately, becoming less popular than few decades ago. In the meantime, at least some other problems related to automated rule extraction (*e.g.*, more accurate treatment of incompatibilities) could be alleviated by publishing more negative results from which machine-learning approaches could benefit. Leaving such considerations to the community (and the funding agencies) to ponder, we reminisce that we ourselves were initially enticed to follow the path of automatic rule extraction – but only until the rules so derived proved inadequate when applied to retrosynthetic planning involving non-trivial targets. When one reaches this conclusion, one must reconsider the entire approach and face the enormity of the task ahead – that is, of coding the rules individually while taking into account reaction mechanism and several physical-organic considerations.

## Mechanism-based rule coding

Per our discussion in the preceding section and also quantification in our previous works,<sup>7</sup> there are on the order of 100 000 distinct reaction classes constituting the body of

modern organic chemistry. The encouraging thing – and contrary to what some authors claimed<sup>12</sup> – is that while the number of specific reactions published in the literature grows exponentially and doubles approximately every 10–15 years,<sup>8,61</sup> the number of new reaction classes/types with distinct mechanisms is not increasing nearly as rapidly (based on our experience and estimates, there are *ca.* 3000–5000 such examples per year). This makes the task of coding the rules manually manageable, at least in principle – in our case, it took over a decade and gave rise to, currently, over 75 000 reaction transforms incorporated into Chematica.

Coding each of these transforms begins with a thorough study and understanding of the reaction mechanism. Assume, for example, that we wish to code diastereoselective Michael addition of terminal vinyl magnesium bromides to cyclic enones (top-left portion of Fig. 11; for clarity, Mg and Br atoms from the Grignard reagent are not shown on the left side of the reaction scheme). The first condition to be met is the intermolecular character of the reaction. This is motivated by the fact that only intermolecular cyclisation can ensure desired *cis* arrangement of  $R_2$  and  $R_3$  groups in the product. In the SMARTS notation in the reaction record shown in the figure, (field ‘Reaction’s SMARTS’), this requirement is indicated by the ‘R0’ sign specifying that atom #9 is

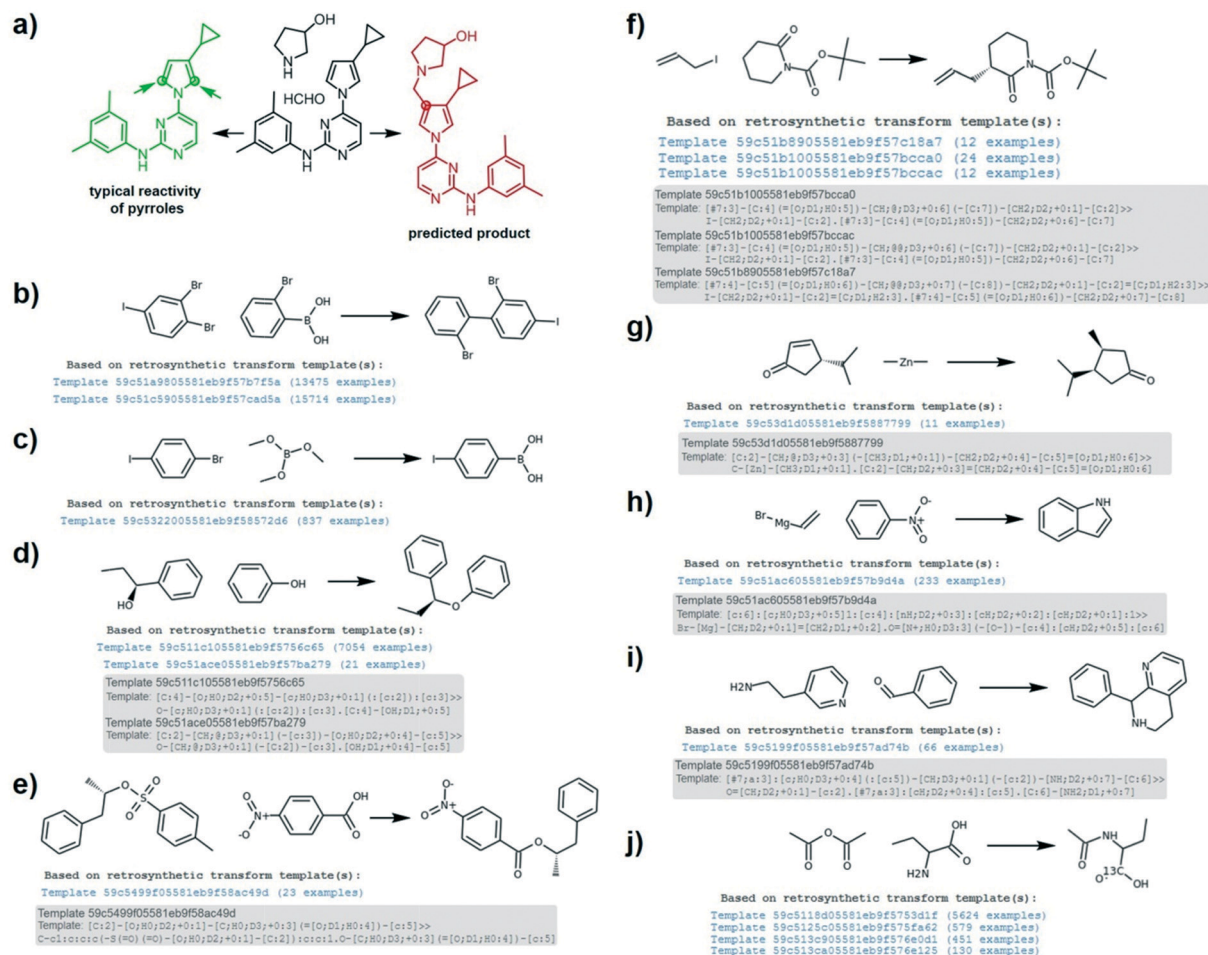


**Fig. 9** Errors in data sources translate into meaningless reaction rules. a) A product in data entry<sup>57</sup> in SciFinder database (red) does not match the one reported in the source publication<sup>58</sup> (green). The database entry corresponds to a tandem *N*-methylation/*C*3-arylation of pyrrolidine (with chlorotoluene being fragmented) while the original reported reaction is an ordinary *N*-arylation. b) Stereochemistry of several stereocenters is mis-assigned in Reaxys database entry 42850901 describing an instance of gold-catalyzed acetalisation.<sup>59</sup> Stereocenters marked red are all incorrectly inverted. c) Reaction template derived from an incorrectly mapped Diels–Alder reaction from the USPTO dataset (top) applied to retrosynthetic planning generates faulty predictions. The reaction and substrates proposed for the synthesis of an adduct coloured red will, in fact, yield the product coloured green.

not allowed to be a part of any ring. On the other hand, the presence of a ring between substituents  $R_2$  and  $R_3$  is allowed. Next, we need to specify the substituents present on the five-membered ring. Available literature on the topic indicates that  $\text{CH}_2$  and oxygen are allowed at position #7 as cyclic ketones and esters could serve as reacting partners. Also, proper chirality of atom #2 has to be specified, as the stereoselective outcome is dictated by orientation of the substituent present at this position. Because atom #7 is indicated as either  $\text{CH}_2$  or oxygen, the substituents at position #1 are limited to those that are admissible for both ketones and lactones serving as Michael acceptors. The substituents at position #8 are limited to unsubstituted alkyl or a hydrogen because: (i) bulky groups on the  $\beta$ -carbon reduce reactivity of Michael acceptor, and (ii) it is necessary to avoid an additional chiral center that might influence stereoselectivity of the reaction, especially in the presence of a ring between sub-

stituents  $R_2$  and  $R_3$ . To prevent bulky groups that may cause steric hindrance and disturb the addition, atom present at position #10 of the vinyl magnesium halide is limited to hydrogen or unsubstituted alkyls. Other substituents, such as vinyl group, need to be specified in separate SMARTS lines. The reaction record also includes additional fields specifying groups that need to be protected (e.g., aldehydes, primary amines and thiols, in total 14 groups), groups that are always incompatible in the reaction (e.g., acid chlorides, other Michael acceptors, etc., in total more than 100 groups of which 47 are listed in the figure), typical reaction conditions, representative literature sources, and 10 other fields (18 in total). We note that different variants of stereoselective Michael addition (e.g. with 6-membered enone serving as a Michael acceptor or other nucleophiles reacting as Michael donors – e.g., aryl Grignard reagents, amines, more substituted vinyl magnesium halides, thiols etc.) are also included in Chematica.

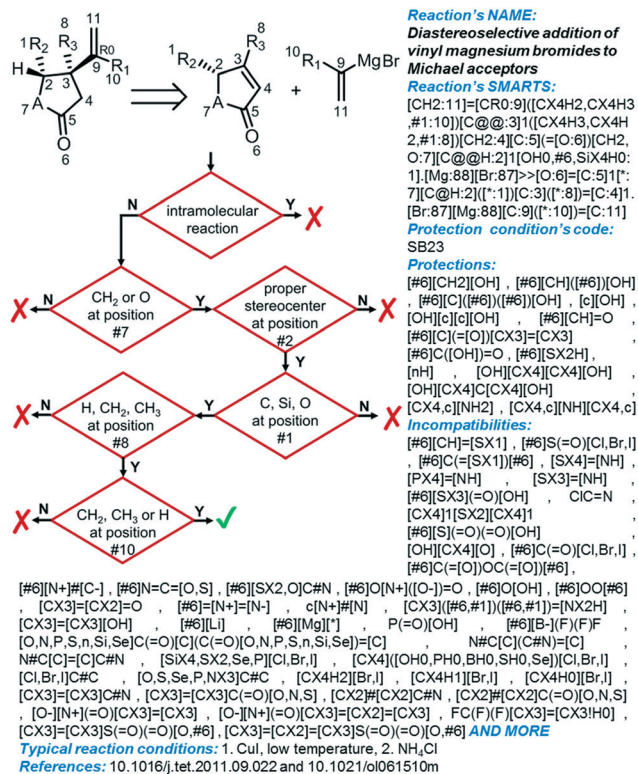
Another example, in Fig. 12, illustrates encoding of a more chemically advanced [2,3]-Wittig rearrangement with chirality transfer. The first requirement is the acyclic character of the ether (variants for cyclic substrates are also known, but they need to be coded in separate lines delineating the scope of substituents on a given ring system). The condition is met by denoting carbon #6 as  $R_0$  (which means it cannot participate in any ring). Another condition describes structure of the migrating group. In our specific example, it is limited to an unsubstituted allyl or allyl with methyl substituent at carbon #7. We note that although many groups can migrate in the [2,3]-Wittig rearrangement, and it is possible to write one, general SMARTS (e.g.,  $[\text{CH}_2:6][\#6, \text{Sn}:7]$ , where #6, Sn – means any carbon or tin atoms), it is not advisable to code the rule in this way since  $\alpha$ -(allyloxy)carbanions need to be generated at low temperatures (usually  $-78^\circ\text{C}$ ) to suppress the competing [1,2]-Wittig shift.<sup>62</sup> Even for the limited SMARTS transcription encompassing only generally accepted migrating groups such as phenyl, propargyl, or allyl, writing one SMARTS line (e.g.  $[\text{CH}_2:6][\text{CX}_3, \text{CX}_2, \text{c}:7]$ ) is not a good solution because these groups differ in terms of incompatibility requirements. Additionally, the line is restricted to the unsubstituted or 2-methyl allyl (and not to any allyl) because (i) it ensures *syn* selectivity of the newly created stereocenters at atoms #1 and #6; and (ii) less substituted allyl is a better migrating group. Another condition describes non-acidifying character of the substituent  $R_2$  (comprised of atoms #5, 11, 12, 13 in the SMARTS line). The chemical rationale is that increasing the acidity of protons adjacent to carbon #4 might result in its deprotonation (instead of deprotonation of atom #6) and migration of the second allyl group, thus resulting in a different product. Another condition is the *E* configuration of the resulting alcohol. [2,3]-Wittig rearrangement is known to be *E*-selective and the *Z* isomer is observed as the minor one. The last requirement that needs to be considered is proper relative stereochemistry at atoms #1 and #6, as *syn* alcohol derives from *Z*-alkene while *anti* is obtained with significantly lower selectivity from the *E*-isomer.<sup>63</sup> As in the example from Fig. 11, the reaction record also contains the list



**Fig. 10** Examples of incorrect predictions based on automatically extracted reaction rules. a) Electrophilic aromatic substitution of pyrroles occurs at positions marked with arrows (left). MCTS algorithm described in ref. 12 incorrectly suggests the possibility of functionalization at a less reactive position (red). b) Incorrect prediction for Suzuki coupling caused by lack of rules accounting for incompatible functional groups. More reactive aryl iodide cannot be a spectator of Suzuki coupling of an aryl bromide. c) Erroneous prediction for the preparation of *p*-iodophenyl boronic acid – this compound cannot be prepared as proposed from an aryl bromide due to higher reactivity of aryl iodide also present in the substrate. d) Improper handling of stereochemistry in Mitsunobu reaction. Proposed stereoretentive process is possible only if an chimeric assistance – missing in the extracted reaction transform – is accounted for. Extracted reaction core (grey frame) is common for primary, secondary (reacting with inversion of configuration) and tertiary (hardly reactive) alcohols. e) Incorrect predictions of stereochemical outcomes are not limited to Mitsunobu displacements. Another incorrectly stereoretentive process is proposed for the simple alkylation of a carboxylic acid with a secondary mesylate. In this case, the algorithmically extracted reaction template (grey frame) is also too general and allows for substrates bearing primary, secondary and tertiary mesylates. f) Faulty predictions *en route* to a valerolactam derivative. The extracted reaction core is too narrow and does not account for the reaction being substrate-controlled (cf. Fig. 8d). Lack of any directing groups in the substrate makes application of such a template to this target molecule incorrect. g) Incorrect prediction for a chiral-catalyst-controlled conjugate addition. The extracted reaction core is too narrow and, in this specific case, allows for mismatched (cf. Fig. 8c and d) substituent. h) To achieve any appreciable yields, the Bartoli indole synthesis requires<sup>60</sup> *ortho*-substitution which is missing in this automatically extracted reaction template. i) Synthesis of tetrahydroisoquinolines *via* Pictet-Spengler cyclisation is feasible only when electron-rich (hetero)arylethylamines are used as substrates. Automatically extracted reaction core is too narrow and allows for annulation of electron-poor pyridine. j) Information related to the presence of isotopically labelled atom is lost during generation of the synthetic precursors. Examples b–j were taken from ASKCOS software.<sup>11,14</sup> For additional examples and details, see Fig. S1–S25.†

of groups that require protection, those that are outright incompatible with the reaction, as well as ten other fields. We note that filling in many of these fields requires careful analysis – for instance, there are some 430 incompatible groups we routinely consider during coding, and the particular selection for a given reaction must reflect reaction mechanism as well as reaction conditions.

While the coding protocols such as those we outlined in this section offer a substantial improvement – in terms of chemical correctness – compared to indiscriminate machine-extraction of rules, they do not yet cover all aspects of rules' applicability, as it is often impossible to capture all the relevant effects just in the SMILES/SMARTS notation. These additional effects generally require augmentation by QM or MM

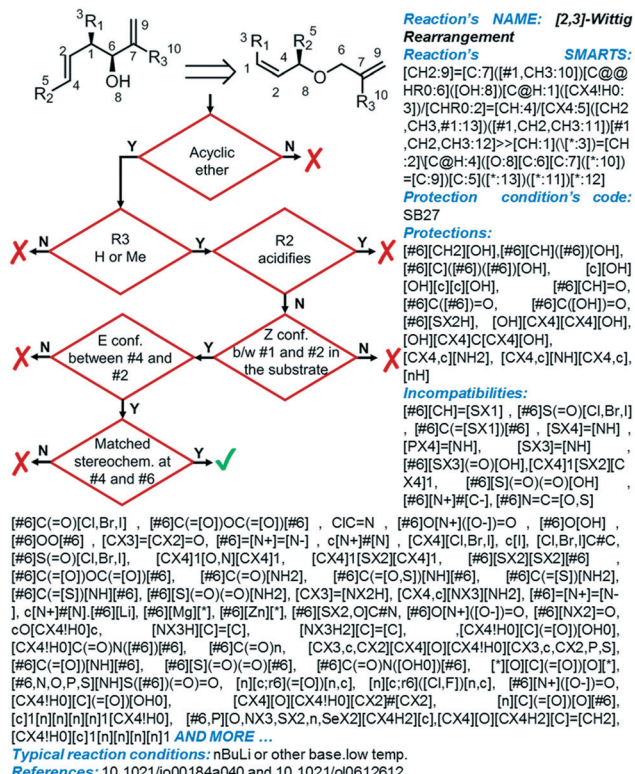


**Fig. 11** Mechanism based translation of diastereoselective Michael addition into a machine readable format. The upper left part of the figure presents the general scheme of the reaction in retrosynthetic direction and the “decision tree” guiding the coding of a corresponding reaction rule. The reaction record also contains information about groups that need protection, incompatibilities, reaction conditions etc.

calculations or by AI methods which we discuss in specific sections below.

## The importance of structural context

In our discussion above, one manifestation of “molecular context” has been the treatment of groups that present cross-reactivity problems or those that need to be protected. In addition, there are context dependencies that relate to the skeletal structure of the retron and/or the synthons. As an example, let's consider a Wittig or a metathesis reaction – these powerful reactions have very broad scopes but cannot proceed – due to strain – to install the double C=C bonds at the bridgehead atoms of small bicyclic systems (the so-called Bredt's rule). Specifying such outcomes at the level of reaction transforms is unfeasible whereas performing detailed calculations on-the-fly would take too much time (in particular, during retrosynthetic planning whereby very large numbers of intermediates are inspected). A more practical option is to prohibit in all molecules considered during planning those motifs that violate the Bredt's rules and, more generally, all those that are excessively strained. Some of such “prohibited motifs” are recognized as impossible by inspection, and some require more careful evaluation of strain *via*



**Fig. 12** Mechanism-based translation of [2,3]-Wittig rearrangement into a machine readable format. The upper-left part of the figure shows general scheme of the reaction in retrosynthetic direction and the “decision tree” guiding the coding of a specific reaction rule. The reaction record also contains information about groups that need protection, incompatibilities, reaction conditions, etc.

molecular mechanics calculations (see Fig. 13a for a small selection of such motifs from our library of *ca.* 600).

A slightly more subtle variation of this problem is what to do with motifs which, in principle, can exist but only under rather extreme conditions – *e.g.*, bicyclo[1.1.0]butanes or cyclopropenes. In such cases, we make the motif prohibited unless it is explicitly present in the target one wishes to synthesize.

A related problem arises when synthon/retron molecules are not themselves strained but the reaction cannot occur because strain develops in the transition state, often during intramolecular cyclization reactions. Some of such cases can be captured by prohibiting pairs of motifs in the retron and in the synthon(s) such that the former cannot form from the latter. In the example shown in Fig. 13b, a *trans*-fused bicyclic system cannot be formed *via* S<sub>N</sub>2 reaction involving base-induced ring opening of epoxide with carbamate. Such clear-cut cases, however, are rare and, in general, one has to quantify strain along the reaction coordinate. This problem is, of course, not a new one but when combined with synthetic planning in which large numbers of molecules are evaluated, it has a distinct flavour – all such calculations have to be performed very rapidly (in Chematica, well below 100 ms to allow inspection of tens of thousands upon thousands of

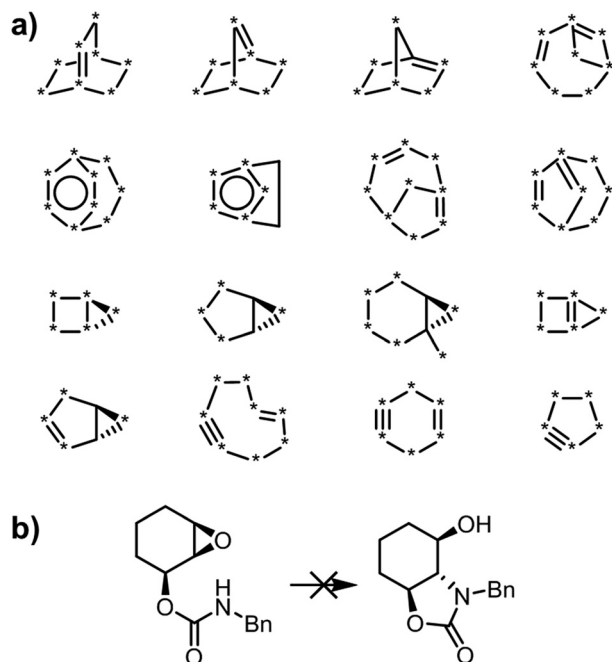


Fig. 13 Strained molecules and reactions. a) A small sample of strained motifs forbidden in retrosynthetic planning. \* denotes any atom. b) A bicyclic system cannot be prepared via base-induced ring opening of epoxide due to high strain along the reaction coordinate.

intermediates during a typical retrosynthetic search<sup>7,9</sup>). In some cases, classic physical organic chemistry knowledge is very helpful as it provides us with the information about the angles at which the cyclizing groups approach one another – in this way, the mutual orientations of the reacting species can be restricted, greatly simplifying the problem. For instance, if a cyclization reaction is between a nucleophile and a carbonyl group, the trajectory of approach is specified by the so-called Bürgi–Dunitz<sup>64</sup> and Flippin–Lodge<sup>65,66</sup> angles. Calculation of energy along such a well-defined coordinate by molecular mechanics is quite rapid and when the threshold strain is calibrated against examples of reactions that are known not to proceed, one can eliminate a sizeable proportion of impossible cyclizations (*cf.* examples in Fig. 14). This being said, we emphasize that we do not yet have a general method available to deal with all cyclizations, since in many cases a unique “angle of approach” is hard to define and one then has to sample more mutual orientations and molecular conformations; these nuances will be described in our upcoming papers on the topic.

## Accounting for non-local electronic effects

In some very popular reaction classes, the number of possible substituents is not only extremely large – too large to enumerate exhaustively – but their mutual placement is essential. The problem here is that one cannot just specify in the reaction transforms the lists of substituents at different positions since not all their combinations are permissible. For in-

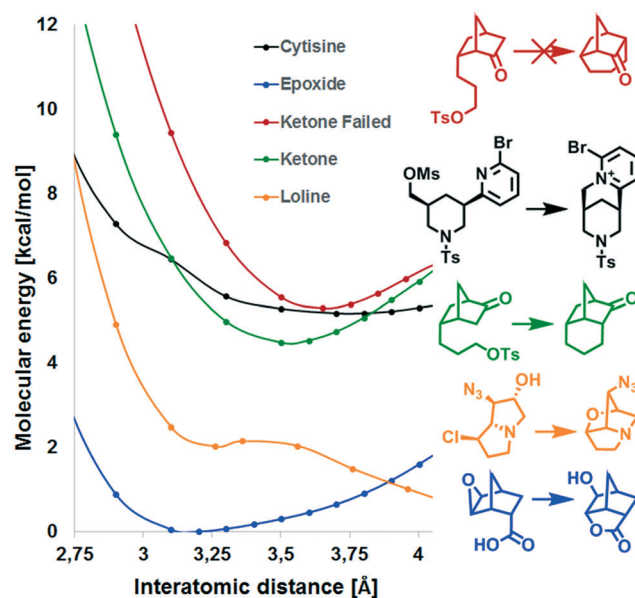


Fig. 14 Energy profiles for several experimentally attempted  $S_N2$  cyclizations with either positive or negative outcomes. The examples span alkylation of a ketone (red and green curves; from ref. 67), opening of an epoxide (blue, ref. 68), cyclisation forming loline's skeleton (orange; ref. 69), and cyclisation executed *en route* to cytosine (black, ref. 70). The admissible energies must be below the red curve which corresponds to a cyclization that is known not to proceed in experiment. The energies were calculated using Merck Molecular Force Field 94 (MMFF) and were close to energies obtained from more precise HF/6-311+G\*\* calculations. Figure reproduced with permission from ref. 71.

stance, for electrophilic aromatic substitutions on a benzene ring, we cannot just specify lists of, say, H, Cl, Br,  $NH_2$ ,  $NO_2$ , *etc.* for every surrounding position because, as is well known from organic chemistry classes, different substituents have different *ortho/para vs. meta*-directing abilities and, depending on a specific arrangement of these groups present on the ring, they might collectively activate or deactivate our position of interest. The problem is further compounded by the fact that the degrees of such influence can be different depending on the aromatic or heteroaromatic system being substituted. In such cases, the only legitimate strategy is to define a reaction core very narrowly (*i.e.*, an electrophile plus an aromatic carbon being substituted) but supplement such a reaction rule by a routine calculating the propensity of this specific carbon atom to undergo the substitution reaction.

In our early works,<sup>7</sup> we used a classical Hückel method to calculate electron densities over aromatic systems and allowed substitutions only at positions for which the densities were above certain threshold values. This method, however, had only ~75% accuracy over diverse aromatic systems. The “diverse” is an important keyword here because methods parameterized only on certain types of aromatic systems (see ref. 72 and 73) are of limited use for generalized synthetic planning (and the more accurate proton/electrophile affinity approaches are prohibitively slow). In the end, we implemented a “hybrid approach” combining Hammett substituent

constants, proton affinities averaged over all aromatic carbons within a specific ring type (pre-calculated at the DFT level of theory using B3LYP functional with 6-31+G\* basis set), the Hückel method (with parameters for heteroatoms taken mostly from ref. 74), as well as some additional heuristics. This model, described in detail in the ESI† to ref. 9, evaluates aromatic substitution reactions with speeds commensurate with synthetic planning (10–100 ms per reaction) and offers accuracy above 90%.

## Augmenting reaction rules by AI models

The most general problem one might face when considering rules' applicability is when both electronic and steric effects are important. In situations when there are multiple reaction precedents/examples available for a given, well defined reaction type, one might fine-tune substituent scope by machine learning, ML, methods. This type of approach has been used recently by Doyle and co-workers<sup>75</sup> for predicting the substituent-dependent outcomes of Buchwald–Hartwig couplings, although the specific implementation of AI methods subsequently received critique<sup>76</sup> for the lack of statistical rigor in model testing.

To illustrate some key aspects of AI rule augmentation, we consider a synthetically powerful<sup>77,78</sup> Diels–Alder reaction for which ~20 000 reaction precedents are readily available in Reaxys. As we discussed in detail in our recent work on this topic,<sup>27</sup> unsymmetrical dienes and dienophiles can react in different orientations to give different regioisomers (Fig. 15a) or diastereoisomers (Fig. 15b) or, when multiple diene/dienophile sites are present in the molecule, altogether different products (Fig. 15c). These outcomes are dictated by the substituents present on the diene (maximum six substituents) and the dienophile (maximum four substituents) – in other words, the reaction core is small and well defined but

much of the game is in the atoms/groups decorating the core. Clearly, enumerating all possible substituent combinations is impractical and, indeed, quite pointless, as the outcomes for the majority of them would have no experimental benchmark to compare with. QM-based calculation are not only slow but, as we showed in,<sup>27</sup> offer an accuracy of only ~80% in terms of predicting correct outcomes. On the other hand, the number of literature precedents is sufficient to train ML models assigning certain “features” to the substituents and learning how the combinations of these features translate into reaction outcome (the leading regio-, diastereo-, or site-isomer). Without repeating the entire discussion from ref. 27, few main points are worth re-emphasizing as their applicability is beyond the Diels–Alder example:

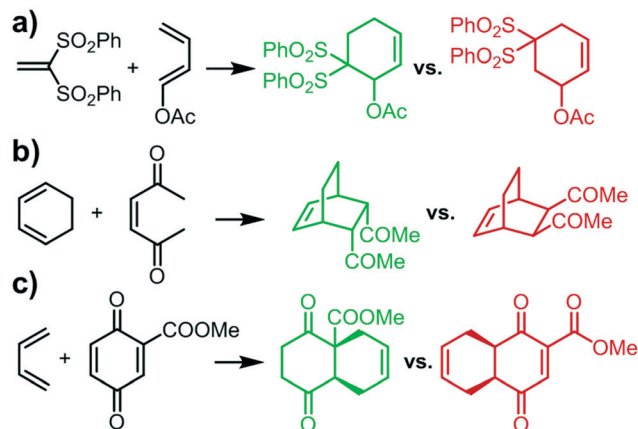
(i) It is essential to use features that capture electron donating/withdrawing propensities of substituents as well as their steric bulk. With, respectively, Hammett constants<sup>79</sup> and the TSEI indices<sup>80</sup> assigned to the substituents, the ML models can offer accuracy well above 90% and are also applicable to structurally diverse examples, including cases not seen during model training.

(ii) Although atom-connectivity-based descriptors (*e.g.*, ECFP4,<sup>81</sup> MACCS, or RDKit<sup>44</sup> fingerprints) or even physically meaningless descriptors (*e.g.* random numbers assigned to substituents) can offer high accuracies when trained on structurally related examples, such models fail when confronted with examples structurally different than those seen during training.

(iii) The AI methods perform significantly better when provided some “insight” about the reaction. When the reaction cores are not specified and the algorithms are learning from just the structures of the products and reagents, their performance is significantly worse.

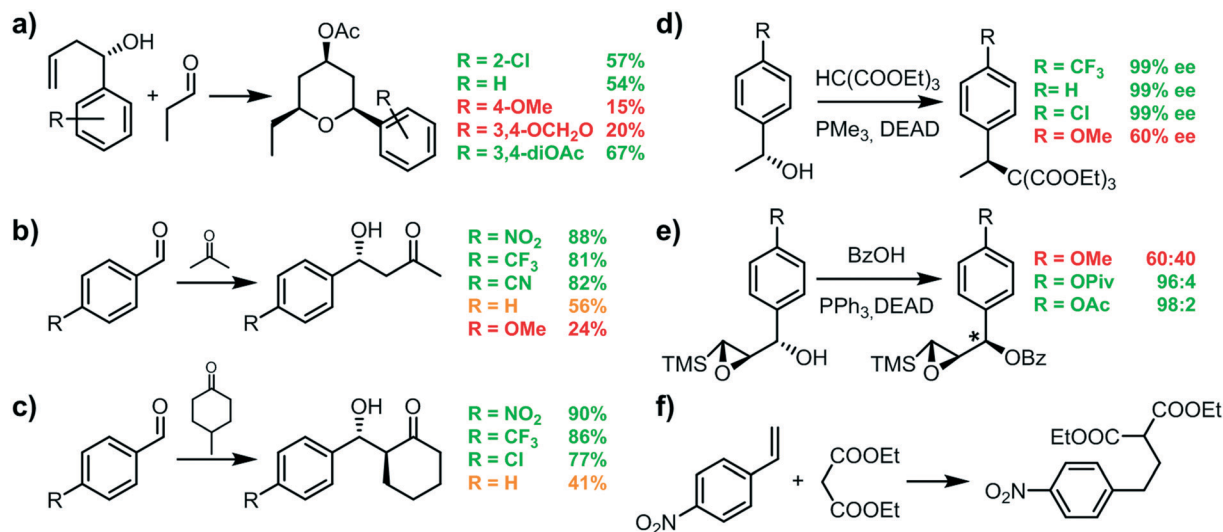
## Ensuring rules' quality and estimating scope

As evidenced by our discussion so far, the number of possible factors that need to be considered during reaction coding is quite substantial. Of course, not all of the aspects we highlighted need to be considered simultaneously: for instance, calculations of electron densities usually are relevant only to systems of delocalized  $\pi$  electrons (but see Fig. 16), whereas AI augmentation routines should be considered only for rules having large numbers (say, more than ~1000) of reliable literature precedents. Still, the coding process is certainly very time-consuming and also requires tremendous care to eliminate possible human errors. In our team, we have implemented several levels of quality control – special scripts checking for rules' proper syntax, scripts testing applicability of rules on some test-molecules, a peer-review cross-checking system (chemists checking each other's transforms), and ultimate verification of the results by a super-user who inputs the reaction rules into Chematica's database. We also have in place an error reporting systems from Chematica's end users and standardized procedures how such errors are fixed. Over



**Fig. 15** Selectivity of Diels–Alder reaction. a) Formation of regioisomers from unsymmetrical dienes. b) Formation of *endo*-/*exo*-diastereoisomers. c) Formation of different products from substrates with multiple reaction sites. Green – observed product, red – possible by-product.





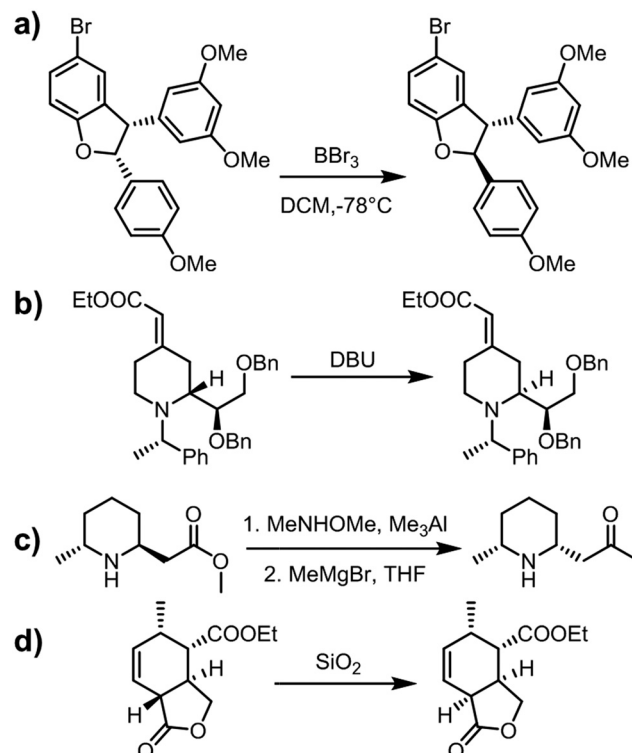
**Fig. 16** Reactions other than electrophilic aromatic substitutions that would benefit from electronic density calculation. a) Prins cyclizations of electron-rich benzylic alcohols suffers from competitive oxonia-Cope rearrangement; b and c) organocatalytic aldol reactions proceed well for electron deficient benzaldehydes; d and e) Mitsunobu reaction of electron-rich secondary alcohols suffers from competitive S<sub>N</sub>1 process leading to erosion of ee; f) addition of nucleophiles to vinylarenes requires electron deficient arene to proceed. Substituents and yields listed in orange and red correspond to cases that one would probably want to eliminate from the reactions' scopes.

the years, we followed these procedures to encode more than 70 000 reaction transforms whose ultimate validation has been their correct performance in experimental execution of computer-planned routes.<sup>9</sup>

In parallel, we have been keeping track of the retro-/synthetic scope of our collection. One illustrative metric has been how many transforms with specific substituents at each position our rules cover – the most recent number stands at ~4.5 million reactions (~70 000 rules/transforms, with a median of distinct 64 specific reactions per transform) expanding to several tens of millions when ‘\*’ symbols denoting ‘any atom’ are considered (with a very conservative assumption that each ‘\*’ admits just one substituent type). This is significantly more than 320 000 unique radius = 2 transforms extracted<sup>29</sup> from the USPTO collection, and more than ~13 million literature precedents in the Reaxys database, altogether attesting to the synthetic latitude of our collection. The bottom-line message of this section is that, contrary to some claims,<sup>12</sup> expert-coding of rules can not only keep up with the ‘current literature’ (as deposited in reaction databases) but can exceed its scope and make high-quality, mechanism-based extensions to reaction variants not yet carried out.

## Some unsolved problems

Of course, there are still rules that need to be added and we estimate that our ultimate collection will comprise some 100 000 transforms to address even the most complex synthetic problems. At least some of these rules will benefit from additional electronic density calculations, probably at levels of theory higher than those used to predict aromatic substitutions (*cf.* earlier in the text and ref. 9). Such calculations



**Fig. 17** Unexpected epimerizations of thermodynamically unstable molecules. a) Epimerization of the *cis*-diaryl tetrahydrofuran via *p*-quinonemethide observed during attempted demethylation of phenols.<sup>89</sup> b) Epimerization of a piperidine derivative through retro 1,6/1,6 addition observed during attempted HWE olefination.<sup>90</sup> c) Epimerization of the *trans*-piperidine via retro 1,4/1,4 addition observed during attempted conversion of an ester to a methyl ketone.<sup>91</sup> d) Epimerization of strained *trans*-fused lactone<sup>92</sup> upon treatment with SiO<sub>2</sub>.

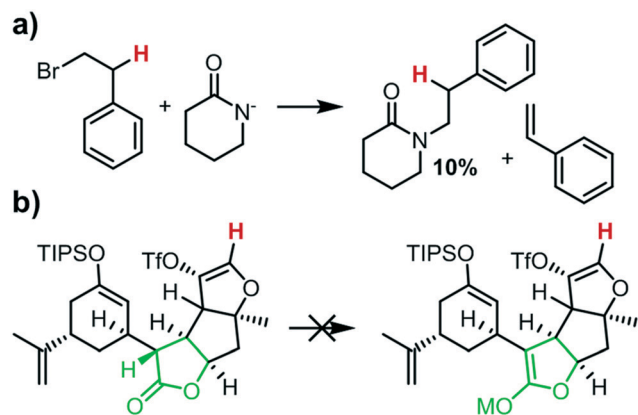


Fig. 18 Presence of acidic H's causing problems in attempted a) alkylation of lactam with phenethyl bromide, and b) enolisation of lactone (green). Interfering, acidic protons are highlighted in red.

could, for instance, better delineate the applicability of Prins cyclisations suffering from competing oxonia-Cope rearrangement for electron rich benzyl-alcohol substrates<sup>82</sup> (Fig. 16a), determine electron density thresholds for benzaldehydes participating in aldol reaction<sup>83,84</sup> (Fig. 16b and c), penalize substitutions occurring at electron rich benzylic positions commonly suffering from competitive  $S_N1$  process leading to erosion of optical purity<sup>85,86</sup> (Fig. 16d and e), or help assure the electron withdrawing character of styrenes alkylated with diesters<sup>87</sup> (Fig. 16f).

Another challenge is to couple the reaction rules with MM calculations of molecules' conformations to ensure that the reaction sites are available and not sterically hindered – such evaluation would be of most relevance to the synthesis of natural products for which there are many examples of conformational effects dictating reactivity.<sup>88</sup> The specific molecules formed as a result of rules' application should also be evaluated for the stability of certain stereochemical motifs. For instance, *cis*-substituted dihydrobenzofuran (Fig. 17a) is known to be unstable during  $BBr_3$  mediated demethylation of phenol,<sup>89</sup> certain piperidines epimerize *via* retro-Michael/Michael addition during HWE olefination<sup>90</sup> or conversion of ester to methyl ketone<sup>91</sup> (Fig. 17b and c), while *trans*-fused Diels–Alder adduct shown in Fig. 17d epimerizes easily to a more stable *cis*-lactone when treated with silica.<sup>92</sup> It is presently unclear to us whether such subtle effects can be reliably calculated/predicted or whether it is better to create and curate a growing, literature-based list of “unstable motifs” that would be matched against the outcomes of reaction rules.

Finally, an important problem to address is the estimation of the  $pK_a$  of CH acids and protic groups. For instance, during alkylation of a valerolactam<sup>93</sup> anion with phenethyl bromide, the low yield can be ascribed to the acidity of benzylic H (red in Fig. 18a) and competing E2 elimination forming styrene by-product. In the same genre but for a more complex target, unexpected acidity of vinyl triflate thwarted the desired deprotonation of lactone during Vandervel's synthesis of Ineleganolide<sup>94</sup> (Fig. 18b). Problems of this sort can be

solved by estimating the  $pK_a$  values within the molecule and inspecting if there are protons more acidic than the one at our desired reaction centre. Although  $pK_a$  values for some specific series of structurally related compounds have been described in the literature,<sup>95</sup> a solution applicable to arbitrary molecules in organic solvents is still missing. There are some models developed internally by large pharma companies or commercial packages from companies<sup>96–99</sup> such as Schrödinger, but it is unclear how accurate these tools are (*e.g.*, Schrödinger's package appears quite accurate for some cases, but in some others – *e.g.*, PhOMe calculated with Jaguar's DFT calculation with empirical corrections – gives predictions missing the experimental values by almost five  $pK_a$  units!). We have been working on developing our own  $pK_a$  predictor but it is too early to estimate its ultimate accuracy.

## Conclusions

In summary, translation of organic-chemical knowledge into machine readable rules is much more than just writing out SMILES/SMARTS strings describing reaction cores. The various types of considerations and calculations that accompany the coding process are, in fact, a modern embodiment of physical-organic chemistry and an interesting junction between this seasoned area of chemical research and contemporary computing methods. Above all, the protocols we strived to illustrate in this Perspective are a cornerstone on which all higher-level routines to find complete synthetic pathways rest – as we mentioned in the introduction and wish to reiterate here, machine's ability to design high-quality synthetic routes will be only as good as the quality of the underlying rules describing individual reactions.

## Conflicts of interest

While the Chematica retrosynthesis platform, mentioned in the text, was originally developed and owned by B.A.G.'s Grzybowski Scientific Inventions, LLC, neither he nor the co-authors no longer hold any stock in this company, which is now property of Merck KGaA, Darmstadt, Germany. The authors continue to collaborate with Merck within the DARPA “Make-It” award. All queries about access options to Chematica (now rebranded as Synthia™), including academic collaborations, should be directed to Dr. Sarah Trice at sarah.trice@sial.com.

## Acknowledgements

The authors thank the U.S. DARPA for generous support under the “Make-It” Award, 69461-CH-DRP #W911NF1610384. B. A. G. also gratefully acknowledges personal support from the Institute for Basic Science Korea, Project Code IBS-R020-D1.

## References

- 1 E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.
- 2 E. Corey, A. Long and S. Rubenstein, *Science*, 1985, **228**, 408–418.

- 3 E. J. Corey, R. D. Cramer and W. J. Howe, *J. Am. Chem. Soc.*, 1972, **94**, 440–459.
- 4 T. H. Varkony, D. H. Smith and C. Djerassi, *Tetrahedron*, 1978, **34**, 841–852.
- 5 I. Ugi, J. Bauer, R. Baumgartner, E. Fontain, D. Forstmeyer and S. Lohberger, *Pure Appl. Chem.*, 1988, **60**, 1573–1586.
- 6 J. B. Hendrickson, D. L. Grier and A. G. Toczko, *J. Am. Chem. Soc.*, 1985, **107**, 5228–5238.
- 7 S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2016, **55**, 5904–5937.
- 8 M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2005, **44**, 7263–7269.
- 9 T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice and B. A. Grzybowski, *Chem*, 2018, **4**, 522–532.
- 10 J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade and H. Y. Ando, *J. Chem. Inf. Model.*, 2009, **49**, 593–602.
- 11 ASKCOS, <http://askcos.mit.edu>, (accessed 04 June 2019).
- 12 M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
- 13 M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski and K. J. M. Bishop, *Angew. Chem., Int. Ed.*, 2012, **51**, 7928–7932.
- 14 C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281–1289.
- 15 S. Lemonick, *Is machine learning overhyped?*, C&EN Glob. Enterp., 2018, 96(34), 16–20.
- 16 Reaxys, <https://www.reaxys.com>, (accessed 04 June 2019).
- 17 SciFinder, <https://scifinder.cas.org>, (accessed 04 June 2019).
- 18 SPRESI, <http://www.infochem.de/products/databases/spresi.shtml>, (accessed 04 June 2019).
- 19 D. M. Lowe, Chemical reactions from US pat., (1976-Sep2016), 2017, [https://figshare.com/articles/Chemical\\_reactions\\_from\\_US\\_patents\\_1976-Sep2016\\_/5104873](https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873), (accessed 04 June 2019).
- 20 F. A. Carey and R. J. Sundberg, *Advanced Organic Chemistry*, Springer, US, Boston, MA, 2007.
- 21 E. V. Anslyn and D. Dennis, *Modern Physical Organic Chemistry*, University Science Books, Herndon, VA, 2005.
- 22 S. Kenis, M. D'hooghe, G. Verniest, T. A. Dang Thi, C. Pham The, T. Van Nguyen and N. De Kimpe, *J. Org. Chem.*, 2012, **77**, 5982–5992.
- 23 K. C. Nicolaou, G. Vassilikogiannakis, W. Mägerlein and R. Kranich, *Angew. Chem., Int. Ed.*, 2001, **40**, 2482–2486.
- 24 W. Jin, C. Coley, R. Barzilay and T. Jaakkola, *NIPS*, 2017, 2607–2616.
- 25 P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9**, 6091–6098.
- 26 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103–1113.
- 27 W. Beker, E. P. Gajewska, T. Badowski and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2019, **58**, 4515–4519.
- 28 W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin and B. A. Grzybowski, *Nat. Commun.*, 2019, **10**, 1434.
- 29 I. A. Watson, J. Wang and C. A. Nicolaou, *Aust. J. Chem.*, 2019, **11**, 1.
- 30 Y. Masaki, H. Arasaki and M. Shiro, *Chem. Lett.*, 2000, **29**, 1180–1181.
- 31 K. Spielmann, R. M. de Figueiredo and J.-M. Campagne, *J. Org. Chem.*, 2017, **82**, 4737–4743.
- 32 S. P. Chavan and H. S. Khatod, *Tetrahedron: Asymmetry*, 2012, **23**, 1410–1415.
- 33 D. Mal, B. Senapati and P. Pahari, *Tetrahedron Lett.*, 2006, **47**, 1071–1075.
- 34 W. R. Roush, M. R. Michaelides, D. F. Tai and W. K. M. Chong, *J. Am. Chem. Soc.*, 1987, **109**, 7575–7577.
- 35 D. Mal and J. Roy, *Tetrahedron*, 2015, **71**, 1247–1253.
- 36 D. Mal and J. Roy, *Org. Biomol. Chem.*, 2015, **13**, 6344–6352.
- 37 J. R. Del Valle and M. Goodman, *J. Org. Chem.*, 2003, **68**, 3923–3931.
- 38 H. Becker, M. A. Soler and K. Barry Sharpless, *Tetrahedron*, 1995, **51**, 1345–1376.
- 39 A. Tanaka, H. Okamoto and M. Bersohn, *J. Chem. Inf. Model.*, 2010, **50**, 327–338.
- 40 R. R. Karimov, D. S. Tan and D. Y. Gin, *Tetrahedron*, 2018, **74**, 3370–3383.
- 41 M. Gao, Y.-C. Wang, K.-R. Yang, W. He, X.-L. Yang and Z.-J. Yao, *Angew. Chem., Int. Ed.*, 2018, **57**, 13313–13318.
- 42 T. Tsukiyama, A. Kinoshita, R. Ichinose and K. Sato, *Biosci., Biotechnol., Biochem.*, 2002, **66**, 1407–1411.
- 43 B. M. Trost, M. R. Machacek and H. C. Tsui, *J. Am. Chem. Soc.*, 2005, **127**, 7014–7024.
- 44 *RDKit: Open-Source Cheminformatics Software*, <http://www.rdkit.org/>, (accessed 04 June 2019).
- 45 *RDChiral*, <https://github.com/connorcoley/rdchiral>, (accessed 04 June 2019).
- 46 J. R. Coombs, F. Haeffner, L. T. Kliman and J. P. Morken, *J. Am. Chem. Soc.*, 2013, **135**, 11222–11231.
- 47 W. R. Roush, A. D. Palkowitz and M. J. Palmer, *J. Org. Chem.*, 1987, **52**, 316–318.
- 48 H. Kim, S. Ho and J. L. Leighton, *J. Am. Chem. Soc.*, 2011, **133**, 6517–6520.
- 49 R. W. Hoffmann, *Chem. Rev.*, 1989, **89**, 1841–1860.
- 50 G. Schmid, T. Fukuyama, K. Akasaka and Y. Kishi, *J. Am. Chem. Soc.*, 1979, **101**, 259–260.
- 51 J. Shi, H. Shigehisa, C. A. Guerrero, R. A. Shenvi, C.-C. Li and P. S. Baran, *Angew. Chem., Int. Ed.*, 2009, **48**, 4328–4331.
- 52 N.-H. Lin, L. E. Overman, M. H. Rabinowitz, L. A. Robinson, M. J. Sharp and J. Zablocki, *J. Am. Chem. Soc.*, 1996, **118**, 9062–9072.
- 53 H. Chiba, S. Oishi, N. Fujii and H. Ohno, *Angew. Chem., Int. Ed.*, 2012, **51**, 9169–9172.
- 54 G. Sabitha, P. AnkiReddy and S. Das, *Synthesis*, 2014, **47**, 330–342.

- 55 J.-F. Brazeau, A.-A. Guilbault, J. Kochuparampil, P. Mochirian and Y. Guindon, *Org. Lett.*, 2010, **12**, 36–39.
- 56 L. C. Dias and A. G. Salles, *J. Org. Chem.*, 2009, **74**, 5584–5589.
- 57 [https://scifinder.cas.org/scifinder/view/link\\_v1/reaction.html?l=T5OKcO0Ri0Vxs5SgbYOjGp9biqO0mXFfn0S569XvHjrkzEryFRL2Nr9NcFsIqYver](https://scifinder.cas.org/scifinder/view/link_v1/reaction.html?l=T5OKcO0Ri0Vxs5SgbYOjGp9biqO0mXFfn0S569XvHjrkzEryFRL2Nr9NcFsIqYver), (accessed 04 June 2019).
- 58 W. Kleist, S. S. Pröckl, M. Drees, K. Köhler and L. Djakovitch, *J. Mol. Catal. A: Chem.*, 2009, **303**, 15–22.
- 59 B. Mishra, M. Neralkar and S. Hotha, *Angew. Chem., Int. Ed.*, 2016, **55**, 7786–7791.
- 60 G. Bartoli, G. Palmieri, M. Bosco and R. Dalpozzo, *Tetrahedron Lett.*, 1989, **30**, 2129–2132.
- 61 K. J. M. Bishop, R. Klajn and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2006, **45**, 5348–5354.
- 62 T. Nakai and K. Mikami, in *Organic Reactions*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1994, pp. 105–209.
- 63 D. J. S. Tsai and M. M. Midland, *J. Org. Chem.*, 1984, **49**, 1842–1843.
- 64 H. B. Bürgi, J. D. Dunitz, J. M. Lehn and G. Wipff, *Tetrahedron*, 1974, **30**, 1563–1572.
- 65 C. H. Heathcock and L. A. Flippin, *J. Am. Chem. Soc.*, 1983, **105**, 1667–1668.
- 66 E. P. Lodge and C. H. Heathcock, *J. Am. Chem. Soc.*, 1987, **109**, 3353–3361.
- 67 J. G. Henkel and L. A. Spurlock, *J. Am. Chem. Soc.*, 1973, **95**, 8339–8351.
- 68 H. Tan and J. H. Espenson, *J. Mol. Catal. A: Chem.*, 1999, **142**, 333–338.
- 69 M. Cakmak, P. Mayer and D. Trauner, *Nat. Chem.*, 2011, **3**, 543–545.
- 70 V. Barát, D. Csókás and R. W. Bates, *J. Org. Chem.*, 2018, **83**, 9088–9095.
- 71 K. Molga, P. Dittwald and B. A. Grzybowski, *Chem*, 2019, **5**, 460–473.
- 72 J. C. Kromann, J. H. Jensen, M. Kruszyk, M. Jessing and M. Jørgensen, *Chem. Sci.*, 2018, **9**, 660–665.
- 73 A. Tomberg, M. J. Johansson and P.-O. Norrby, *J. Org. Chem.*, 2019, **84**, 4695–4703.
- 74 F. A. Van-Catledge, *J. Org. Chem.*, 1980, **45**, 4801–4802.
- 75 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
- 76 K. V. Chuang and M. J. Keiser, *Science*, 2018, **362**, eaat8603.
- 77 K. C. Nicolaou, S. A. Snyder, T. Montagnon and G. Vassilikogiannakis, *Angew. Chem., Int. Ed.*, 2002, **41**, 1668–1698.
- 78 M. Juhl and D. Tanner, *Chem. Soc. Rev.*, 2009, **38**, 2983–2992.
- 79 C. Hansch, A. Leo and R. W. Taft, *Chem. Rev.*, 1991, **91**, 165–195.
- 80 C. Cao and L. Liu, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 678–687.
- 81 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
- 82 S. R. Crosby, J. R. Harding, C. D. King, G. D. Parker and C. L. Willis, *Org. Lett.*, 2002, **4**, 577–580.
- 83 L. Li, S. Gou and F. Liu, *Tetrahedron: Asymmetry*, 2014, **25**, 193–197.
- 84 S. Luo, H. Xu, J. Li, L. Zhang, X. Mi, X. Zheng and J.-P. Cheng, *Tetrahedron*, 2007, **63**, 11307–11314.
- 85 M. C. Hillier, J.-N. Desrosiers, J.-F. Marcoux and E. J. J. Grabowski, *Org. Lett.*, 2004, **6**, 573–576.
- 86 R. F. C. Brown, W. R. Jackson and T. D. McCarthy, *Tetrahedron Lett.*, 1993, **34**, 1195–1196.
- 87 K. K. Gnanasekaran, J. Yoon and R. A. Bunce, *Tetrahedron Lett.*, 2016, **57**, 3190–3193.
- 88 E. M. Carreira and L. Kvaerno, *Classics in Stereoselective Synthesis*, Wiley – VCH, Weinheim, 2009.
- 89 Y. Natori, M. Ito, M. Anada, H. Nambu and S. Hashimoto, *Tetrahedron Lett.*, 2015, **56**, 4324–4327.
- 90 P. Etayo, R. Badorrey, M. D. Díaz-de-Villegas and J. A. Gálvez, *Chem. Commun.*, 2006, 3420–3422.
- 91 K. Csatayová, I. Špánik, V. Ďurišová and P. Szolcsányi, *Tetrahedron Lett.*, 2010, **51**, 6611–6614.
- 92 J. Wu, H. Yu, Y. Wang, X. Xing and W.-M. Dai, *Tetrahedron Lett.*, 2007, **48**, 6543–6547.
- 93 T. Fujii, S. Yoshifuji and K. Yamada, *Chem. Pharm. Bull.*, 1978, **26**, 2071–2080.
- 94 E. J. Horn, J. S. Silverston and C. D. Vanderwal, *J. Org. Chem.*, 2016, **81**, 1819–1838.
- 95 F. G. Bordwell, *Acc. Chem. Res.*, 1988, **21**, 456–463.
- 96 R. Fraczekiewicz, M. Lobell, A. H. Göller, U. Krenz, R. Schoenneis, R. D. Clark and A. Hillisch, *J. Chem. Inf. Model.*, 2015, **55**, 389–397.
- 97 C. Liao and M. C. Nicklaus, *J. Chem. Inf. Model.*, 2009, **49**, 2801–2812.
- 98 J. C. Shelley, A. Cholleti, L. L. Frye, J. R. Greenwood, M. R. Timlin and M. Uchimaya, *J. Comput.-Aided Mol. Des.*, 2007, **21**, 681–691.
- 99 A. D. Bochevarov, M. A. Watson, J. R. Greenwood and D. M. Philipp, *J. Chem. Theory Comput.*, 2016, **12**, 6001–6019.

## Supplementary Information for Manuscript

**The logic of translating chemical knowledge into machine-processable forms: A modern playground for physical-organic chemistry.**

*Karol Molga<sup>1</sup>, Ewa P. Gajewska<sup>1</sup>, Sara Szymkuć<sup>1</sup>, Bartosz A. Grzybowski<sup>1,2,\*</sup>*

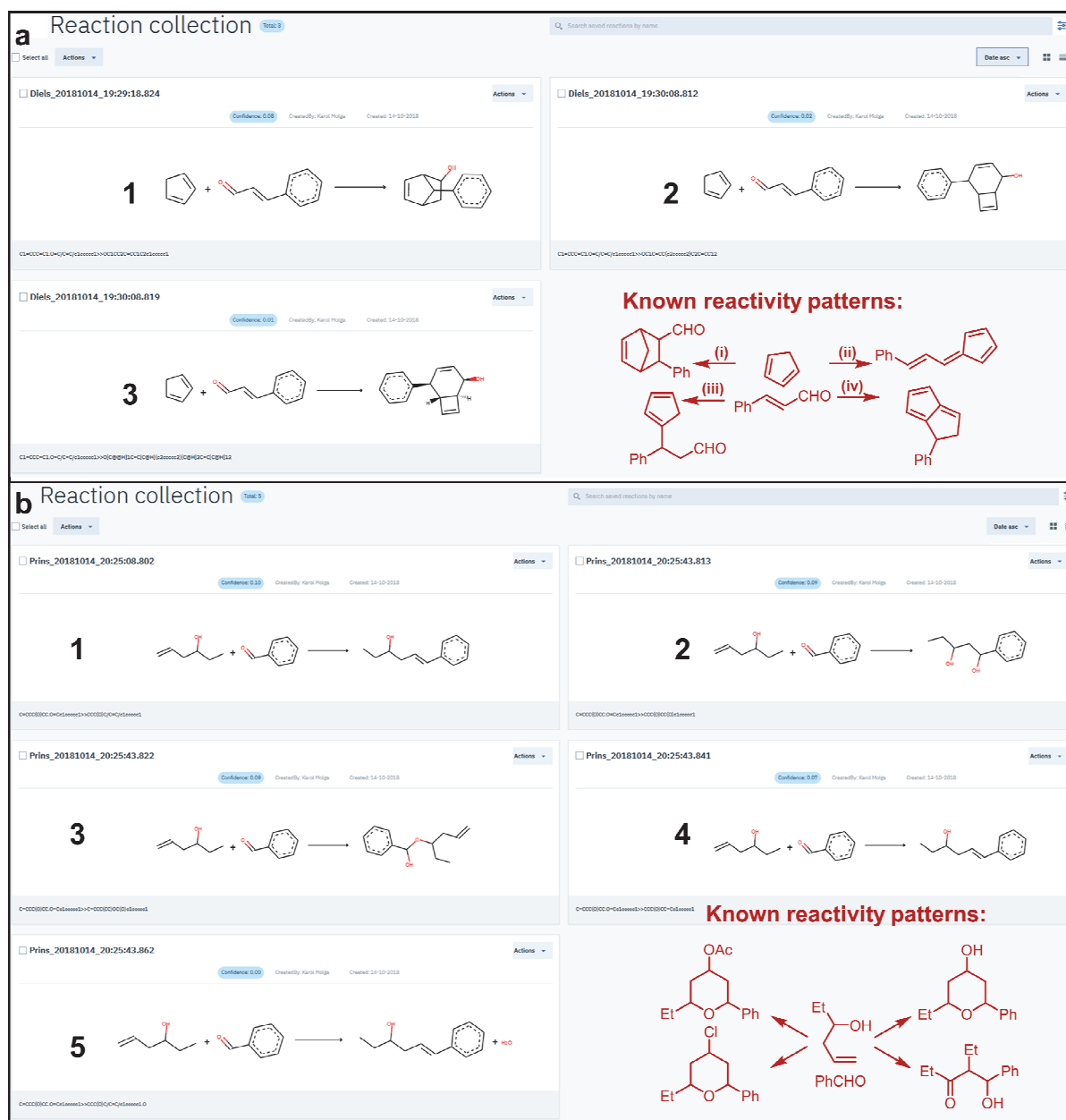
<sup>1</sup> Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland

<sup>2</sup> IBS Center for Soft and Living Matter and Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

\*Correspondence to: nanogrzybowski@gmail.com

### Contents:

Examples of synthetic predictions by linguistics-based (**Figure S1**) and machine-extracted, rule-based (**Figures S2-S25**) algorithms vs. predictions by Chematica (**Figures S2-S25**) using mechanism-based reaction rules. The examples are chosen to emphasize the difference in rule quality when dealing with even simple targets. For more complex synthetic examples using Chematica, see ref. 7 and the experimentally validated syntheses in ref. 9 in the main text. The examples are based on the versions of the softwares as of late October 2018.



**Figure S1.** IBM's tool (<https://rxn.res.ibm.com>) for predicting outcomes of organic reactions using neural sequence-to-sequence models (based on Schwaller *et. al.*, *Chem. Sci.*, **2018**, 9, 6091) fails even for simple examples. **a)** Neural network predictions for the reaction of cinnamaldehyde with cyclopentadiene are far from the known and realistic reactivity patterns. These two compounds are known to participate in Diels-Alder (i), aldol (ii), ene (iii) or tandem ene-aldol (iv) reactions shown in the red inset. None of these reactivity patterns are proposed. Instead, formation of products returned by the neural network requires extrusion of carbon atom from enal (#1) or migration of carbon atom from cyclopentadiene to enal to form 6-4 ring system in #2 and #3. Both processes proposed by the platform would raise an eyebrow (mildly speaking...) of any practicing organic chemist. **b)** Allyl alcohols are well known to react with aldehydes under acidic conditions yielding tetrahydropyrans in Prins reaction or under metal catalysis to deliver reductive aldol adducts

(reactions shown in red in the inset). None of these reactivity patterns are rediscovered by IBM's neural network. Instead, the proposed products are made via a highly improbable (unknown in chemistry!) coupling of alkene and aldehyde with the extrusion of carbon atoms from the unactivated alkyl chain yielding products with five (#2) and six carbon (#1,#4,#5) chains. (Additionally, #1, #4 and #5 return the same main product with different confidence levels). The only reasonable suggestion (albeit predicted as only third) is the formation of acyclic acetal. For this example, comparison to Chematica is not feasible since Chematica works in the retrosynthetic, not "forward" direction.





the program produced two variants (#14 is identical with #23) of the same method – Leimgruber-Batcho cyclisation – with only minute differences in leaving groups (diethylamine vs. pyrrolidine). The proposed Bartoli indole synthesis requires *ortho*-substitution (which is missing in the examined case and in the extracted reaction core, see **Figure S3**) to proceed with any satisfactory yield – we note that such small nuances of chemical reactivity are extremely important as the small difference in structure may cause dramatic change in reactivity and virtually impossible to be covered by any automatic rule extraction system due to insufficient number of reported examples. The program also does not account for the pricing of starting materials and scores higher cyclisation of more expensive (4\$/g) nitrophenylacetonitrile (#2) over similar reaction (#6) using aniline (1\$/g) and tris-*tert*-butylamine (1\$/g) or reduction (#4) of indolinone (1\$/g). Top 25 out of 32 returned results are shown. Top predictions returned by Chematica (bottom) take advantage of inexpensive and commercially available derivatives (indoline, methyl indole-3-carboxylate or chloroindole). Alternatively, the indole skeleton is created from the phenylacetonitrile or in the Fischer, Leimgruber-Batcho, or Mandelung processes.

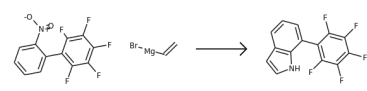
#### Template 59c51ac605581eb9f57b9d4a

Template: [c:6]:[c:1;R0;D3;+0:5]:[c:4]:[NH;D2;+0:3]:[cR;D2;+0:2]:[cR;D2;+0:1]:>>Br-[Mg]-[CR;D2;+0:1]=[CR2;D2;+0:2].O=[H;R0;D3;3](-[O-])-[c:4]:[cR;D2;+0:5]:[c:6]

233 total references

Export Reaxys query for precedents

[c:6][c:1;R0;D3;+0:5]:[c:4]:[NH;D2;+0:3]:[cR;D2;+0:2]:[cR;D2;+0:1]:>>Br-[Mg][CH;D2;+0:1]=[CH2;D1;+0:2]O=[H;R0;D3;3]([O-])[c:4]:[cR;D2;+0:5]:[c:6]

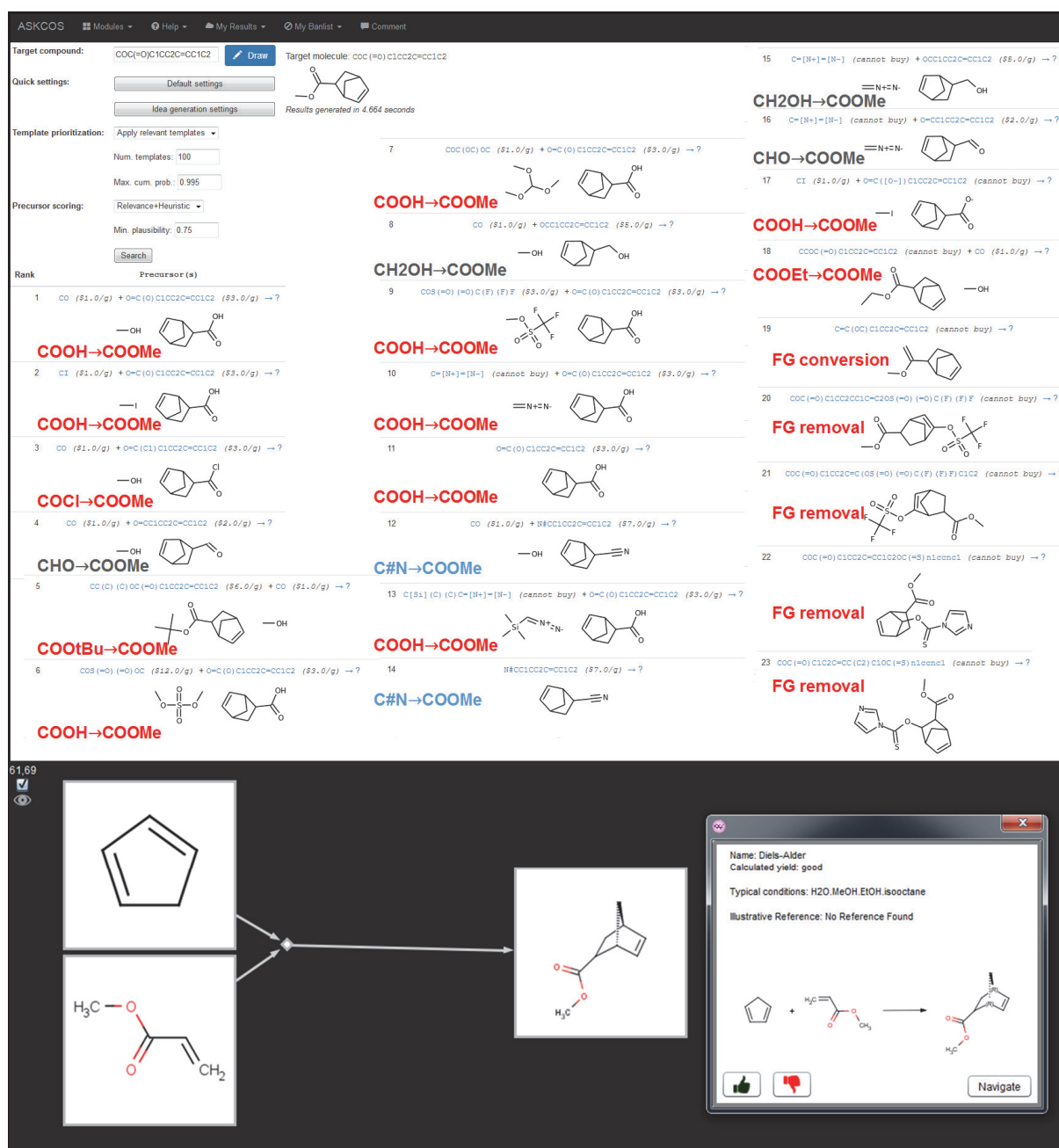
Rxn ID	Instance	Reaction							Entry Date
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other	
38633656	1 of 1		80.0	ammonium chloride	(none)	(none)	-40.0	unk	2014/11/08

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccooley@mit.edu](mailto:ccooley@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

1738106, 1758144, 1989625, 1994179, 2017081, 2098898, 2123051, 2208567, 2209646, 2259411, 2259412, 2259413, 2259415, 3421629, 4879502, 4879506, 5196777, 5196961, 5199500, 5199510, 8679608, 8679643, 8680230, 8681668, 8682007, 8682129, 8682383, 8755391, 8939936, 8945593, 8945596, 8945598, 8945602, 8945610, 8945628, 9183201, 9193082, 9193190, 9396193, 9589325, 9556268, 10041751, 10292193, 10294317, 10295018, 10297915, 10637305, 11285447, 11289566, 23751012, 25760364, 25763630, 25763632, 25944429, 27774647, 27904686, 27913769, 27966783, 28403372, 28403382, 28417654, 28417656, 28616938, 28624936, 28720243, 28720244, 28848961, 28959473, 29059363, 29077303, 29184391, 29184400, 29261665, 29268423, 29269426, 29290259, 29540067, 30199687, 30425522, 30425525, 30425526, 30983862, 31061238, 32702868, 32939036, 33168897, 33278706, 33278708, 33278709, 33278711, 33278715, 33278716, 33278772, 33420586, 33533869, 33850190, 33978844, 34899904, 35927625, 35927629, 35945001, 35945002, 35945003, 35945004, 36020129, 36402101, 36455507, 36455515, 37773937, 38300304, 38440058, 38501745, 38693656, 38754764, 39386238, 39915813, 40692349, 42779103, 42779104, 42779110, 42979929, 43323759, 43323769, 43323812, 43323845, 43323846, 43323876, 43384667, 43438289, 43588578, 43749789, 43770023, 44549236, 44757896

**Figure S3.** Automatically extracted core of the Bartoli indole synthesis from the ASKCOS software. The necessary *ortho*-substitution is not taken into account.



**Figure S4.** Retrosynthetic planning of methyl bicyclo[2.2.1]hept-5-ene-2-carboxylate in MIT's ASKCOS system (<http://askcos.mit.edu/retro/>). As a reference, this compound can be prepared in *one* step from commercially available methyl acrylate and cyclopentadiene in the must-know, synthetically powerful Diels-Alder cycloaddition – this approach is easily found by Chematica (screenshot of Chematica's top-scoring solution shown at the bottom, on black background) but was not suggested by the neural network. Instead, several disconnections of the simple methyl ester (hydrolyses, transesterifications and modifications of oxidation states) dominated the top-20 results. Additionally, nonproductive removals of functional groups were suggested. All ASKCOS-proposed results for this target are shown.

ASKCOS Modules Help My Results My Banlist Comment

Target compound: COC(=O)N1CCN(C2=CC(=O)C=C2)CC1 Target molecule: COC(=O)N1CCN(C2=CC(=O)C=C2)CC1

Quick settings: Default settings Idea generation settings

Template prioritization: Apply relevant templates Results generated in 0.691 seconds

Num. templates: 100 Max. cum. prob.: 0.995

Precursor scoring: Relevance+Heuristic Min. plausibility: 0.75 Search

Rank Precursor(s)

- Carbamoylation**  
COC(=O)Cl (\$1.0/g) + Ic1ccc(N2CCNCC2)cc1 (\$26.0/g) → ?
- Aryl incompatible**  
Hc1ccc(Br)cc1 (\$1.0/g) + COC(=O)N1CCNCC1 (\$10.0/g) → ?
- Aryl incompatible**  
COC(=O)N1CCNCC1 (\$10.0/g) + Clc1ccc(I)cc1 (\$1.0/g) → ?
- Aryl incompatible**  
COC(=O)N1CCNCC1 (\$10.0/g) + Ic1ccc(I)cc1 (\$1.0/g) → ?
- Low reactivity in S<sub>N</sub>Ar**  
COC(=O)N1CCNCC1 (\$10.0/g) + Fc1ccc(I)cc1 (\$1.0/g) → ?
- Selectivity issues**  
COC(=O)N1CCNCC1 (\$10.0/g) + OB(O)c1ccc(I)cc1 (\$9.0/g) → ?
- FG conversion**  
COC(=O)N1CCN(C2=CC(=O)C=C2)CC1 (\$42.0/g) → ?
- Redundant with #1, FG removal**  
COC(=O)Cl (\$1.0/g) + Ic1ccc(N2CCNCC2)CC1 (cannot buy) → ?
- Iodination**  
COC(=O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- FG conversion**  
H2Nc1ccc(N2CCNCC2)cc1 (cannot buy) + COC(=O)Cl (\$1.0/g) → ?
- Redundant with #1, FG removal**  
COC(=O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- FG removal**  
COC(=O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- FG removal**  
COC(=O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- Redundant with #1**  
COC(=O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- Redundant with #1**  
COC(=O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- Redundant with #1**  
COC(=O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- Unstable substrate**  
Cl (\$1.0/g) + O=C(O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- Unstable substrate**  
Cl (\$1.0/g) + O=C(O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- Selectivity issues**  
C(=N)=N (cannot buy) + O=C(O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- Unstable substrate**  
C(=N)=N (cannot buy) + O=C(O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?
- Unstable substrate**  
O=C(O)N1CCN(C2=CC(=O)C=C2)CC1 (cannot buy) → ?

Template 59c51b9e05581eb9f57c29ea  
Template: C#N2D1+O+1-[O]R2D2+O+4-[C]3=[O]D1;R2+2>>C-[S]1(-C)-[CH]D2;+O+1=[N]=N-1.[O]D1;R2+2=[C]3-[OR]D1+O+4  
4056 total references

Rxn ID Instance Yield [%] Reagent(s) Catalyst(s) Solvent(s) Temp. [C] Time [h] Other Entry Date

10034269 1 of 3 100.0 (none) (none) methanol, diethyl ether, and hexane 20.0 3.0 2007/12/16

ASKCOS 163.00 163.26

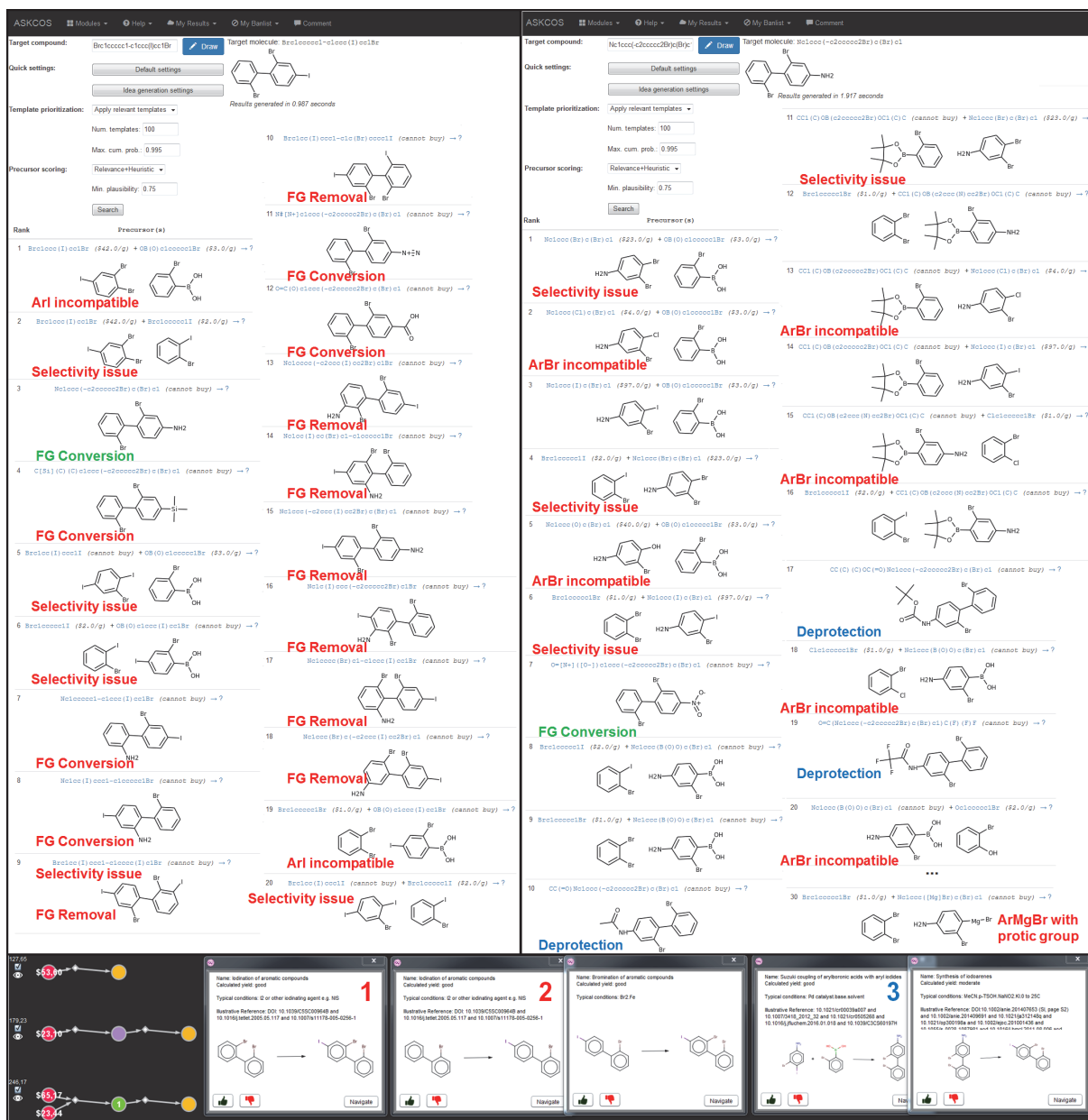
ASKCOS 163.00 163.26

ASKCOS 163.00 163.26

ASKCOS 163.00 163.26

**Figure S5.** Retrosynthetic planning of methyl 4-(4-iodo-phenyl) piperazine-1-carboxylate in MIT's ASKCOS system. As a reference, the synthesis of this class of compounds is usually finalized by the formation of carbamate or electrophilic aromatic iodination – this approaches are easily found by Chematica (screenshots of the top-scoring solutions at the bottom). Although the formation of carbamate was the top ASKCOS' prediction, the electrophilic aromatic iodination was only its ninth suggestion. In addition, several other suggestions from top20 results raise reactivity concerns. For instance, Buchwald-type amination of an aryl chloride or bromide (reactions ranked #2 and #3) is virtually impossible in the presence of a more reactive aryl iodide (see main text, Figure 6).

Analogous reaction of an aryl fluoride ranked as #5 (occurring via  $S_NAr$  mechanism) has marginal chances to succeed due to lack of electron withdrawing groups activating the ring towards the nucleophilic substitution. Additionally, reactions #6 and #21 raise chemoselectivity concerns due to the presence of an aryl iodide. Moreover, ASKCOS proposed (#19,#20,#22,#23) unstable and decarboxylation-prone carbamic acid as a substrate. Inspection of the reaction template responsible for the outcome shown in #19 clearly demonstrates that automatically generated reaction rule cannot distinguish groups (here, unstable carbamic and stable carboxylic acids) of quite different reactivities. All returned results are shown.



**Figure S6.** Attempted multistep retrosynthetic planning of 2,2'-dibromo-iodobiphenyl. This class of polyhalogenated building blocks is usually prepared via iodination, Suzuki coupling, and Sandmeyer reaction. ASKCOS suggestions for this compound rise several serious selectivity and cross-reactivity concerns. For example, Suzuki coupling (#1, #19) with aryl bromide will not proceed if a more reactive aryl iodide is present (see main text, Figure 6). Additionally, an aryl halide cross coupling shown in (#2, #20), Suzuki coupling (#5,#6), and selective removal of one of the iodides (#9 and #10) will lead to a mixture of products (due to a presence of multiple nonequivalent groups). As in previous examples, multiple FG removals (here: deaminations) were also present in top20 results. Expanding one of the proposed intermediates – a substrate for Sandmeyer reaction – leads to the set of second-step suggestions (right panel) often suffering from the abovementioned selectivity and reactivity concerns: arylation of an aryl chloride

(#2,#13,#15) or phenol (#5,#20) is proposed in the presence of an incompatible aryl bromide while cross coupling of aryl halides (#4) or Suzuki coupling with dibromoaniline (#1,#11) will lead to a mixture of products. Finally, in #30 arylmagnesium reagent with protic NH<sub>2</sub> group is used as a substrate while these functional groups simply cannot exist together. Left panel: top 20 out of 28 returned results are shown, right panel: top 20 and #30 out of 41 returned results are shown. Full synthetic pathways generated autonomously by Chematica (top 3 shown in the bottom part) rely either on the electrophilic aromatic halogenations (#1, #2) or take advantage of Suzuki coupling and Sandmeyer reaction (#3). This is one example we found where some of Chematica's suggestions raise concerns: namely aromatic substitutions of substrates suggested in #1 and the second step of #2 are known to yield mainly regioisomeric products (here: 2,2'-dibromo-5-iodobiphenyl and 2,5-dibromo-4'-iodobiphenyl). This inaccuracy is due to the Hammett/Hueckel/Proton-affinity based filter evaluating aromatic substitutions (90% correct, the example here is obviously in the 10% of erroneous predictions). However, the third solution returned by Chematica does not rise any reactivity and selectivity concerns.

ASKOOS Modules Help My Results My Basket Comment

Target compound: Nc1ccc(N2CCOCC2)c(F)c1 Draw Target molecule: Nc1ccc(N2CCOCC2)c(F)c1

Quick settings: Default settings Idea generation settings

Template prioritization: Apply relevant templates Num. templates: 100 Max. cum. prob.: 0.995 Relevance+Heuristic Min. plausibility: 0.75 Search

Precursor scoring: Search

Rank Precursor(s)

1 O=[N+]([O-])c1ccc(N2CCOCC2)c(F)c1 (\$12.0/g) → ?  
**FG Conversion**

2 C1COCCN1 (\$1.0/g) + Nc1ccc(Br)c(F)c1 (\$2.0/g) → ?  
**Selectivity issues**

3 C1COCCN1 (\$1.0/g) + O=[N+]([O-])c1ccc(F)c(F)c1 (\$1.0/g) → ?  
**2-steps**

4 C1COCCN1 (\$1.0/g) + Nc1ccc(F)c(F)c1 (\$2.0/g) → ?  
**Selectivity issues**

5 C1COCCN1 (\$1.0/g) + Nc1ccc(Cl)c(F)c1 (\$5.0/g) → ?

6 C1COCCN1 (\$1.0/g) + Nc1ccc(I)c(F)c1 (\$23.0/g) → ?

7 O=C(Nc1ccc(N2CCOCC2)c(F)c1)Oc1ccccc1 (\$10.0/g) → ?  
**Deprotection**

8 Fc1ccc(Br)cc1N1CCOCC1 (\$92.0/g) → ?  
**FG Conversion**

9 O=C(O)c1ccc(N2CCOCC2)c(F)c1 (\$45.0/g) → ?  
**FG Conversion**

10 BrCCOCCBr (\$2.0/g) + Nc1ccc(N)c(F)c1 (cannot buy) → ?  
**Selectivity issues**

11 C1COCCCl1 (\$1.0/g) + Nc1ccc(N)c(F)c1 (cannot buy) → ?  
**Selectivity issues**

12 CC(=O)Nc1ccc(N2CCOCC2)c(F)c1 (cannot buy) → ?  
**Deprotection**

13 O=Nc1ccc(N2CCOCC2)c(F)c1 (cannot buy) → ?  
**FG Conversion**

14 CC(C)(C)[Si](C)(C)Nc1ccc(N2CCOCC2)c(F)c1 (cannot buy) → ?  
**Deprotection**

15 O=C(Nc1ccc(N2CCOCC2)c(F)c1)C(F)F (cannot buy) → ?  
**Deprotection**

16 CC(C)(C)OC(=O)Nc1ccc(N2CCOCC2)c(F)c1 (cannot buy) → ?  
**Deprotection**

17 Fc1ccc(Cl)cc1N1CCOCC1 (cannot buy) → ?  
**FG Conversion**

18 O=C(Nc1ccc(N2CCOCC2)c(F)c1)Cl (cannot buy) → ?  
**Deprotection**

19 Fc1ccc(N(Cc2ccccc2)C(=O)OCC1)cc1N1CCOCC1 (cannot buy) → ?  
**Deprotection**

20 Fc1ccc(N1CCOCC1)cc1N1CCOCC1 (cannot buy) → ?  
**2-steps**

21 Nc1ccc(F)c(N2CCOCC2)cc1Br (cannot buy) → ?  
**FG removal**

22 Nc1ccc(F)c(N2CCOCC2)c(Br)c1 (cannot buy) → ?  
**FG removal**

23 Nc1ccc(N2CCOCC2)c(F)c1Br (cannot buy) → ?  
**FG removal**

24 Nc1ccc(N2CCOCC2)c(F)c(Cl)c1 (cannot buy) → ?  
**FG removal**

25 Nc1ccc(F)c(N2CCOCC2)cc1Cl (cannot buy) → ?  
**FG removal**

26 Nc1ccc(F)c(N2CCOCC2)c(Cl)c1 (cannot buy) → ?  
**FG removal**

27 [N-]=[N+]=Nc1ccc(N2CCOCC2)c(F)c1 (cannot buy) → ?  
**FG Conversion**

28 Fc1ccc(N=C(c2ccccc2)c2ccccc2)cc1N1CCOCC1 (cannot buy) → ?  
**Deprotection**

29 Nc1ccc([N+]2([O-])CCOCC2)c(F)c1 (cannot buy) → ?  
**FG Conversion**

30 Nc1ccc(N2CCOCC2=O)c(F)c1 (cannot buy) → ?  
**FG Conversion**

31 Nc1ccc(F)c(N2CCOCC2)c(F)c1 (cannot buy) → ?  
**FG removal**

32 Fc1ccc(NC(=O)OCC1)cc1N1CCOCC1 (cannot buy) → ?  
**Deprotection**

33 C[Si](C)(C)N(c1ccc(N2CCOCC2)c(F)c1)[Si](C)(C)C (cannot buy) → ?  
**Deprotection**

34 CC(C)(C)OC(=O)N(C(=O)OC(C)(C)C)c1ccc(N2CCOCC2)c(F)c1 (cannot buy) → ?  
**Deprotection**

35 O=[N+]([O-])c1ccc(F)c(N2CCOCC2)c(Br)c1 (cannot buy) → ?  
**FG removal**

Reaction 1: Reduction of nitro group  
 Name: Reduction of nitro group  
 Calculated yield: good  
 Typical conditions: Zn, eq. H4, EtOH (2h, HCl)  
 Illustrative Reference: DOI: 10.1002/anie.201112005 and 10.1002/anie.20114681 and 10.2396/molecules.19022055 and 10.1021/ab003344 (page 3) and an eReference with a title

Reaction 2: Cleavage of benzoyloxycarbamates  
 Name: Cleavage of benzoyloxycarbamates  
 Calculated yield: moderate  
 Typical conditions: HBr, AcOH  
 Illustrative Reference: W020447248 o 79 and W020445670 p 128 and US201078209 s 64 and 10.3987/COOL-16-13336

Reaction 3: Annulation of aryl iodides  
 Name: Annulation of aryl iodides  
 Calculated yield: good  
 Typical conditions: PdG or CuI base solvent  
 Illustrative Reference: 10.1016/j.t.2013.02.040 AND 10.1021/ab000919 AND 10.1021/ab000919

Reaction 4: Annulation of aryl bromides  
 Name: Annulation of aryl bromides  
 Calculated yield: good  
 Typical conditions: Pd ligand base or CuI ligand base  
 Illustrative Reference: 10.1021/ab000944 AND 10.1021/ab000944 AND 10.1021/ab000944 AND 10.1021/ab000944

Reaction 5: Annulation of aryl bromides  
 Name: Annulation of aryl bromides  
 Calculated yield: good  
 Typical conditions: Pd ligand base or CuI ligand base  
 Illustrative Reference: 10.1021/ab000944 AND 10.1021/ab000944 AND 10.1021/ab000944 AND 10.1021/ab000944

Reaction 6: Coupling of Armona with Aryl Halides  
 Name: Coupling of Armona with Aryl Halides  
 Calculated yield: good  
 Typical conditions: Pd(PPh3)4/SiO2 Na2O2 dioxane heat  
 Illustrative Reference: DOI: 10.1021/ab000944

Reaction 7: Cleavage of benzoyloxycarbamates  
 Name: Cleavage of benzoyloxycarbamates  
 Calculated yield: moderate  
 Typical conditions: HBr, AcOH  
 Illustrative Reference: W020447248 o 79 and W020445670 p 128 and US201078209 s 64 and 10.3987/COOL-16-13336

Reaction 8: Annulation of aryl iodides  
 Name: Annulation of aryl iodides  
 Calculated yield: good  
 Typical conditions: PdG or CuI base solvent  
 Illustrative Reference: 10.1016/j.t.2013.02.040 AND 10.1021/ab000919 AND 10.1021/ab000919

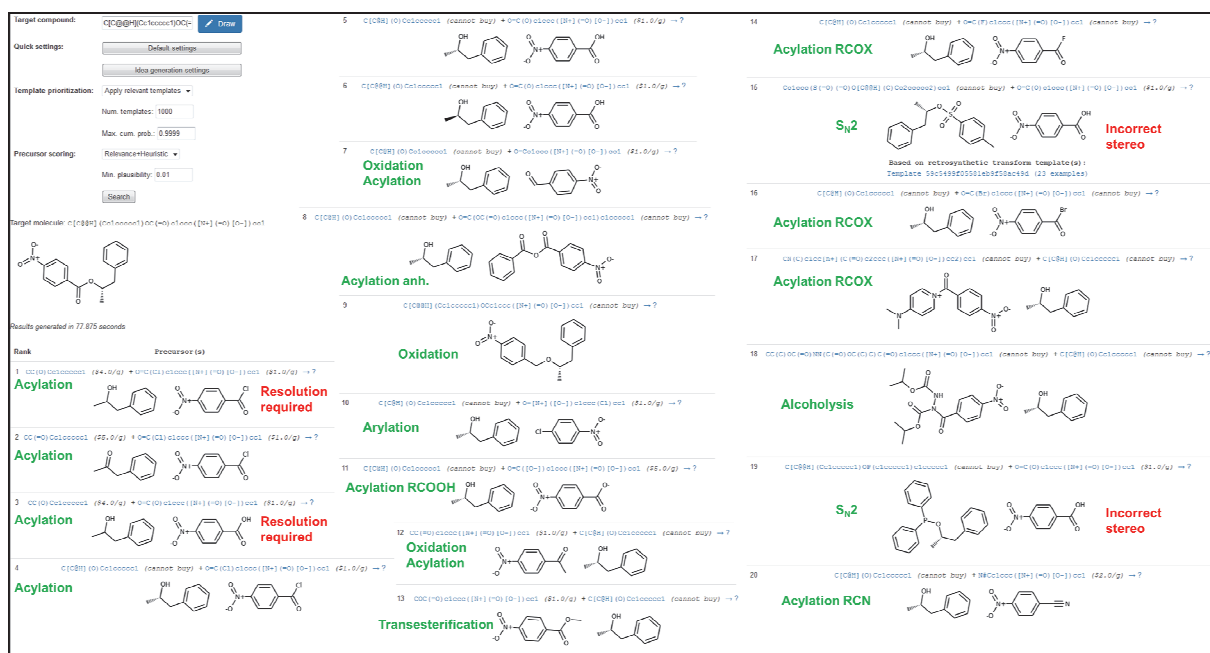
Reaction 9: Annulation of aryl bromides  
 Name: Annulation of aryl bromides  
 Calculated yield: good  
 Typical conditions: Pd ligand base or CuI ligand base  
 Illustrative Reference: 10.1021/ab000944 AND 10.1021/ab000944 AND 10.1021/ab000944 AND 10.1021/ab000944

Reaction 10: Annulation of aryl bromides  
 Name: Annulation of aryl bromides  
 Calculated yield: good  
 Typical conditions: Pd ligand base or CuI ligand base  
 Illustrative Reference: 10.1021/ab000944 AND 10.1021/ab000944 AND 10.1021/ab000944 AND 10.1021/ab000944

Reaction 11: Coupling of Armona with Aryl Halides  
 Name: Coupling of Armona with Aryl Halides  
 Calculated yield: good  
 Typical conditions: Pd(PPh3)4/SiO2 Na2O2 dioxane heat  
 Illustrative Reference: DOI: 10.1021/ab000944

**Figure S7.** Retrosynthetic planning of Linezolid's intermediate in MIT's ASKCOS system. The target is of interest since it is found in many patented routes we describe in <sup>69</sup>. Several of ASKCOS' suggestions for this target rise serious concerns of low selectivity and/or insufficient chemical reactivity. Selective functionalization of unsymmetrical diamine (#10,#11) has never been reported and will deliver a mixture of products. Presence of the electron-donating amine group in #4 deactivating the ring towards S<sub>N</sub>Ar reaction is highly problematic and renders this transformation unfeasible to perform – even more reactive 3,4,5-trifluoroaniline required heating to 180 °C for 36 h in neat morpholine to give the product in moderate yield (*Eur. J. Org. Chem.* **2012**, 7048–7052). Selective substitution of difluoroaniline (#4) is also a currently an unknown process – in fact, the attempted reaction with alcohol (*Eur. J. Org. Chem.* **2012**, 7048–7052) led to a 1:1 mixture of products. As in previous examples, several (19/35) unproductive removals of protecting groups (blue) or halogen atoms (grey) were present in top35 results. Additionally, #3 (substitution/reduction) and #20 (nitration/reduction) are overlapping with the top prediction. Top 35 out of 39 returned results are shown. In contrast, synthetic pathways returned by Chematica (bottom) do not suffer from any chemical reactivity issues.





**Figure S8.** Retrosynthetic planning of a chiral acylated alcohol in the ASKCOS system. Displacement of mesylate (#15) or alkoxydiphenylphosphine (#19) with carboxylic acid occurs with inversion of configuration via S<sub>N</sub>2 mechanism. However, the reaction templates for these reactions generate substrates corresponding to a chemically incorrect stereoretentive process. See **Figure S9** for details of ASKCOS' template for #15.

#### Template 59c5499f05581eb9f58ac49d

Template: [C:2]-[O:R0:D2+0:1]-[C:R0:D2+0:1]([O:D1:R0+4])-[C:5]>>C-c1ccccc1-c(O)(=O)O.[C:2]:[O:R0:D2+0:1]-[C:2]:[O:R0:D2+0:1]([O:D1:R0+4])-[C:5]  
 23 total references  
 Export Reaxys query for precedents  
 [C:2][O:R0:D2+0:1][C:R0:D2+0:3][O:D1:R0:4][C:5]>>C-c1ccccc1-c(O)(=O)O.[C:2]:[O:R0:D2+0:1][C:2]:[O:R0:D2+0:3][O:D1:R0:4][C:5]

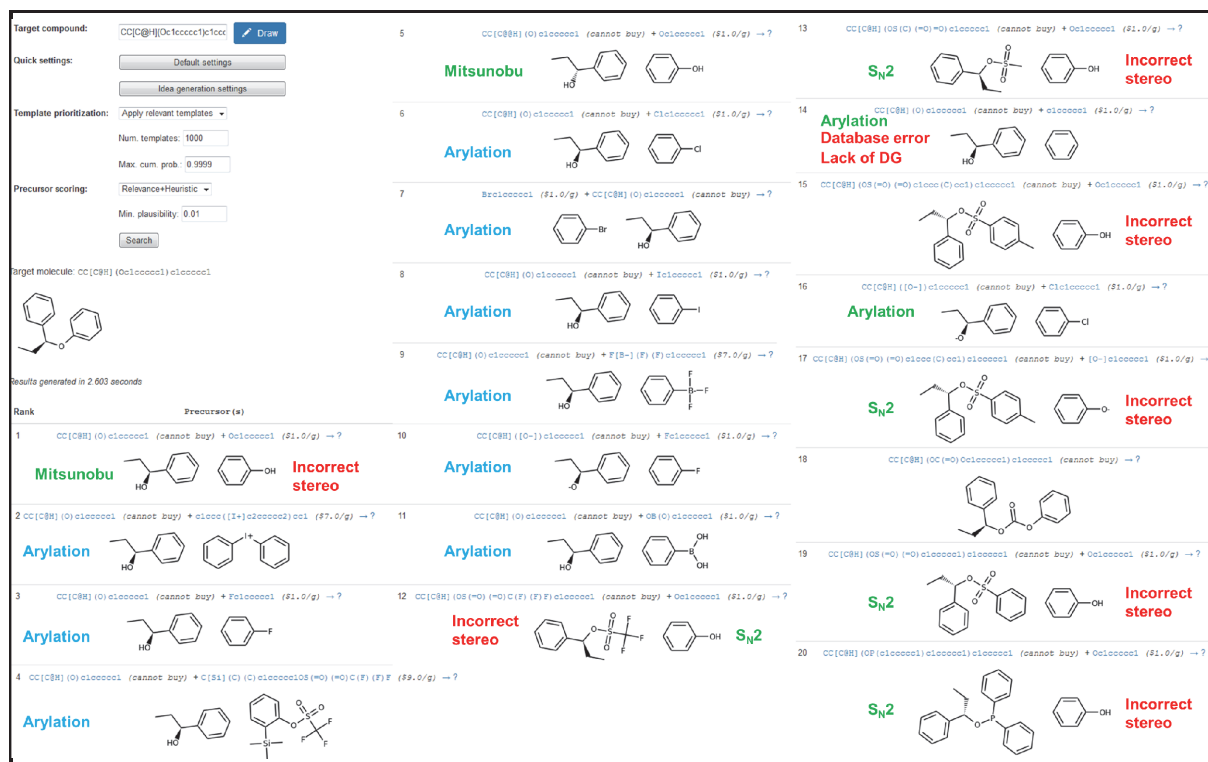
Rxn ID	Instance	Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [°C]	Time [h]	Other	Entry Date
9903675	1 of 2	95.0	 -R[O] R[GT] [P<sub>sub</sub>=65614</sub>] [N<sub>sub</sub>=2</sub>] -R[O] R[GT]- and N-ethyl-N,N-disopropylamine	(none)	(none)	80.0	unk		2007/12/14

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact ccoley@mit.edu

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

9903675, 9903682, 9903685, 10292150, 10436788, 23149443, 24287194, 24580041, 28327466, 29321312, 29321323, 29321333, 34239389, 40419151, 40419281, 43506408, 44216537, 44216538, 44216539

**Figure S9.** Automatically extracted core for the substitution of a mesylate with a carboxylic acid ignores inversion of configuration and generates substrates with improper stereochemistry. The same template is not limited to primary and secondary mesylates and allows for substrates bearing hardly reactive, tertiary mesylates.



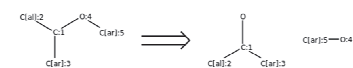
**Figure S10.** Retrosynthetic planning of a chiral aryl ether in MIT's ASKCOS system. Several suggestions relying on  $S_N2$  displacements (#1, #12, #13, #15, #17, #19 and #20) use substrates corresponding to chemically incorrect stereoretentive process. Additionally, the oxidative alkoxylation proposed as #14 is feasible only when appropriate metal-coordinating group is present in the substrate to facilitate C-H activation. The reaction template for this transformation (see **Figure S12**) does not take into account this requirement.

## Template 59c51ace05581eb9f57ba279

Template: [c@H]1[C@@H](O)C(C)(C)C(=O)N1C >> [c@H]1[C@@H](O)C(C)(C)C(=O)N1C

21 total references

Export Reaxys query for precedents



Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Reaction ID	Instance	Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other	Entry Date
9778947	1 of 1	92.0	triphenylphosphine and 1,1'-azobiscarbonyl-diipiperidine	(none)	benzene	20.0	12.0		2007/12/14

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccooly@mit.edu](mailto:ccooly@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

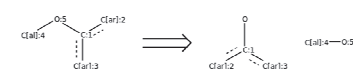
1742856, 1983371, 1983372, 1983373, 1983374, 2054500, 2054501, 2054502, 3829844, 3829845, 3829846, 4822003, 9778950, 9778947, 28193355, 30759700, 30759704

## Template 59c511c105581eb9f5756c65

Template: [C@@H]1[C@H](O)C(C)(C)C(=O)N1C >> [C@@H]1[C@H](O)C(C)(C)C(=O)N1C

7054 total references

Export Reaxys query for precedents



Reaction ID	Instance	Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other	Entry Date
10345251	1 of 1	100.0	triphenylphosphine and diethylazodicarboxylate	(none)	tetrahydrofuran and toluene	20.0	unk		2007/12/15

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccooly@mit.edu](mailto:ccooly@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

51706, 14231, 2599171, 271092, 235127, 2951320, 2951619, 2951645, 3189191, 3911490, 3910100, 3924371, 3953651, 4122868, 4124221, 4124610, 4134410, 4134411, 4134412, 4134413, 4516431, 4662609, 4667407, 4986311, 5096671, 5115828, 5115835, 5211201, 8642488, 8746174, 8765389, 8998410, 9171953, 9175099, 9176091, 9177380, 9207888, 9259878, 927168, 927968, 9295878, 9295879, 9295880, 9334877, 9475295, 9573845, 9580814, 9645188, 9647720, 9663814, 9680696, 9678977, 972926, 976968, 9776928, 9776929, 9776930, 9776931, 9780514, 9782013, 9782014, 9782015, 9782016, 9782017, 9782018, 9782019, 9782021, 9782022, 9847476, 9875396, 9875397, 9875398, 9875399, 9875400, 9893813, 9933945, 10054807, 10184209, 10182037, 10239098, 10241311, 10279517, 10284521, 1031298, 10345211, 1039161, 1039214, 1039285, 1039286, 1039287, 1039288, 1042614, 1042615, 1042616, 1042617, 1042618, 1042619, 1079248, 1079249, 1084618, 11206314, 11237112, 11237116, 11238644, 11241429, 11255388, 11255390, 12848096, 12848148, 12961946, 12962006, 12962010, 12962014, 12962018, 14295616, 14295617, 14295618, 14295619, 14295620, 14295621, 14295622, 14295623, 14295624, 14295625, 14295626, 14295627, 14295628, 14295629, 14295630, 14295631, 14295632, 14295633, 14295634, 14295635, 14295636, 14295637, 14295638, 14295639, 14295640, 14295641, 14295642, 14295643, 14295644, 14295645, 14295646, 14295647, 14295648, 14295649, 14295650, 14295651, 14295652, 14295653, 14295654, 14295655, 14295656, 14295657, 14295658, 14295659, 14295660, 14295661, 14295662, 14295663, 14295664, 14295665, 14295666, 14295667, 14295668, 14295669, 14295670, 14295671, 14295672, 14295673, 14295674, 14295675, 14295676, 14295677, 14295678, 14295679, 14295680, 14295681, 14295682, 14295683, 14295684, 14295685, 14295686, 14295687, 14295688, 14295689, 14295690, 14295691, 14295692, 14295693, 14295694, 14295695, 14295696, 14295697, 14295698, 14295699, 14295700, 14295701, 14295702, 14295703, 14295704, 14295705, 14295706, 14295707, 14295708, 14295709, 14295710, 14295711, 14295712, 14295713, 14295714, 14295715, 14295716, 14295717, 14295718, 14295719, 14295720, 14295721, 14295722, 14295723, 14295724, 14295725, 14295726, 14295727, 14295728, 14295729, 14295730, 14295731, 14295732, 14295733, 14295734, 14295735, 14295736, 14295737, 14295738, 14295739, 14295740, 14295741, 14295742, 14295743, 14295744, 14295745, 14295746, 14295747, 14295748, 14295749, 14295750, 14295751, 14295752, 14295753, 14295754, 14295755, 14295756, 14295757, 14295758, 14295759, 14295760, 14295761, 14295762, 14295763, 14295764, 14295765, 14295766, 14295767, 14295768, 14295769, 14295770, 14295771, 14295772, 14295773, 14295774, 14295775, 14295776, 14295777, 14295778, 14295779, 14295780, 14295781, 14295782, 14295783, 14295784, 14295785, 14295786, 14295787, 14295788, 14295789, 14295790, 14295791, 14295792, 14295793, 14295794, 14295795, 14295796, 14295797, 14295798, 14295799, 14295800, 14295801, 14295802, 14295803, 14295804, 14295805, 14295806, 14295807, 14295808, 14295809, 14295810, 14295811, 14295812, 14295813, 14295814, 14295815, 14295816, 14295817, 14295818, 14295819, 14295820, 14295821, 14295822, 14295823, 14295824, 14295825, 14295826, 14295827, 14295828, 14295829, 14295830, 14295831, 14295832, 14295833, 14295834, 14295835, 14295836, 14295837, 14295838, 14295839, 14295840, 14295841, 14295842, 14295843, 14295844, 14295845, 14295846, 14295847, 14295848, 14295849, 14295850, 14295851, 14295852, 14295853, 14295854, 14295855, 14295856, 14295857, 14295858, 14295859, 14295860, 14295861, 14295862, 14295863, 14295864, 14295865, 14295866, 14295867, 14295868, 14295869, 14295870, 14295871, 14295872, 14295873, 14295874, 14295875, 14295876, 14295877, 14295878, 14295879, 14295880, 14295881, 14295882, 14295883, 14295884, 14295885, 14295886, 14295887, 14295888, 14295889, 14295890, 14295891, 14295892, 14295893, 14295894, 14295895, 14295896, 14295897, 14295898, 14295899, 14295900, 14295901, 14295902, 14295903, 14295904, 14295905, 14295906, 14295907, 14295908, 14295909, 14295910, 14295911, 14295912, 14295913, 14295914, 14295915, 14295916, 14295917, 14295918, 14295919, 14295920, 14295921, 14295922, 14295923, 14295924, 14295925, 14295926, 14295927, 14295928, 14295929, 14295930, 14295931, 14295932, 14295933, 14295934, 14295935, 14295936, 14295937, 14295938, 14295939, 14295940, 14295941, 14295942, 14295943, 14295944, 14295945, 14295946, 14295947, 14295948, 14295949, 14295950, 14295951, 14295952, 14295953, 14295954, 14295955, 14295956, 14295957, 14295958, 14295959, 14295960, 14295961, 14295962, 14295963, 14295964, 14295965, 14295966, 14295967, 14295968, 14295969, 14295970, 14295971, 14295972, 14295973, 14295974, 14295975, 14295976, 14295977, 14295978, 14295979, 14295980, 14295981, 14295982, 14295983, 14295984, 14295985, 14295986, 14295987, 14295988, 14295989, 14295990, 14295991, 14295992, 14295993, 14295994, 14295995, 14295996, 14295997, 14295998, 14295999, 14300000

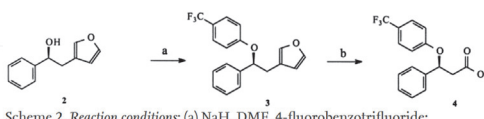
**Figure S11.** Automatically extracted templates for Mitsunobu displacement with a phenol nucleophile in the ASKCOS software. Top: This reaction occurs with inversion of configuration. Stereoretentive process is feasible only when anchimeric assistance (here, from a tertiary amine present in the specific literature precedent but missing in the extracted reaction transform) is possible. Bottom: The reaction template extracted from a large number of examples (>7000) also does not account for inversion of configuration for secondary alcohols.

# Template 59c518c205581eb9f57a55fb

Template: [C:1]~([O]H;D2;+O:2)~([O]H;D2;+O:4) (:[O:3]) : [O:5]>>[C:1]~([O]H;D1;+O:2).[O:3] : [OH;D2;+O:4] : [O:5]

188 total references

Export Reaxys query for precedents



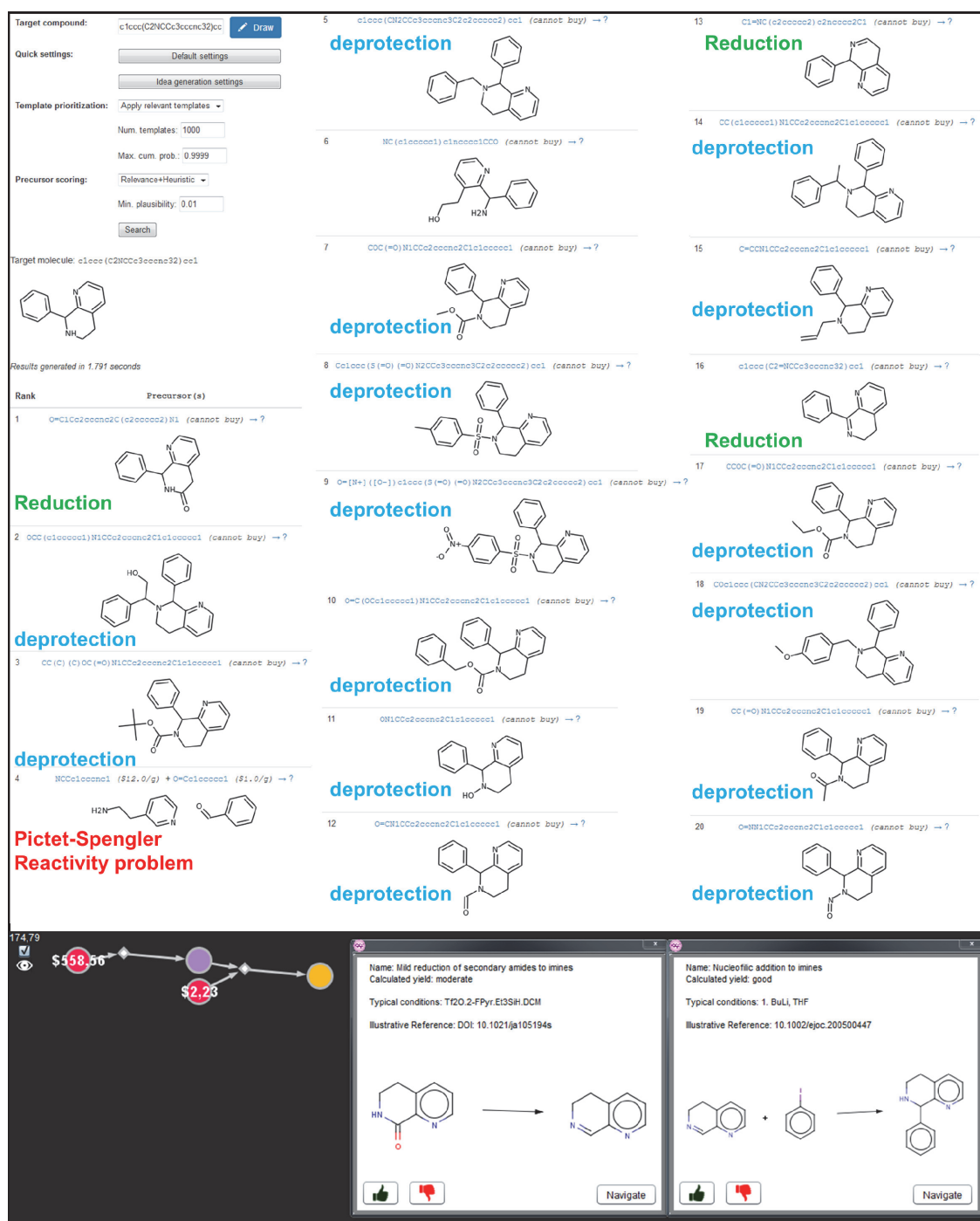
Rxn ID	Instance	Reaction							Entry Date
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other	
8948496	1 of 1		92.0	sodium hydride	(none)	N,N-dimethyl acetamide and N,N-dimethyl acetamide	50 - 55	0.75	2007/12/09

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [cocoley@mit.edu](mailto:cocoley@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

1441513, 2938196, 8940496, 8940496, 22823762, 26905353, 30456426, 30456448, 30456454, 30832519, 30940064, 32052956, 33169618, 33169619, 33169620, 34105606, 34105608, 34105609, 34105613, 34254344, 34254349, 34254355, 34254356, 34254357, 34254360, 34254363, 34254366, 34254369, 34254370, 34254372, 34254376, 34254381, 34254382, 34254383, 34254384, 34254386, 34254387, 34442272, 34442275, 34442278, 34442279, 34442281, 34442291, 35422213, 36196876, 36200019, 36200020, 36200021, 36200022, 36200024, 36200026, 36200028, 36200029, 36200030, 36200031, 36200032, 36200034, 36200036, 36200038, 36200039, 36200040, 36200042, 36200044, 36200055, 36424152, 36481952, 36481953, 36481954, 36481955, 36481956, 36481957, 36481959, 36481960, 36481961, 36481962, 36481963, 36481965, 36481966, 36481967, 36481969, 36481970, 36481971, 36481972, 36481973, 38141340, 38141344, 38141346, 38141357, 38141359, 38141361, 38141362, 38141363, 38141364, 38141365, 38141367, 39761826, 39761827, 39761828, 39761829, 39761830, 39761831, 39761832, 39761833, 39906458, 39906459, 39906460, 39906463, 39906464, 39906465, 39906467, 39906468, 39906469, 39906471, 39906473, 39906474, 39906475, 39906477, 39906478, 39906511, 39906512, 39906514, 39906515, 39906517, 39906520, 39906522, 39906523, 39906526, 39906528, 39906531, 41153009, 41480222, 41480223, 41480226, 41480227, 41480230, 41480231, 41480234, 41480235, 41480236, 41480239, 41480241, 41480242, 41480243, 42748060, 42748066, 42748068, 42748069, 42748070, 42748071, 44077753, 44077754, 44077755, 44484223, 44484224, 44484225, 44484226, 44484227, 44484228, 44484229, 44484230, 44484231, 44484232, 44484233, 44484234, 44484235, 44484236, 44484237, 44484238, 44484239, 44749673

**Figure S12.** Automatically extracted core of oxidative alkoxylation in the ASKCOS software. The proposed process is feasible only if appropriate metal-coordinating group (usually pyridine, N-heterocycle, or amide) is present in the substrate. Additionally, the literature precedent used to support this reaction template (bottom) was erroneously deposited in the database and misses fluoride atom being substituted in  $S_NAr$  process in the original publication (inset in top-right part).



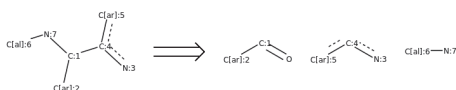
**Figure S13.** Retrosynthetic planning of tetrahydroisoquinoline in MIT's ASKCOS system. Several of the top-scoring solutions are limited to removals of different protecting groups or reductions of an amide (#1) or cyclic imines (#13,#16). The automatically extracted reaction template (see **Figure S14**) for Pictet-Spengler cyclisation (#4) is too general and lacks information regarding the type of the reacting aromatic system. This reaction is limited to electron-rich arenes and heteroarenes while ASKCOS allows for the annulation of electron-poor pyridine. Bottom: In contrast, Chematica's top-scoring solution commences with reduction of a commercially available amide to imine and subsequent addition of an organometallic reagent derived from iodobenzene.

## Template 59c5199f05581eb9f57ad74b

Template: [R7;A:3]:[C;R0;D3;+0:4]([C:5])-[CH;D3;+0:1]-[C:2]-[NH;D2;+0:7]-[C:6]>>[CH;D2;+0:1]-[C:2]-[R7;A:3]:[CH;D2;+0:4]:[C:5]-[NH2;D1;+0:7]

65 total references

Export Reaxys query for precedents



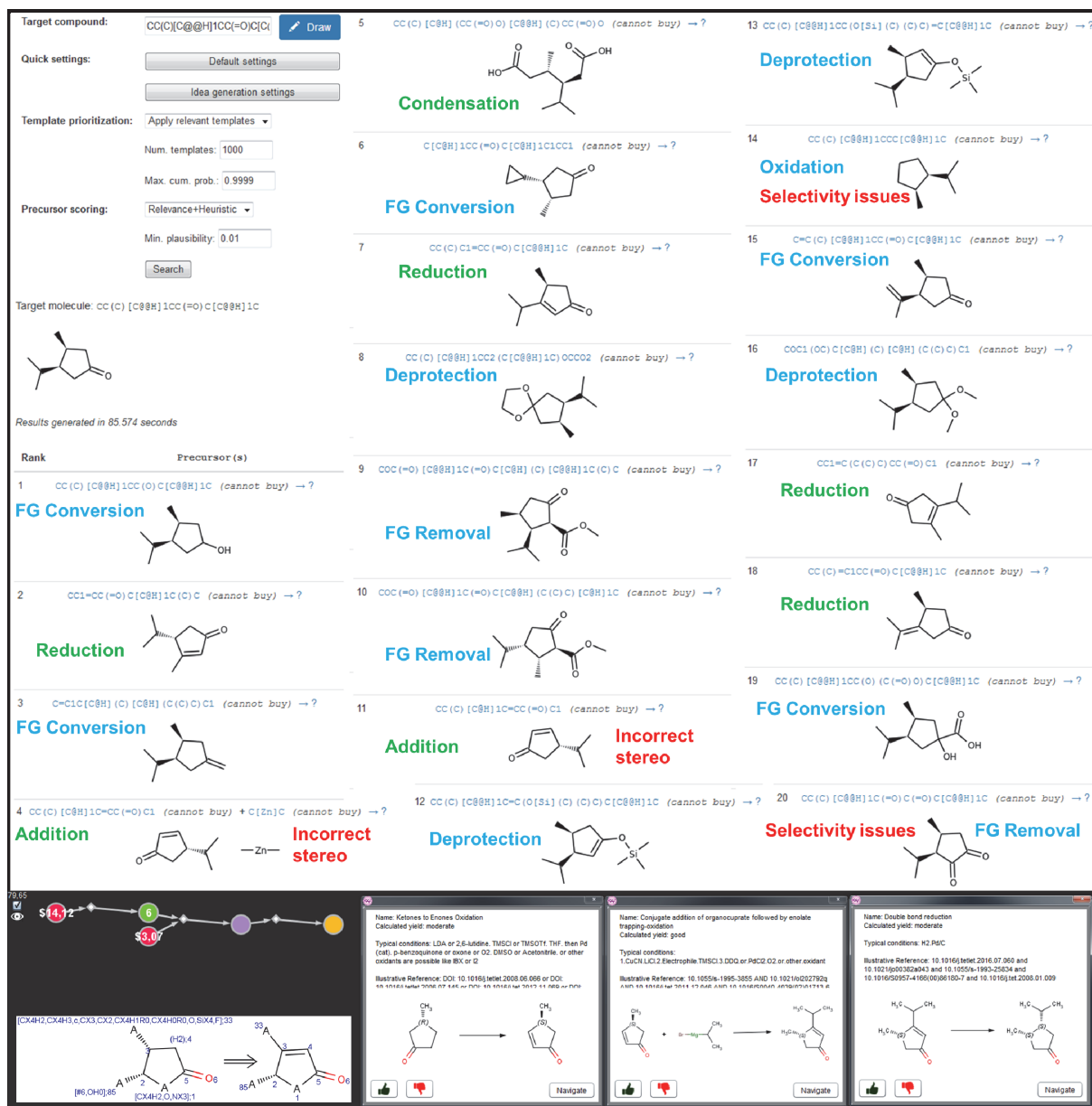
Rxn ID	Instance	Reaction						Entry Date
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	
38844682	1 of 1							2015/01/07
		99.0	acetic acid	(none)	(none)	80.0	2.0	

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccooley@mit.edu](mailto:ccooley@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

1564321, 1643674, 1777560, 2689786, 4071528, 4084182, 8929453, 8951554, 8959154, 8963266, 9241255, 9960153, 9960763, 9960862, 9961622, 9963717, 9964398, 8968243, 10136789, 10142064, 10414357, 11039429, 11039432, 23575211, 23826411, 23833324, 27821469, 27821470, 28549585, 32260235, 32260236, 33257272, 33257273, 33257307, 33257308, 33257309, 33779926, 33779943, 36736637, 36736641, 38844682, 38844683, 38994860, 41421347, 41421352, 41421408, 41421412, 41421413, 41578911, 42376735

**Figure S14.** Automatically extracted core of Pictet-Spengler cyclisation from the ASKCOS software incorrectly allowing for the annulation of electron poor heteroarenes (here, pyridine in example from **Figure S13**).



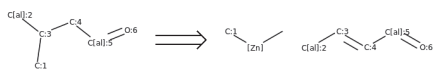
**Figure S15.** Retrosynthetic planning of chiral *cis* cyclopentanone in MIT's ASKCOS system. The reaction template (see **Figure S16**) for chiral conjugated addition of an organometallic reagent to an enone (#4) does not account for the presence of mismatched substituents and allows for chemically incorrect *syn*-selective process. None of the 14 reaction templates (see **Figures S18-S22**) for substrate-controlled reductions of alkenes (#2, #7, #17, #18) accounts for the necessary structural features controlling the reaction's outcome. Bottom: In Chematica's solution, the desired 1,2-*cis* cyclopentanone is constructed via addition of an organometallic reagent followed by trapping and oxidation of an enolate and substrate-controlled reduction of an alkene. In sharp contrast to automatically extracted rules (cf. **Figures S18-S22**), the expert-coded template (shown in the inset in the bottom-left part of the Figure) accounts for the presence of substituents dictating the stereoselective reaction's outcome.

## Template 59c53d1d05581eb9f5887799

Template: [C:2]-[CH:8;D2:+0:3]-[CH3;D1:+0:1]-[CH2;D2:+0:4]-[C:5]=[O;D1;R0:6]>>C-[Zn]-[CH3;D1:+0:1]-[C:2]-[CH;D2:+0:3]=[CH;D2:+0:4]-[C:5]=[O;D1;R0:6]

11 total references

Export Reaxys query for precedents



Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Rxn ID	Instance	Reaction							Entry Date	
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other		
8874509	1 of 6		99.0	copper acetylacetonate, 1-[2-(S)-1-hydroxy-2-butanylamino]-3-isopropyl-3-methylbenzimidazolium iodide, and caesium carbonate	(none)	tetrahydrofuran, hexane, tetrahydrofuran, and hexane	20.0	0.25		2012/06/15

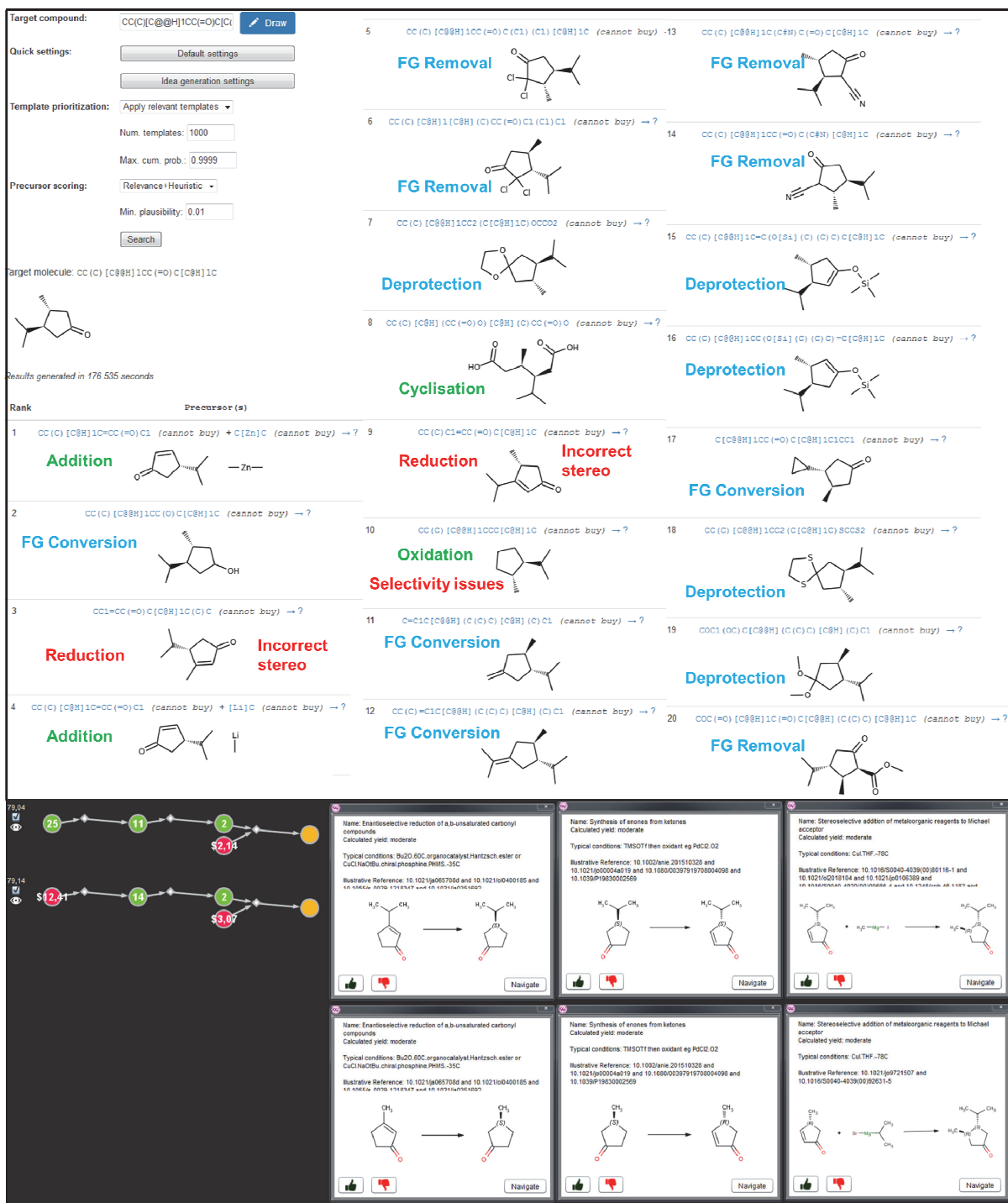
Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccoley@MIT.edu](mailto:ccoley@MIT.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

8874509, 9193652, 9193655, 28323166, 35617502

**Figure S16.** ASKCOS' automatically extracted core of a chiral-catalyst-controlled conjugated addition allows for mismatched substrates.





**Figure S17.** Retrosynthetic planning of chiral *trans*-cyclopentanone in MIT's ASKCOS system. None of the 14 reaction templates (see **Figures S18-S22**) for substrate-controlled reductions of alkenes (#2, #7, #17, #18) accounts for the necessary structural features controlling the reaction's outcome and allows for chemically incorrect process leading to the *trans* product. Bottom: In Chematica's top-scoring suggestions, the desired 1,2-*trans* configuration is achieved enantioselectively via reduction of an enone, subsequent Saegusa type reoxidation, and substrate-controlled addition of an organometallic reagent.





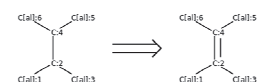


## Template 59c5134405581eb9f5769aeb

Template: [C:1]-[C:8;D3;+0;2](-[C:3])-[C:1;R;D3;+0;4](-[C:5])-[C:6]>>[C:1]-[C:2;D3;+0;2](-[C:3])=[C:2;D3;+0;4](-[C:5])-[C:6]

16 total references

Export Reaxys query for precedents



Note: this template should be used for intramolecular reactions only

Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Rxn ID	Instance	Reaction							Entry Date
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other	
3510437	1 of 2		100.0	hydrogen	platinum(IV) oxide	acetic acid	unk	unk	2007/11/22

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccoley@mit.edu](mailto:ccoley@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

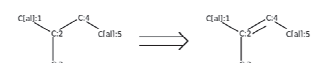
426927, 446346, 450459, 1080922, 2668383, 2824517, 3510437, 4234968, 4314821, 4315270, 4367021, 4407613, 7968037, 8103318, 8812886

## Template 59c5134505581eb9f5769b91

Template: [C:1]-[C:8;D3;+0;2](-[C:3])-[C:2;D3;+0;4](-[C:5])>>[C:1]-[C:2;R;D3;+0;2](-[C:3])=[C:2;D3;+0;4](-[C:5])

70 total references

Export Reaxys query for precedents



Note: this template should be used for intramolecular reactions only

Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Rxn ID	Instance	Reaction							Entry Date
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other	
11187129	1 of 1		100.0	hydrogen	palladium on activated charcoal	methanol	20.0	1.0	2008/03/12

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccoley@mit.edu](mailto:ccoley@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

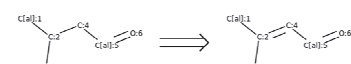
427654, 816216, 1645362, 1670670, 1674842, 1187263, 1191963, 1273204, 1322564, 1368777, 2009841, 2009842, 2197819, 2529253, 2546519, 2549450, 2611691, 2631118, 2655892, 2780960, 3140237, 3193320, 3453221, 3571598, 3601511, 3681312, 3703527, 3705520, 3726396, 4046021, 4141163, 4191707, 4233051, 4493067, 4596338, 4741783, 4888349, 5263860, 5706290, 7156864, 8093676, 8093880, 8107010, 8988107, 10588012, 10513970, 11187129, 2952936, 29794684, 30812866, 32218338, 33244589, 33244594, 33244716, 36599883, 37164296, 37196184, 38677631, 39347965, 40734178, 40734366, 42047736, 42428952

## Template 59c5135405581eb9f576a28d

Template: [C:1]-[C:8;D3;+0;2](-[C:3])-[C:2;D3;+0;4](-[C:5])>>[C:1]-[C:2;D3;+0;2](-[C:3])=[C:2;D3;+0;4](-[C:5])=[C:2;D3;+0;4](-[C:5])

134 total references

Export Reaxys query for precedents



Note: this template should be used for intramolecular reactions only

Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Rxn ID	Instance	Reaction							Entry Date
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other	
2844347	1 of 1		100.0	hydrogen	(none)	(none)	unk	unk	2007/11/18

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccoley@mit.edu](mailto:ccoley@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

438970, 1145529, 1380029, 1389943, 2017299, 2158227, 2305481, 2531916, 2550316, 3553237, 2571252, 2749113, 2841347, 2982227, 3040180, 3108247, 3108248, 3123745, 3128188, 3223737, 3235949, 3235960, 3237810, 3275533, 3618536, 3722299, 4004191, 4058795, 4221837, 4221632, 4232225, 4236877, 4306514, 4325727, 4468391, 4658763, 4803468, 4908875, 4917720, 4917721, 5018972, 5216636, 6676275, 8588257, 8819552, 9168346, 9189882, 9540435, 9642705, 10262153, 10403459, 24273414, 28128093, 28177753, 28220778, 29375011, 29699507, 30620159, 30620225, 33284481, 35961152, 41151640, 42279138, 43805931

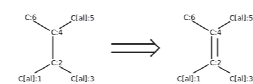
**Figure S21.** Reaction templates extracted from substrate-controlled reductions of alkenes. None of the 14 templates (with the remaining ones shown in **Figure S22**) used to generate precursors for the chiral cyclopentanone from **Figure S15** accounts for necessary substituents dictating the reaction's stereoselective outcome.

## Template 59c5163b05581eb9f578a7c0

Template: [C:1]-[C:8;D:2;+0:2](-[C:3])-[C:8;D:2;+0:4](-[C:5])-[C:1;D:3;+6]>>[C:1]-[C:8;D:2;+0:2](-[C:3])=[C:8;D:2;+0:4](-[C:5])-[C:1;D:3;+6]

8 total references

[Export Reaxys query for precedents](#)



Note: this template should be used for intramolecular reactions only

Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Rxn ID	Instance	Reaction							Entry Date	
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other		
2965844	1 of 1		91.0	(tricyclohexylphosphine)(1,5-cyclooctadiene)dicyclohexylidenehexafluorophosphate and hydrogen	(none)	dichloromethane	23.0	12.0		2010/09/12

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccooley@mit.edu](mailto:ccooley@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

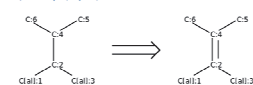
1657404, 2145035, 4220540, 4329569, 8969386, 9963768, 29518786, 29658044

## Template 59c5135a05581eb9f576a5d9

Template: [C:1]-[C:8;D:2;+0:2](-[C:3])-[C:8;D:2;+0:4](-[C:5])-[C:1;D:3;+6]>>[C:1]-[C:8;D:2;+0:2](-[C:3])=[C:8;D:2;+0:4](-[C:5])-[C:1;D:3;+6]

20 total references

[Export Reaxys query for precedents](#)



Note: this template should be used for intramolecular reactions only

Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Rxn ID	Instance	Reaction							Entry Date	
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other		
1107582	1 of 2		99.0	hydrogen	platinum(IV) oxide	tetrahydrofuran and acetic acid	unk	unk		2007/11/13

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccooley@mit.edu](mailto:ccooley@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

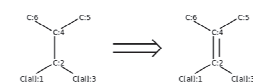
442949, 1167582, 1301156, 1382604, 3516316, 3691309, 3691309, 4049929, 4608056, 4844289, 5121372, 7909567, 8721532, 24271127, 43451726

## Template 59c5165105581eb9f578b8b4

Template: [C:1]-[C:8;D:2;+0:2](-[C:3])-[C:8;D:2;+0:4](-[C:5])-[C:1;D:3;+6]>>[C:1]-[C:8;D:2;+0:2](-[C:3])=[C:8;D:2;+0:4](-[C:5])-[C:1;D:3;+6]

14 total references

[Export Reaxys query for precedents](#)



Note: this template should be used for intramolecular reactions only

Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

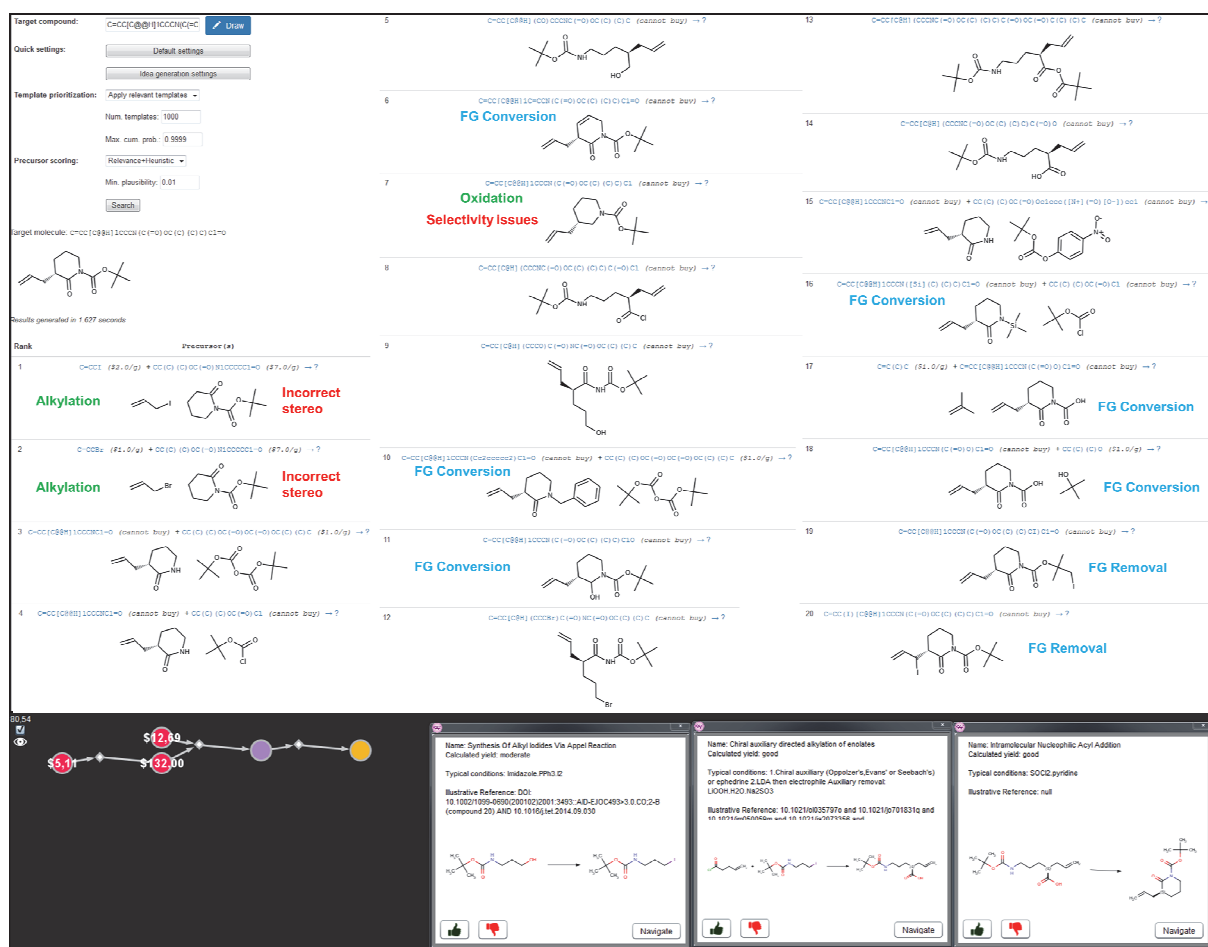
Rxn ID	Instance	Reaction							Entry Date	
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other		
3998331	1 of 1		100.0	hydrazine	(none)	dichloromethane	-10.0	unk		2007/11/22

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact [ccooley@mit.edu](mailto:ccooley@mit.edu)

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

1672564, 2123394, 2580594, 2047577, 3498123, 3498124, 3998331, 4326917, 4376131, 4852629, 4852630, 7156967, 7156968

**Figure S22.** Reaction templates extracted from substrate-controlled reductions of alkenes. None of the 14 templates used to generate precursors for the chiral cyclopentanone from **Figure S15** accounts for necessary substituents dictating reaction's stereoselective outcome.



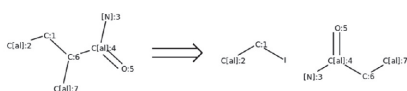
**Figure S23.** Retrosynthetic planning of a chiral lactam in MIT's ASKCOS system. Templates of stereoselective alkylations of the lactam (see **Figure S24**) do not account for necessary structural features dictating the reaction's stereoselective outcome. Bottom: In contrast, in Chematica's solution, the necessary stereocenter is created in acyclic system via alkylation controlled by a chiral auxiliary (cf. typical conditions for the second step). Subsequent removal of the auxiliary and intramolecular acylation yield the target molecule.

## Template 59c51b1005581eb9f57bcc0

Template: [#7:3]-[C:4] (=O;D1;R0:5)-[CH;R;D3;+0:6] (-[C:7])-[CH2;D2;+0:1]-[C:2]>>1-[CH2;D2;+0:1]-[C:2].[#7:3]-[C:4] (=O;D1;R0:5)-[CH2;D2;+0:6]-[C:7]

24 total references

Export Reaxys query for precedents



Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Rxn ID	Instance	Reaction							Entry Date	
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other		
27850789	1 of 1		100.0	n-butyllithium, diisopropylamine, and lithium chloride	(none)	tetrahydrofuran and tetrahydrofuran	-78 -20	1.5		2008/12/07

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact ccoley@mit.edu

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

1781832, 1852591, 1916346, 2040489, 2542041, 3951727, 4476766, 4676620, 8529026, 8541945, 9027256, 9516370, 11277603, 23825867, 27850789, 27938976, 29897068, 33275286, 36189073, 36189074, 36897762

## Template 59c51b1005581eb9f57bccac

Template: [#7:3]-[C:4] (=O;D1;R0:5)-[CH;R;D3;+0:6] (-[C:7])-[CH2;D2;+0:1]-[C:2]>>1-[CH2;D2;+0:1]-[C:2].[#7:3]-[C:4] (=O;D1;R0:5)-[CH2;D2;+0:6]-[C:7]

12 total references

Export Reaxys query for precedents



Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Rxn ID	Instance	Reaction							Entry Date	
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other		
36189076	1 of 1		96.0	n-butyllithium, diisopropylamine, and lithium chloride	(none)	tetrahydrofuran, cyclohexane, tetrahydrofuran, and cyclohexane	-20 -10	0.5	Inert atmosphere	2013/10/08

Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact ccoley@mit.edu

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

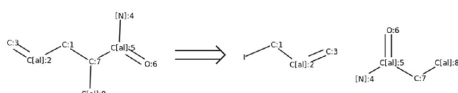
1781895, 1852635, 1852652, 1852653, 3468191, 9516369, 10392413, 11232920, 26020325, 36189075, 36189076

## Template 59c51b8905581eb9f57c18a7

Template: [#7:4]-[C:5] (=O;D1;R0:6)-[CH;R;D3;+0:7] (-[C:8])-[CH2;D2;+0:1]-[C:2]=[C:2;D2;R2:3]>>2-[CH2;D2;+0:1]-[C:2]=[C:2;D2;R2:3].[#7:4]-[C:5] (=O;D1;R0:6)-[CH2;D2;+0:7]-[C:8]

12 total references

Export Reaxys query for precedents



Note: this template looks like it might contain chiral specifications - tetrahedral chirality is not depicted in the template drawing currently

Rxn ID	Instance	Reaction							Entry Date	
		Yield [%]	Reagent(s)	Catalyst(s)	Solvent(s)	Temp. [C]	Time [h]	Other		
38042280	1 of 1		85.0	sodium hexamethyldisilazane	(none)	tetrahydrofuran and tetrahydrofuran	-78.0	1.0	Inert atmosphere	2014/07/22

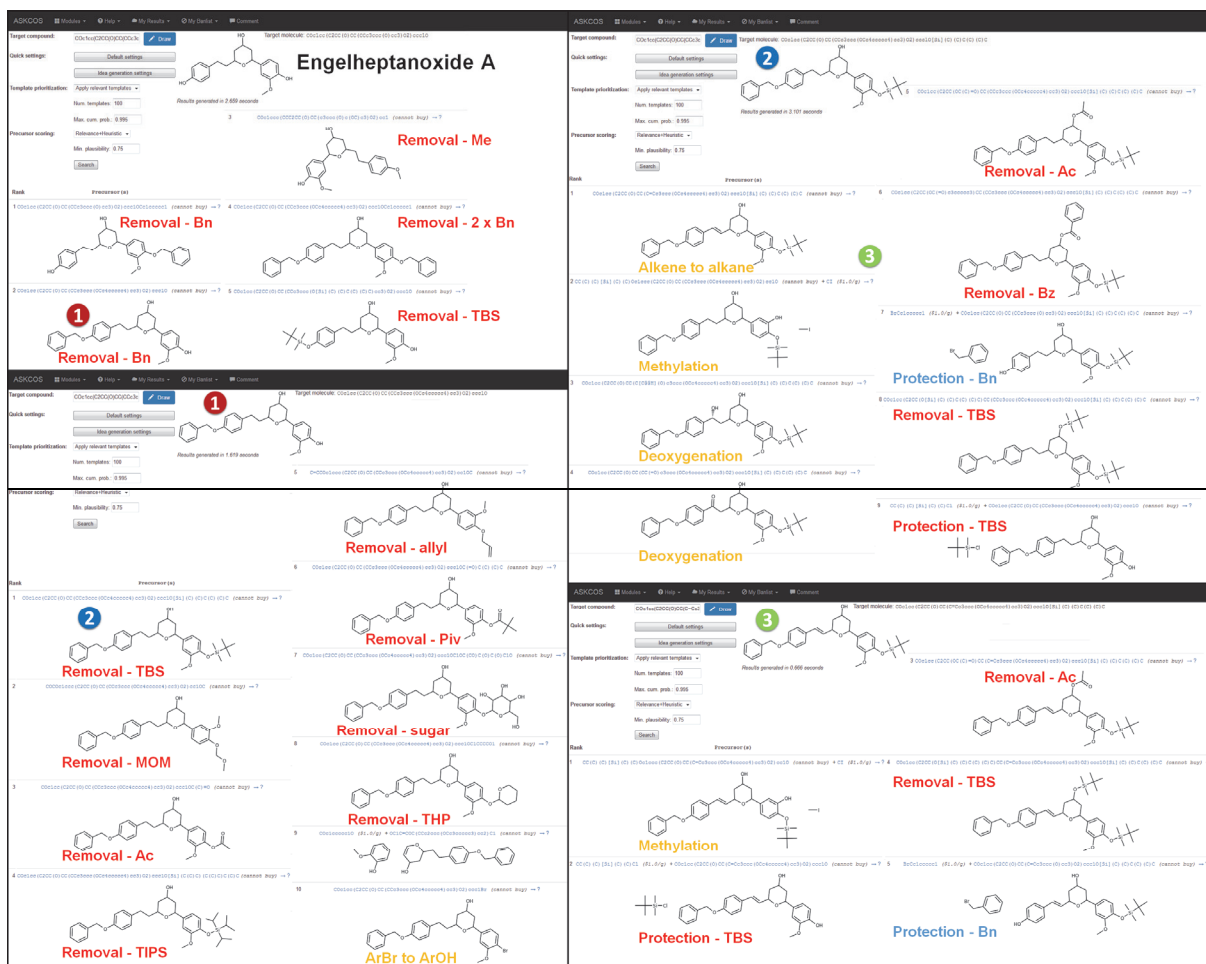
Note: You do not have access to view more than a single precedent for this transformation. If you believe this is in error, please contact ccoley@mit.edu

Reaction IDs corresponding to the 500 highest-yielding examples of this template:

1853647, 4864749, 9395522, 10390252, 20397820, 28397821, 28999269, 33157147, 38042280, 38534108, 40596420, 43913882

**Figure S24.** ASKCOS' automatically extracted cores of substrate-controlled alkylation of imides do not account for the necessary structural features controlling the reaction's outcome.





**Figure S25.** Attempted multistep retrosynthetic planning of Engelheptanoxide A in MIT's ASKCOS system. In this final exercise, we attempted to design a synthetic pathway for Engelheptanoxide A, for which a synthetic plan was previously predicted by the Chematica software and executed experimentally (see *Chem* 4, 522, 2018). The suggestions obtained from ASKCOS system for this target molecule and a few proposed predecessors are nonproductive and limited to deprotections, protections, and nonproductive functional group interconversions. None of the proposed disconnections allowed for the formation of the key tetrahydropyran fragment – either via Prins cyclisation utilized previously in the synthesis of this class of compounds or even via a must-know  $S_N2$  alkylation of alcohol. Top left: top 10 out of 88 returned results are shown, bottom left: top 10 out of 47 returned results are shown, top right: top 9 of 37 returned results are shown, bottom right: top 5 out of 30 returned results are shown. Note: The statement that no proper reactions were found pertains to the full sets of results (88/47/37/30), not only the top examples shown.

# Network search algorithms and scoring functions for advanced-level computerized synthesis planning

Bartosz A. Grzybowski<sup>1,2,3</sup>  | Tomasz Badowski<sup>1</sup> | Karol Molga<sup>1</sup> | Sara Szymkuć<sup>1</sup>

<sup>1</sup>Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland

<sup>2</sup>Center for Soft and Living Matter, Institute for Basic Science (IBS), Ulsan, Republic of Korea

<sup>3</sup>Department of Chemistry, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

## Correspondence

Bartosz A. Grzybowski, Department of Chemistry, Ulsan National Institute of Science and Technology (UNIST), 50 UNIST-gil, Ulsan 44919, Republic of Korea.

Email: [nanogrzybowski@gmail.com](mailto:nanogrzybowski@gmail.com)

**Edited by:** Raghavan Sunoj, Associate Editor

## Abstract

In 2020, a “hybrid” expert-AI computer program called Chematica (a.k.a. Synthia) was shown to autonomously plan multistep syntheses of complex natural products, which remain outside the reach of purely data-driven AI programs. The ability to plan at this level of chemical sophistication has been attributed mainly to the superior quality of Chematica’s reactions rules. However, rules alone are not sufficient for advanced synthetic planning which also requires appropriately crafted algorithms with which to intelligently navigate the enormous networks of synthetic possibilities, score the synthetic positions encountered, and rank the pathways identified. Chematica’s algorithms are distinct from *prêt-à-porter* algorithmic solutions and are product of multiple rounds of improvements, against target structures of increasing complexity. Since descriptions of these improvements have been scattered among several of our prior publications, the aim of the current Review is to narrate the development process in a more comprehensive manner.

This article is categorized under:

Data Science > Computer Algorithms and Programming

Data Science > Artificial Intelligence/Machine Learning

Quantum Computing > Algorithms

## KEYWORDS

artificial intelligence, Chematica, expert systems, networks, synthesis

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY



DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY



DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

## Chemist Ex Machina: Advanced Synthesis Planning by Computers

Published as part of the Accounts of Chemical Research special issue "Data Science Meets Chemistry".

Karol Molga, Sara Szymkuć, and Bartosz A. Grzybowski\*



Cite This: *Acc. Chem. Res.* 2021, 54, 1094–1106



Read Online

DOSTĘP OGRANICZONY



DOSTEP OGRANICZONY

DOSTEP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTEP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTEP OGRANICZONY

DOSTEP OGRANICZONY

DOSTEP OGRANICZONY



DOSTEP OGRANICZONY

DOSTEP OGRANICZONY

DOSTEP OGRANICZONY

DOSTĘP OGRANICZONY

Cite this: *Chem. Sci.*, 2019, 10, 4640

All publication charges for this article have been paid for by the Royal Society of Chemistry

## Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans†

Tomasz Badowski,<sup>‡a</sup> Karol Molga<sup>‡a</sup> and Bartosz A. Grzybowski<sup>‡\*ab</sup>

As the programs for computer-aided retrosynthetic design come of age, they are no longer identifying just one or few synthetic routes but a multitude of chemically plausible syntheses, together forming large, directed graphs of solutions. An important problem then emerges: how to select from these graphs and present to the user manageable numbers of top-scoring pathways that are cost-effective, promote convergent vs. linear solutions, and are chemically diverse so that they do not repeat only minor variations in the same chemical theme. This paper describes a family of reaction network algorithms that address this problem by (i) using recursive formulae to assign realistic prices to individual pathways and (ii) applying penalties to chemically similar strategies so that they are not dominating the top-scoring routes. Synthetic examples are provided to illustrate how these algorithms can be implemented – on the timescales of ~1 s even for large graphs – to rapidly query the space of synthetic solutions under the scenarios of different reaction yields and/or costs associated with performing reaction operations on different scales.

Received 16th December 2018

Accepted 24th February 2019

DOI: 10.1039/c8sc05611k

rsc.li/chemical-science

### Introduction

Recent years have witnessed the revival of interest in computer-assisted retrosynthetic planning, which has been an elusive goal since the late 1960s.<sup>1–11</sup> With foundational work on the representation of large collections of chemical reactions as networks<sup>12–14</sup> and the so-called bipartite graphs<sup>15–18</sup> and with modern hardware and algorithms allowing for rapid searches for synthetic pathways, synthetic planning by computers has finally become a tangible possibility. Indeed, several software platforms<sup>7–11,19,20</sup> have been developed differing in the details of search algorithms and also in the origin of synthetic rules (expert-coded<sup>19,20</sup> based on reaction mechanisms vs. automatically extracted from the literature<sup>7–11</sup>). The year 2018 also marked the first demonstration<sup>20</sup> – on our Chematica platform – of autonomous computer design and subsequent experimental validation of multiple efficient syntheses leading to medically important targets. Despite this undeniable progress, however, several challenges remain and need to be considered, especially if the programs are to be adopted by practicing organic chemists. One of the challenges we consider here is how to present to the program's user synthetic

solutions that are not only viable but also economical and chemically diverse.

In the early stages of its development, Chematica was able to identify relatively small numbers of viable syntheses which were often variations of a similar synthetic theme. With the increasing knowledge base of reactions and with improved algorithms for the exploration of synthetic options,<sup>19,20</sup> however, the searches started to identify increasingly large numbers of chemically correct solutions which themselves formed large synthetic networks (*cf.* Fig. 1). The question then arose how to estimate and rank the realistic costs of these possible pathways, taking into account not only the absolute number of steps and the costs of starting materials but also the path structure – that is, its linearity vs. convergence, the placement of the convergence points within the pathway, or the optimal “timing” to use the most expensive reagents (see examples in Fig. 2). In addition, because organic molecules can be made in different ways and the ultimate choice of a pathway often reflects practical considerations (ranging from the availability of certain reagents or equipment to the familiarity of a given chemist with particular types of reactions/procedures), it is important to present to the user multiple choices differing in the key reactions they entail. We note that although the problems of (i) finding a desired number of the best/lowest-cost solutions within the so-called directed graphs with weighted nodes (*e.g.*, in random time-dependent networks,<sup>21–23</sup> transit networks,<sup>21,24,25</sup> or reaction networks<sup>19,20,26</sup>) and also (ii) identifying qualitatively different pathways (*e.g.*, within transportation networks<sup>27</sup>) have been individually studied in graph theory, the specific approaches are not easily extendable to realistic synthetic-organic planning (*cf.* Discussion in Section S3†). Curiously,

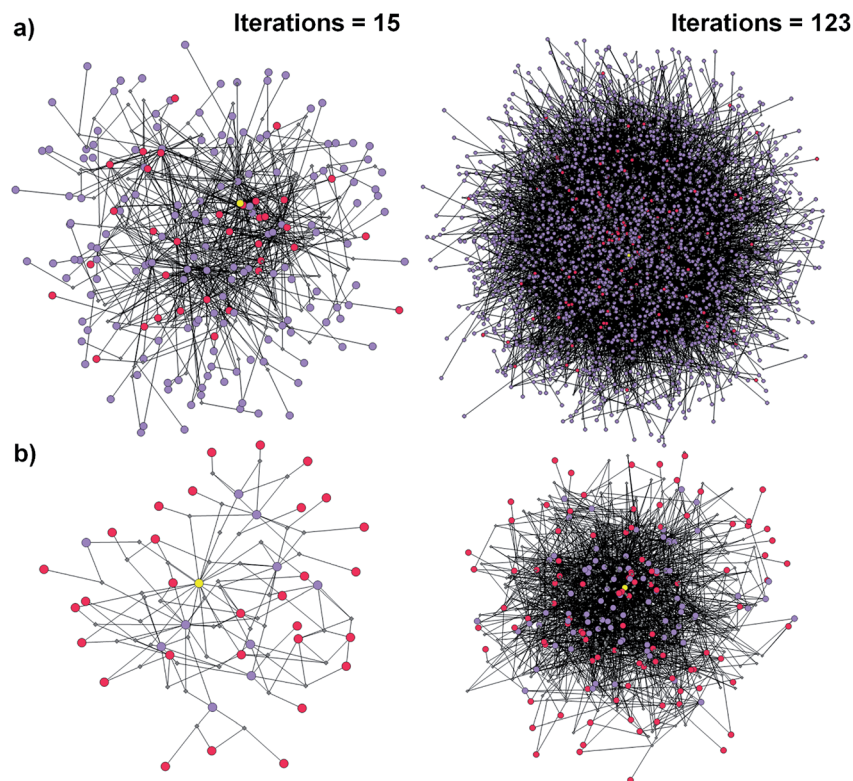
<sup>a</sup>Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland. E-mail: nanogrybowski@gmail.com

<sup>b</sup>IBS Center for Soft and Living Matter, Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulsan, 689-798, South Korea

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c8sc05611k

‡ Authors contributed equally.





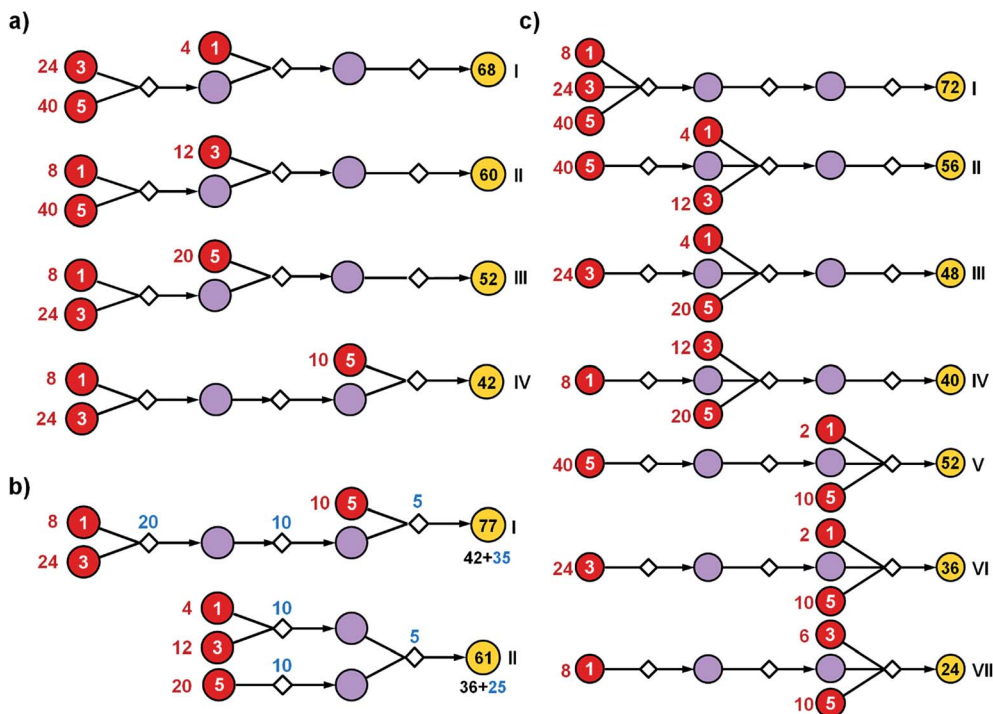
**Fig. 1** Reaction networks/graphs inspected during synthetic searches and the subgraphs of viable solutions. (a) The graphs of all molecule nodes visited during Chematica's retrosynthetic searches for the syntheses of the triarylamine target (same as in Fig. 6). The network on the left is for the early stage of the search (15 iterations of expanding retrons into progeny synthons; overall within  $\sim 20$  s computing time on a 64-core machine) and contains 456 nodes (*i.e.*, all molecules and reactions considered during the search). The graph on the right is later in the search (123 iterations within  $< 2$  min of computing time on a 64-core machine) and contains 5300 nodes. (b) The corresponding subgraphs of the networks from (a) contain only the viable syntheses. Note that as the search progresses, the subgraphs of solutions are themselves becoming quite complex, here increasing from 90 nodes after 15 iterations to 779 nodes after 123 iterations. Color coding of the nodes: yellow = target; violet = intermediates; red = commercially available starting materials; small grey diamonds = reaction nodes.

neither cost effectiveness nor diversity has been addressed in the existing retrosynthetic platforms, which might explain why the relevant publications usually describe just one top-scoring solution (diversity issue) and why the published pathways are often linear rather than convergent sequences (lack of realistic treatment of costs and yields). In our earlier versions of Chematica,<sup>16,19,20</sup> the selection algorithms were also rudimentary and the cost-calculation schemes were not only extremely slow (running for thousands of seconds for large graphs of solutions) but also did not properly capture the efficiencies of individual steps and the overall path structure (linearity *vs.* convergence), translating into unrealistic costs of starting materials consumed and/or reaction operations performed (*cf.* Section S3.1†). In light of these considerations, we see the improved approaches – reflecting true synthetic costs and operating within just seconds – described in the current paper as an important advance not only for Chematica but also for other efforts in this exciting area of research.

Computer-assisted retrosynthetic searches rely on iterative expansion of the parent/retron nodes into daughter/synthon nodes and on navigating the thus-created synthetic space (with the help of various scoring functions) to ultimately reach simple and commercially available substrates. Since the

search procedures are not the subject of our current work (for details, see ref. 9, 19 and 20), the starting point for our analyses is an already existing large graph of molecules considered/“visited” during synthetic planning. In a more technical parlance, we consider a large directed bipartite graph (Fig. 1a) composed of two types of nodes: molecules represented in all figures as circular nodes and reactions represented as smaller diamond-shaped nodes. The molecule nodes are of three types: the target (marked *yellow* in the figures), its progeny nodes (in specific chemical examples in Fig. 6–10 colored *green* if a molecule is known in the literature<sup>19,20</sup> and *violet* otherwise) corresponding to synthetic intermediates, and commercially available starting materials (*red*). To enable meaningful cost estimates of the synthetic pathways, the starting materials must have realistic prices standardized to a certain common quantity – in Chematica, there are over 200 000 such nodes from the Sigma-Aldrich catalog and their prices are all standardized to “per gram,” which is easily convertible to “per mmol” we use here. The reaction nodes carry with them some “fixed cost” of performing a reaction operation *r* to obtain some unit quantity of the product – this cost can be loosely construed as a cost of labor plus equipment/solvent/purification and does not yet account for





**Fig. 2** Effects of path structure on the cost of syntheses. The schemes illustrate hypothetical synthetic plans in which 1 mmol of the final product (yellow nodes on the right) is made from the same substrates (red nodes with prices per mmol given by the white numbers, *i.e.*, 1, 3 or 5 of some currency). All syntheses use three reactions. The reaction operations in this bipartite representation are indicated by small diamond nodes and are assumed to proceed in 50% yield each. Intermediates are denoted by violet circles. Black numbers within yellow circles give the calculated cost of making 1 mmol of the target while red numbers next to the substrate nodes are costs of the starting materials that need to be used for this purpose. (a) More expensive substrates should be introduced later in the synthesis. Here, placing the substrate with cost “3” at the second step (II) lowers the cost of materials needed for the synthesis of 1 mmol of the target from 68 to 60. Likewise, as the substrate with cost “5” is introduced in the first (I), the second (III), or the third step (IV), the overall cost is progressively lowered, from 60 to 52 to 42. (b) The realistic overall cost of synthesis should also incorporate the cost of reaction operations per some unit scale (here “5” for making 1 mmol of a given reaction product). Blue numbers next to reaction nodes denote the scale required for each 50%-yield reaction to ultimately produce 1 mmol of the target. Given this 50% yield of each step and with reference to the optimal solution (IV) from panel (a), 4 mmol of the product of the first step has to be prepared (cost  $5 \times 4 = 20$ ), 2 mmol of the product of the second step (cost  $5 \times 4 = 10$ ), and ultimately 1 mmol of the target (cost  $5 \times 1 = 5$ ). Overall, the cost of the pathway is 77 and is the sum of materials’ costs (42) and reaction operations’ cost (35). In contrast, with the same yields, the more convergent approach marked as (II) reduces both the labor cost (25 vs. 35) and the cost of starting materials (36 vs. 42). (c) A multicomponent reaction (MCR) used *en route* to a given target offers most significant cost savings if this “convergence point” is placed later within the pathway and when it uses the most expensive substrates. Here, these conditions are met for (VII).

the prices of specific substrates and/or reaction yield that are considered only when evaluating specific pathways. The algorithms we describe below do not change if the “fixed costs” are the same or different for different reaction types (as in the example in Fig. 3). Finally, the reactions are assumed to proceed with a certain yield – although yields of each reaction in the network can be estimated individually (by machine-learning<sup>28</sup> or by thermodynamic models<sup>29</sup>), we assume here, without losing generality of the algorithms, that the yields of all reactions in the graph are the same. In Chematica, the specific value of such a “global”/average yield can be set by the user allowing him/her to query the graph of synthetic solutions under different yield scenarios.

## Results and discussion

### Scoring and selecting cost-optimal pathways

With algorithmic details described in the ESI, Section S1,<sup>†</sup> the general procedure for pathway selection is illustrated in Fig. 3.

Within the initial network in Fig. 3a, we define (i) chemical nodes as “synthesizable” if they are targets of at least one synthetic pathway tracing back to commercially available substrates and (ii) reaction nodes as “viable” if all their substrates are synthesizable. In the first step, the algorithm finds all synthesizable nodes in the network in a depth-first-search-like manner and using the fact that a chemical is synthesizable only if it is commercially available or is a product of some viable reaction. If the target is not among the synthesizable nodes, then the selection algorithm stops without returning any pathways. Otherwise, it proceeds as follows. A subgraph of the network induced by synthesizable nodes is computed and retained (Fig. 3b). This step removes all substance nodes that are not synthesizable and reactions that are not viable. Then, the remaining subgraph is further restricted to one induced by ancestors of the target and the target itself (Fig. 3c). This step removes nodes which do not belong to any pathways leading to the target. Over the remaining subgraph, called the solution graph, the cost of each chemical node is taken as the smallest



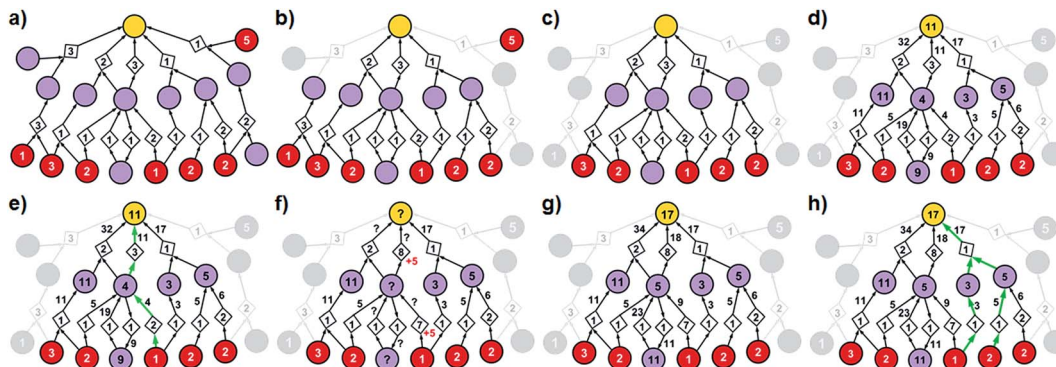


Fig. 3 Stages of selecting cost-effective yet diverse pathways from a synthetic graph. Parameters used are 50% yield for all reactions and penalty  $P = 5$ . (a) A hypothetical chemical reaction network created during a retrosynthetic search. Hypothetical costs of substrates per mmol are given over red nodes and fixed costs of reaction operations are indicated inside the diamond-shaped reaction nodes. Note that not all pathways terminate in commercially available starting materials (red nodes) as the search algorithm visited/probed some intermediates that did not lead to complete synthetic solutions. Such nodes and the pathways they are involved in are removed from consideration in (b) and (c). (d) The costs of all nodes in the remaining subgraph are computed by propagating from starting materials to the target as described in detail in the main text (see also Fig. 4). (e) The lowest-cost synthesis of the target is selected and here indicated in green. (f) Penalty  $P$  is added to the reactions from the selected pathway (here,  $P = +5$ , red numbers). Nodes whose costs increase due to such penalization are marked with question marks and are recalculated as in (g). The new “best” synthetic pathway is selected in (h) and the penalization-selection cycle can be again repeated as needed.

cost of all syntheses that can produce this chemical. The cost of any reaction node in the network is the smallest cost of all synthetic pathways containing this particular reaction and giving this reaction product. Accordingly, for each non-starting-material chemical,  $c$ , in the network, its cost is prescribed recursively by  $\text{cost}(c) = \min_{r \in \text{pred}(c)} (\text{cost}(r))$ , while for each reaction node we have  $\text{cost}(r) = \text{fixed\_cost}(r) + \sum_{c \in \text{pred}(r)} \frac{\text{cost}(c)}{\text{yield}(r)}$ ,

where  $\text{fixed\_cost}(r)$  was discussed in the preceding paragraph (cost of performing synthetic operations on some unit scale) and  $\text{pred}$  denotes the set of predecessors of a given node in the network. In the subsequent step, the costs of all nodes in the network are calculated bottom-up (*i.e.*, from the starting materials to the target) using a Dijkstra-like algorithm similar to the one for finding minimum-weight B-paths in weighted hyper-graphs<sup>24</sup> (see also ref. 26).

To illustrate how these operations work, let us first consider a simple tree in Fig. 4 in which each of the intermediates can be made in only one way and all reactions have, say, 50% yield. For the left branch, the substrate with price “3” enters in the reaction with a fixed cost of “1”. Because of the 50% yield, making 1 mmol of this reaction product requires 2 mmol of the substrate, and the total reaction cost is  $1 + 3/50\% = 7$ . For the reaction in the right branch, the unit cost is different (“2”; this reaction may be just harder to perform experimentally) but the cost calculation is analogous,  $2 + (1 + 2)/50\% = 8$ . These costs are assigned to the intermediates and propagated to the target in another 50%-yield reaction – the overall cost of making 1 mmol of the target will be  $1 + (7 + 8)/50\% = 31$ . The result of this recursive procedure agrees with the overall chemical balance – indeed, to make 1 mmol of the target, we used 4 mmols of each substrate (cost  $4 \times 3 + 4 \times 2 + 4 \times 1 = 24$ ) and performed the initial two reactions (from the substrates) on twice the scale of the final reaction making the target – hence,

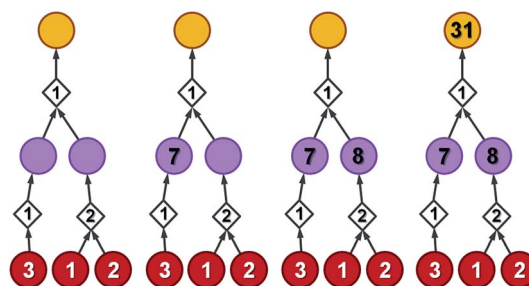


Fig. 4 A simple example illustrating bottom-up propagation of costs. All reactions have the same yield of 50%. Costs of starting materials (per mmol) are given by the numbers in the red nodes. Fixed costs of synthetic operations are indicated inside the diamond-shaped reaction nodes. For details of the calculations, see the main text.

the cost of reaction operations is  $(2 \times 1 + 2 \times 2) + 1 \times 1 = 7$  and the total cost of making 1 mmol of the target is  $24 + 7 = 31$ . We note that such calculations can be performed rapidly for arbitrary graphs including those that contain cycles (see the small cycle involving the violet node in the bottom row of networks in Fig. 3) – the cycles, however, are chemically unproductive and the costs they entail are always higher than for acyclic pathways (compare the costs of paths  $1 \rightarrow 4$  vs.  $1 \rightarrow 4 \rightarrow 9 \rightarrow 4$  in Fig. 3d).

Coming back to Fig. 3d, we observe that in realistic networks, there is generally more than one pathway to make a given chemical – for instance, the second-from-the-left intermediate can be made in three ways, *via* reactions with costs of 5, 19, and 4. Of these, we chose the least expensive option and assign to the intermediate the cost of 4, as prescribed by the formula  $\text{cost}(c) = \min_{r \in \text{pred}(c)} (\text{cost}(r))$ . Having scored all nodes within the graph, we then easily identify the most cost-effective pathway by subsequent choices (from target “down”) of the lowest-scoring reactions at each synthetic generation (in our example, “11”





followed by “4”; Fig. 3e). The information about other pathways (*i.e.*, their fragments and estimated costs) is kept in a priority queue, like in an A\* algorithm, and the graph is re-searched *via* a greedy-descent-type algorithm to find the second, third, *etc.* best pathways (see algorithmic details in the ESI, Section S1.4†).

Note that if one wishes to find pathways composed of minimal numbers of steps – which is a common situation in small-scale pharmaceutical synthesis whereby time is of essence and one might not even care about the prices of substrates or yields but just focus on synthesizing the target as rapidly as possible in amounts adequate for the upcoming assays – then the algorithm's parameters should be set to 100% yields, zero cost for all starting materials, and all fixed costs set to some common value. Under such assumption, the overall pathway score is simply the sum of the  $\text{fixed\_cost}(r)$  over all  $r$ 's (with the exception of some pathways which are not trees; see ESI, Section S1.1†). In another limiting case, when the fixed costs (labor costs) are negligible ( $\text{fixed\_cost}(r) = 0$ ), the cost is equal to the total cost of starting materials needed to synthesize 1 mmol of the target (taking into account the loss of mass for realistic yields <100%). The full scoring scheme we consider takes into account not just the number of steps (through the  $\text{fixed\_cost}$  cost term) or costs of starting materials but also both of these factors simultaneously along with yields and most optimal placement of convergence points within a pathway.

### Assigning penalties and ensuring synthetic diversity

The selection algorithm described so far can return  $n$  best-scoring pathways but does not guarantee in any way that these pathways are structurally diverse. For instance, two top-scoring solutions for the synthesis of triarylamine in Fig. 6d rely on the key Buchwald–Hartwig-type amination of the bromopyrimidine and differ only in the method of preparation of the diarylamine. In the same spirit, negligible modifications such as changing an aryl bromide to an iodide are formally different pathways to the computer but are pretty much equivalent to a user chemist. To avoid these and other unproductive repetitions and to select cost-effective yet chemically diverse pathways, we proceed as follows. After finding the best pathway (*cf.* above), the algorithm repeats the following sequence of steps until it finds the requested number of pathways or discovers that there are no more pathways left in the network:

- (i) A penalty  $P$  is added to the fixed costs of each reaction from the most-recently-found pathway (Fig. 3f) and, to avoid reusing similar synthetic solutions in other pathways, also to other reactions in the network that have the same product and non-trivial (*i.e.*, having at least four carbon atoms) substrates;
- (ii) A depth-first-search-like algorithm is used to identify the nodes (both reaction and molecule nodes) whose cost is affected due to the newly imposed penalization (nodes marked with question marks in Fig. 3f);
- (iii) The costs of all affected nodes are recalculated by a modified Dijkstra algorithm (Fig. 3g);
- (iv) Finally, a new lowest-cost pathway is identified and cycles (i)–(iv) are repeated. For all other algorithmic details, see the ESI, Sections S1.5 and S1.6.†

### Algorithms' performance

One of our key motivations for developing the selection and diversity routines has been to allow queries of the solution space on timescales much shorter than those involved in the initial retrosynthetic planning creating this space. During retrosynthetic planning, Chematica has to perform multiple operations ranging from relatively rapid matching of the reaction-rule templates (such matching is common to all retrosynthesis platforms) to much slower and Chematica-peculiar assignments of proper stereo- and regiochemistry, calculations of electronic populations for some reaction types, and several more (for details, see ref. 19, 20 and 30). In effect, searches for the solutions take from minutes for medicinal-chemistry targets to hours for complex natural products, in the end presenting to the user a given number (on the order of 100) of top-scoring solutions and, at this point, discarding the remaining ones. Retaining (*e.g.*, saving on disk) the entire space of solutions allows the user to query it multiple times under different scenarios (costs of reactions, average yields, and magnitudes of imposed diversity penalties). Importantly, querying a solution graph does not involve all the slow routines of retrosynthetic planning and should thus be possible on much shorter time scales – indeed, typical times for assigning costs and selecting pathways are on the order of 1 s, even for large solution graphs and for different target molecules. Specifically, Fig. 5a shows the times  $t_{100}$  to select (on a machine with 2.5 GHz AMD Opteron 6380 processors) 100 lowest-cost pathways from solution graphs ranging in size from 90 to *ca.* 12 000 nodes – these solution graphs are for the actual synthetic examples we discuss later in the text (triarylamine, Fig. 6; Bayer's Clofedanol, Fig. 7; Amgen's AMG641 modulator of the calcium sensing receptor, Fig. 8). As seen, these  $t_{100}$  times are on the order of 0.25 s without any diversity penalties and  $\sim 0.5$  s when diversity penalties  $P$  are added and costs of nodes need to be recalculated as new pathways are being selected. We note that the times to select  $n$  lowest cost pathways,  $t_n$ , scale approximately linearly with  $n$  and are also below 0.5 s for the largest solution graphs (Fig. 5b).

### Illustrative synthetic examples

To illustrate how the above procedures work in practice, we considered several realistic synthetic-design examples in which the solution graphs were created by Chematica within 2–10 minutes using its standard scoring functions (see ref. 19 and 20) and comprised pathways terminating in commercially available starting materials (with prices in USD per gram, converted by the program to per mmol). We queried the solution graphs varying the average yields, the fixed costs of individual reactions on a 1 mmol scale, and the diversity penalties (henceforth denoted, respectively,  $Y$ ,  $RxC$ , and  $P$ ).

(i) **Pathway ordering under various yield scenarios.** In the first example, Chematica designed routes to an unsymmetrical triarylamine used previously in the context of photochemical synthesis of complex carbazoles in continuous flow.<sup>31</sup> Within *ca.* 2 min the program searched the graph of 17 881 nodes (6826 intermediates, 293 starting materials, and 10 761 reactions; Fig. 6a), from which a solution graph composed of 3176 nodes



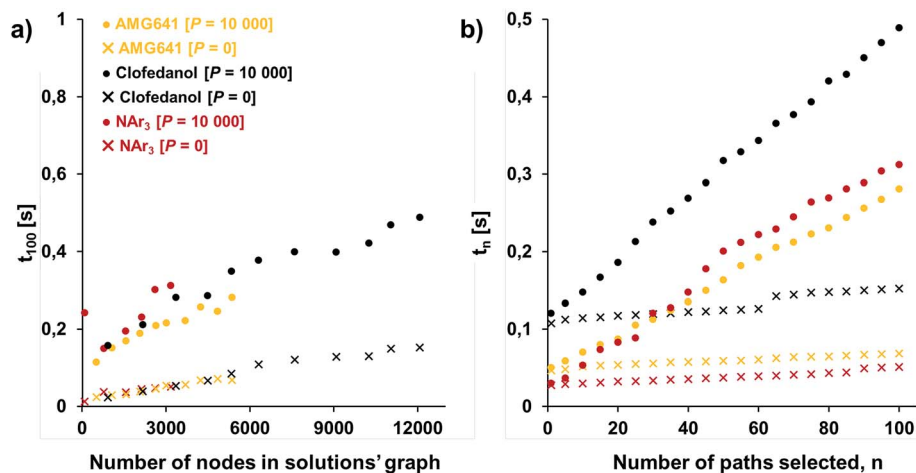


Fig. 5 Typical times to select pathways from the solution graphs. (a) Times  $t_{100}$  to select 100 lowest-cost pathways from graphs of different sizes (graph size increases as the search algorithm identifies new solutions); (b) times  $t_n$  to select  $n$  lowest-cost pathways from the maximum-size solution graphs considered here ( $\sim 3000$  nodes for triarylamine from Fig. 6;  $\sim 12\ 000$  nodes for Bayer's Clofedanol from Fig. 7; and  $\sim 5400$  nodes for Amgen's AMG641 modulator of the calcium sensing receptor from Fig. 8). Cross markers are for selection without diversity penalty  $P$ ; solid circles are for selection with penalty  $P = 10\ 000$ .

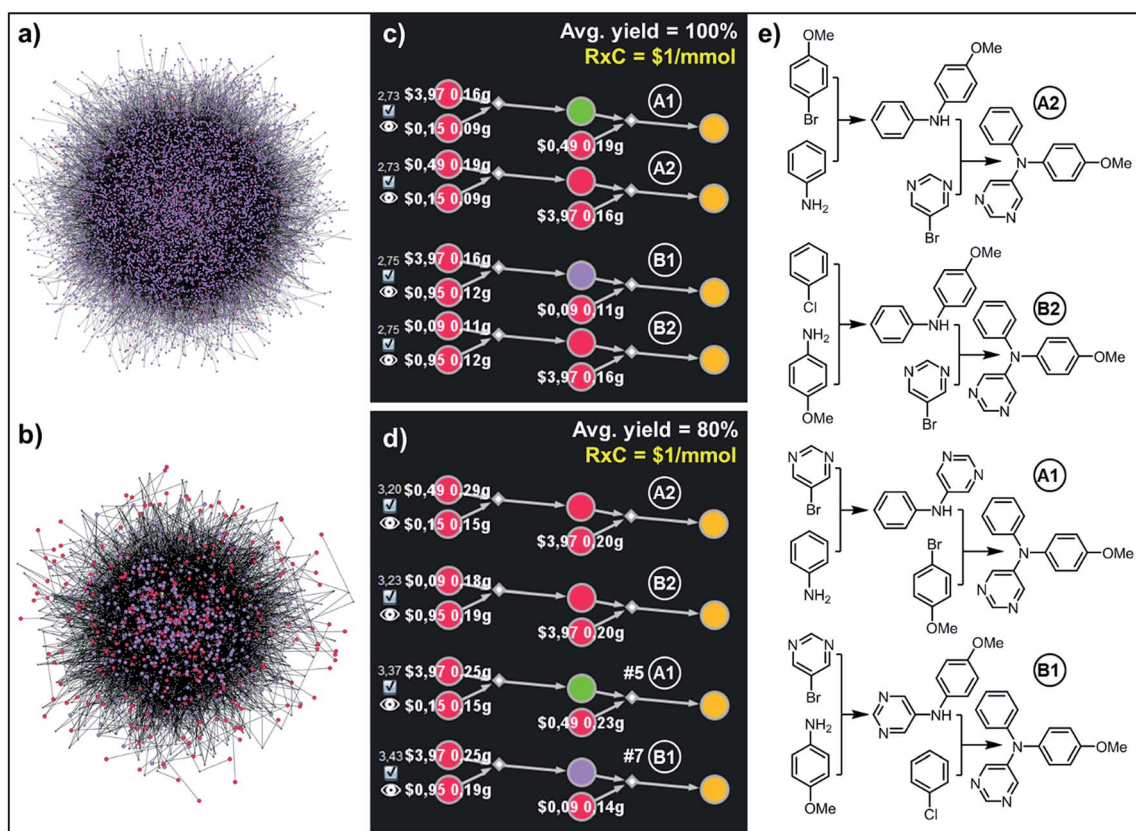


Fig. 6 Top-scoring syntheses of unsymmetrical triarylamine<sup>31</sup> proposed by Chematica under different yield scenarios. (a) Full graph searched during the retrosynthetic planning and (b) the solution graph. (c) Chematica's screenshots of the four top-scoring syntheses obtained with yields of all steps set to 100%. (d) Ordering of these top-scoring solutions changes when the yield is set to a more realistic 80%. (e) Chemical details of the pathways. The top scoring pathway A2 is identical to the one performed experimentally in ref. 31. For further details including reaction conditions proposed by Chematica, see the ESI, Section S4.† In (c and d),  $RxC$  specifies the fixed cost of performing each reaction on a 1 mmol scale (\$1) while the color coding of the nodes in Chematica's pathway miniatures is as follows: yellow = target; green = intermediates whose syntheses have been already reported in the literature and are stored in the Network of Organic Chemistry, NOC;<sup>12,14</sup> violet = intermediates not known in the NOC; red = starting materials commercially available from Sigma-Aldrich. Pairs of white numbers over the starting materials specify the costs and amounts of these starting materials necessary to make 1 mmol of the target.



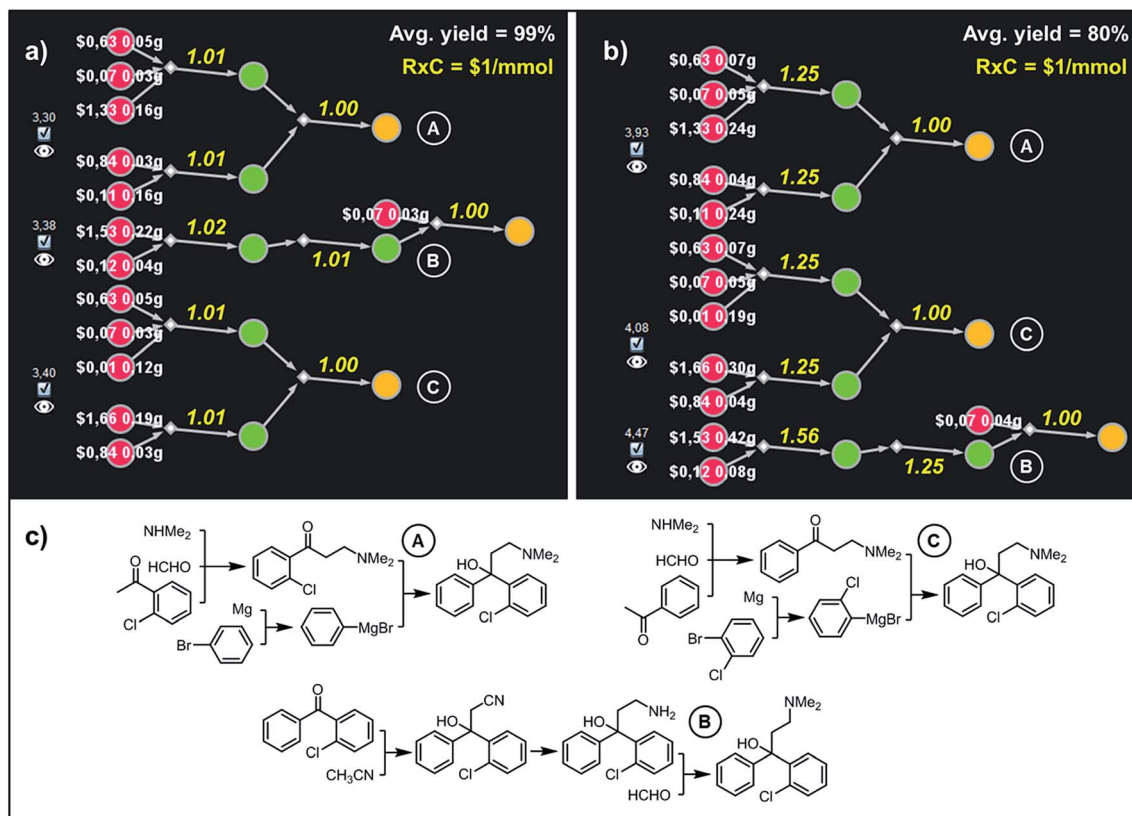
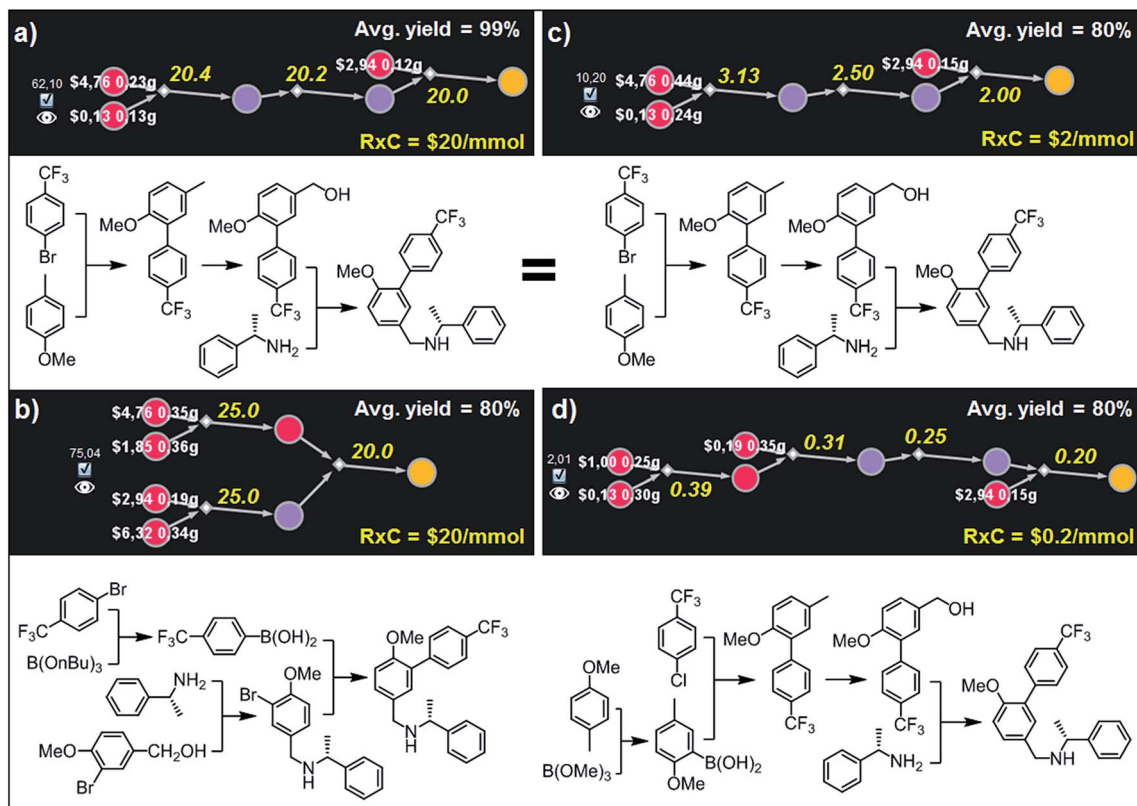


Fig. 7 Top-scoring syntheses of Clofedanol proposed by Chematica under different yield scenarios. (a) Chematica's screenshots of the three top-scoring syntheses obtained with yields of all steps set to 99%. (b) Ordering of these top-scoring solutions changes when the yield is set to a more realistic 80%. (c) Chemical details of the pathways.  $RxC$  specifies the fixed cost of performing each reaction on a 1 mmol scale (\$1). Yellow numbers over reaction arrows are fixed costs rescaled to the scale required to ultimately make 1 mmol of the target. Colors of the nodes and all legends are explained in the caption to Fig. 6. For further details including reaction conditions suggested by Chematica, see the ESI, Section S5.†

(392 intermediates, 293 starting materials, and 2490 reactions) was selected (Fig. 6b). When this solution graph was queried with the fixed cost of each reaction operation per mmol,  $RxC = \$1$ , and with an average yield of  $Y = 100\%$  – that is, naively omitting mass losses at each step – the costs of the top-scoring pathways in Fig. 6c and e were simple sums of costs of performing reactions (here, \$1 per mmol step  $\times$  2 steps = \$2 per mmol) plus the costs of starting materials. While all these solutions, relying on Buchwald–Hartwig amination, were chemically correct, the algorithm was not able to capture the differences in the costs of various starting materials being used in the first vs. second steps. In particular, the costs of syntheses utilizing the most expensive reagent, bromopyrimidine, in the first (A1) vs. the second steps (A2) were exactly the same (“\$2.73,” numbers to the left of the pathways in Fig. 6c), which is in contrast to the considerations from Fig. 2 showing that most expensive substrates should come in “closer to the target”. This problem was avoided by specifying a more realistic average yield ( $Y = 80\%$ , close to the average yield of all known reactions, see ref. 28) such that the pathway costs now reflected mass loss at each step – under this condition, the top-scoring pathways A2 and B2 (Fig. 6d and e) used the expensive bromopyrimidine in the second step. We note that the top-scoring pathway A2 was actually validated experimentally in ref. 31 which inspired this example.

In the second example, more relevant to pharmaceutical chemistry, Chematica designed pathways leading to Clofedanol, a dry cough suppressant. Choosing from the solution graph created within 10 min search time and comprising 12 074 nodes in total, the cost-optimal pathways were sought with the same fixed per-mmol cost of each reaction ( $RxC = \$1$ ) but under two different average-yield scenarios,  $Y = 99\%$  and  $Y = 80\%$ . Under the first scenario, the lowest-cost pathway – marked as (A) in Fig. 7 – commences with the three component Mannich reaction of 2-chloroacetophenone, followed by addition of phenylmagnesium bromide to the obtained ketone. This solution resembles the route patented in 2009 by Zhejiang Hisoar Pharma.<sup>32</sup> The second-scoring synthetic plan, (B), starts with the addition of acetonitrile to appropriate benzophenone, reduction of the nitrile to an amine, and reductive dimethylation with formaldehyde. This strategy is, in fact, the same as the method of preparation described in Bayer's initial (1962) patent covering Clofedanol.<sup>33</sup> Finally, the third-best solution, (C), also relies on the Mannich reaction of acetophenone, followed by the addition of Grignard reagent derived from *o*-bromochlorobenzene.<sup>34</sup> In contrast, when the average yield is  $Y = 80\%$ , Bayer's pathway is disfavored. In re-evaluating it, the algorithm recalculates the amounts and costs of necessary starting materials (e.g., one now needs 0.42 g of benzophenone vs. 0.22 g under 99%-yield





**Fig. 8** Top-scoring syntheses of AMG641 proposed by Chematica under different yield and reaction-cost scenarios. Chematica's miniatures and the pathways shown below them were selected from the solution graph (created in 7 min of retrosynthetic planning; 5363 nodes in total) assuming the following values of parameters: (a)  $Y = 99\%$ ,  $RxC = \$20$ ; (b)  $Y = 80\%$ ,  $RxC = \$20$ ; (c)  $Y = 80\%$ ,  $RxC = \$2$ ; (d)  $Y = 80\%$ ,  $RxC = \$0.2$ . Yellow numbers over reaction arrows are fixed reaction costs rescaled to the scale required to ultimately make 1 mmol of the target. Pairs of white numbers over the starting materials specify the costs and amounts of these starting materials necessary to make 1 mmol of the target. Colors of the nodes are explained in the caption to Fig. 6. For further details including reaction conditions suggested by Chematica, see the ESI, Section S6.†

assumption) and scales the costs of performing synthetic steps on larger scales (compare yellow numbers in Fig. 7a and b; e.g., the addition of acetonitrile to benzophenone must now yield over 1.5 mmol of the adduct if 1 mmol of Clofedanol is expected at the end). Consequently, pathway (B) appears to be less economically feasible and is ranked lower than both approaches taking advantage of the Mannich reaction. Of course, when making such comparisons in industrial reality, it would be essential to use substrate catalogs with wholesale prices available to a specific organization, not catalog prices of Sigma-Aldrich focusing on the sales of small quantities of specialty chemicals. Fortunately, connecting a requisite catalog to Chematica or any other retrosynthetic program is a technically trivial task.

**(ii) Pathway ordering under various yield and fixed-reaction-cost scenarios.** The example in this section is intended to illustrate how the optimal pathways vary when both the average reaction yields and the fixed costs of performing individual reactions on a given scale change. Specifically, we query the graph of synthetic solutions leading to Amgen's AMG641 (ref. 35) – an orally efficacious, positive allosteric modulator of the calcium sensing receptor – varying  $Y$  from 99% down to 80% and  $RxC$  from \$20 (expensive, probably small-scale synthesis) to \$0.2

(relatively inexpensive, probably larger scale production). For  $Y = 99\%$  and  $RxC = \$20$ , the best-scoring solution in Fig. 8a is a three step linear sequence initialized by an elegant one-pot *ortho*-lithiation/Pd-mediated coupling<sup>36</sup> with 4-trifluoromethylbromobenzene; subsequent oxidation of the benzylic position<sup>37</sup> and alkylation of commercially available chiral amines yield the target molecule. When reaction costs remain the same but one accounts for the mass loss at each step ( $Y = 80\%$ ; Fig. 8b), the best solution is a convergent three-step sequence, mirroring the original Amgen's route and using benzylic alcohol to alkylate the amine in one-pot oxidation-reductive amination<sup>38</sup> and boronic acid (prepared in one step from appropriate bromoarene) to construct the biphenyl part of AMG641.

Although both of these routes are chemically correct, they might not be optimal if AMG641 goes into large-scale production characterized by lower reaction-operation costs (e.g., achieved by solvent recycling, use of crystallization rather than chromatography, etc.). To emulate such hypothetical scale-up, we kept  $Y = 80\%$  but decreased  $RxC$  to \$2 and then to \$0.2. In the first case, the best-scoring solution (Fig. 8c) is actually the same as in Fig. 8a for  $Y = 99\%$ . We note, however, that the overall cost of this plan recalculated with  $Y = 80\%$ ,  $RxC = \$2$  constrains is, as expected, very different (\$10.2 per mmol vs.



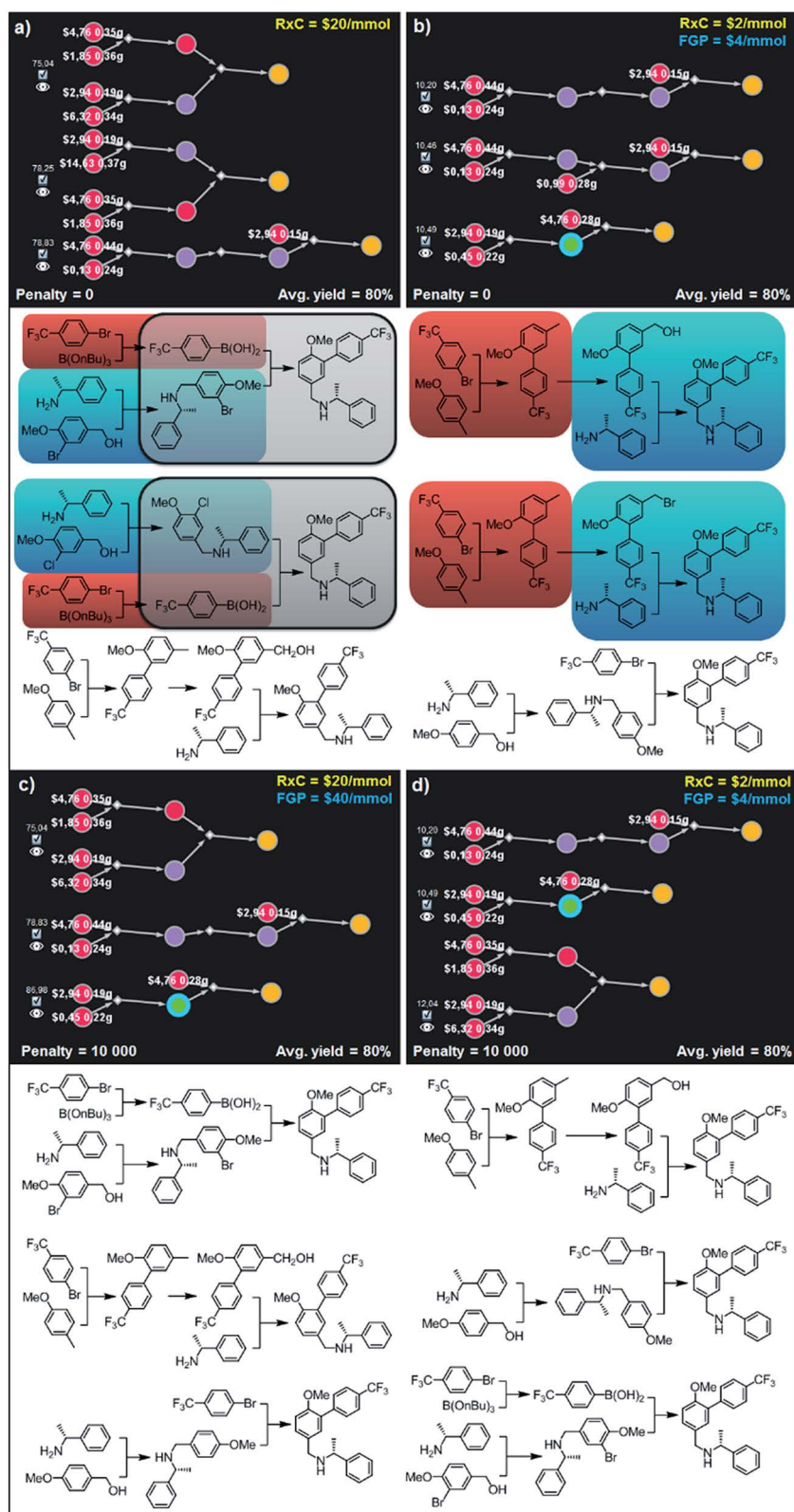
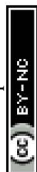


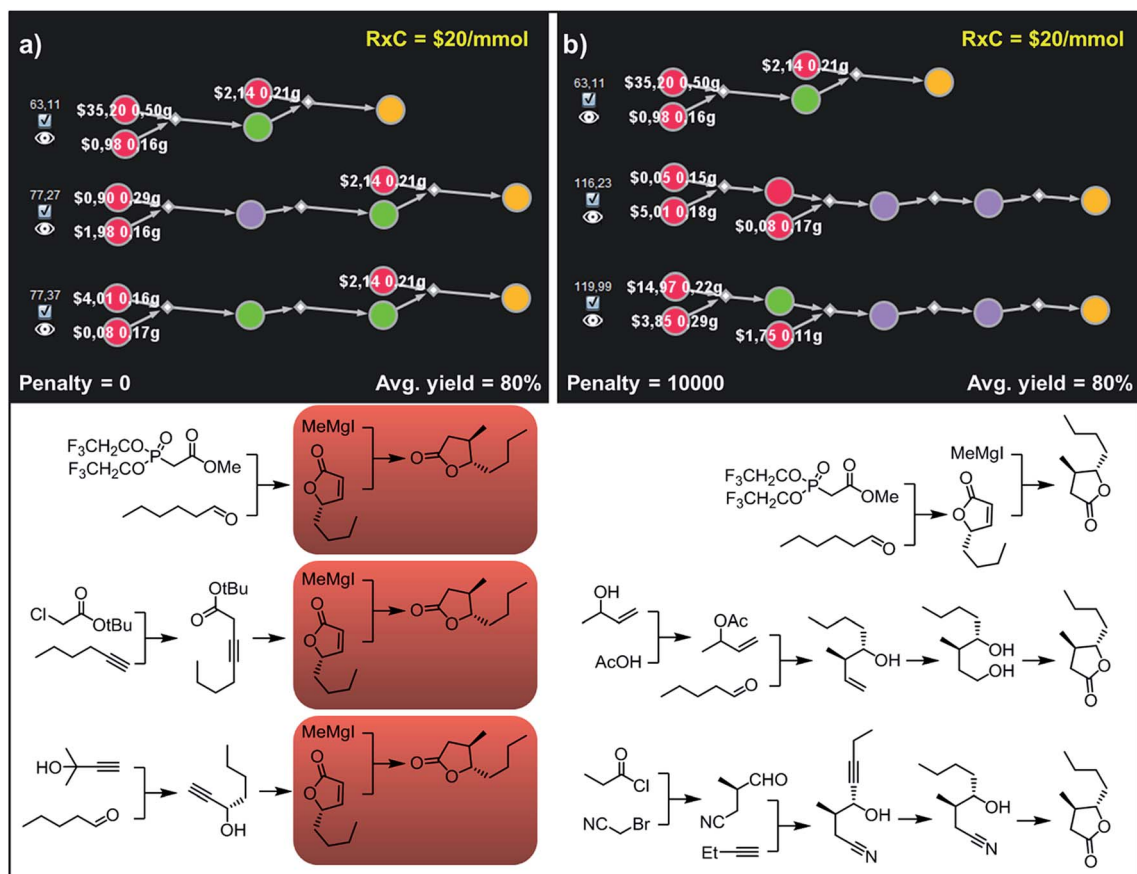
Fig. 9 Top-scoring syntheses of AMG641 proposed by Chematica without and with the application of diversity penalties. Chematica's sets of solutions shown in (a and b) have several identical (red) or very similar (blue and grey) steps when the pathways are selected from the solution graph with no penalty for reuse of already used reactions. When such a penalty is added (c and d) diversity of the synthetic plans returned is improved and each transformation is used only once. FGP is the cost of reactions requiring protection (of substances whose nodes are surrounded by blue halos). In Chematica, the cost of protection is set by the user, typically at twice the cost of reactions not requiring protection (here,  $FGP = 2 \text{ RxC}$ ). For further details including reaction conditions suggested by Chematica, see the ESI, Section S7.†



\$62.1 per mmol in Fig. 8a), reflecting slightly higher quantities and costs of starting materials (\$2.56 vs. \$1.38) but much lower costs of reaction operations (7.63\$ vs. 60.6\$). Finally, further decrease of  $RxC$  to \$0.2 adds an extra step (labor/operations are now cheap!) but sources the synthesis from very inexpensive starting materials (4-methylanisole and chloroarene). This four step linear sequence is shown in Fig. 8d and begins with the Suzuki coupling of aryl chloride and boronic acid prepared *via ortho*-lithiation and trapping of the obtained aryllithium with trimethyl borate. Subsequent oxidation of the benzylic position and junction with an appropriate amine leads to the target molecule. Taken together, the examples we discussed in this section illustrate that by varying the  $Y$  and  $RxC$  parameters, the machine makes pathway selections that reflect the economical differences between medicinal chemistry and manufacturing operations.

**(iii) Selection of diverse pathways.** The pathways we described in previous examples were all chemically viable and the selection algorithm adapted to different scoring/pricing scenarios, but within each scenario, the variability among the  $n$  best-scoring pathways was far from satisfactory – in other words, the  $n$  top-scoring pathways selected for given values of  $Y$  and  $RxC$  could rely on the same or chemically equivalent

transformations. This limits the menu of solutions the user is presented with. To illustrate the problem and how to remedy it by imposing penalties  $P$  on the reuse of equivalent transforms (see Fig. 3 and accompanying algorithm described earlier in the text), we required that Chematica returns three top scoring syntheses of the AMG641 target under the two  $Y$ - $RxC$  scenarios from Fig. 8b,c but with vs. without diversity penalization. With no penalization ( $P = 0$ ) the selection algorithm proposed sets of synthetic plans in which the same or similar transformations were used several times. For example, the first and second solutions shown in Fig. 9a ( $Y = 80\%$ ,  $RxC = \$20$  per mmol) rely on the Suzuki coupling of *p*-trifluoromethylbenzeneboronic acid with either bromo- or chloroarene (grey). Additionally, the necessary haloarene is prepared *via* alkylation of the same benzylamine with either appropriate bromobenzyl bromide or chlorobenzyl alcohol (blue) while the preparation of the boronic acid in both plans starts from the same bromobenzene (red) undergoing Br/Li exchange and trapping with tributyl borate. Similar redundancy was observed in results obtained under different  $Y$ - $RxC$  scenarios ( $Y = 80\%$ ,  $RxC = \$2$  per mmol) and is illustrated in Fig. 9b. Here, the only minor difference between the top and the second-best solutions is, in fact, the leaving group of the benzylating agent used in the  $N$ -alkylation. In both



**Fig. 10** Top-scoring syntheses of trans whisky lactone proposed by Chematica without and with the application of diversity penalties. (a) Without any diversity penalties, all pathways use the same method to install the C3 stereocenter (red frames). (b) When  $P = 10\,000$  penalties are imposed, all three top-scoring plans are substantially different and rely on different methodologies. For further details including reaction conditions suggested by Chematica, see the ESI, Section S8.†



pathways, the last step (*blue*) requires the same amine undergoing alkylation while the construction of the biphenyl part of AMG641 takes advantage of the identical lithiation–arylation (*red*). In sharp contrast, results obtained after applying large diversity penalty ( $P = 10\,000$ ) are chemically diverse. In particular, sets of synthetic plans shown in Fig. 9c and d rely on the (i) alkylation of the amine with the *m*-bromobenzyl alcohol and subsequent Suzuki coupling, (ii) *ortho*-lithiation/arylation, followed by hydroxylation and *N*-alkylation, or (iii) alkylation of the amine with *p*-methoxybenzyl alcohol and late-stage lithiation/arylation. All of the transformations used in these sets of plans are unique and used only once in the entire series of solutions – though, we observe, these relatively simple syntheses still bear some “thematic” similarity.

Accordingly, to allow for more synthetic latitude and diversity, our final example deals with more complex enantioselective syntheses of *trans*-whisky lactone (3-methyl-4-octanolide) isolated from oak wood and responsible for the taste of aged spirits.<sup>39</sup> With no penalization applied ( $P = 0$ ) each of the three top-scoring synthetic plans relies on the formation of butenolides and subsequent *trans*-selective 1,4-addition of organocuprate (derived from methylmagnesium iodide; Fig. 10a, red frames) to set the C3 stereocenter, mimicking previous literature approaches.<sup>40,41</sup> The necessary enantioenriched butenolide can be obtained from hexanal *via* proline-mediated amination-olefination<sup>42</sup> (Fig. 10a, top path). We note that this approach was demonstrated experimentally during preparation of the structurally similar *trans*-cognac lactone.<sup>42</sup> Alternatively, the butenolide can be prepared *via* enantioselective isomerization-cyclisation<sup>43</sup> of  $\beta,\gamma$ -alkynoic ester which is available in one step from hexyne and chloroacetate (Fig. 10a, middle), or *via* enantioselective addition<sup>44</sup> of protected acetylene to pentanal, followed by carbonylative cyclisation<sup>45,46</sup> (Fig. 10a, bottom). In contrast, after applying diversity penalty ( $P = 10\,000$ ), the alternative pathways no longer hinge on the 1,4-addition and both contiguous stereocenters are forged prior to the formation of the lactone. In particular, the second-best solution (Fig. 10b, middle path) now takes advantage of the Krische's crotylation<sup>47</sup> of pentanal setting both stereocenters. Hydroboration of the homoallylic alcohol thus obtained yields a 1,4-diol undergoing oxidative cyclisation<sup>48</sup> to the target molecule. Finally, the third-plan (Fig. 10b, bottom) commences with a chiral-auxiliary-controlled cyanomethylation of the enolate with bromoacetonitrile.<sup>49</sup> Subsequent addition of butynal controlled by a chiral catalyst<sup>50,51</sup> yields hydroxynitrile, which then undergoes reduction of alkyne and intramolecular alcoholysis to the whisky lactone target.

## Conclusions

In summary, we described a family of algorithms that select and score the most economical and diverse synthetic pathways from large graphs of synthetically viable solutions. This problem has not been addressed in detail in previous literature on computer-assisted retrosynthesis likely because – until now – few solutions were produced during retrosynthetic searches and *any* chemically viable outcome has been deemed a success. Now,

with much improved algorithms and modern computing power, the situation has changed and one faces the *embarras de richesse* problem, with very large numbers of potential solutions, all chemically plausible. With the algorithms like the ones we described, one can save the entire solution space and then query it rapidly, within seconds, for pathways meeting desired cost scenarios (instead of re-running the slow retrosynthetic search with different parameters). As we mentioned in the text, to truly reflect the realistic costs of specific organizations, the algorithm should be interfaced with catalogs of starting materials with prices peculiar to these organizations. In the future, one could also think of augmenting the penalization schemes – here, used to ensure chemical diversity – to downplay the use of reagents that are undesirable (toxic, volatile, *etc.*) or reaction types known to be particularly difficult or finicky.

## Author contributions

T. B. designed and implemented the selection algorithms. K. M. validated the algorithms for synthetic correctness and provided examples of syntheses described in the text. B. A. G. conceived Chematica in graduate school and has directed the development of its various aspects – including the current work – ever since. All authors contributed to the writing of the manuscript.

## Conflicts of interest

While Chematica was originally developed and owned by B. A. G.'s Grzybowski Scientific Inventions, LLC, neither he nor the co-authors currently hold any stock in this company, which is now a property of Merck KGaA, Darmstadt, Germany. The authors continue to collaborate with Merck KGaA, Darmstadt, within the DARPA “Make-It” award. All queries about access options to Chematica (now rebranded as Synthia™), including academic collaborations, should be directed to Dr Sarah Trice at sarah.trice@sial.com.

## Acknowledgements

This work was supported by the U.S. DARPA under the “Make-It” Award, 69461-CH-DRP #W911NF1610384. B. A. G. also gratefully acknowledges personal support from the National Science Center, NCN, Poland (Symfonia Award #2014/12/W/ST5/00592) and from the Institute for Basic Science Korea, Project Code IBS-R020-D1. We would like to thank Dr Piotr Dittwald for generating images of reaction networks.

## References

- 1 E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.
- 2 E. J. Corey, W. T. Wipke, R. D. Cramer and W. J. Howe, *J. Am. Chem. Soc.*, 1972, **94**, 421–430.
- 3 E. J. Corey, W. T. Wipke, R. D. Cramer and W. J. Howe, *J. Am. Chem. Soc.*, 1972, **94**, 431–439.
- 4 H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer and J. E. Searleman, *Science*, 1977, **197**, 1041–1049.



- 5 S. Hanessian, J. Franco and B. Larouche, *Pure Appl. Chem.*, 1990, **62**, 1887–1910.
- 6 J. B. Hendrickson, *J. Am. Chem. Soc.*, 1977, **99**, 5439–5450.
- 7 J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade and H. Y. Ando, *J. Chem. Inf. Model.*, 2009, **49**, 593–602.
- 8 A. Bøgevig, H.-J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Löw, C. Oppawsky, T. Rein and H. Saller, *Org. Process Res. Dev.*, 2015, **19**, 357–368.
- 9 M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
- 10 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 1237–1245.
- 11 C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281–1289.
- 12 M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2005, **44**, 7263–7269.
- 13 K. J. M. Bishop, R. Klajn and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2006, **45**, 5348–5354.
- 14 B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk and C. E. Wilmer, *Nat. Chem.*, 2009, **1**, 31–36.
- 15 C. M. Gothard, S. Soh, N. A. Gothard, B. Kowalczyk, Y. Wei, B. Baytekin and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2012, **51**, 7922–7927.
- 16 M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski and K. J. M. Bishop, *Angew. Chem., Int. Ed.*, 2012, **51**, 7928–7932.
- 17 P. E. Fuller, C. M. Gothard, N. A. Gothard, A. Weckiewicz and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2012, **51**, 7933–7937.
- 18 C. Chaouiya, *Briefings Bioinf.*, 2007, **8**, 210–219.
- 19 S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2016, **55**, 5904–5937.
- 20 T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice and B. A. Grzybowski, *Chem*, 2018, **4**, 522–532.
- 21 L. R. Nielsen, K. A. Andersen and D. Pretolani, *Comput. Oper. Res.*, 2005, **32**, 1477–1497.
- 22 E. Miller-Hooks, *Networks*, 2001, **37**, 35–52.
- 23 D. Pretolani, *Eur. J. Oper. Res.*, 2000, **123**, 315–324.
- 24 G. Gallo, G. Longo, S. Pallottino and S. Nguyen, *Discrete Appl. Math.*, 1993, **42**, 177–201.
- 25 S. Nguyen and S. Pallottino, *Eur. J. Oper. Res.*, 1988, **37**, 176–186.
- 26 R. Fagerberg, C. Flamm, R. Kianian, D. Merkle and P. F. Stadler, *J. Cheminf.*, 2018, **10**, 19.
- 27 V. Akgün, E. Erkut and R. Batta, *Eur. J. Oper. Res.*, 2000, **121**, 232–246.
- 28 G. Skoraczynski, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski and A. Gambin, *Sci. Rep.*, 2017, **7**, 3582.
- 29 F. S. Emami, A. Vahid, E. K. Wylie, S. Szymkuć, P. Dittwald, K. Molga and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2015, **54**, 10797–10801.
- 30 W. Beker, E. P. Gajewska, T. Badowski and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2019, **58**, 4515–4519.
- 31 A. C. Hernandez-Perez, A. Caron and S. K. Collins, *Chem.–Eur. J.*, 2015, **21**, 16673–16678.
- 32 T. He and Y. Wenqiu, CN101844989, 2009.
- 33 R. Lorenz, R. Gosswald and H. Henecka, US3031377A, 1962.
- 34 R. S. Sulake, C. Chen, H.-R. Lin and A.-C. Lua, *Bioorg. Med. Chem. Lett.*, 2011, **21**, 5719–5721.
- 35 P. E. Harrington, D. J. St. Jean, J. Clarine, T. S. Coulter, M. Croghan, A. Davenport, J. Davis, C. Ghiron, J. Hutchinson, M. G. Kelly, F. Lott, J. Y.-L. Lu, D. Martin, S. Morony, S. F. Poon, E. Portero-Larragueta, J. D. Reagan, K. A. Regal, A. Tasker, M. Wang, Y. Yang, G. Yao, Q. Zeng, C. Henley and C. Fotsch, *Bioorg. Med. Chem. Lett.*, 2010, **20**, 5544–5547.
- 36 M. Giannerini, V. Hornillos, C. Vila, M. Fañanas-Mastral and B. L. Feringa, *Angew. Chem., Int. Ed.*, 2013, **52**, 13329–13333.
- 37 A. D. Cort, L. Mandolini and S. Panaioli, *Synth. Commun.*, 1988, **18**, 613–616.
- 38 C. Guérin, V. Bellosta, G. Guillamot and J. Cossy, *Org. Lett.*, 2011, **13**, 3534–3537.
- 39 K. Otsuka, Y. Zenibayashi, M. Itoh and A. Totsuka, *Agric. Biol. Chem.*, 1974, **38**, 485–490.
- 40 P. Koschker, M. Kähny and B. Breit, *J. Am. Chem. Soc.*, 2015, **137**, 3131–3137.
- 41 B. Mao, K. Geurts, M. Fañanas-Mastral, A. W. van Zijl, S. P. Fletcher, A. J. Minnaard and B. L. Feringa, *Org. Lett.*, 2011, **13**, 948–951.
- 42 D. A. Devalankar, P. V. Chouthaiwale and A. Sudalai, *Tetrahedron: Asymmetry*, 2012, **23**, 240–244.
- 43 H. Liu, D. Leow, K.-W. Huang and C.-H. Tan, *J. Am. Chem. Soc.*, 2009, **131**, 7212–7213.
- 44 D. Boyall, F. López, H. Sasaki, D. Frantz and E. M. Carreira, *Org. Lett.*, 2000, **2**, 4233–4236.
- 45 W.-Y. Yu and H. Alper, *J. Org. Chem.*, 1997, **62**, 5684–5687.
- 46 W. P. Gallagher and R. E. Maleczka, *J. Org. Chem.*, 2003, **68**, 6775–6779.
- 47 X. Gao, I. A. Townsend and M. J. Krische, *J. Org. Chem.*, 2011, **76**, 2350–2354.
- 48 M. Ito, A. Osaku, A. Shiibashi and T. Ikariya, *Org. Lett.*, 2007, **9**, 1821–1824.
- 49 M. T. Crimmins, M. Shamszad and A. E. Mattson, *Org. Lett.*, 2010, **12**, 2614–2617.
- 50 J. A. Marshall and M. P. Bourbeau, *Org. Lett.*, 2003, **5**, 3197–3199.
- 51 R. Takita, K. Yakura, T. Ohshima and M. Shibasaki, *J. Am. Chem. Soc.*, 2005, **127**, 13760–13761.





**Supplementary Information** for Manuscript titled “*Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans*” by Tomasz Badowski<sup>1+</sup>, Karol Molga<sup>1+</sup>, Bartosz A. Grzybowski<sup>1,2\*</sup>

<sup>1</sup> Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland

<sup>2</sup> IBS Center for Soft and Living Matter and Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

+ Authors contributed equally

\*Correspondence to: [nanogrzybowski@gmail.com](mailto:nanogrzybowski@gmail.com)

## **Section S1. Description of the algorithms.**

This section provides descriptions of the key aspects of our pathway selection algorithms. In **Section S1.1** we formalize the definitions of a chemical reaction network, of synthetic pathways within it, and the costs of such pathways. In **Section S1.2**, we discuss a procedure for restricting the initial retrosynthetic graph to its subgraph containing all syntheses of the target, and called a solutions' graph. In **Section S1.3**, we outline the algorithm for computing costs of nodes in the solutions' graph and in **Section S1.4**, the algorithm for finding the cheapest pathways in the solutions' graph and the path retrieving part of the algorithm for finding both cheap and diverse pathways. In **Section S1.5**, we focus on the penalization of reactions and on finding the nodes whose costs increase due to such penalization. Finally, in **Section S1.6** we discuss how such costs are recomputed.

### **S1.1. Definitions.**

A chemical network is represented as a finite directed bipartite graph comprised of chemical nodes and reaction nodes. We assume that for each reaction node in the network there is exactly one edge from it to some chemical node, which is called the product of the reaction. Chemical nodes from which there are edges to a reaction node are called substrates of this reaction. We assume that each reaction in the network has at least one substrate. Some chemical nodes in the network are considered to be starting materials and are not products of any reactions in the network.

A synthetic pathway leading to the target chemical node is an acyclic subgraph of the network, containing the target, such that:

- (i) each reaction in the pathway has all the same substrates and product as in the network,
- (ii) each chemical node in the pathway which is not a starting material is a product of exactly one reaction in the pathway,
- (iii) the target is not a substrate of any reaction in the pathway and the remaining chemical nodes in the pathway are substrates of at least one reaction in it.

We define the cost of a synthetic pathway (per millimole of its target) as follows. If the pathway consists of a single node (which is a starting material), we assume its cost to be given and equal to the cost of a millimole of this starting material. Otherwise, the pathway's cost is defined recursively as:

$$\text{cost}(S) = \text{fixed\_cost}(r) + \sum_{c \in \text{pred}(r)} a_{r,c} \text{cost}(\text{subpath}(S, c)),$$

where  $r$  is the only reaction in  $S$  producing the target,  $\text{fixed\_cost}(r)$  is a nonnegative fixed cost of the reaction as discussed in the main text,  $\text{pred}(r)$  is the set of predecessors of  $r$  in  $S$  (i.e., its substrates),  $\text{subpath}(S, c)$  is the only (sub)pathway in  $S$  having  $c$  as its target, and  $a_{r,c} \geq 1$  are some coefficients denoting the number of millimoles of substrate  $c$  needed to synthesize one millimole of the product of reaction  $r$ . As discussed in the main text, we implemented our algorithms for the special case of  $a_{r,c} = 1/\text{yield}$ , for  $\text{yield} \in (0,1]$  denoting the average/global yield, though our algorithms also apply to other definition of  $a_{r,c}$ . The computation of costs of subpathways of  $S$  according to the above formula takes place in the topological order of their targets, i.e., the cost of a chemical node being pathway's target is computed only after such costs for all its predecessor chemical nodes in  $S$  being the targets have been computed. The assumption  $a_{r,c} \geq 1$  ensures that our Dijkstra-like algorithms for computing costs of nodes in the network work, as discussed in **Section S1.3** (see also <sup>26</sup> and references therein for a related definition of a synthesis plan and its cost).

For coefficients  $a_{r,c}$  equal to 1 (yield equal to 100% for  $a_{r,c} = 1/\text{yield}$ ) and costs of starting materials equal zero, the cost of a pathway is equal to the sum of  $\text{fixed\_cost}(r)$  over all reaction nodes  $r$  in the pathway if the pathway is a tree (i.e., a directed tree rooted in the target) or at least if the subgraph of the pathway induced by its nodes which are not starting materials is a tree. If such a subgraph is not a tree, then  $\text{fixed\_cost}(r)$  for some reactions  $r$

will appear several times in a sum obtained by unfolding the above recursive formula for  $cost(S)$ .

### **S1.2. Restriction of the retrosynthetic graph to its solutions' subgraph.**

A chemical node  $c$  in a reaction network  $G$  is called synthesizable if there exists a synthetic pathway in  $G$  of which  $c$  is a target. A reaction  $r$  in  $G$  is called viable if all substrates of  $r$  are synthesizable, or equivalently if in  $G$  there exists a synthesis pathway of  $r$ 's product containing  $r$ . An algorithm described by pseudocode in **Figure S1** updates the set of previously found synthesizable chemical and viable reaction nodes (together called synthesizable nodes) considering newly discovered synthesizable chemicals. The algorithm performs a DFS-like search of the reaction network, beginning with the newly discovered synthesizable chemicals and using the definition of a reaction being viable and the fact that a chemical node which is not a starting material is synthesizable only if it is product of some viable reaction. This algorithm is similar to the one for finding nodes B-connected to a source node in a hypergraph from <sup>24</sup> (if one identifies reactions with hyperarcs in such hypergraphs analogously as in <sup>26,S1</sup>).

To find all synthesizable nodes in the network, it is sufficient to run the procedure from **Figure S1** with the argument *synthFound* (denoting the initial set of found synthesizable nodes) being an empty set, *newSynthChems* (i.e., the list of newly discovered synthesizable chemical nodes) consisting of all the starting materials in the network, and with the dictionary *numNonsynthSubs* (mapping reaction nodes to the numbers of their substrates not in *synthFound*) initialized to store the total numbers of substrates for each reaction in the network. After the procedure finishes, *synthFound* consists of all the synthesizable nodes in the network.

---

```

1: procedure UPDATESYNTHFOUND(G, synthFound, newSynthChems,
  numNonsynthSubs)
  ▷ Updates the set synthFound of synthesizable nodes found considering
  newly discovered synthesizable chemical nodes newSynthChems.
  ▷ Arguments:
  ▷ G: reaction network
  ▷ synthFound: set of previously found synthesizable nodes
  ▷ newSynthChems: list of newly found synthesizable chemical nodes
  ▷ numNonsynthSubs: dictionary mapping reactions to numbers of their
  substrates not in synthFound
2:   Add nodes from newSynthChems to synthFound.
3:   while newSynthChems is nonempty do
4:     chem ← newSynthChems.pop()
5:     for rx ∈ G.successors(chem) do
6:       numNonsynthSubs[rx] ← numNonsynthSubs[rx] − 1
7:       if numNonsynthSubs[rx] = 0 then
8:         synthFound.add(rx)
9:         product ← G.successor(rx)
10:        if product ∉ synthFound then
11:          synthFound.add(product)
12:          newSynthChems.add(product)

```

---

**Figure S1. Pseudocode of an algorithm for updating the set of synthesizable nodes in a reaction network.** Triangles denote comments. The method *pop()* removes and returns the last element from a list. The method *successors* of a graph returns the set of successor reactions of a chemical node given as the argument (i.e., reactions of which it is a substrate), while the method *successor* returns the single successor product of a given reaction node. The method *add* adds the element given as its argument to a set or a list. For a dictionary *d*, *d[x]* denotes the value stored in *d* corresponding to a key *x*.

We note that if the graph grows with the progress of retrosynthetic searches (cf. main text), it is more effective to update the set of synthesizable nodes each time a new reaction and its substrates are added to the reaction network. In this way, one can immediately find out when the target becomes synthesizable and one can proceed with selecting its synthetic pathways. Such an update of the set of found synthesizable nodes *synthFound* can be realized as follows. After a reaction *rx* and its substrates are added to the network, *rx*'s substrates which are new starting materials are added to *synthFound* and information about the number of *rx*'s substrates not in *synthFound* is recorded in the dictionary *numNonsynthSubs*. Next, if

all the *rx*'s substrates are in *synthFound* (i.e., *rx* is viable), *rx* is also added to *synthFound*, and if further the *rx*'s product is not in *synthFound*, the procedure from **Figure S1** is run with the list *newSynthChems* comprising only this product.

If the target belongs to the set of synthesizable nodes found, we further proceed as follows. We find the set of ancestors of the target in the subgraph of the original retrosynthetic graph induced by its synthesizable nodes. We do this without actually computing such a subgraph – we perform a DFS-like search of the original network, starting from the target and exploring nodes which are yet undiscovered synthesizable ancestors of the already visited nodes. Once such a set of ancestors is determined, we compute a subgraph of the original network induced by such ancestors and the target. Note that the resulting subgraph is a reaction network containing all the synthesis pathways of the target present in the original network *G* (this follows from the fact that all nodes of every synthetic pathway of the target in a reaction network are synthesizable and are either ancestors of the target or the target itself). Thus, we call such a subgraph a solutions' graph. Because the solutions' graph is not larger (and typically much smaller) than the original graph, it uses significantly less memory (e.g., when saved on a disk). It is also not more computationally expensive (and typically cheaper) to perform computation of initial costs on it (discussed **Section S1.3**) or to find nodes whose costs are affected by penalization and to recompute these costs (discussed in **Sections S1.5** and **S1.6**).

Assuming that the number of substrates of each reaction (i.e. its in-degree) in the network *G* is bounded by a given constant (e.g., in our numerical experiments, all reactions had no more than four substrates), and that the set *synthFound* and dictionary *numNonsynthSubs* are implemented using hash tables, our algorithms for computing the solutions' graph discussed in this section run in time  $O(\text{number of nodes in } G)$ .

### **S1.3. Computing the initial costs in solutions' graph.**

The cost of a chemical node in a reaction network is defined as the cost of its lowest-cost synthetic pathway in the network, while the cost of a reaction node in a network is defined as the cost of the cheapest synthetic pathway of the reaction's product and containing this reaction. Such costs fulfill the following generalized Bellman's equations<sup>24</sup>: for each chemical node *c* in the network which is not a starting material we have

$$\text{cost}(c) = \min_{r \in \text{pred}(c)}(\text{cost}(r))$$

and for each reaction *r* in the network

$$cost(r) = fixed\_cost(r) + \sum_{c \in pred(r)} a_{r,c} cost(c).$$

We compute the costs of all the nodes in the network which are not starting materials using a Dijkstra-like algorithm. The algorithm is similar to the one for finding minimum weight B-paths in weighted hypergraphs described in <sup>24</sup> (see also <sup>26</sup>) for a binary heap used as a priority queue. This algorithm can be used in our case because the fixed costs of reactions we consider are nonnegative and, because  $a_{r,c} \geq 1$ , the so-called gain-free condition (which guarantees that cycles in the network are nondecreasing) is satisfied<sup>24</sup>. One difference between our algorithm and the one from <sup>24</sup> is that rather than starting from a single source node, our algorithm begins with computing the costs of reactions whose all substrates are starting materials, and pushing the minimum cost of their products and such products themselves onto the priority queue. Assuming boundedness of reactions' in-degrees (i.e., number of substrates in each reaction, see above), our algorithm for computing the costs runs in time  $O(n \log(n))$  for  $n$  denoting the number of nodes in the solutions' graph.

#### **S1.4. Finding the pathways.**

We shall first discuss our algorithm for finding a desired number of the target's minimal-cost syntheses and then the algorithm for finding both economical and diverse routes. The number of pathways found by each of these algorithms is equal to the minimum of a user specified positive integer  $k$  and the number of all synthesis pathways of the target that exist in the solution's graph  $G$ .

Our algorithm for finding the minimal-cost pathways constructs them recursively starting from the target. The function used for expanding a pathway selects a reaction's product provided as the function's argument and calls itself recursively for each substrate of this reaction which is not a starting material, given as an argument. During the pathway's construction, the function maintains a set of argument products with which it was called on the recursively processed path from the target (i.e. the argument product is added to this set at the beginning of the function and removed at its end) and a dictionary mapping products with which it was called to the selected reactions yielding such products. The algorithm maintains a directed graph, called a sequence graph, whose nodes are unique integer identifiers representing different sequences of reactions that can be consecutively selected during the recursive construction of pathways. A node  $n_1$  in the sequence graph has an edge to node  $n_2$ , only if  $n_2$  is an identifier of a sequence of reactions represented by  $n_1$  followed by a reaction that can be chosen next by the recursive function. For each identifier in the sequence graph,

the last reaction in the corresponding sequence is remembered in a dictionary. The algorithm also keeps a priority queue (implemented using a binary heap) containing sequence identifiers. The score of a sequence identifier in the queue is the minimum possible cost of pathways comprising the reactions from the corresponding sequence. The queue is initialized to contain an identifier of an empty sequence of reactions with a score equal to the cost of the target (computed as discussed in **Section S1.3**).

The algorithm keeps performing the following procedure in a loop until  $k$  pathways are returned or the priority queue becomes empty (meaning that all synthetic pathways of the target in  $G$  have been returned). First, it pops from the priority queue the sequence identifier with the lowest score  $v$ . It then retrieves all the ancestors of the sequence identifier from the sequence graph and, using them and the dictionary mapping identifiers to the last reactions in their sequences, it reconstructs a list of consecutive reactions to make during pathway's construction. Then, the abovementioned recursive function is called with the target and the list of reactions given as arguments. The function tries to construct a pathway containing reactions from the list and with cost equal to  $v$  as follows. If the list of reactions is nonempty, the function pops a next reaction to perform from the list. Otherwise, it proceeds as follows. It selects a lowest-cost reaction  $r_{min}$  in  $G$  producing the argument product  $p$ . The further operations made by the function depend on whether it is called with  $p$  provided as its argument for the first time during the pathway's construction. If this is the case, then the function finds reactions in  $G$  producing  $p$  which do not create a cycle (we say that a reaction  $r$  creates a cycle if there is a cycle in the subgraph induced by reactions selected so far by the function,  $r$ , as well as substrates and products of these reactions). Reactions creating a cycle cannot be chosen during pathway expansion, since, by definition, synthetic pathways cannot contain cycles. To verify if a reaction  $r$  creates a cycle, the function checks if any of  $r$ 's substrates is present in the maintained set of products from the recursively processed path from the target to  $p$ . The function adds to the sequence graph new identifiers representing the sequences of reactions selected so far followed by each of the found reactions producing  $p$  that does not create a cycle. Each such new identifier corresponding to some last reaction  $r$  other than the selected cheapest one  $r_{min}$  is also added to the priority queue. The score of such an identifier in the queue is computed as the popped identifier's score  $v$  plus a product of  $cost(r) - cost(r_{min})$  and the product of coefficients  $a_{s,c}$  over the reactions  $s$  and their substrates  $c$  encountered on the recursively processed path from the target to  $p$  (which for  $a_{s,c} = 1/yield$  is equal to the inverse yield raised to the power of the number of reactions in such a path). If  $r_{min}$  creates a cycle, then the pathway expansion function is terminated without

a success and the algorithm starts expanding another pathway from the beginning (i.e. starting from popping a new identifier from the priority queue). In a situation when the function is called with the argument product  $p$  for the second or later time during the pathway's construction, it proceeds as follows. It checks if the selected reaction  $r_{min}$  is equal to the previously chosen one  $r$  for this product (using the maintained dictionary mapping products to the selected reactions producing them). If not, then the function adds to the sequence graph a new identifier corresponding to choosing  $r$  again and to the priority queue this identifier with a score computed identically as discussed above. Furthermore, the function is terminated without a success (by definition, synthesis pathways can contain only one reaction with a given product), and the algorithm starts expanding another pathway. If, on the other hand,  $r = r_{min}$ , then the function adds an identifier corresponding to choosing this reaction to the sequence graph. When the function called with the target as an argument finishes successfully, a pathway comprised of the selected reactions is returned.

The algorithm for finding economically feasible and diverse pathways performs the following steps in a loop until it stops. It runs a procedure like the one above for finding the lowest-cost pathways until a pathway that was not returned yet is retrieved or the procedure discovers that there are no more synthetic pathways left in  $G$  (i.e. the priority queue becomes empty). If such a new pathway is found, it is returned. When the  $k$  pathways requested by the user are returned or the procedure discovers that there are no more pathways left in  $G$ , the algorithm stops. Otherwise, it penalizes appropriate reactions in the solutions' graph and recomputes the costs of nodes affected by such a penalization as discussed in **Sections S1.5** and **S1.6**. In our implementation, different runs of the procedure for finding lowest-cost pathways in the above loop reuse the same sequence graph (but, of course, the priority queue is reinitialized each time at the beginning of the procedure). Note that in this algorithm, our procedure for finding a sequence of lowest-cost pathways could be replaced by an alternative one (see, e.g.,<sup>21</sup> and its discussion in<sup>26</sup>).

### **S1.5. Penalization of reactions and identification of nodes whose costs increase due to such penalization.**

To promote finding diverse pathways, we add a penalty  $p > 0$  to (i) fixed costs of reactions from the previously found pathway and (ii) fixed costs of other, similar reactions in the network. We consider a reaction  $s$  to be similar to reaction  $r$  if  $s$  has the same product as  $r$  and at least one of the substrates of  $s$  belongs to the set of main substrates of  $r$  (main substrates are those with at least four carbon atoms or the largest number of carbon atoms).



Let  $cost$  denote the cost function defined as in **Section S1.3** before penalization of fixed costs of reactions, and  $cost'$  – after the penalization. We are interested in finding the set  $S_{incr}$  of nodes  $n$  in the solutions' graph  $G$  for which  $cost(n) < cost'(n)$ , i.e., whose costs increase due to penalization. From the generalized Bellman's equations in **Section S1.3**,  $S_{incr}$  satisfies the following two conditions (for  $S$  replaced by  $S_{incr}$ ).

**Condition 1.** A reaction node  $r$  (from  $G$ ) belongs to  $S$  only if it is one of the penalized reactions or some of  $r$ 's substrates belong to  $S$ .

**Condition 2.** A chemical node  $c$  belongs to  $S$  only if all reactions  $r$  producing  $c$  and such that  $cost(r) = cost(c)$  form a nonempty subset of  $S$ .

The algorithm described by the pseudocode in **Figure S2** finds and returns the smallest set  $S_{found}$  satisfying the above two conditions, i.e., such that for any other  $S$  satisfying them, we must have  $S_{found} \subset S$ . In particular, we have  $S_{found} \subset S_{incr}$ , i.e., all the nodes found by this algorithm increase costs due to penalization.

---

```

1: function GETNODESINCREASINGCOST( $G$ ,  $penalizedRxs$ ,  $cost$ )
  ▷ Returns a set of nodes in  $G$  whose costs increase due to penalization.
  ▷ Arguments:
  ▷  $G$ : solutions' graph
  ▷  $penalizedRxs$ : list of penalized reactions
  ▷  $cost$ : dictionary mapping nodes in  $G$  to their costs before penalization
  ▷ set of found nodes in  $G$  whose costs increase due to penalization
2:    $nodesIncrCost \leftarrow set(penalizedRxs)$ 
  ▷ list of reactions whose descendants need to be checked for increasing costs
3:    $rxsCheckDescIncr \leftarrow penalizedRxs$ 
  ▷ dictionary mapping chemical nodes to the number of their cheapest pre-
  predecessor reactions which are not known to increase cost
4:    $numCheapestRxs \leftarrow$  empty dictionary
5:   while  $rxsCheckDescIncr$  is nonempty do
6:      $rx \leftarrow rxsCheckDescIncr.pop()$ 
7:      $product \leftarrow G.successor(rx)$ 
8:     if  $product \notin nodesIncrCost$  then
  ▷ if  $rx$  is the cheapest predecessor reaction of  $product$ 
9:       if  $cost[rx] = cost[product]$  then
10:        if  $product \notin numCheapestRxs$  then
11:           $numCheapestRxs[product] \leftarrow$  number of predecessor re-
  actions of  $product$  with cost equal  $cost[product]$ 
12:           $numCheapestRxs[product] \leftarrow numCheapestRxs[product] - 1$ 
  ▷ if all the cheapest reactions leading to  $product$  are known to increase cost
13:          if  $numCheapestRxs[product] = 0$  then
14:             $nodesIncrCost.add(product)$ 
15:            for  $rx \in G.successors(product)$  do
16:              if  $rx \notin nodesIncrCost$  then
17:                 $nodesIncrCost.add(rx)$ 
18:                 $rxsCheckDescIncr.add(rx)$ 
19:   return  $nodesIncrCost$ 

```

---

**Figure S2. Pseudocode of an algorithm identifying nodes of the solutions' graph whose costs increase due to penalization.** For a list  $l$ ,  $set(l)$  creates and returns a set with the same elements as  $l$ . The remaining notations used are the same as in **Figure S1**.

We will show that under the additional **Assumption 1** below, we also have  $S_{incr} = S_{found}$ , i.e., this algorithm returns exactly the nodes whose costs increase due to penalization.

**Assumption 1:** For each reaction  $r$  in  $G$  and its substrate  $c$ ,  $cost(c) < cost(r)$ . This assumption holds, e.g., if all the fixed costs of reactions are positive or if the costs of

starting materials are positive and  $a_{r,c} > 1$  (which for  $a_{r,c} = 1/yield$  is equivalent to  $yield < 100\%$ ).

Let us now make **Assumption 1**. We will show that  $S_{incr} \setminus S_{found}$  is empty, which (along with  $S_{found} \subset S_{incr}$ ) implies that  $S_{incr} = S_{found}$ . Let  $C$  be the set of nodes in  $S_{incr} \setminus S_{found}$  with minimum value of  $cost$ , i.e., for

$$m = \min(cost(n): n \in S_{incr} \setminus S_{found}),$$

we have

$$C = \{n \in S_{incr} \setminus S_{found}: cost(n) = m\}.$$

We will show that  $C$  is empty, which will imply that  $S_{incr} \setminus S_{found}$  is empty.

Assume, aiming at a contradiction, that for some reaction node  $r$ ,  $r \in C$ . Then, due to **Condition 1** for  $S = S_{incr}$  (and the fact that  $r$  cannot be penalized since  $r \notin S_{found}$ ), some substrate  $c$  of  $r$  must belong to  $S_{incr}$ . From **Assumption 1**, for this substrate it holds that  $cost(c) < cost(r)$ . We must have  $c \in S_{found}$  as otherwise it would hold  $c \in S_{incr} \setminus S_{found}$  and  $cost(c) < cost(r) = m = \min(cost(n): n \in S_{incr} \setminus S_{found})$ , which is impossible. Thus, from **Condition 1** for  $S = S_{found}$ , we must also have  $r \in S_{found}$ . We received a contradiction with  $r \in C \subset S_{incr} \setminus S_{found}$ . Thus,  $C$  cannot contain any reaction nodes.

Assume now, again aiming for a contradiction, that for some chemical node  $c$ ,  $c \in C$ . Then, from **Condition 2** for  $S = S_{incr}$ , all reactions  $r$  of which  $c$  is a product and for which  $cost(r) = cost(c)$ , fulfill  $r \in S_{incr}$ . For such reactions we must have  $r \in S_{found}$ , as otherwise we would have  $r \in S_{incr} \setminus S_{found}$  and from  $cost(r) = cost(c) = m$ , it would hold  $r \in C$ , which we just proved to be impossible. Therefore, from **Condition 2** for  $S = S_{found}$ , we have  $c \in S_{found}$ , which is in contradiction with  $c \in C \subset S_{incr} \setminus S_{found}$ . Thus,  $C$  also cannot contain any chemical nodes, i.e. it is indeed empty.

Assuming the boundedness of reactions' in-degrees, and that dictionaries and sets used in the algorithm described by pseudocode in **Figure S2** are implemented using hash tables, this algorithm runs in time  $O(\text{number of nodes in solutions' graph})$ .

We note that if **Assumption 1** does not hold, then, instead of finding nodes according to the above algorithm and recomputing their costs as discussed in **Section S1.6**, one can recompute the cost of all nodes which are not starting materials from scratch as discussed in **Section S1.3**.

### **S1.6. Recomputing the costs which increase due to penalization.**

To recompute the costs of nodes whose costs increase (due to penalization), we use a Dijkstra-like algorithm similar to the one described in **S1.4**. The algorithm starts with computing the new costs (i.e. after penalization) of penalized reactions whose substrates do not increase their costs. Then, it finds chemical nodes whose cost increases and which are products of at least one reaction with known new cost and pushes the minimum new costs of such products and the products themselves onto the priority queue.

## Section S2. Performance experiments.

Our algorithms were implemented in Python (without any parallelization) and run on a computer with AMD Opteron 6380 processors with 2.5 GHz clockspeed. In all performance tests, 80% yield was used. For clofedanol and  $\text{NAr}_3$ , fixed reaction cost \$1/mmol was used, while for AMG641 it was set to \$20/mmol. For reactions requiring protections, additional penalty (AMG641: \$40/mmol;  $\text{NAr}_3$ : \$1/mmol; Clofedanol: \$2/mmol) was added. For each molecule, we saved retrosynthesis graphs of various sizes from a single search and ran the full path selection algorithm on them either with (i) no diversity penalty or (ii) with such penalty equal to 10,000. CPU times of various stages of the algorithm and of finding the consecutive pathways were recorded. In **Figure 5** in the main text, only the times  $t_n$  of computing the cost of solutions' graph and finding first  $n$  pathways were considered for graphs which contained at least 100 different synthesis pathways of the target.

The table below summarizes information about CPU times of all the stages of the full algorithm for selecting 100 pathways from the largest retrosynthetic graphs for each molecule. The time to find synthesizable chemical nodes and viable reactions in the graph (using the procedure from **Figure S1** with *newSynthChems* consisting of starting materials as discussed in **Section S1.2**, which could be executed after the retrosynthetic graph was constructed as opposed to the alternative updating approach) is denoted as  $t_{synth}$ ; the time to find the ancestors of the target in the subgraph induced by synthesizable nodes is  $t_{ancestors}$ ; the time to restrict the retrosynthetic graph to the solutions' subgraph induced by the target and such ancestors as  $t_{subgraph}$ ; and the time to compute the initial costs in the solutions' graph as  $t_{icost}$ . Since these parts are identical with and without diversity penalties applied, the table lists averages of CPU times for both of these scenarios. The time of finding paths is denoted as  $t_{paths}$  and of penalizing reactions, finding the nodes changing costs, and recomputing their costs as  $t_{rcost}$  (for  $p$  equal to zero such operations are not performed and thus their CPU time is zero). The sum of CPU times of all stages is denoted as  $t_{total}$ .

Molecule	p	$t_{\text{synth}}$ [s]	$t_{\text{ancestors}}$ [s]	$t_{\text{subgraph}}$ [s]	$t_{\text{icost}}$ [s]	$t_{\text{paths}}$ [s]	$t_{\text{rcost}}$ [s]	$t_{\text{total}}$
Clofedanol	0					0,045	0	0,311
	10000	0,033	0,032	0,088	0,113	0,081	0,289	0,635
AMG641	0					0,022	0	0,136
	10000	0,014	0,015	0,038	0,047	0,073	0,159	0,346
NAr <sub>3</sub>	0					0,024	0	0,086
	10000	0,007	0,008	0,019	0,028	0,068	0,216	0,345

As seen in the Table, the total CPU time with  $p = 0$  is less than 0.32 sec and less than 0.65 sec for  $p = 10,000$ . Note that  $t_{\text{rcost}}$  for experiments in which nonzero penalties were used was in all cases much smaller than  $99 \cdot t_{\text{icost}}$ , which demonstrates that finding and recomputing only the costs of nodes increasing due to penalization is much faster than recomputing the costs of all nodes from scratch.

## Section S3. Differences with prior approaches.

**S3.1. Comparison with Chematica’s early path-selection algorithms.** Previous versions of Chematica included a rudimentary path-selection algorithm described briefly in the SI Section S6.3 of our 2018 *Chem* publication<sup>20</sup>. This prior method differed from the current one in the several important ways – both in terms of unrealistic chemical assumptions and also much less efficient algorithms, together translating into chemically sub-optimal solutions being found and into painfully long path retrieval times. These differences are detailed below:

**S.3.1.1. Chemical differences.** Previous version of the algorithm used a different, less realistic definition of cost of a synthetic pathway. The cost of a pathway was based on the grams of starting materials rather than millimoles and, more importantly, did not take into account **reaction yields**, and it was assumed that each step produces one gram of the product from one gram of each of reaction’s substrates:

$$\text{cost}(S) = \text{fixed\_cost}(r) + \sum_{c \in \text{pred}(r)} \text{cost}(\text{subpath}(S, c)),$$

This formulation translated into unrealistic cost estimates – for instance, a ten step linear pathway would score on par with a convergent 5+5 synthesis starting from the same number of similarly priced materials, although it is evident that in practice, the latter route is significantly more economical. The new implementation, taking into accounts yields and per-millimole conversions is much more chemical and can discriminate between such cases (see also main-text Figure 2).

Next, the penalties assigned to avoid repetition of **similar reactions** are now improved to select really diverse pathways. As detailed in Section **S1.5**, we penalize reactions that were already present in the previously found pathways and also those that use similar reactions. We consider a reaction *s* to be similar to reaction *r* if *s* has the same product as *r* and *at least one* of the substrates of *s* belongs to the set of main substrates of *r* (main substrates are those with at least four carbon atoms or the largest number of carbon atoms). In contrast, in the previous version of the algorithm, reaction *s* was considered to be similar to reaction *r* if it had the same product and the substrate with the highest number of carbon atoms (for several substrates having the same, largest number of carbon atoms, the one with the lexicographically longest SMILES string<sup>S2</sup> was considered). This condition for similarity was narrower in scope than the new one and, consequently, resulted in smaller set of reactions being penalized. For example, analogous steps marked in grey in Figure 9a and blue in Figure

9b in the main text are similar according to the new definition, but not according to the previous one.

**S.3.1.2. Algorithmic differences.** Our new selection algorithms are much more time and memory efficient. In particular, for realistic networks of solutions, they now execute in a fraction of a second vs. thousands of seconds in the previous version of Chematica. To achieve these improvements, most of algorithm's routines have been thoroughly changed; some of the key changes are in the modules responsible for:

**(i) Updating synthesizable nodes.** A more efficient algorithm for updating the set of synthesizable nodes (discussed in Section S1.2) is now implemented. Notably, to verify if a reaction is viable, we now check if the number of its substrates not yet found to be synthesizable, maintained in a dictionary *numNonsynthSubs*, is equal to zero (as in line 7 of Figure S1). Before, this was achieved by iteration over all of reaction's substrates and checking if they are synthesizable.

**(ii) Extraction of the solution's graphs from the entire network of nodes visited during retrosynthetic searches.** In the previous version of the algorithm for finding the solutions' graph, the subgraph of the original retrosynthetic network induced by synthesizable nodes was computed before the ancestors of the target in this subgraph were found. As discussed in Section S1.2, in the current version, such ancestors are found using a DFS-like search of the original network without the time-consuming computation of this subgraph.

**(iii) Computing and re-computing of costs.** Previously, to compute the costs in the solutions' graph, the algorithm began with finding a graph of strongly connected components of the solutions' graph. Then, such components were visited in the topological order and costs of nodes within each of these components were calculated using a Dijkstra-like algorithm. This approach was also used to recompute the costs of all nodes in the whole solutions' graph after penalization of the fixed costs of reactions.

In the new version, we compute initial costs of nodes in the solutions' graph using a Dijkstra-like algorithm (discussed in Section S1.3), which is not only faster than the previously used approach but also much simpler to implement. Furthermore, in the algorithm for finding diverse pathways, we find and recompute only the costs changing due to penalization, which is typically significantly faster than recomputing all costs from scratch (see sections S1.5, S1.6, and S2).

**(iv) Retrieval of the minimal-cost and diverse pathways.** In the old and new implementation, this algorithm tried to further expand parts of pathways corresponding to



different sequences of reactions chosen in the initial stage of pathway construction. The information about such sequences was stored in a list of tuples consisting of a list of nodes visited up to a given point during pathway expansion, the list of substrates to be expanded (consisting of unexpanded substrates of visited reactions which were not starting materials), the so-called “accumulated costs” (equal to sums of fixed costs of visited reactions and costs of visited starting materials), and the “total costs”, equal to sums of the accumulated costs and the computed costs (in the solutions’ graph) of substrates to be expanded. Such a list consumed much more memory and time to construct than the priority queue with sequence identifiers and scores (corresponding to the abovementioned total costs) as well as the sequence graph used to reconstruct the sequences of reactions from such identifiers, both of which are used in our new algorithm.

In the old algorithm, before the pathway expansion phase, the total and accumulated costs of all elements of the list of tuples were recomputed (using the information about the visited nodes and substrates to be expanded in the tuples) and the element with the minimum total cost (corresponding to the part of the pathway to be further expanded) was found in the list and removed from it. Such recomputing of total and accumulated costs was very time consuming but was needed in cases when the costs of visited reactions or substrates to be expanded changed as a result of penalization and recomputing of costs in the solutions’ graph. Also, in the pathway expansion process, as long as the list of substrates to expand was nonempty, the algorithm proceeded as follows. A chemical node  $p$  was popped from this list and reactions from the solutions’ graph producing this chemical that did not create a cycle were found (see Section **S1.4** for the definition of reactions creating a cycle) by computing subgraphs of the solutions’ graph induced by visited nodes and reactions, and by checking if such subgraphs were directed acyclic graphs. Then, the cheapest reaction  $r_{min}$  producing  $p$  was found and the tuples corresponding to reactions not creating a cycle other than  $r_{min}$  were computed (using accumulated cost to compute their total costs) and added to the list of tuples. If  $r_{min}$  created a cycle, then pathway expansion was terminated and the algorithm moved on to expanding another pathway. Otherwise, the algorithm selected  $r_{min}$  as the next reaction during the pathway expansion and updated the accumulated cost, set of visited nodes, and substrates to expand. Once the list of substrates to expand became empty, the pathway corresponding to visited nodes was returned. Then, if fewer than the required number of pathways were returned, the algorithm penalized appropriate reactions and recomputed the costs in solutions’ graph (as discussed above) and moved on to identify another pathway. Unlike our current algorithm, this old implementation did not have any mechanisms ensuring that the found

pathways had only one reaction with a given product. Thus, it sometimes returned as “pathway” graphs containing several reactions producing a given chemical.

In our new algorithm, there is no need to recompute the scores in the priority queue after the costs in solutions’ graph change due to penalization – this is so because after the costs change, a new priority queue is constructed. The new algorithm is also much more efficient in checking if a reaction creates a cycle (see Section **S1.4**).

Finally, we note that with our implementation, we added the possibility of saving a solutions’ graph *during* retrosynthetic search to later load it and select pathways from it multiple times under different scenarios (i.e., different costs of reactions, average yields, magnitudes of imposed diversity penalties). In the previous version of Chematica, only the diversity penalties could be changed during retrosynthetic search using the “select diverse” slider in the Chematica’s main window. This affected the diversity of the next set of pathways selected from the continuously expanding retrosynthetic graph. The cost of reactions, however, was fixed and specified by the user *before* search – any change in this parameter required the user to restart the entire, slow retrosynthetic search.

## **S3.2. Comparison with other relevant works in the area.**

In this Section, we narrate briefly other publications in which problems and algorithms related to our work have been addressed, albeit not in the context of chemically realistic retrosynthetic design or even not in the context of chemistry at all.

**S.3.2.1. Differences from methods for finding the best K synthesis plans<sup>26</sup> and K shortest hyperpaths<sup>21</sup>.** In reference<sup>26</sup>, the authors reformulate the problem of finding the K lowest-cost synthesis plans in a reaction network in terms of the problem of finding K lowest-cost hyperpaths in a hypergraph. They also apply an algorithm from ref<sup>21</sup> (for the special case of the latter problem for acyclic hypergraphs) to find K synthesis plans with the lowest total weight of starting materials, assuming fixed reaction yields. Unfortunately, they consider a completely unrealistically simple mathematical model of a reaction network, in which the molecules are represented as carbon skeletons and reactions rely on forming bonds between arbitrary carbon atoms of different substrates to join them, or between the atoms of the same substrate to form rings. Even the authors themselves admit that real reactions can differ significantly from the ones in their model and the “skeleton plans” resulting from their model

may not correspond to any feasible syntheses. There is also no mention in their work of any selection based on synthetic diversity.

In contrast, we demonstrate that our algorithm is applicable to realistic, large reaction networks, possibly containing cycles, from which it can rapidly select chemically viable syntheses. In fact, the synthetic examples we provide are the first demonstration of computer-generated plans that are not only chemically correct but also scored realistically against (simultaneously!) prices of the starting materials, reaction operation costs, and yields, and selected according to synthetic diversity criteria.

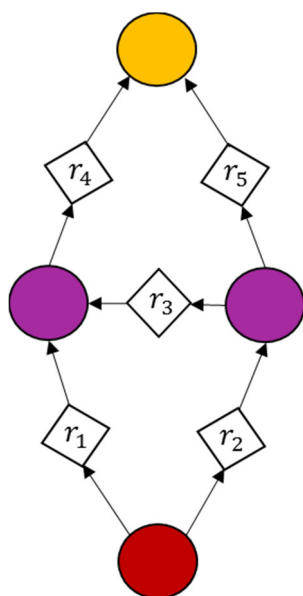
Down to some more technical detail, we note that the authors of ref <sup>26</sup> mention – but do not demonstrate – that a more complicated version of the algorithm from <sup>21</sup> *could* also be applied to more general reaction networks, like the ones admitting cycles (though, as opposed to our work, they do not provide sufficient conditions for the algorithm to be applicable to such networks). They also suggest that this algorithm could be used with more general synthesis plan costs, having the recursive form of so called “additive weighting functions” (see <sup>21</sup>), e.g., allowing to consider fixed-reaction-costs and costs of consumed starting materials similar as in our work. Note, however, that even if implemented, the algorithm from <sup>21</sup> is expected to be much slower than the version of our algorithm for finding lowest-cost pathways (both run on solutions’ graphs similar as in our performance experiments). To show this, consider the following argument. Recall that for a given solutions’ graph, our algorithm for selecting a given number of the lowest-cost pathways first computes the initial cost of nodes in the graph (as discussed in Section **S1.3**), and then finds the pathways in it (see Section **S1.4**). Note also that, in all our performance experiments in Section **S2** for computing 100 lowest-cost pathways on the largest solutions’ graphs, the time  $t_{icost}$  of computing the initial costs was higher than the time  $t_{paths}$  of finding all the pathways. The algorithm from <sup>21</sup> requires the computation of costs of nodes in a modified graph using a Dijkstra-like method from <sup>24</sup> (i.e. similar as in our work) at least once for each pathway found. This is the case both for the slower and the improved versions of this algorithm called, respectively, *Yen* and *LB<sub>Yen</sub>* in <sup>21</sup>. Furthermore, the CPU time of both versions of the algorithm in numerical experiments in <sup>21</sup> was roughly proportional to the number of such cost computations made. Thus, even if the algorithm from <sup>21</sup> required only one computation of costs for each of the 100 pathways found in our solutions’ graphs, it can still be expected to run much slower than our algorithm (i.e., at least 50 times slower).

### S.3.2.2. Differences from methods for finding dissimilar paths in graphs.

There has been some work in the non-chemical literature on the problem of finding dissimilar but possibly short paths between a given origin and a destination in weighted graphs – for instance, in the context of finding spatially dissimilar paths in transportation networks<sup>27,S3</sup>. We note that this problem is significantly less general than considered in our work. First, a weighted directed graph can be identified only with a reaction network in which graph's nodes are represented by chemical nodes and its edges, by unary reactions with fixed costs equal to the edges' weights. Furthermore, for the network containing a single starting material whose cost is zero and for yield equal to 100%, synthetic pathways of a target in the network have cost equal to the length of the corresponding paths from origin to destination in the graph setting.

The algorithm for finding short but dissimilar pathways in graphs that is most related to the approach in our work is a so-called Iterative Penalty Method (IPM) (see<sup>27</sup> and references therein). It relies on the repetitive application of finding the shortest path (e.g., using the Dijkstra algorithm) and then adding penalties, e.g., to the edges from such a path. An approach analogous to IPM in the context of our reaction networks could rely on repetitively computing the costs in the network (or efficiently re-computing only the changing costs), finding the lowest-cost pathway, and penalizing the fixed costs of reactions from this pathway. One of the differences between our algorithm and such an IPM analogue is that, after finding a pathway, we penalize not only the reactions from this pathway but also appropriately defined similar reactions (for example, pairs of analogous reactions marked in blue and grey in **Figure 9a** and **9b** in the main text are similar according to our definition though not identical). Another important difference is that our algorithm does not return the lowest-cost pathway in the graph with recomputed cost, but the lowest-cost pathway not returned before (using our method for generating the consecutive lowest-cost pathways until a new pathway is discovered). This ensures that our method cannot return the same pathway several times and that it returns all the existing pathways when their total number is not higher than the number of pathways requested by the user. The IPM-like algorithm, on the other hand, can return repeated pathways and may never return some existing pathways no matter for how many iterations it is run. An IPM version for finding K distinct diverse paths was used in ref<sup>27</sup> whereby, when a repeated path is found, the algorithm rejects it (but applies penalties to its edges) and goes to another iteration of the method. Note that a similar idea could be used in the IPM analogue for reaction networks. Unfortunately, such an algorithm

will never finish in the case when  $K$  is greater than the number of existing pathways in the network (which the user does not know a priori when specifying  $K$ ) and even in some cases when there exist at least  $K$  distinct paths in the graph. To illustrate this, consider a simple reaction network in **Figure S3** below and assume that the fixed costs of all reactions, diversity penalty, and yield are all equal to 1, as well as that the cost of the only starting material is zero. This network contains three pathways:  $p_1$  containing reactions  $r_1$  and  $r_4$  and with cost 2,  $p_2$  with reactions  $r_2$  and  $r_5$  and cost also 2, and  $p_3$  with reactions  $r_2$ ,  $r_3$ , and  $r_4$  and cost 3.



**Figure S3.** An example of an extremely simple reaction network used to compare our algorithm against an IPM-type approach. Red node is the starting material, violet nodes are intermediates, and the yellow dot is the target molecule.

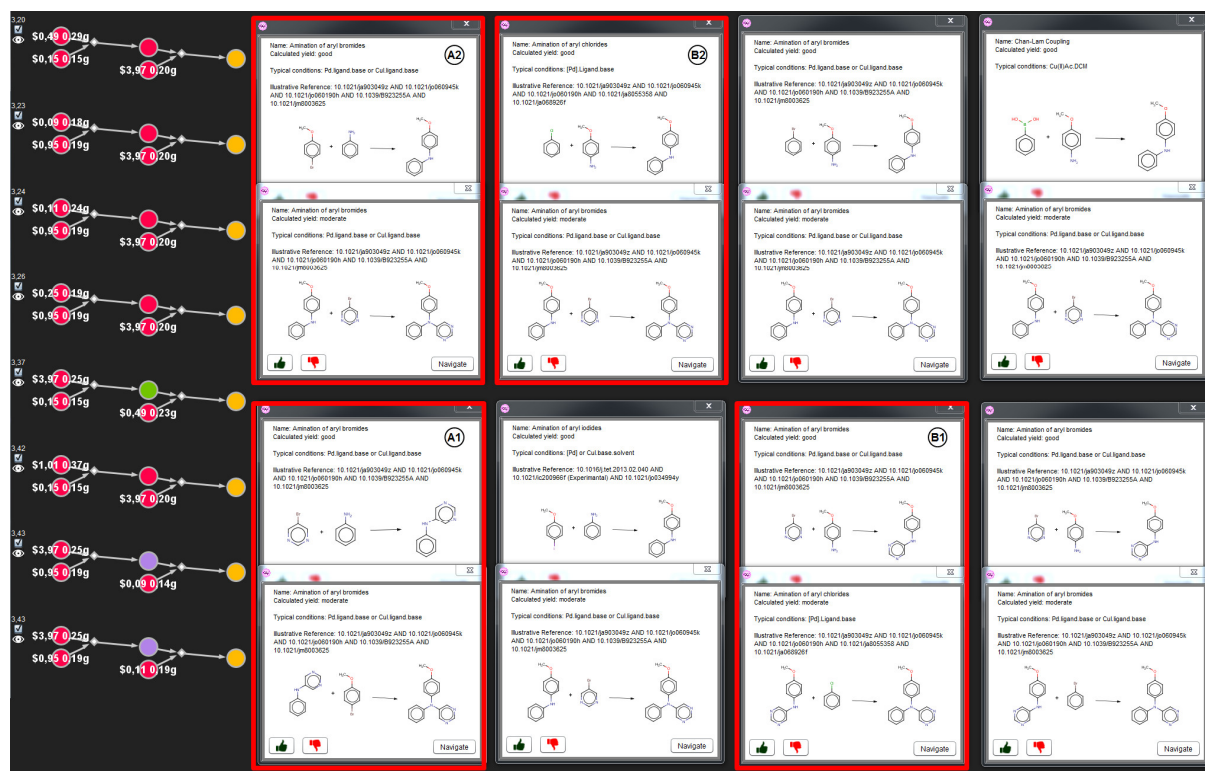
For this network, both our and the IPM-like algorithm could first return pathway  $p_1$  and then  $p_2$ . When queried for more pathways, our algorithm would next return pathway  $p_3$  and then discover that there are no more pathways left in the network. The IPM analogue, on the other hand, would again return pathway  $p_1$ , then again  $p_2$ , and so on, never returning pathway  $p_3$ . Thus, if the technique of rejecting repeated pathways were used, when queried for three or more pathways, the IPM-like algorithm would get stuck in an infinite loop. The same problems can occur with the original IPM algorithm in the graph setting (e.g., the above example can be easily reformulated in the directed graph setting) and an approach similar to ours could be used to overcome them, i.e., instead of finding the shortest path in the penalized graph, one could generate a sequence of shortest paths (e.g., using Yen' algorithm<sup>S4</sup>) until a

new path is found or the algorithm discovers that there are no more paths left in the graph. However, to our knowledge, this has never been done in the literature.

### Supplementary references.

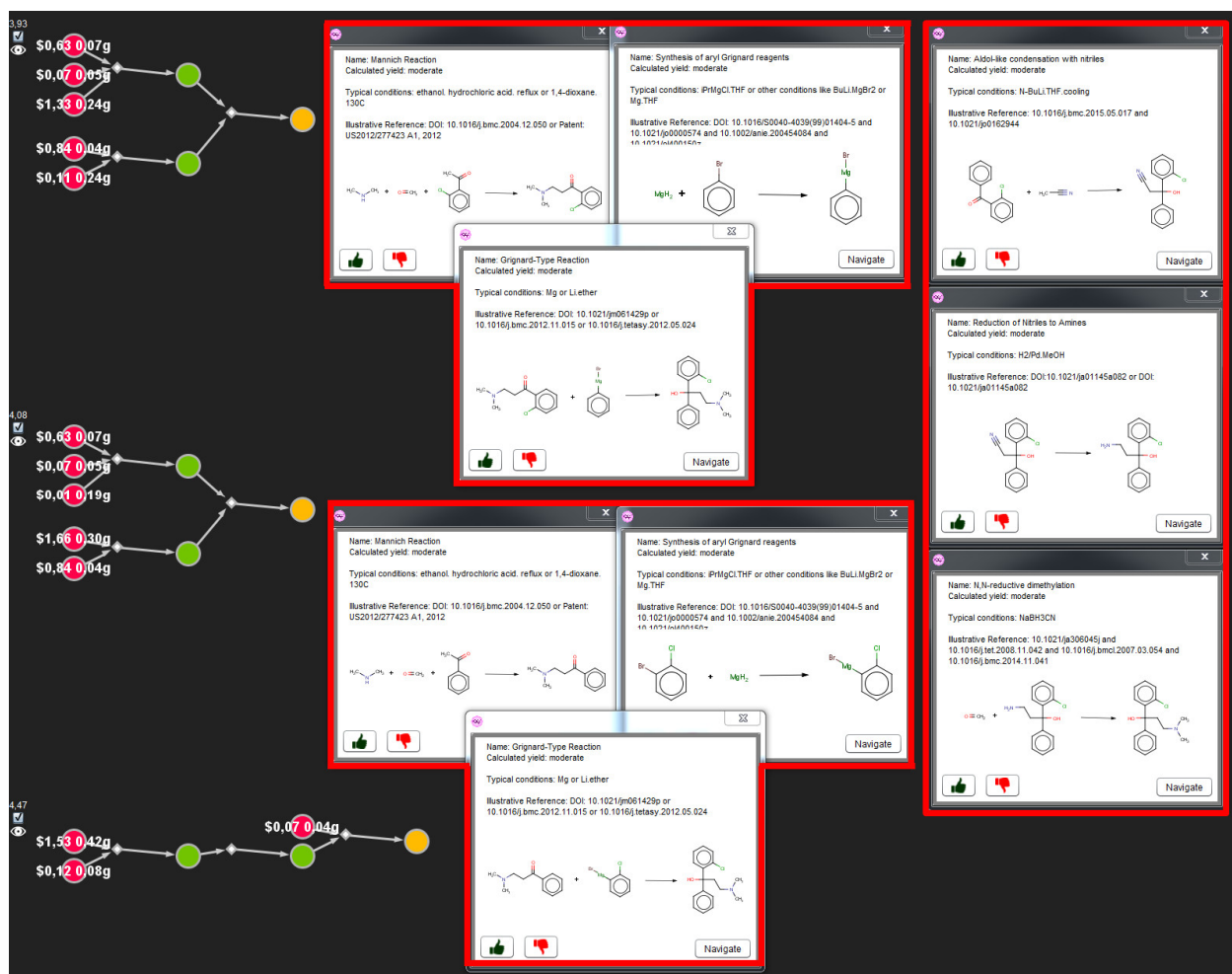
- S1 P. Carbonell, D. Fichera, S. B. Pandit and J.-L. Faulon, *BMC Syst. Biol.*, 2012, **6**, 10.  
 S2 D. Weininger, *J. Chem. Inf. Model.*, 1988, **28**, 31–36  
 S3 H. Liu, C. Jin, B. Yang and A. Zhou, *IEEE Trans. Knowl. Data Eng.*, 2018, **30**, 488–502.  
 S4 J. Y. Yen, *Manage. Sci.*, 1971, **17**, 712–716.

### Section S4. Details of Chematica's syntheses of triarylamine.



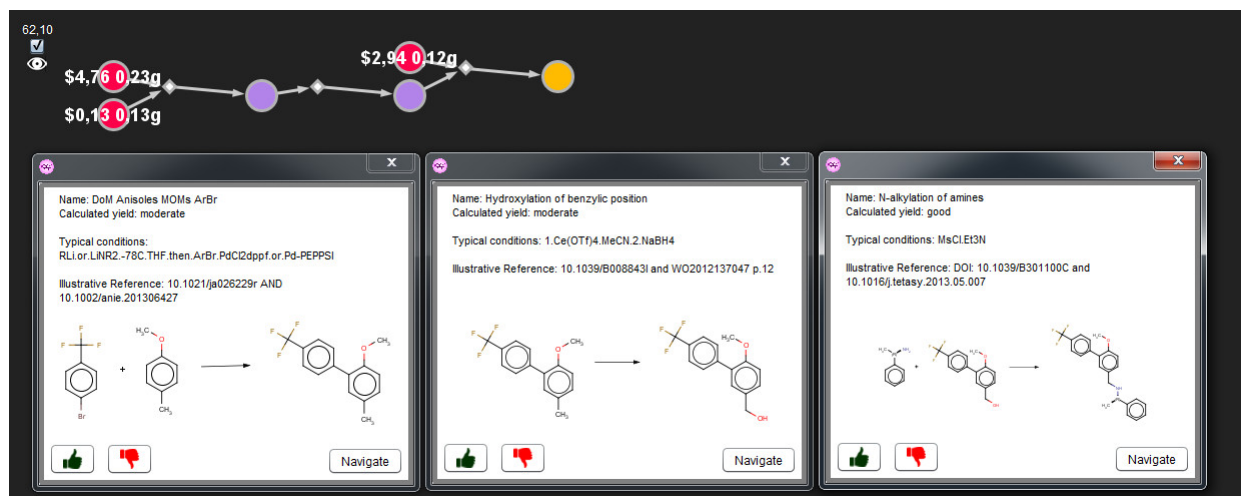
**Figure S4.** Details of top ten synthetic pathways obtained for triarylamine with  $RxC = \$1/\text{mmol}$ ,  $Y = 80\%$ . Pathways depicted in **Figure 6c,d** are marked with red frames.

## Section S5. Details of Chemica's syntheses of Clofedanol.

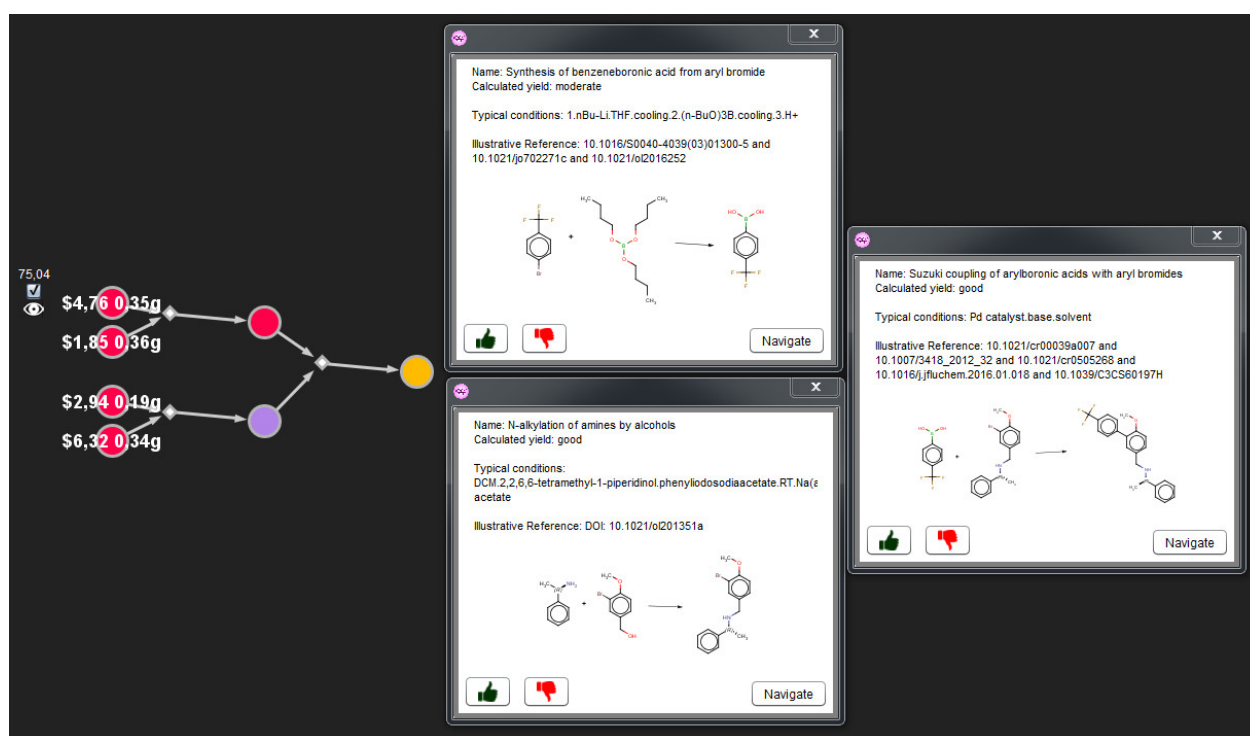


**Figure S5.** Details of top three synthetic pathways obtained for clofedanol. Paths are arranged in the order obtained with  $R_{\text{x}C} = \$1/\text{mmol}$ ,  $Y = 80\%$ .

## Section S6. Details of Chemica's syntheses of AMG641 with different *RxC-Y* settings.

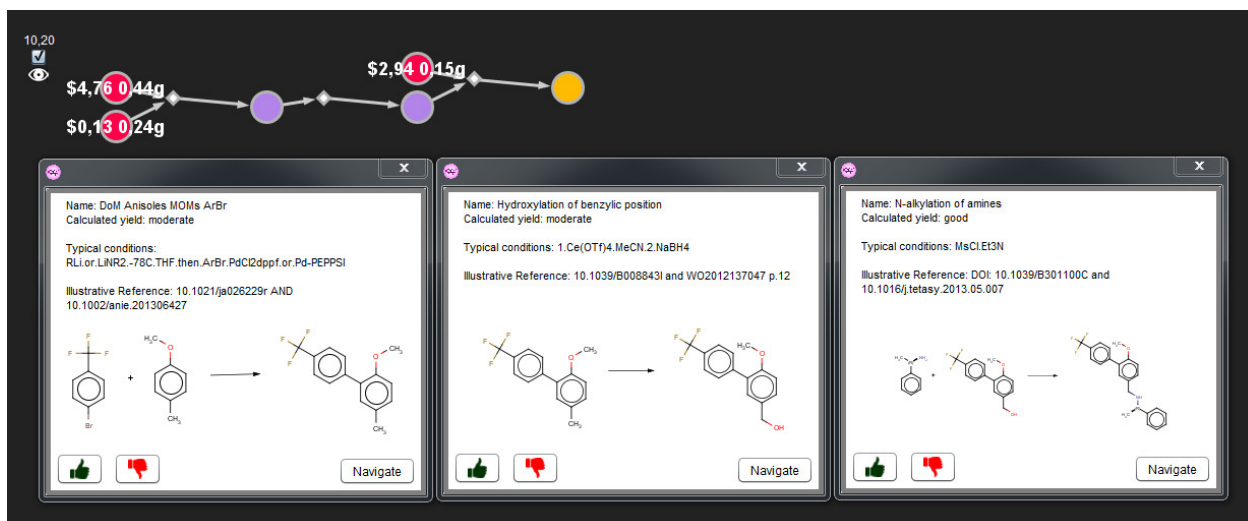


**Figure S6.** Details of the top-scoring synthetic pathway obtained for AMG641 with  $RxC = \$20/\text{mmol}$ ,  $Y = 99\%$  discussed in **Figure 8a**.

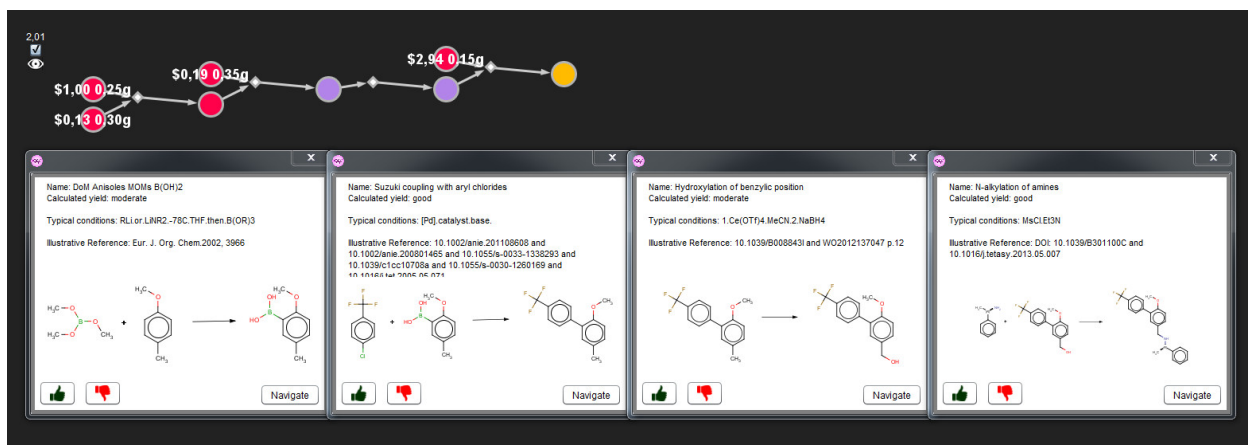


**Figure S7.** Details of the top-scoring synthetic pathway obtained for AMG641 with  $RxC = \$20/\text{mmol}$ ,  $Y = 80\%$  discussed in **Figure 8b**.



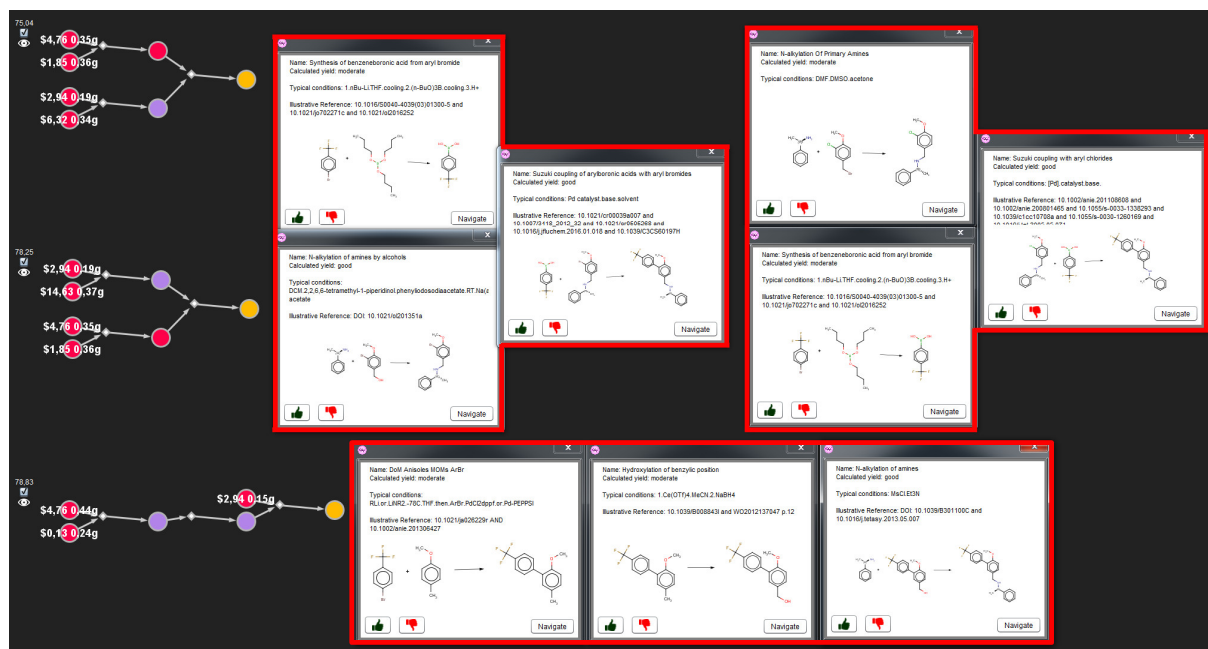


**Figure S8.** Details of the top-scoring synthetic pathway obtained for AMG641 with  $RxC = \$2/\text{mmol}$ ,  $Y = 80\%$  discussed in **Figure 8c**.

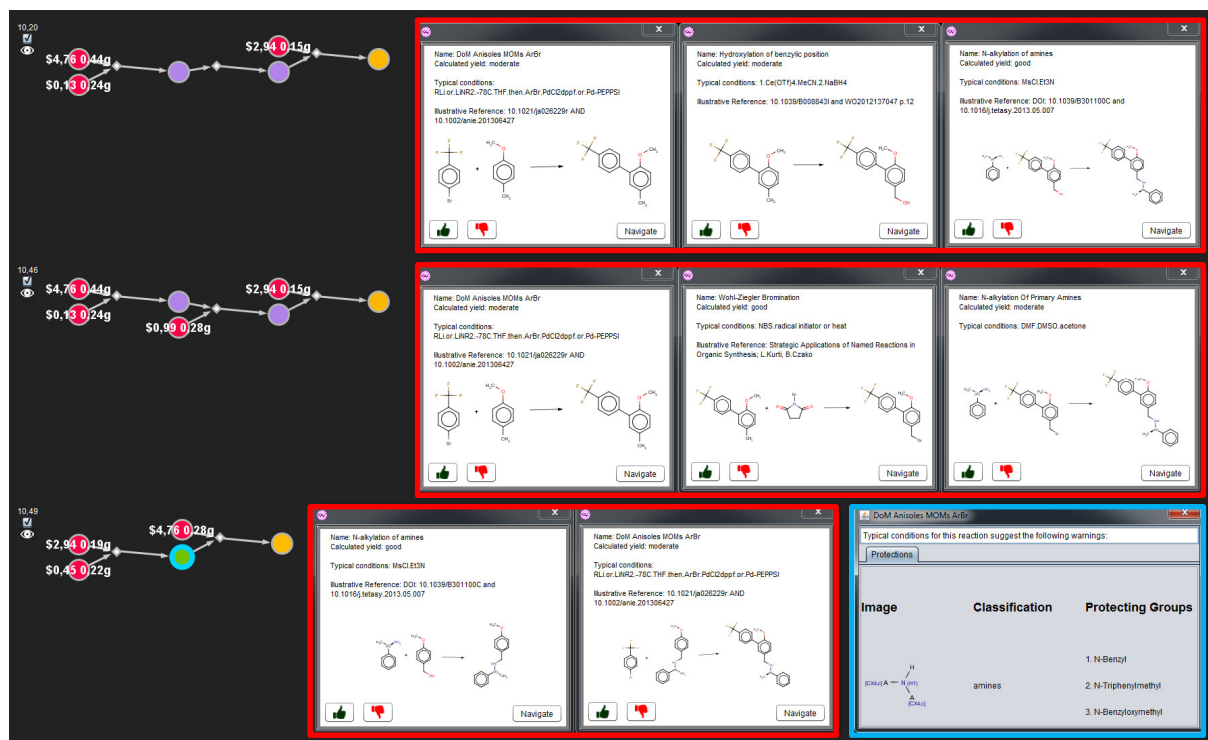


**Figure S9.** Details of the top-scoring synthetic pathway obtained for AMG641 with  $RxC = \$0.2/\text{mmol}$ ,  $Y = 80\%$  discussed in **Figure 8d**.

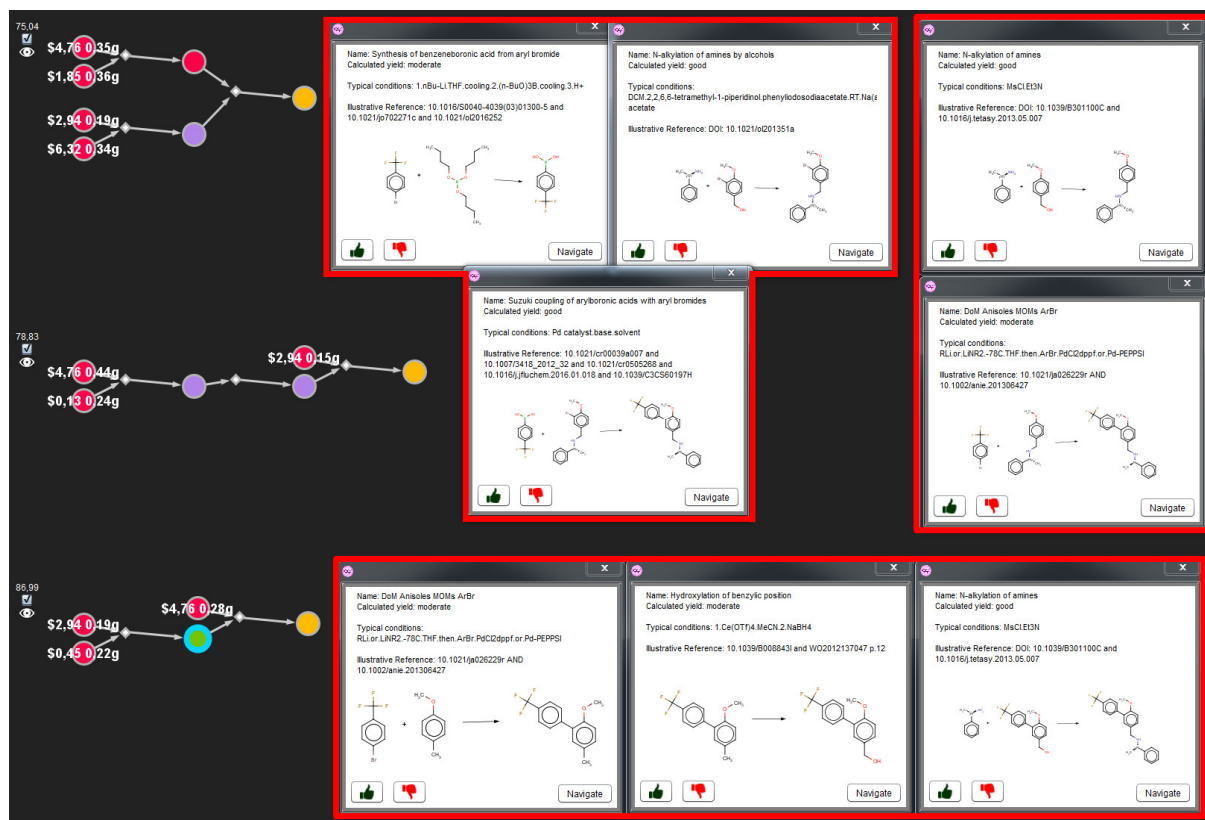
## Section S7. Details of Chematica's syntheses of AMG641 with different *P* settings.



**Figure S10.** Details of the top three synthetic pathways obtained for AMG641 with  $RxC = \$20/\text{mmol}$ ,  $Y = 80\%$ ,  $P = 0$  discussed in **Figure 9a**.

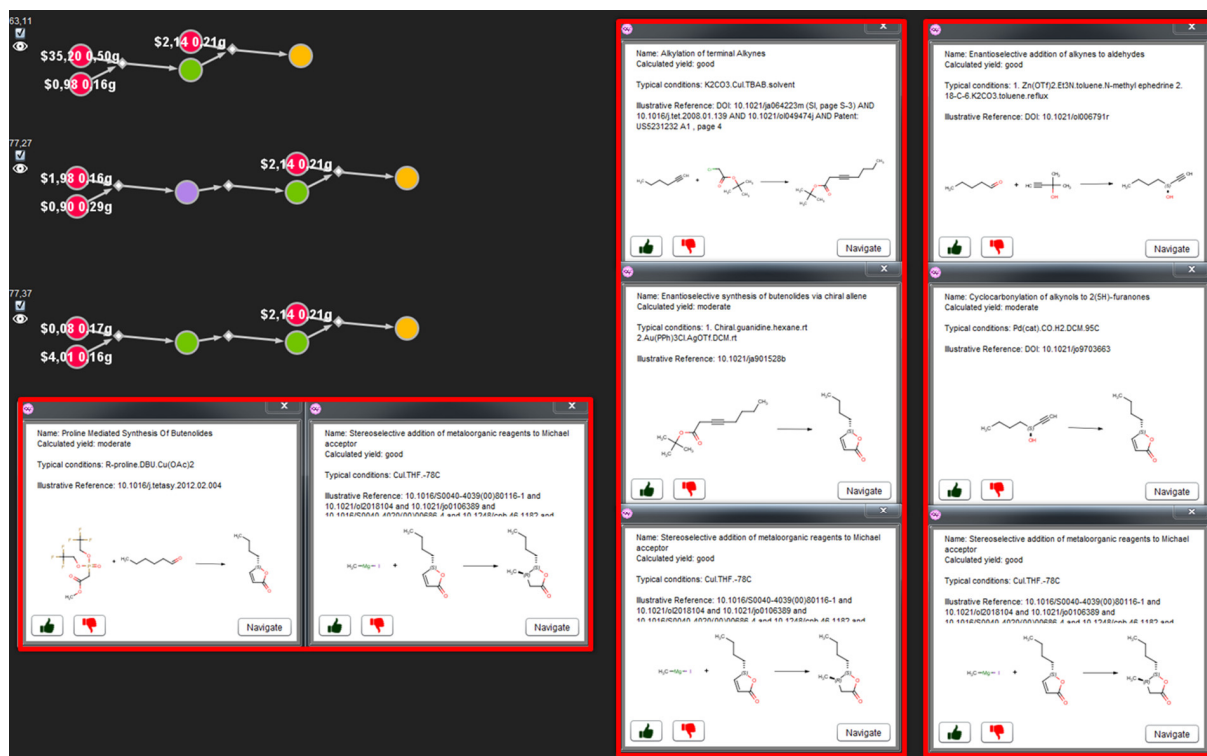


**Figure S11.** Details of the top three synthetic pathways obtained for AMG641 with  $RxC = \$2/\text{mmol}$ ,  $Y = 80\%$ ,  $P = 0$  discussed in **Figure 9b**. Chematica's proposed protecting group for the last step is shown in blue frame.

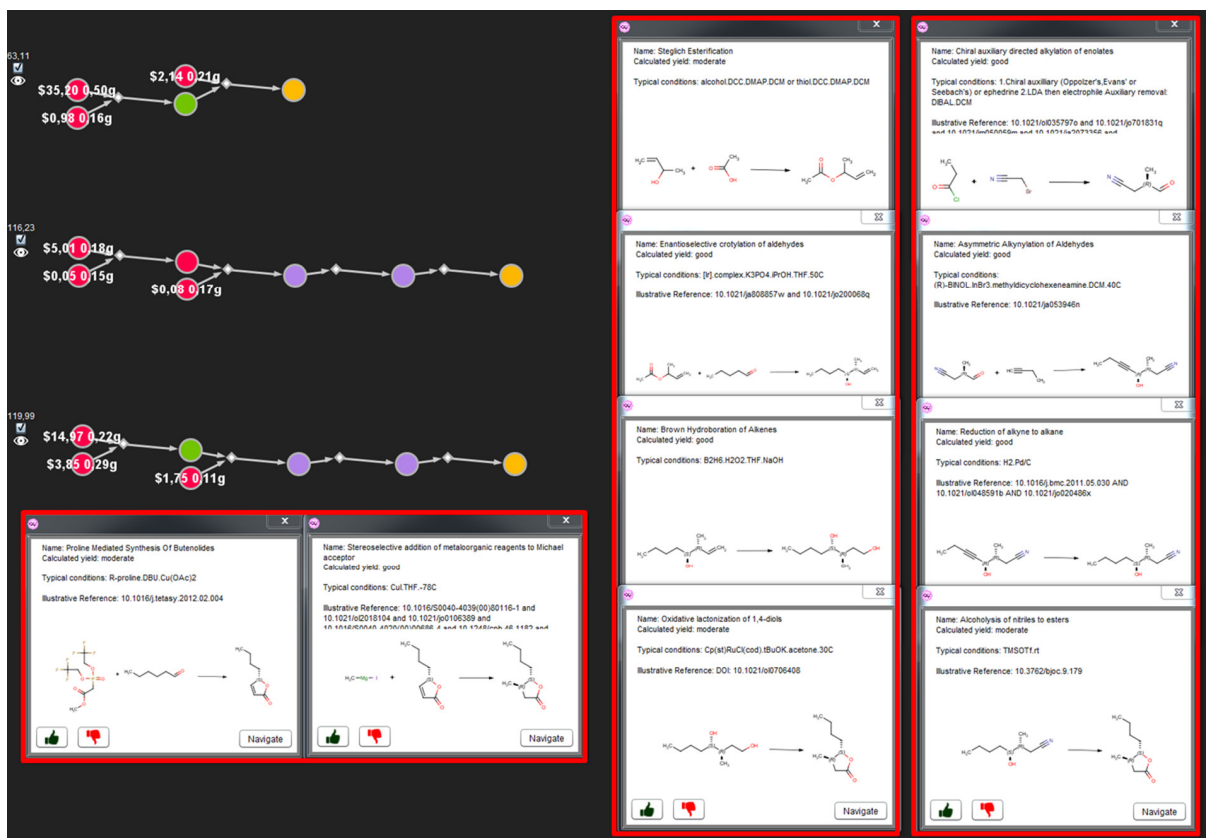


**Figure S12.** Details of the top three synthetic pathways obtained for AMG641 with  $Y = 80\%$ ,  $P = 10\ 000$  discussed in **Figure 9c,d**. Paths are arranged in the order obtained with  $RxC = \$20/\text{mmol}$ ,  $Y = 80\%$ .

## Section S8. Details of Chematica's syntheses of the whisky lactone with different $P$ settings.



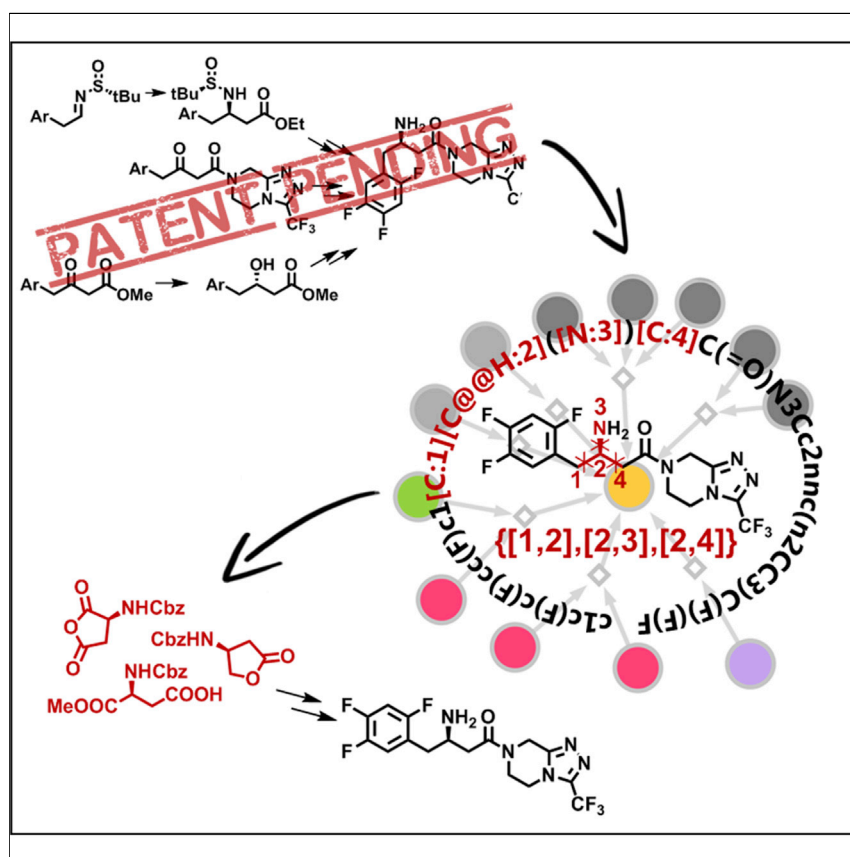
**Figure S13.** Details of the top three synthetic pathways obtained for whisky lactone with  $Y = 80\%$ ,  $P = 0$  discussed in **Figure 10a**.



**Figure S14.** Details of the top three synthetic pathways obtained for whisky lactone with  $Y = 80\%$ ,  $P = 10\ 000$  discussed in **Figure 10b**.

## Article

# Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways



By identifying and keeping track of key disconnections essential to patent-protected syntheses, a retrosynthetic computer program can autonomously design synthetic routes, “navigating around” previously published or patented approaches. This new modality of *in silico* synthetic design is illustrated by examples of patent-evading syntheses leading to three blockbuster drugs.

Karol Molga, Piotr Dittwald,  
Bartosz A. Grzybowski

piotr.dittwald@gmail.com (P.D.)  
nanogrzybowski@gmail.com (B.A.G.)

## HIGHLIGHTS

Computer autonomously designs syntheses avoiding desired key disconnections

By doing so, it navigates around patent-protected syntheses of blockbuster drugs

The algorithm can be useful to both identify and prevent such patent “bypasses”



Molga et al., Chem 5, 460–473  
February 14, 2019 © 2018 Published by Elsevier Inc.  
<https://doi.org/10.1016/j.chempr.2018.12.004>



## Article

# Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways

Karol Molga,<sup>1</sup> Piotr Dittwald,<sup>1,\*</sup> and Bartosz A. Grzybowski<sup>1,2,3,\*</sup>

## SUMMARY

By keeping track of lists of specific bonds one wishes to preserve, a computer program is able to identify the key disconnections used in the patented syntheses and design synthetic routes that circumvent these approaches. Here, we provide examples of computer-designed syntheses relevant to medicinal chemistry, in which the machine avoids “strategic” disconnections common to industrial patents and is forced to use different starting materials. The ability of modern retrosynthetic planners to navigate around patented solutions may have significant implications for the ways in which intellectual property related to multistep syntheses is protected and/or challenged.

## INTRODUCTION

Teaching the computer to plan multistep chemical syntheses leading to non-trivial targets has been an elusive goal for over five decades,<sup>1–6</sup> and it has only been recently that the first comprehensive validation of *in silico* synthetic predictions has been provided. Specifically, in Klucznik et al.,<sup>7</sup> we described how the Chematica program (see [Computational Methods](#) for a synopsis and Szymkuć et al.<sup>6</sup> and Klucznik et al.<sup>7</sup> for a detailed description) designed, without any human supervision, complete pathways leading to eight structurally diverse and medically relevant targets and how these pathways were subsequently executed in the laboratory, offering substantial improvements over previous approaches or providing the first documented routes to a given target. With such reassuring examples at hand, one can consider expanding the scope of automated retrosynthetic design modalities. One of the interesting and important possibilities is to challenge the machine to search for pathways significantly different than those already published or patented. In principle, this can be done by excluding specific intermediates or reaction types along the route (see Supplementary Section 6.4 in Klucznik et al.<sup>7</sup>). In practice, however, creating lists of “excluded” substances or reaction types is not only cumbersome for the software’s user but can also be of limited value—indeed, it does not prevent the machine from using intermediates chemically equivalent to those present in original routes or alternative methodologies resulting in identical retrosynthetic disconnections. Here, we describe a more convenient and robust approach in which the machine helps identify the target’s “key” bonds (whose disconnections are most common in a patent portfolio and also most structure simplifying) and performs synthetic planning with these bonds “preserved” to ultimately find qualitatively different, patent-circumventing synthetic plans. The computer can succeed in this task because it has a vast knowledge base of methodologies (in Chematica, ca. 60,000 reaction rules) that can substitute for reactions used in the patents and has access to diverse collections of starting materials (in the current study, ~200,000 commercial chemicals from

## The Bigger Picture

Although intelligent algorithms can now reliably design synthetic pathways leading to arbitrary targets, to date no methods are available to ensure that machine-designed syntheses are substantially different from prior approaches. Developing such methods could not only boost the general synthetic creativity but also be of practical value, enabling navigation around syntheses protected by seemingly unbreakable patents. Here, we describe a family of algorithms that can first identify the key disconnections underlying known and/or patented routes and then—by preserving the “key” bonds during *in silico* retrosynthetic planning—seek alternative, patent-evading syntheses, including those leading to blockbuster drugs. Although such algorithms will help inventors of new drugs enumerate and preemptively patent large numbers of viable synthetic routes, ensuring that their products are safe from generic copies, they may also ultimately yield more efficient and economical syntheses of pharmaceuticals, driving costs down—meaning that once expensive and exclusive medications are more readily available to all.

Sigma-Aldrich with prices per gram and >7 million literature-known substances). We illustrate the strength of this bond-preservation approach by designing patent-evading syntheses (each constructed within a few minutes) of three commercial drugs: linezolid (Figure 3), sitagliptin (Figure 4), and panobinostat (Figure 5). The computer-generated plans are chemically sensible and, in many cases, are substantiated by literature examples of similar methodologies used for the syntheses of other targets. Although Chematica is now a commercial product,<sup>8</sup> the bond-preservation algorithm we introduce here is generic and may benefit and extend the scope of other retrosynthetic software,<sup>3–5</sup> especially those that are trained on collections of reaction precedents and thus inclined to find solutions that closely resemble previously published syntheses.<sup>5</sup>

## RESULTS AND DISCUSSION

### Identification of Essential Disconnections in the Networks of Patented Syntheses

As long as there are just a few patented syntheses of a given target, identification of the “key” disconnection(s) on which these patents hinge can be achieved by inspection. In more realistic situations, however, syntheses of blockbuster drugs can be protected by tens to hundreds of patents claiming hundreds of synthetic steps and altogether forming quite complex reaction networks, such as those shown in Figure 1 for Pfizer’s linezolid (Figure 1A: patent network comprising 138 unique molecules and 156 unique reactions used in all patents 303 times), Merck & Co.’s sitagliptin (Figure 1B: 410 molecules and 469 unique reactions used 1,081 times), or Novartis’s panobinostat (Figure 1C: 28 molecules and 21 unique reactions used 42 times). Accordingly, to help guide the design of patent-evading routes, it could be useful to task the machine with identifying the bonds in the target whose formation is essential to the patents. By essential, we mean not only those bonds whose disconnections are most popular in the patents (e.g., some trivial protections or de-protections or the formation of some common substrates may be quite popular but non-essential chemically) but also those that offer the most structural simplification during synthesis. Syntheses in which at least some of such bonds were not disconnected could be qualitatively different from the patented routes (e.g., would start from substantially different substrates, use different methodologies, etc.).

Reactions from patented syntheses are easily available by querying repositories such as Reaxys or SciFinder; however, they come as individual steps rather than complete synthetic plans, and the atoms in them are not numbered. Therefore, our first step is to assign atom mappings to each of these reactions. We do so by using Chematica’s SMARTS reaction templates and our in-house atom-mapping codes (commercial mappers in ChemDraw or Marvin can also be used) to match both substrates and products and also unambiguously assign the reaction type. Such processed reactions are assembled into a network over which atom numbering is then unified, starting from the target and proceeding from retrons to synthons (see Supplemental Information Section S3 for a pseudocode). This unification of atom numbers is important to keep track within the network of which bonds (denoted by atom pairs) of the target are being disconnected or made in each specific reaction  $r$ .

Importantly, for each reaction  $r$  in the network, the set of disconnected bonds is assigned a complexity weight  $w_r$ , quantifying how essential this transform is. This weight is the product of two components: (1) one that is akin to Chematica’s chemical scoring function<sup>6,7</sup> and scores highly the reactions that disconnect the retron into equally sized synthons (or into multiple substrates, as in multicomponent reactions; see the Figure 1 legend for the formulas) and (2) one that scales with the number of

<sup>1</sup>Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland

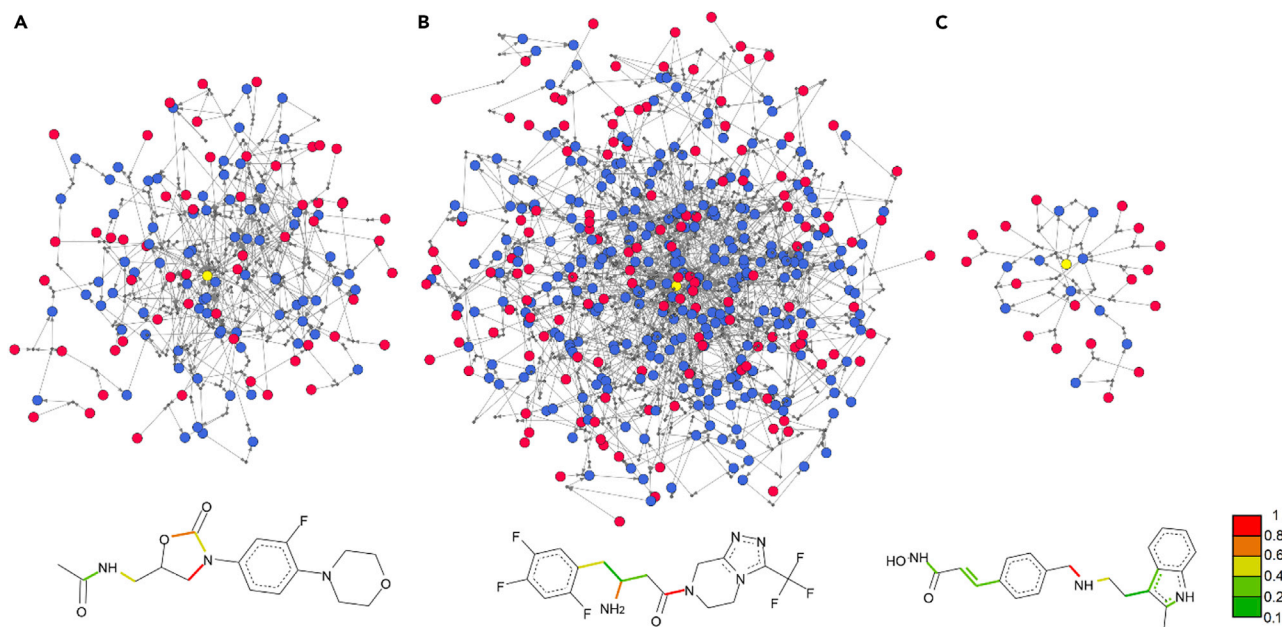
<sup>2</sup>IBS Center for Soft and Living Matter and Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulsan-gun, Ulsan 689-798, South Korea

<sup>3</sup>Lead Contact

\*Correspondence:  
piotr.dittwald@gmail.com (P.D.),  
nanogrzybowski@gmail.com (B.A.G.)

<https://doi.org/10.1016/j.chempr.2018.12.004>





**Figure 1. Network Representation of Patented Syntheses and the “Essential” Bonds Underlying the Patented Approaches**

The top row has networks unifying all reactions in the patents protecting the syntheses of linezolid (A), sitagliptin (B), and panobinostat (C). Color coding is as follows: yellow, target molecule; blue, intermediates; and red, starting materials. The bottom row shows the target molecules with bonds color coded according to their importance in the patented routes. Scores closer to unity (red color) indicate the most essential bonds; bonds with scores < 0.1 or with no patent-reported disconnections are not colored. These scores were calculated for each bond by summing and normalizing the complexity weights  $w_r$  of all reactions in which a particular bond was disconnected. The individual weights reflected the structural simplification a given reaction offered as well as number of atoms common to this reaction’s product and the network’s ultimate target,  $w_r = \text{simplification}(r) \cdot \text{common\_atoms}(r)$ . The

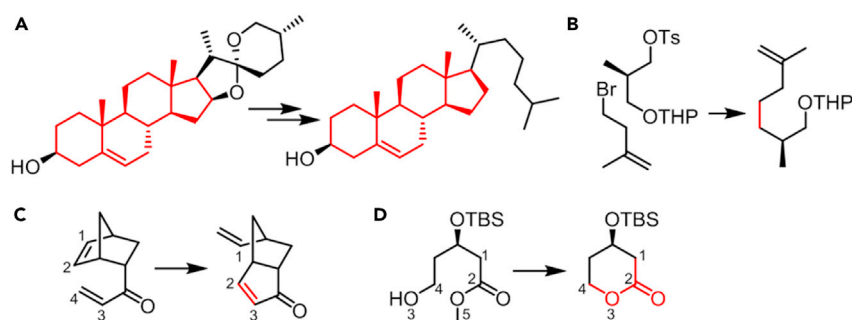
“simplification” term was defined akin to Chematica’s simplest scoring function,  $\text{simplification}(r) = \frac{\sum A_{\text{prod},i}^n}{\sum A_{\text{substr},i}^n}$ , where A stands for the number of atoms in

the product(s)/substrate(s), and power  $n$  is greater than 1 (here, as in one of Chematica’s standard scoring functions,  $n = 3$ ). This measure promotes “central” disconnections into same-size fragments. For instance, if the number of atoms in the reaction’s (sole) product is  $A_p = 10$  and the two substrates each have five atoms,  $A_{s1} = A_{s2} = 5$ , then the value of the simplification factor is  $10^3 / (5^3 + 5^3) = 4$ . For less central disconnections, say, 7:3, the value is always smaller (e.g.,  $10^3 / (7^3 + 3^3) = 2.7$ ). Note: for compounds with large numbers of rings or stereocenters, the simplification factor might need to take these measures into account (in Chematica’s scoring functions, this can be expressed by linear combinations of  $A^n$  (or SMILES\_length<sup>n</sup>), number of rings, and number of stereocenters terms; for further details on scoring functions, see Szymkuć et al.<sup>6</sup> and Klucznik et al.<sup>7</sup>).

atoms common to the retron of a specific reaction and the ultimate target of the entire synthesis (this measure gives higher weight to disconnections performed on more advanced intermediates). We note that for each of the target’s bonds, the algorithm sums the weights  $w_r$  of reactions in which this bond was disconnected, normalizes these sums, and ultimately color codes the final scores into the target’s bonds, as in the structures shown in the bottom row of Figure 1. The closer the score is to unity, the more essential the formation of a particular bond is to the patented approaches. Consequently, to maximize our chances of finding patent-evading routes, these “key” bonds or even the structural motifs they belong to are recommended to be preserved (i.e., never disconnected or altered) during Chematica’s synthetic planning.

### Bond Preservation Algorithm to Accompany Retrosynthetic Planning

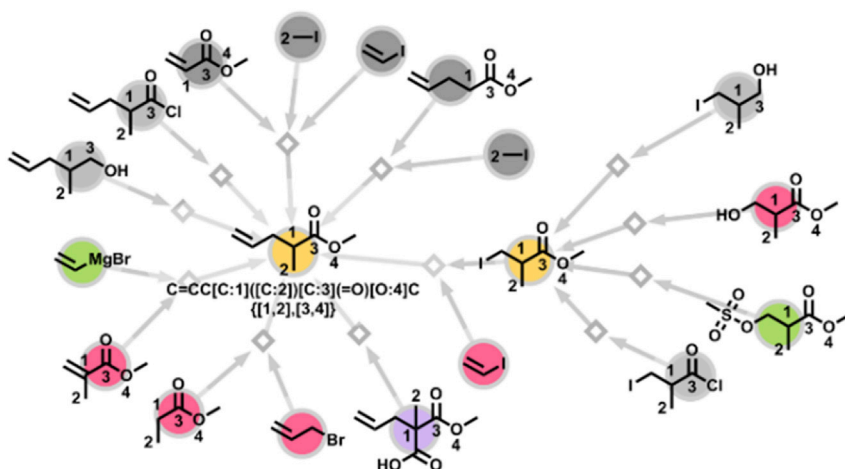
At first sight, tracking of not-to-be-altered motifs might be implemented easily by checking whether the desired substructure remains intact in the synthons of each reaction the computer considers. However, this approach works well only when the substructures are larger and unique (Figure 2A). With smaller motifs or individual bonds, one rapidly runs into problems associated with their non-uniqueness (e.g., which of the seven C–C bonds in Figure 2B should really be preserved). In addition,



**Figure 2. Shortcomings of Naive Preservation of Structural Motifs**

A naive algorithm that inspects if the specified not-to-be-disconnected motifs (marked in red) are preserved in the synthons (A) works well if these motifs are large and unique but (B–D) fails when (B) several non-unique bonds are present (here, seven single C–C bonds in the main skeleton<sup>9</sup>) or (C and D) when actual changes of the bonds within motifs are not detected. For instance, in (C), the red-colored double bond C<sub>2</sub>=C<sub>3</sub> might seem to be preserved, but in reality it is disconnected during olefin metathesis,<sup>10</sup> as indicated by the atom numbers. Similarly, in (D) the ester motif appears intact but in reality is not preserved during lactonization reaction.<sup>11</sup>

a motif might appear to remain intact but, in reality, changes during the reaction (see examples of olefin metathesis in Figure 2C or lactonization in Figure 2D). To avoid such problems, we number the atoms within the motif(s) to be preserved and then, during all generations of retrosynthetic planning, keep track of the pairs of atoms corresponding to bonds not to be disconnected. Importantly, while doing so, we minimize search times and memory usage by merging the solution space for identical synthons along different putative routes. Down to technical level (see also Supplemental Information Sections S1–S3), the user's graphical input of the target molecule, *t*, with "preserved" bonds marked (Figure S1), is first translated into Extended Molfile format and then into a SMILES string. Atom labels are stored as SMILES atom index properties, and a list of pairs indicating protected bonds is created and stored as a "bond set" *B*(*t*) (see text below the central node corresponding to the target in Figure 3). As the retrosynthetic search commences, the matching reaction templates are applied, and the first generation of synthon sets is created. For each candidate retron-to-synthon(s) transformation,  $r \rightarrow s_1, s_2, \dots, s_N$  (where  $r = t$  in the first generation), the labels of marked atoms are propagated from the retron to the synthons. The algorithm then checks whether the set of bonds marked in the target, *B*(*t*), is preserved among the synthons. Specifically, defining the subset of these bonds in a synthon *s<sub>i</sub>* as *B*(*s<sub>i</sub>*), we require that  $B(r = t) = B(s_1) \cup \dots \cup B(s_N)$ , where  $\cup$  is a union set operator. Only reactions fulfilling this condition are further considered and evaluated (e.g., in Chematica, by its scoring functions quantifying simplification of the structure, assigning penalties for protection requirements, non-selectivities, etc.; for details, see Szymkuć et al.<sup>6</sup> and Klucznik et al.<sup>7</sup>). The most promising options are further "expanded" into subsequent generations for which the same procedure of atom labeling is used and to which the same criteria of bond-set conservation are applied. During consecutive expansions, the searches strive to keep the search space as compact as possible—for instance, in a relatively frequent scenario where the same synthons are found within different pathways, they are stored as one molecule within the search graph. This being said, if identical synthons contain different marked bonds, they can possibly have different retrosynthetic histories and are thus stored as separate entities distinguished not by the molecular structure but also by the list of "protected" bonds (see Figure S2). The search continues until all synthons are commercially available or, if the user chooses so, described in literature. All in all, the algorithm has the following desired



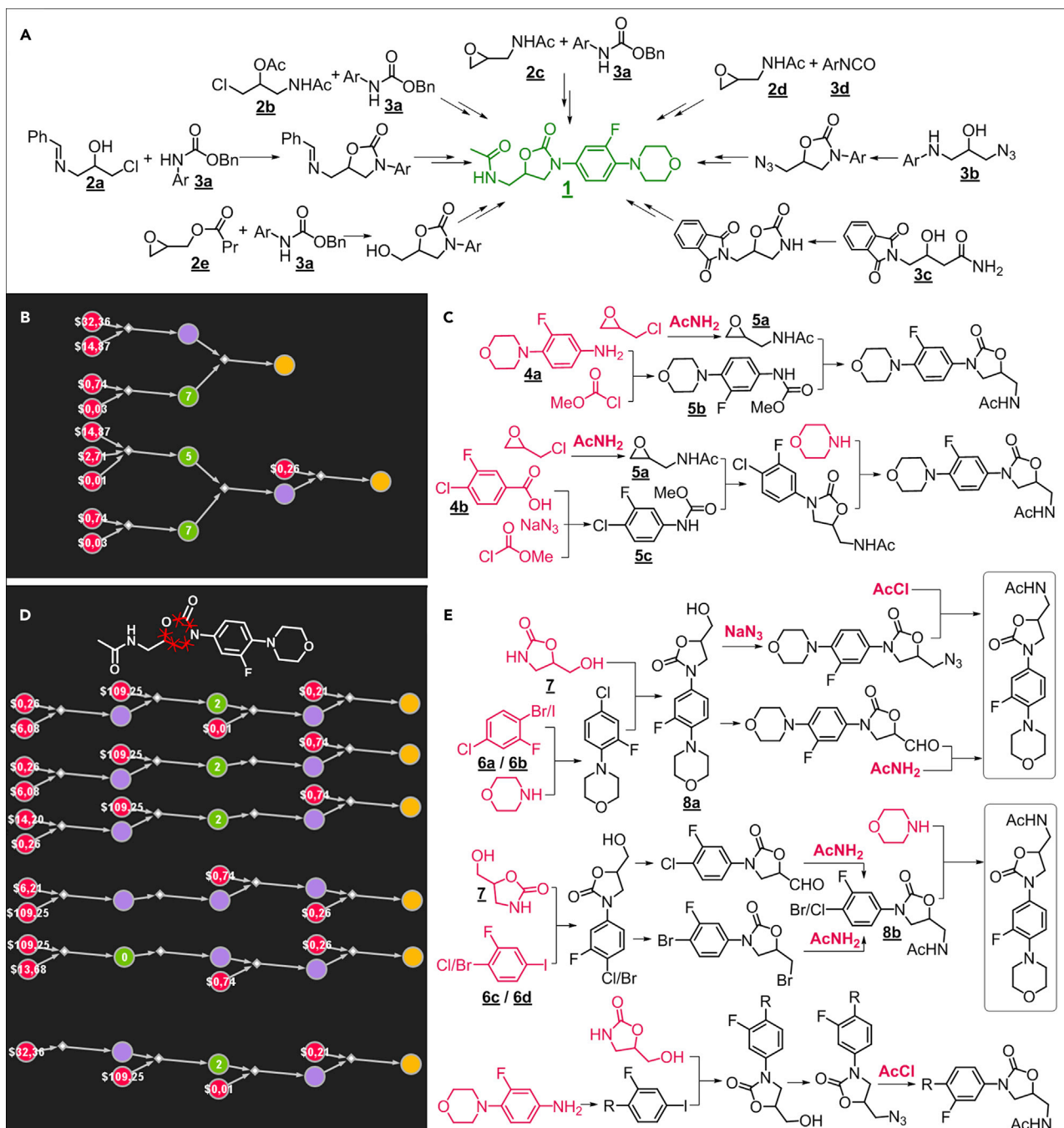
**Figure 3. Illustration of the Retrosynthetic Search Algorithm with Bond and Motif Preservation**

Numbering of atoms in the target molecule  $t$  is encoded in its SMILES (here, C=CC[C:1]([C:2])[C:3](=O)[O:4]C) and the accompanying set of atom pairs (denoting bonds not to be disconnected; here,  $B(t) = \{[1,2], [3,4]\}$ ). Each reaction operation (represented by an open diamond) generates a set of synthons (circles). If the union of bond sets over the synthons is different than in the target, then such a reaction candidate is removed from further consideration (nodes colored in gray). The remaining synthon nodes can be further expanded (e.g., second-generation expansion on the right). The searches terminate when reaching buyable chemicals (red nodes) or, if allowed by the user, substances whose syntheses have already been published (green nodes). The violet node denotes a new or unknown substance and cannot be a stop point for the search. For realistic, more complex targets, the number of expanded “spiders” is in the thousands. From such graphs, the program selects and ranks a user-specified number (usually 50) of qualitatively different (i.e., without trivial variations), viable pathways. Such pathways are displayed as illustrated in Figures 4, 5, 6, 7, and S7–S15.

characteristics: (1) it preserves the not-to-be-altered bonds along entire pathways it identifies; (2) it can preserve motifs that are disjoint in the target—in such a case, at a given generation, more than one bond set  $B(s_i)$  is not empty, meaning that the motifs are split between different synthons; and (3) it can be implemented to prevent either complete bond disconnections or changes in bond order (the latter, by adding bond-order labels to atom labels). For the pseudocode of the algorithm, see Supplemental Information Section S3.

### Syntheses of Linezolid

In our first example, we charged Chematica with finding pathways that would navigate around multiple patented routes—visualized as a network in Figure 1A and listed in Figure S5—leading to Pfizer’s linezolid (tradename Zyvox), antibiotic, 1. As we have seen in Figure 1A, the key disconnections in the patents are within the oxazolidinone ring, and this ring is chosen as a motif to be preserved. In the patented syntheses, the oxazolidinone ring is formed via (1) base-induced cyclization of hydroxylin 2a or 2b or of epoxide 2c, 2d, or 2e with *N*-aryl carbamate 3a or isocyanate 3d; (2) cyclization of 3b; or (3) Curtius rearrangement of 3c (Figures 4A and S5 for complete list). Without any bond-preservation constraints imposed on the target, Chematica proposed similar plans, where the top-scoring pathways (Figures 4B and 4C) construct oxazolidinone via the opening of a known oxirane 5a with carbamate 5b (prepared from appropriate amine 4a) or 5c (prepared via Curtius rearrangement of benzoic acid 4b) and subsequent *N*-arylation of morpholine<sup>12,13</sup> (Figures 4B and 4C; for further reaction details, see Chematica’s raw output in Figure S7). In contrast, after specifying the bonds within the oxazolidinone ring as



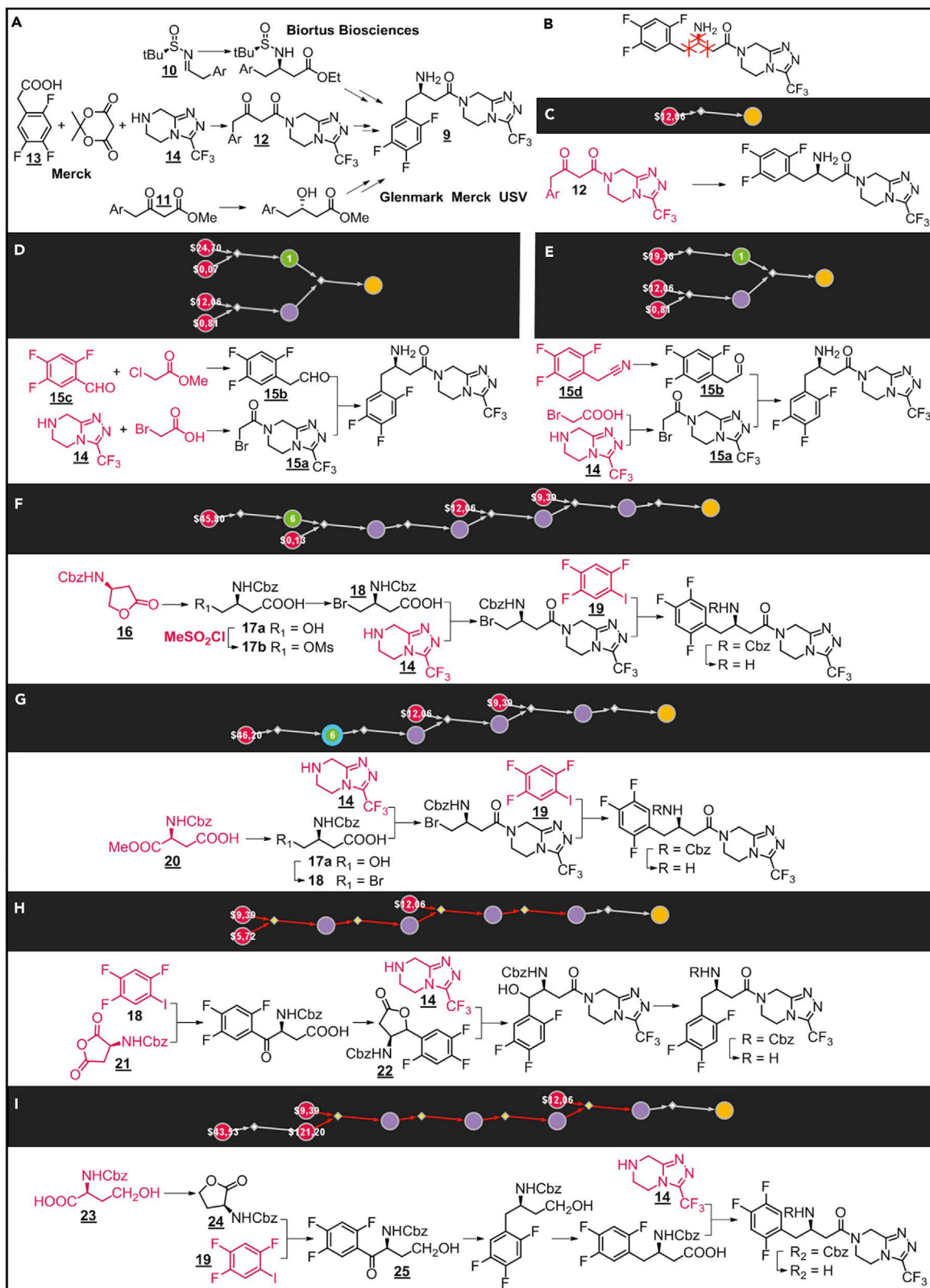
**Figure 4. Patented versus Patent-Evading Routes to Linezolid**

Figure360> For a Figure360 author presentation of Figure 4, see <https://dx.doi.org/10.1016/j.chempr.2018.12.004#mmc3>.

(A) Most patent-protected syntheses of linezolid rely on the formation of the oxazolidinone ring.

(B and C) Similar solutions identified by Chematica without any bond-preservation constraints.

(D and E) Qualitatively different synthetic plans found by the program when the bonds within the oxazolidinone rings are preserved. In the miniatures of Chematica's pathways, color coding is as in Figure 3. In addition, the numbers in the red nodes are prices from Sigma-Aldrich catalog (in \$US/g), and the numbers in the green nodes are synthetic popularities (i.e., the number of literature-reported syntheses in which a particular molecule was used as a substrate). The total search time for all syntheses shown in the figure was ~5 min. For details of Chematica's pathways, including suggested reaction conditions, illustrative references, and more, see Chematica's raw output in Section S6.



**Figure 5. Patented versus Patent-Evading Routes to Sitagliptin**

(A) Key disconnections in patent-protected syntheses of sitagliptin.

(B) Bonds surrounding the stereocenter marked as not to be broken during synthesis planning.

(C–E) Chematica's solutions (with no constraints imposed) are similar to the patented ones.

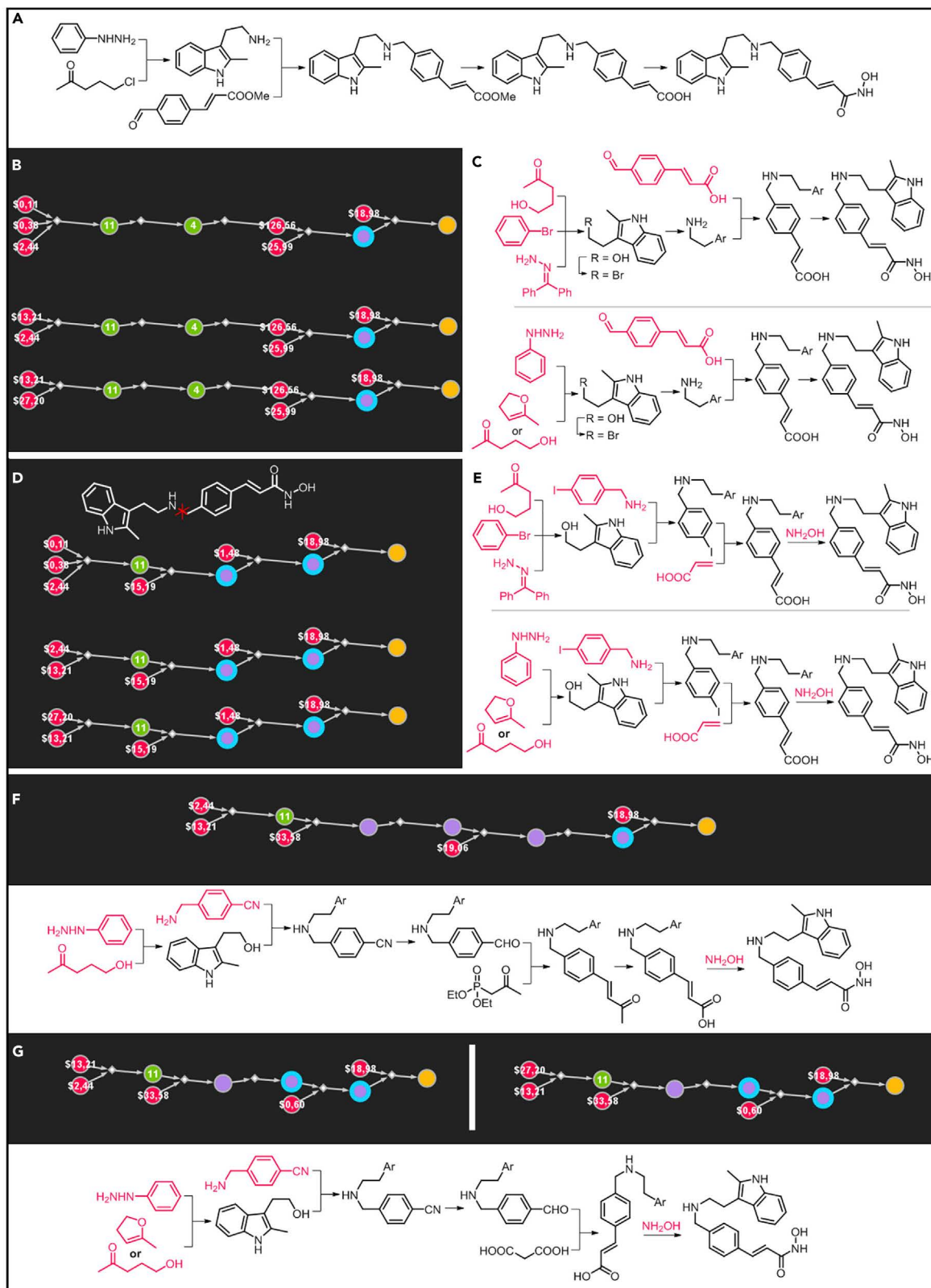
(F–I) Top-scoring syntheses found by the program with bond-preservation constraints from (B). Pathways (H) and (I) capitalize on Chematica's two-step strategic sequences (arrows colored red), in which the first step in the retrosynthetic direction does not offer any immediate structural simplification but sets up the scene for a subsequent complexity-reducing step (see Supplementary Section 7.3 in Klucznik et al.<sup>7</sup>). All searches shown in the figure took ca. 15 min of computer time. For further details, see Chematica's raw output in [Supplemental Information Section S7](#).

not to be broken (bonds marked red in the target structure in [Figure 3D](#); also see [Video S1](#)), the algorithm is forced to avoid the abovementioned key steps from the patents, and its three top-scoring solutions (top portion of [Figures 4D](#) and [4E](#)) start from commercially available halobenzenes **6a** and **6b** undergoing copper-catalyzed amination with morpholine under Zhang's or Nolan's conditions.<sup>12,13</sup> Subsequent arylation of the commercially available **7** with remaining, less-reactive aryl chloride<sup>14,15</sup> yields the desired *N*-aryl oxazolidinone **8a**. The four-step sequence is completed by either (1) formation of the azide under Mitsunobu conditions and subsequent one-pot reduction and acylation or (2) oxidation of the alcohol to the aldehyde followed by reductive amidation.<sup>16</sup> Another family of Chematica's synthetic plans (middle part of [Figures 4D](#) and [4E](#)) utilizes an "opposite" reactivity pattern whereby the more reactive aryl iodide **6c** or **6d** is allowed to react with **7**. Subsequent (1) conversion to alkyl bromide and reaction with acetamide anion or (2) oxidation to aldehyde and reductive amidation lead to **8b** used in the Buchwald-Hartwig amination of morpholine to complete the synthesis. Finally, the solution shown in the lower portion of [Figures 4D](#) and [4E](#) starts from the commercially available fluoroaniline. Conversion via diazonium salt to iodoarene (incidentally, previously obtained in 85% yield and used for functionalization of cytoxzone<sup>17</sup>) followed by *N*-arylation of **7**, formation of azide, and conversion to acetamide yields the product in four steps. We note that although Sigma-Aldrich' catalog price of **7** (>\$100/g) used as a common intermediate in Chematica's plans is rather high, this compound can be prepared in one step (not claimed in any patent) from orders-of-magnitude less expensive 3-amino-1,2-propanediol and diethyl carbonate in 60% yield according to Steckhan's procedure.<sup>18</sup>

**Chiral-Pool-Oriented Syntheses of Sitagliptin**

In the second example, the target was sitagliptin **9**, which is an oral antihyperglycemic agent marketed by Merck & Co. as Januvia, with annual sales in 2017 approaching \$4 billion. The current patented approaches ([Figures 5A](#) and [S6](#)) rely on (1) addition of enolate<sup>19</sup> derived from bromoester to chiral sulfinylimine **10** ([Figure 5A](#), top), (2) enantioselective reduction<sup>20–22</sup> of ketoester **11** ([Figure 5A](#), bottom), or (3) formation of ketoamide **12** via one-pot condensation of meldrum acid with arylacetic acid **13** and secondary amine **14** with subsequent reductive amination setting the stereocenter with chiral rhodium catalyst,<sup>23</sup> engineered transaminase,<sup>24</sup> or a chiral auxiliary.<sup>25</sup> Without any additional constraints, Chematica suggests ([Figures 5C–5E](#)) solutions closely mirroring these approaches, relying either on the stereoselective reduction of commercially available ketoamide **12** or on the addition of enolate derived from amide **15a** (prepared in one step from **14** and bromoacetic acid) to chiral imine derived from aldehyde **15b**, which, in turn, is accessible via Darzensen-type homologation<sup>26</sup> of benzaldehyde **15c** or from nitrile **15d**.

Analysis from [Figure 1B](#) suggests that the bonds one might consider preserving are either the amide bond or the bonds next to the stereocenter. "Locking" all of them at the same time is problematic because it simply does not leave any realistic synthetic choices open—indeed, with such maximalist constraints, Chematica does not find



**Figure 6. Patented versus Patent-Evading Routes to Panobinostat**

(A) Patented synthesis of panobinostat relies on reductive amination of tryptamine.

(B and C) Similar solutions identified by Chematica without any bond-preservation constraints.

(D and E) Qualitatively different synthetic plans found by the program when the key disconnection in the patented route (bond marked red) was forbidden.

(F and G) Synthetic plans obtained when the same key disconnection was forbidden and, additionally, reactions requiring Pd catalysts were penalized. All searches shown in the figure took ca. 15 min of computer time (longest individual search was 8 min). For further details, see Chematica's raw output in [Supplemental Information Section S8](#).

any sensible syntheses. When the amide bond alone is locked, the program follows Merck & Co. and related patents. Accordingly, we decided to preserve the bonds surrounding the stereocenter (Figure 5B), in effect forcing Chematica to design routes that would source the synthesis from chiral starting materials. With this constraint, the algorithm returns several plausible and qualitatively different routes. For example, the top-scoring pathway in Figure 5F sources the stereocenter from protected lactone 16. In the first step, hydrolysis under basic conditions yields the hydroxyacid 17a. Sulfonylation in the presence of unprotected carboxylic acid<sup>27</sup> prepares mesylate 17b, which is then converted to bromocarboxylic acid 18. Subsequent amidation with triazolamine 14 gives alkyl bromide participating in Pd-mediated Kumada-type coupling, after which Cbz is removed to yield the target. Interestingly, this approach is virtually identical with the Esteve Quimica's synthesis of sitagliptin save that the order of two steps is exchanged (see Bartra-Sanmarti et al.<sup>28</sup> and Figure S6E). In another, related variant (Figure 5G), Chematica begins with *N*-Cbz-protected  $\alpha$ -methyl (*S*)-aspartate 20. Chemoselective reduction of the ester and subsequent Appel reaction (note blue halo indicating the need for protection of carboxylic acid<sup>29</sup>) prepare the bromide 18, which then follows steps as in Figure 5F. Interestingly, a similar approach relying on the alkylation of a Grignard reagent with a bromide derived from protected aspartate was, in fact, demonstrated experimentally by Liu et al.<sup>30</sup> In yet another route in Figure 5H, the program begins with Cbz-protected aspartic anhydride 21, which is opened regioselectively<sup>31</sup> with a Grignard reagent derived from 19. Subsequent tandem reduction-lactonization<sup>32</sup> gives lactone 22 opened with amine 14. The target molecule is then obtained after deoxygenation of benzylic alcohol and deprotection. Finally, the pathway in Figure 5I commences from commercially available *N*-Cbz (*S*)-homoserine 23. Formation of lactone 24 eliminates the need for OH protection in the reaction<sup>33</sup> with Grignard reagent derived from 19. Deoxygenation of benzylic ketone 25, oxidation of alcohol to carboxylic acid, coupling with 14, and deprotection follow, yielding sitagliptin in six steps. Of note, similar approaches (Figure S6) relying on the addition of a Grignard reagent to derivatives of aspartic acid<sup>34</sup> (either Weinreb amide, or mixed anhydride, or oxazolidinone) and subsequent deoxygenation of benzylic ketone were demonstrated experimentally.

**Syntheses of Panobinostat with Additional "Process" Constraints**

In the third and final example, Chematica was challenged with designing routes to panobinostat (trade name Farydak), which is marketed by Novartis and approved for treatment of multiple myeloma. The current patented approach<sup>35</sup> starts with the formation of 2-methyltryptamine in tandem Fischer indolization- $S_N2$  reaction, followed by reductive amination and conversion of methyl ester to the desired hydroxamic acid (Figure 6A). Without any additional constraints, Chematica produces similar plans (Figures 6B and 6C). The synthesis of 2-methyltryptamine starts with (1) three-component Buchwald synthesis of indole<sup>36</sup> with benzophenone hydrazone as phenylhydrazine surrogate, (2) Fischer indolization of 5-hydroxypentan-2-one, or (3) Fischer-type<sup>37</sup> indolization of 5-methyl-2,3-dihydrofuran. Hydroxyl derivative is converted to tryptamine via Appel reaction and Gabriel amination. Subsequent

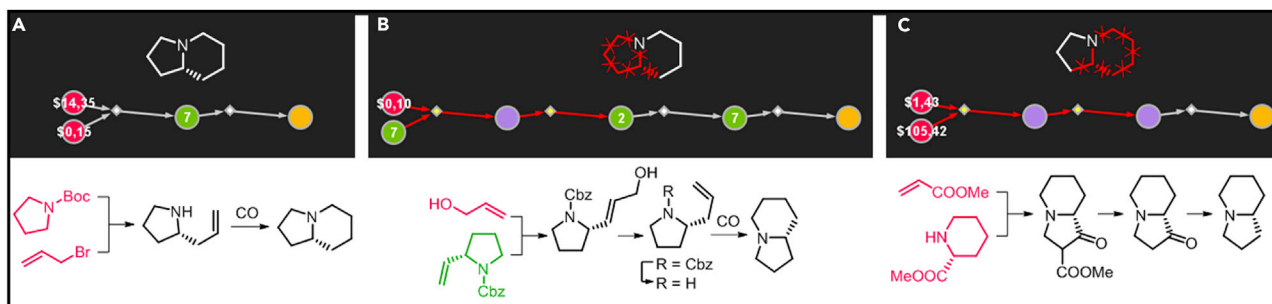


reductive amination with 4-formyl cinnamic acid and formation of hydroxamic acid follow the Novartis' route to panobinostat. Note that in the last step, the program correctly warns the user about potential selectivity issues and the need to protect nucleophilic secondary amine (blue halo in the pathway miniatures); in Novartis' route, this problem is circumvented by brute force via excess hydroxylamine.

To navigate around such paths, the central C–N bond (marked red in Figure 6D) is preserved—this bond is also the top-scoring suggestion from the patent-network analysis in Figure 1C. With this additional constraint, the top-scoring syntheses (Figures 6D and 6E) again start with the formation of 3-hydroxyethylindole but subsequently, instead of making tryptamine, couple it with *p*-iodobenzylamine (either via mesylate or via one-pot oxidation-reductive amination<sup>38</sup>). Heck coupling with acrylic acid<sup>39</sup> (notice the blue halo indicating the need for protection of secondary alkylamine interfering with palladium catalyst) introduces the side chain, and reaction with hydroxylamine gives the target. Although chemically correct, pathways from Figure 6D might be problematic in industrial settings as a result of the use of Pd catalysts because both the catalyst and supporting phosphine ligands are expensive and occasionally difficult to remove from the products (on the other hand, it should be noted that the use of Pd catalysis is becoming increasingly popular in pharmaceutical industry because of the efficiency of bond disconnection offered by this metal). Accordingly, in our last analysis, we combined the do-not-cut-bonds constraint with Chematica's "avoid" option whereby the user can eliminate reaction types that contain certain keywords in the reaction record (see Section S6.4 in Klucznik et al.<sup>7</sup>). With the "Pd" and "palladium" keywords to be avoided, the top-scoring syntheses are those shown in Figures 6F and 6G. In the route in Figure 6F, hydroxyethylindole is now prepared via the Fischer process rather than Buchwald synthesis. Heck coupling is also no longer used to construct the cinnamic acid side chain. Instead, hydroxyethylindole activated as a mesylate is used to alkylate commercially available 4-cyanobenzylamine. Subsequent reduction of nitrile to aldehyde and Horner-Wadsworth-Emmons olefination yields the unsaturated ketone oxidatively degraded to acid,<sup>40</sup> which is then converted to the target molecule. The routes in Figure 6G share similar initial steps and are shorter but are scored lower on account of two protections needed (blue halos in pathway miniatures; for details, see Figure S14). The cinnamic side chain is constructed directly from appropriate aldehyde and malonic acid under Knoevenagel-Doebner conditions.<sup>41</sup>

## Conclusions

To sum up, the above examples illustrate how the combination of automated retrosynthetic design with bond preservation modality enables navigation around patented routes. Another aspect of the algorithm we have not focused on—but that is already available with the methodology described—is that by marking "not-to-be-cut" bonds defining larger structural fragments, the searches can be made to seek and terminate only in certain desired families of substrate scaffolds (Figure 7). The fact that some of the bond-constrained routes proposed by Chematica actually mirror patent-circumventing syntheses that were carried out experimentally (e.g., Figure 5G versus Pan et al.<sup>30</sup> and Figure 5I versus Liu et al.<sup>34</sup>) or syntheses in which entire structural motifs were kept intact (Figure 7A versus Coldham and Leonori<sup>42</sup> and Figure 7C versus Clemo and Ramage<sup>43</sup>) implies that our approach emulates the ways in which human experts think about such problems. In a broader context, we believe that the methods we outlined not only will be useful in research practice but also will have a significant impact on how the intellectual property related to synthetic routes is protected or challenged—the former by harnessing the computer's power to enumerate and preemptively patent very large numbers



**Figure 7. Syntheses of (R)-Coniceine Starting from User-Preferred Chiral-Pool Scaffolds**

(A) When no bonds are protected, Chematica suggests a two-step synthesis starting from stereoselective lithiation-allylation of *N*-Boc pyrrolidine, followed by deprotection and Rh-mediated hydroaminomethylation. This sequence is virtually identical with Coldham's approach.<sup>42</sup>

(B) When the five-membered ring and bonds adjacent to the stereocenter are preserved, Chematica sources the synthesis from a known *N*-Cbz proline derivative. Olefin cross-metathesis followed by reductive transposition of allylic alcohol yields, after deprotection, the same intermediate as in (A) and then the target. To the best of our knowledge, this target has not been synthesized with this type of bond-preservation constraint.

(C) Preserving the six-membered ring and bonds adjacent to the stereocenter forces the algorithm to use a more expensive methyl pipecolate as the starting material (still, even this synthesis is shorter than the majority of routes to (R)-coniceine described in the literature). Initial tandem Michael condensation with methyl acrylate followed by decarboxylation of the obtained ketoester and reduction of ketone yield the target molecule. We note that this sequence resembles one of the first syntheses of coniceine reported by Clemo and Ramage.<sup>43</sup>

For further details of the synthetic plans, see [Supplemental Information Section S9](#).

of viable synthetic routes and the latter by navigating around patented methodologies (especially those leaving "an analytical fingerprint" in the target; for interesting discussion of this and related patent-law topics, see Pohl,<sup>44</sup> Heuer,<sup>45</sup> and Lowe<sup>46</sup>). Of course, when applied in industrial settings, the catalogs of the starting materials should be updated to reflect large-scale pricing model and suppliers of a specific company; in practice, such a replacement is a trivial substitution of one text file supporting the program's search algorithms and chemical knowledge base. In addition, it might be interesting to supplement the algorithm we discussed here with other metrics of process efficiency<sup>47</sup> or to link the searches with lists of environmentally hazardous materials that should be avoided during synthetic planning.

## EXPERIMENTAL PROCEDURES

### Computational Methods

Chematica is a software platform that unites network theory, modern high-power computing, artificial intelligence, and expert chemical knowledge to rapidly design synthetic pathways leading to arbitrary (i.e., previously made or never attempted) targets. The program uses over 70,000 rules describing different reaction classes, where each rule coded by synthetic experts takes into account mechanistic considerations, providing substituent scope as well as contextual information about potential reactivity conflicts, protection requirements, selectivity issues, etc. Some of the popular rules spanning more than the reaction "core" (e.g., substitutions in polycyclic aromatic systems) are augmented by machine-learning models. The rules are the "basic moves" that are used by graph-search algorithms to navigate enormous trees of synthetic possibilities in intelligent ways. The "synthetic positions" on these trees are evaluated according to the so-called reaction- and chemical-scoring functions. The exploration of the trees is further guided by multistep strategies comprising steps that initially increase complexity but then decrease it, in effect overcoming local molecular-complexity barriers,<sup>6,7</sup> as well as a host of quantum-mechanical and molecular-mechanics routines that inspect the structures of the intermediates created during planning (see [Supplemental Information Section S10](#)). Chematica's code is parallelized and deployed on multicore machines. Input and output are supported by a graphical user interface (see [Figures S7–S15](#) and [Video S1](#)). For a

thorough discussion of the algorithmic basis of Chematica, see Szymkuć et al.<sup>6</sup> and, especially, the supplementary information in Klucznik et al.<sup>7</sup>

### Synthetic Design

All syntheses discussed in the current work were designed by Chematica running on a 64-core machine. The pathways were found and post-processed for display within, on average, 5 min per search. The longest design (of panobinostat with additional constraints) took 8 min.

### SUPPLEMENTAL INFORMATION

Supplemental Information includes further algorithmic details, details of the syntheses described in the main text, 22 figures, and one video and can be found with this article online at <https://doi.org/10.1016/j.chempr.2018.12.004>.

### ACKNOWLEDGMENTS

K.M., P.D., and B.A.G. thank the US DARPA for generous support under the “Make-It” Award 69461-CH-DRP #W911NF1610384. B.A.G. also gratefully acknowledges personal support from the Institute for Basic Science Korea, project code IBS-R020-D1. The authors thank Ewa Gajewska, Wiktor Beker, and Rafał Roszak for providing brief descriptions of computational routines used in Chematica and Sara Szymkuć for helpful discussions.

### AUTHOR CONTRIBUTIONS

K.M. and P.D. were key developers of Chematica. B.A.G. conceived Chematica in graduate school and has directed its development ever since. All authors contributed to the writing of the manuscript.

### DECLARATION OF INTERESTS

Although Chematica was originally developed and owned by B.A.G.’s Grzybowski Scientific Inventions, LLC, neither he nor the co-authors hold any stock in this company, which is now property of Merck KGaA, Darmstadt, Germany. The authors continue to collaborate with Merck within the DARPA “Make-It” award. All queries about access options to Chematica (now rebranded as Synthia), including academic collaborations, should be directed to Dr. Sarah Trice at [sarah.trice@sial.com](mailto:sarah.trice@sial.com).

Received: July 23, 2018

Revised: September 23, 2018

Accepted: December 5, 2018

Published: January 17, 2019

### REFERENCES AND NOTES

1. Corey, E.J., and Wipke, W.T. (1969). Computer-assisted design of complex organic syntheses. *Science* 166, 178–192.
2. Gelernter, H.L., Sanders, A.F., Larsen, D.L., Aganwal, K.K., Boivie, R.H., Spritzer, G.A., and Searleman, J.E. (1977). Empirical explorations of SynChem. *Science* 197, 1041–1049.
3. Ravitz, O. (2013). Data-driven computer aided synthesis design. *Drug Discov. Today Technol.* 10, e443–e449.
4. Bøgevig, A., Federsel, H.-J., Huerta, F., Hutchings, M.G., Kraut, H., Langer, T., Löw, P., Oppawsky, C., Rein, T., and Saller, H. (2015). Route design in the 21st century: the IC SYNTH software tool as an idea generator for synthesis prediction. *Org. Process Res. Dev.* 19, 357–368.
5. Segler, M.H.S., Preuss, M., and Waller, M.P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610.
6. Szymkuć, S., Gajewska, E.P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., Bajczyk, M., and Grzybowski, B.A. (2016). Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* 55, 5904–5937.
7. Klucznik, T., Mikulak-Klucznik, B., McCormack, M.P., Lima, H., Szymkuć, S., Bhowmick, M., Molga, K., Zhou, Y., Rickershauser, L., Gajewska, E.P., et al. (2018). Efficient syntheses of diverse, medically relevant targets planned by computer and executed in the laboratory. *Chem* 4, 522–532.
8. Krämer, K. (2018). AI-invented syntheses prove a hit as they make their lab debut. <https://www.chemistryworld.com/news/ai-invented-syntheses-are-a-hit-in-their-lab-debut/3008730.article>.
9. Klar, U., Buchmann, B., Schwede, W., Skuballa, W., Hoffmann, J., and Lichtner, R.B. (2006).

- Total synthesis and antitumor activity of ZK-EPO: the first fully synthetic epothilone in clinical development. *Angew. Chem. Int. Ed.* **45**, 7942–7948.
- Holub, N., and Blechert, S. (2007). Ring-rearrangement metathesis. *Chem. Asian J.* **2**, 1064–1082.
  - Botubol-Ares, J.M., Durán-Peña, M.J., Hernández-Galán, R., Collado, I.G., Harwood, L.M., and Macías-Sánchez, A.J. (2015). Diastereoselective and enantioselective preparation of nor-mevaldic acid surrogates through desymmetrisation methodology. Enantioselective synthesis of (+) and (–) nor-mevalonic lactones. *Tetrahedron* **71**, 7531–7538.
  - Xie, X., Zhang, T.Y., and Zhang, Z. (2006). Synthesis of bulky and electron-rich MOP-type ligands and their applications in palladium-catalyzed C–N bond formation. *J. Org. Chem.* **71**, 6522–6529.
  - Marion, N., Ecarnot, E.C., Navarro, O., Amoroso, D., Bell, A., and Nolan, S.P. (2006). (IPr)Pd(acac)Cl: an easily synthesized, efficient, and versatile precatalyst for C–N and C–C bond formation. *J. Org. Chem.* **71**, 3816–3821.
  - Ghosh, A., Sieser, J.E., Riou, M., Cai, W., and Rivera-Ruiz, L. (2003). Palladium-catalyzed synthesis of *N*-aryloxazolidinones from aryl chlorides. *Org. Lett.* **5**, 2207–2210.
  - DeAngelis, A.J., Gildner, P.G., Chow, R., and Colacot, T.J. (2015). Generating active “L-Pd(0)” via neutral or cationic  $\pi$ -allylpalladium complexes featuring biaryl/bipyrazolylphosphines: synthetic, mechanistic, and structure–activity studies in challenging cross-coupling reactions. *J. Org. Chem.* **80**, 6794–6813.
  - Fache, F., Jacquot, L., and Lemaire, M. (1994). Extension of the Eschweiler-Clarke procedure to the *N*-alkylation of amides. *Tetrahedron Lett.* **35**, 3313–3314.
  - Naresh, A., Venkateswara Rao, M., Kotapalli, S.S., Ummanni, R., and Venkateswara Rao, B. (2014). Oxazolidinone derivatives: Cytosaxone–Linezolid hybrids induces apoptosis and senescence in DU145 prostate cancer cells. *Eur. J. Med. Chem.* **80**, 295–307.
  - Danielmeier, K., and Steckhan, E. (1995). Efficient pathways to (*R*)- and (*S*)-5-hydroxymethyl-2-oxazolidinone and some derivatives. *Tetrahedron Asymmetry* **6**, 1181–1190.
  - Zhang, H., Zhang, S., and Bian, W. (2016). A method for the synthesis of sitagliptin. Patent CN 103483340 B, filed July 29, 2013, and granted September 7, 2016.
  - Sathe, D.G., Damle, S.V., Arote, N.D., Ambre, R.R., Sawant, K.D., and Naik, T.A. (2012). Sitagliptin, salts and polymorphs thereof. Patent WO 2012/025944 A2, filed August 26, 2011, and published March 1, 2012.
  - Badgujar, S., Gharpure, M.M., and Yadav, R.S. (2012). Processes for the preparation of *R*-sitagliptin and intermediates thereof. Patent WO 2012/042534 A2, filed September 21, 2011, and published April 5, 2012.
  - Hansen, K.B., Balsells, J., Dreher, S., Hsiao, Y., Kubryk, M., Palucki, M., Rivera, N., Steinhuebel, D., Armstrong, J.D., Askin, D., et al. (2005). First generation process for the preparation of the DPP-IV inhibitor sitagliptin. *Org. Process Res. Dev.* **9**, 634–639.
  - Xiao, Y., Armstrong, J.D., Krska, S.W., Njolito, E., Rivera, N.R., Sun, Y., Rosner, T., and Clausen, A.M. (2006). Process to chiral beta amino acid derivatives by asymmetric hydrogenation. Patent WO 2006/081151 A1, filed January 20, 2006, and published August 3, 2006.
  - Savile, C.K., Janey, J.M., Mundorff, E.C., Moore, J.C., Tam, S., Jarvis, W.R., Colbeck, J.C., Krebber, A., Fleitz, F.J., Brands, J., et al. (2010). Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science* **329**, 305–309.
  - Padi, P.R., Ireni, B., Polavarapu, S., Padamata, S., Nerella, K., Ramasamy, V.A., and Vangala, R.R. (2009). Processes for the preparation of sitagliptin and pharmaceutically acceptable salts thereof. Patent WO 2009/085990 A2, filed December 18, 2008, and published July 9, 2009.
  - Yang, X., Yang, S., Xiang, L., Pang, X., Chen, B., Huang, G., and Yan, R. (2015). Synthesis of 3-arylpiperidines via palladium/copper-catalyzed annulation of allylamine/1,3-propanediamine and aldehydes. *Adv. Synth. Catal.* **357**, 3732–3736.
  - Nussbaumer, C., and Frater, G. (1987). Stereoselective synthesis of ( $\pm$ )-*cis*- $\alpha$ -Irene. *J. Org. Chem.* **52**, 2096–2098.
  - Bartra-Sanmarti, M., Rustullet-Oliver, A., Fernandez-Hernandez, S., and Monsalvatje-Llagostera, M. (2010). Process for the preparation of a chiral beta amino acid derivative and intermediates thereof. Patent WO 2010/097420 A1, filed February 24, 2010, and published September 2, 2010.
  - Pachamuthu, K., Zhu, X., and Schmidt, R.R. (2005). Reversed approach to *S*-farnesylation and *S*-palmitoylation: application to an efficient synthesis of the C-terminus of lipidated human *N*-Ras hexapeptide. *J. Org. Chem.* **70**, 3720–3723.
  - Pan, X., Yu, W., Ou, W., Tao, X., Wan, J., and Liu, F. (2011). Synthesis of a chiral  $\beta$ -amino acid derivative by a cobalt-catalysed coupling reaction. *J. Chem. Res. (S)* **35**, 545–546.
  - Deguest, G., Bischoff, L., Fruit, C., and Marsais, F. (2006). Regioselective opening of *N*-Cbz glutamic and aspartic anhydrides with carbon nucleophiles. *Tetrahedron Asymmetry* **17**, 2120–2125.
  - Seki, M., and Matsumoto, K. (1996). A facile synthesis of (*R*)-3-Amino-4-phenylbutyric acid from L-aspartic acid. *Biosci. Biotechnol. Biochem.* **60**, 916–917.
  - Gündoğdu, Ö., Turhan, P., Köse, A., Altundaş, R., and Kara, Y. (2017). Reaction of (*S*)-homoserine lactone with Grignard reagents: synthesis of amino-keto-alcohols and  $\beta$ -amino acid derivatives. *Tetrahedron Asymmetry* **28**, 1163–1168.
  - Liu, F., Yu, W., Ou, W., Xu, X., Ruan, L., Wang, X., Li, Y., Peng, X., Tao, X., Mao, J., et al. (2010). The synthesis of a chiral  $\beta$ -amino acid derivative by the Grignard reaction of an aspartic acid equivalent. *J. Chem. Res. (S)* **34**, 517–519.
  - Acemoglu, M., Bajwa, J.S., Parker, D.J., and Slade, J. (2009). Process for making *N*-hydroxy-3-[4-[[[(2-methyl-1*H*-indol-3-yl)ethyl]amino]methyl]phenyl]-2*E*-2-propenamide and starting materials therefor. Patent US 2009/0306405 A1, filed June 12, 2006, and published December 10, 2009.
  - Wagaw, S., Yang, B.H., and Buchwald, S.L. (1999). A palladium-catalyzed method for the preparation of indoles via the Fischer indole synthesis. *J. Am. Chem. Soc.* **121**, 10251–10263.
  - Campos, K.R., Woo, J.C.S., Lee, S., and Tillyer, R.D. (2004). A general synthesis of substituted indoles from cyclic enol ethers and enol lactones. *Org. Lett.* **6**, 79–82.
  - Guérin, C., Bellosta, V., Guillamot, G., and Cossy, J. (2011). Mild nonpiperizing *N*-alkylation of amines by alcohols without transition metals. *Org. Lett.* **13**, 3534–3537.
  - Bumagin, N.A., More, P.G., and Beletskaya, I.P. (1989). Synthesis of substituted cinnamic acids and cinnamionitriles via palladium catalyzed coupling reactions of aryl halides with acrylic acid and acrylonitrile in aqueous media. *J. Organomet. Chem.* **371**, 397–401.
  - Aravinda Kumar, K., Venkateswarlu, V., Vishwakarma, R., and Sawant, S. (2015). A metal-free approach to carboxylic acids by oxidation of alkyl, aryl, or heteroaryl alkyl ketones or arylalkynes. *Synthesis* **47**, 3161–3168.
  - Bao, X., Song, D., Qiao, X., Zhao, X., and Chen, G. (2016). The development of an effective synthetic route of Belinostat. *Org. Process Res. Dev.* **20**, 1482–1488.
  - Coldham, I., and Leonori, D. (2010). Regioselective and stereoselective copper(I)-promoted allylation and conjugate addition of *N*-Boc-2-lithiopyrrolidine and *N*-Boc-2-lithiopiperidine. *J. Org. Chem.* **75**, 4069–4077.
  - Clemons, G.R., and Ramage, G.R. (1932). 456. Octahydropyrrocoline. *J. Chem. Soc.* **28**, 2969–2973.
  - Pohl, M. (2008). How to control the United States pharmaceutical API market using patents on new synthetic intermediate compounds. *J. Intellect. Prop. Rights* **13**, 473–479.
  - Heuer, L. (2018). AI could threaten pharmaceutical patents. *Nature* **558**, 519.
  - Lowe, D. (2018). AI will not threaten pharma patents – not this way. <http://blogs.sciencemag.org/pipeline/archives/2018/06/27/ai-will-not-threaten-pharma-patents-not-this-way>.
  - Dach, R., Song, J.J., Roschangar, F., Samstag, W., and Senanayake, C.H. (2012). The eight criteria defining a good chemical manufacturing process. *Org. Process Res. Dev.* **16**, 1697–1706.

**Chem, Volume 5**

**Supplemental Information**

**Navigating around Patented Routes  
by Preserving Specific Motifs along  
Computer-Planned Retrosynthetic Pathways**

**Karol Molga, Piotr Dittwald, and Bartosz A. Grzybowski**

Supplementary Information for manuscript entitled “***Navigating around patented routes by preserving specific motifs along computer-planned retrosynthetic pathways***” by Karol Molga, Piotr Dittwald\*, Bartosz A. Grzybowski\*

## **CONTENTS:**

**Section S1.** Overview of the bond protecting algorithm.

**Section S2.** Handling the search space in the presence of repeating intermediates.

**Section S3.** Pseudocodes of atom-number-unification and bond-protection algorithms.

**Section S4.** Patented syntheses of Linezolid.

**Section S5.** Patented syntheses of Sitagliptin.

**Section S6.** Screenshots showing Chematica’s syntheses of Linezolid.

**Section S7.** Screenshots showing Chematica’s syntheses of Sitagliptin.

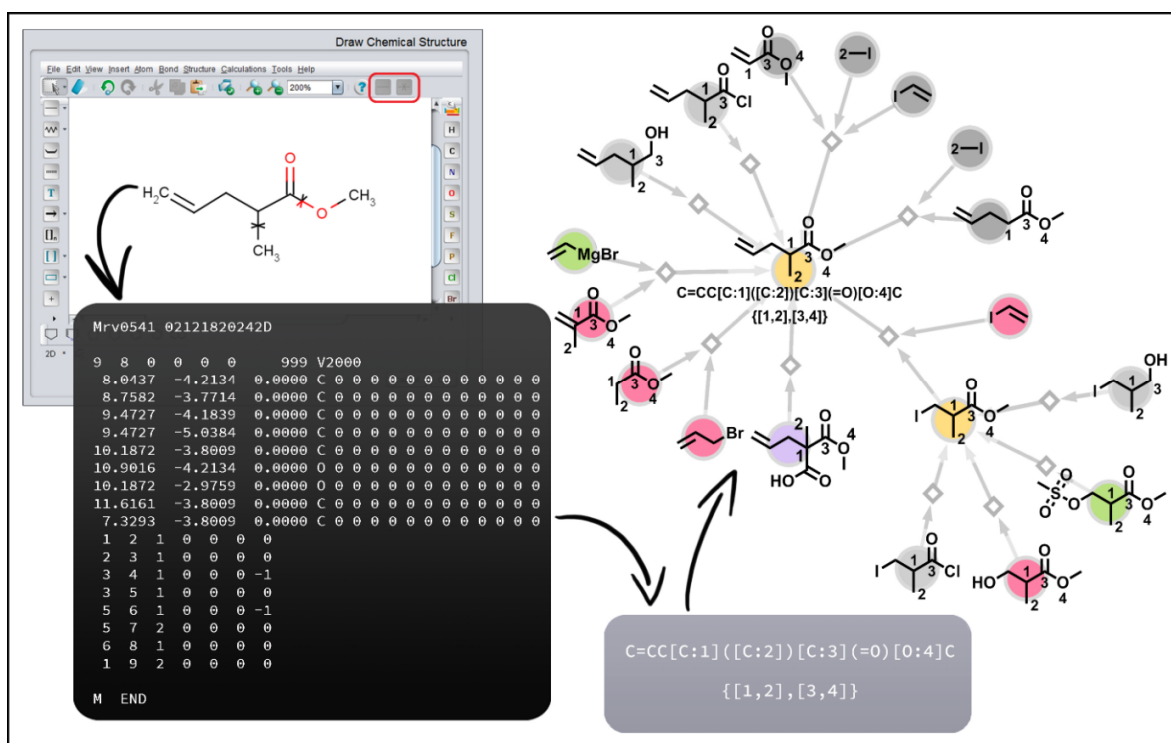
**Section S8.** Chematica’s syntheses of Panobinostat.

**Section S9.** Details of Chematica’s enantioselective syntheses of coniceine.

**Section S10.** Outline of some important computational routines underlying Chematica.

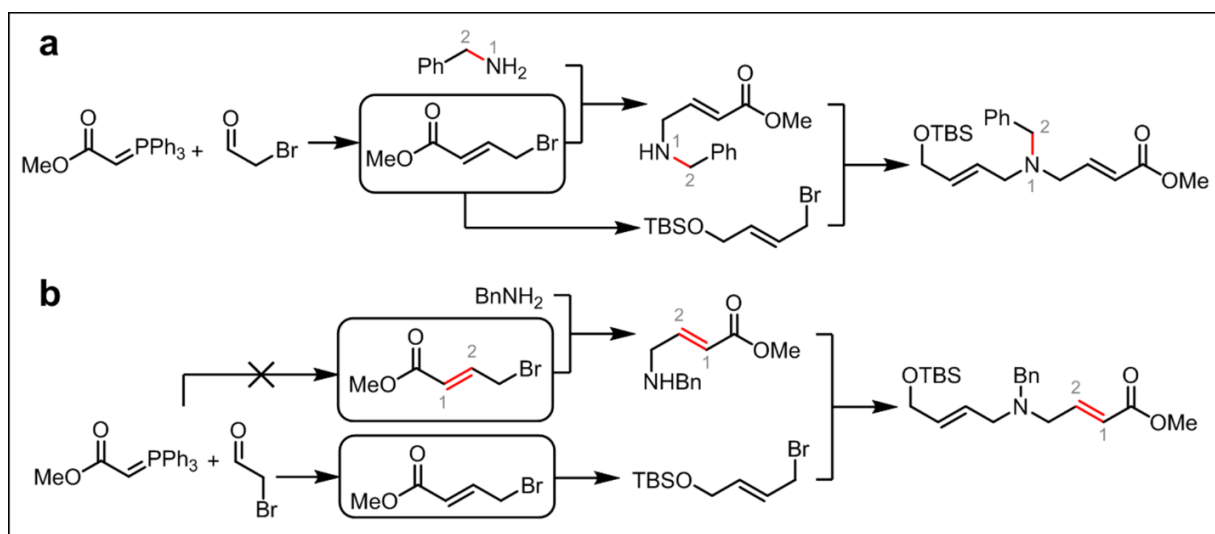
**Section S11.** Caption to Supplementary Movie S1.

## Section S1. Overview of the bond protecting algorithm.



**Figure S1.** Scheme illustrating the functioning of the bond protecting algorithm (counterclockwise from upper left panel). Bonds not-to-be-disconnected are marked by the user. Marking tool is available in the upper-right menu circled in red. The so-called Extended Molfile (black panel) with added bond information (-1 in the last column) is generated and sent to the server. The file is translated into a SMILES string (gray panel) with atoms belonging to the selected bonds numbered; a list of atom number pairs (“bond list”) denoting the marked bonds is also generated. When the retrosynthetic search commences, the algorithm inspects if none of the “preserved” bonds (here, 1-2 and 3-4) is disconnected in the synthons. If any marked bond is disconnected and the bond set changes, such synthetic options (gray nodes) are no longer considered. Otherwise, next-generation nodes are expanded and the search continues until stop conditions are fulfilled (when reaching commercially available or previously made chemicals; red and green nodes, respectively). The graph shown is a vast simplification of the full graphs generated in Chematica in which each generations has, on average, ~100 synthetic options.

## Section S2. Handling the search space in the presence of repeating intermediates.



**Figure S2. Schematic illustration of searches with the same repeating intermediate.** **a)** When, during the same search, an identical intermediate is encountered in several putative pathways (here, methyl bromocrotonate, framed) but does not contain any bonds marked as not-to-be-disconnected, it is considered as only one node common to different pathways. **b)** If, however, one of bromocrotonates contains protected bonds while the other does not, then synthetic histories for these two molecules may be different and they need to be kept as separate nodes in the search space. In this specific example, Wittig reaction can be applied to one molecule from the pair but not to the other since it would affect bonds 1-2 marked as not-to-be-disconnected.



### Section S3. Pseudocodes of atom-number-unification and bond-protection algorithms.

---

```
1: function GETTARGETNUMBERED(G)
  ▷ G reaction network (with exactly one target)
2:   for node in G.nodes() do
3:     if G.outDegree(node) = 0 then break
4:   idx ← 0
5:   for atom in node.atoms() do
6:     atom.assignMapping(idx)
7:   idx ← idx + 1
  return node

8: function GETSUBSTRATESNUMBERED(G, node, rxn)
  ▷ G reaction network
  ▷ node node with target-based atom numbering
  ▷ rxn reaction, where node is a product
  ▷ this function transmits atom numbers from node to substrates of rxn

9:   rxnMapping ← rxn.getRxnMapping()
  //dictionary mapping substrates' atoms to products' atoms
10:  substrates ← G.predecessors(node)
11:  for substrate in substrates do
12:    for atom in substrate.atoms() do
13:      atomProd ← rxnMapping[atom]
14:      if atomProd.HasNoMapping() then continue
  //skip numeration of atoms not referring to target
15:      atom.assignMapping(atomProd.getMMapping())
  return substrates

16: function TRANSMITTARGETBASEDATOMNUMBERINGS(G)
  ▷ G reaction network
  ▷ this function propagates mapping trough entire network
  ▷ for pseudocode brevity unambiguous target-based atom numbering is
  assumed for each molecule

17:  nodeNumberedSet ← {getTargetNumbered(G)}
18:  visited ← ∅
19:  while true do
20:    if nodeNumberedSet = ∅ then break
21:    nodeNumberedSetNew ← ∅
22:    for node in nodeNumberedSet do
23:      if node ∈ visited then continue
24:      visited.add(node)
25:      for rxn in G.predecessors(node) do
26:        substrates ← getSubstratesNumbered(G, node, rxn)
27:        nodeNumberedSetNew.addElements(substrates)
28:    nodeNumberedSet ← nodeNumberedSetNew
```

---

**Figure S3.** The algorithm unifies atom numbering over the entire network of reactions  $G$  (function *transmitTargetBasedAtomNumberings*; lines 16-28 in pseudocode). To begin with, unique numbers are assigned to each atom in the target (function *getTargetNumbered*; lines 1-7 in pseudocode). Then  $G$  is traversed starting from the target and the encountered reaction nodes are iteratively analyzed. At each iteration, target-based atom numeration is transmitted from the reaction's product to its substrates (function *getSubstratesNumbered*, lines 8-15 in pseudocode), while atoms not mapping onto the target remain unnumbered. Finally, for each molecule in  $G$ , its atom numbering indicates the region of the target where this molecule is "transferred" through (one- or multi-step) synthetic sequence. Consequently, bonds disconnected in reactions considered within the network can be accurately mapped onto the corresponding bonds in the target.

---

```

1: function CALCULATEB(mol,  $B_t$ )
  ▷ mol - molecule to be checked
  ▷  $B_t$  set of pairs of atom labels adjacent to protected bonds in the target
2:    $B \leftarrow \emptyset$ 
3:   for bond in mol.bonds() do
4:      $a1, a2 \leftarrow bond.getAdjacentAtoms()$ 
5:      $idx1 \leftarrow a1.getLabel()$ 
6:      $idx2 \leftarrow a2.getLabel()$ 
7:     if  $Null \in [idx1, idx2]$  then continue
8:      $idx1, idx2 \leftarrow sorted(idx1, idx2)$ 
9:     if  $(idx1, idx2) \in B_t$  then  $B \leftarrow B \cup \{[idx1, idx2]\}$ 
10:  return  $B$ 

11: function CHECKIFTRANSFORMATIONAPPLICABLE(retron, synthons,  $B_t$ )
  ▷ retron, synthons molecules in considered transformations
  ▷  $B_t$  set of pairs of atom labels adjacent to protected bonds in the target
12:   $B_r \leftarrow calculateB(retron, B_t)$ 
13:   $B_s \leftarrow \emptyset$ 
14:  for synthon in synthons do
15:     $B_s \leftarrow B_s \cup calculateB(synthon, B_t)$ 
16:  return ( $B_r = B_s$ )

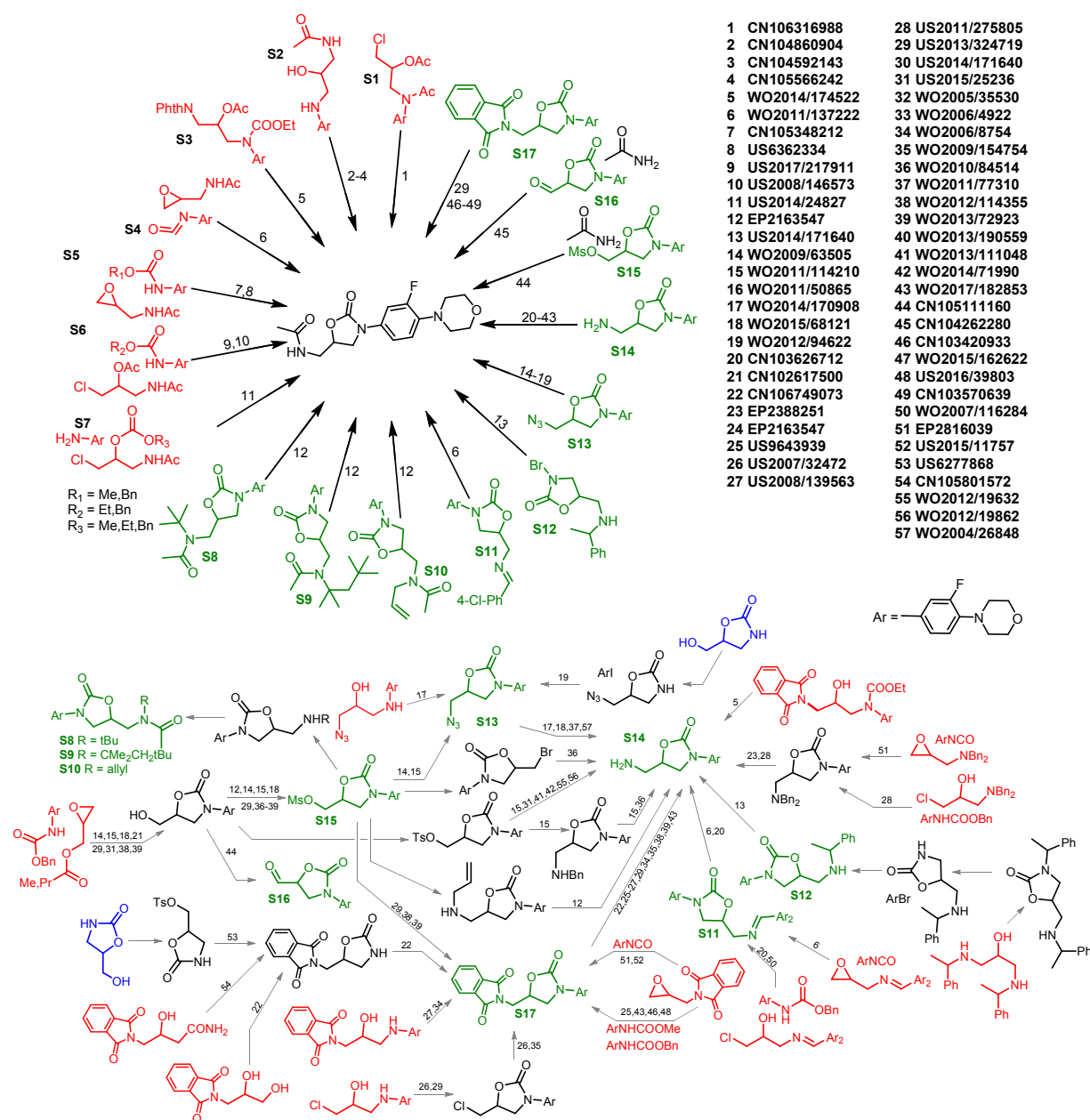
```

---

**Figure S4.** At each retrosynthetic step,  $r \rightarrow s_1, s_2, \dots, s_N$ , the algorithm applies function *checkIfTransformApplicable* (pseudocode, lines 11-16) to appropriate retron  $r$ , set of synthons  $s_1, s_2, \dots, s_N$ , and set of not-to-be-altered bonds as defined by the user. The transform is accepted if and only if the following condition is satisfied (\*):  $B(r) = B(s_1) \cup \dots \cup B(s_N)$ , where  $B(m)$  is a subset of protected bonds preserved in molecule  $m$  calculated by function *calculateB* (pseudocode, lines 1-10). To explain how the protected bonds are conserved during retrosynthetic search, let us consider the pathway with the following generations  $R_0, R_1, \dots, R_k$ , corresponding to sets of synthons available after each step. For the initial generation,  $R_0 = \{t\}$ , i.e., we start from single target molecule. On the other hand,  $R_k$ , the final generation,

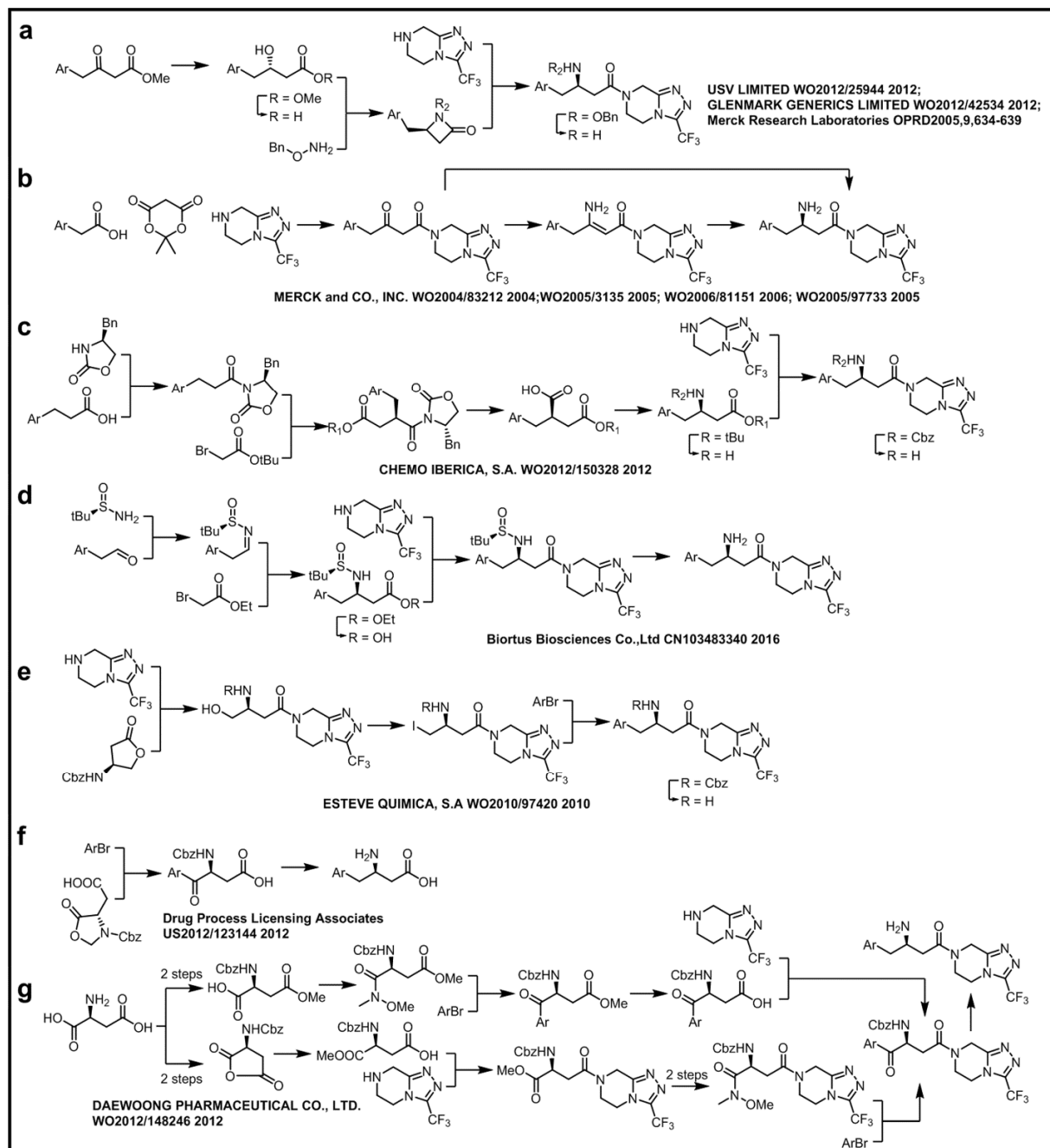
is composed of synthons that are fulfilling user-defined stop criteria (e.g., all are commercially available). For retrosynthetic step  $r \rightarrow s_1, s_2, \dots, s_N$  leading from  $R_{i-1}$  to  $R_i$ , we have  $R_i = R_{i-1} \cup \{s_1, s_2, \dots, s_N\} \setminus \{r\}$  (where  $\setminus$  is a minus operator on sets), namely retron is replaced by the set of synthons. By applying condition (\*) as a step constraint, we obtain that  $\cup_{s \in R_k} B(s) = \cup_{s \in R_{k-1}} B(s) = \dots = \cup_{s \in R_0} B(s) = B(t)$ , i.e., the algorithm preserves the not-to-be-altered bonds along entire pathways it identifies.

## Section S4. Patented syntheses of Linezolid.



**Figure S5. Syntheses of linezolid described in patents.** Only two pathways (starting from hydroxymethyloxazolidinone (blue)) do not rely on the formation of oxazolidinone ring. Red = compounds used in the ring forming steps.

**Section S5. Patented syntheses of Sitagliptin.**



**Figure S6.** Selected approaches leading to Sitagliptin relying on stereoselective reductions (a,b), chiral-auxiliary-directed additions of bromoesters (c,d), or chiral starting materials (e-g). a) Merck's first generation / Glenmark's process relying on the stereoselective reduction of ketoester. Subsequent hydrolysis,

Mitsunobu reaction, coupling with triazoloamine and reductive cleavage of hydroxylamine are used to finalize the synthesis. **b)** Merck's second generation process relies on the stereoselective hydrogenation of enamine amide derived from ketoamide formed in a three component reaction of Meldrum acid, carboxylic acid and triazoloamine. **c)** Chemo Iberica's approach to Sitagliptin takes advantage of Evans' auxiliary directed alkylation of enolate and Curtius rearrangement to set the necessary stereocenter. **d)** Biotus' methodology exploits Reformatsky-type addition of zinc enolate to Elmann's sulfinylimine to prepare the  $\beta$ -amino acid. **e)** Esteve Quimica relied on the chiral starting material. *S*-homoserine lactone is opened with triazoloamine and transformed to primary iodide. Trifluorophenyl moiety is installed in Kumada-Corriu type coupling. One of the Chematica's approaches is virtually identical with this solution except for the order of two steps being exchanged. **f,g)** Syntheses of Sitagliptin sourcing the stereocenter from aspartic acid and relying on the Grignard reaction and subsequent deoxygenation of the ketone obtained.

## Section S6. Screenshots showing Chematica's syntheses of Linezolid.

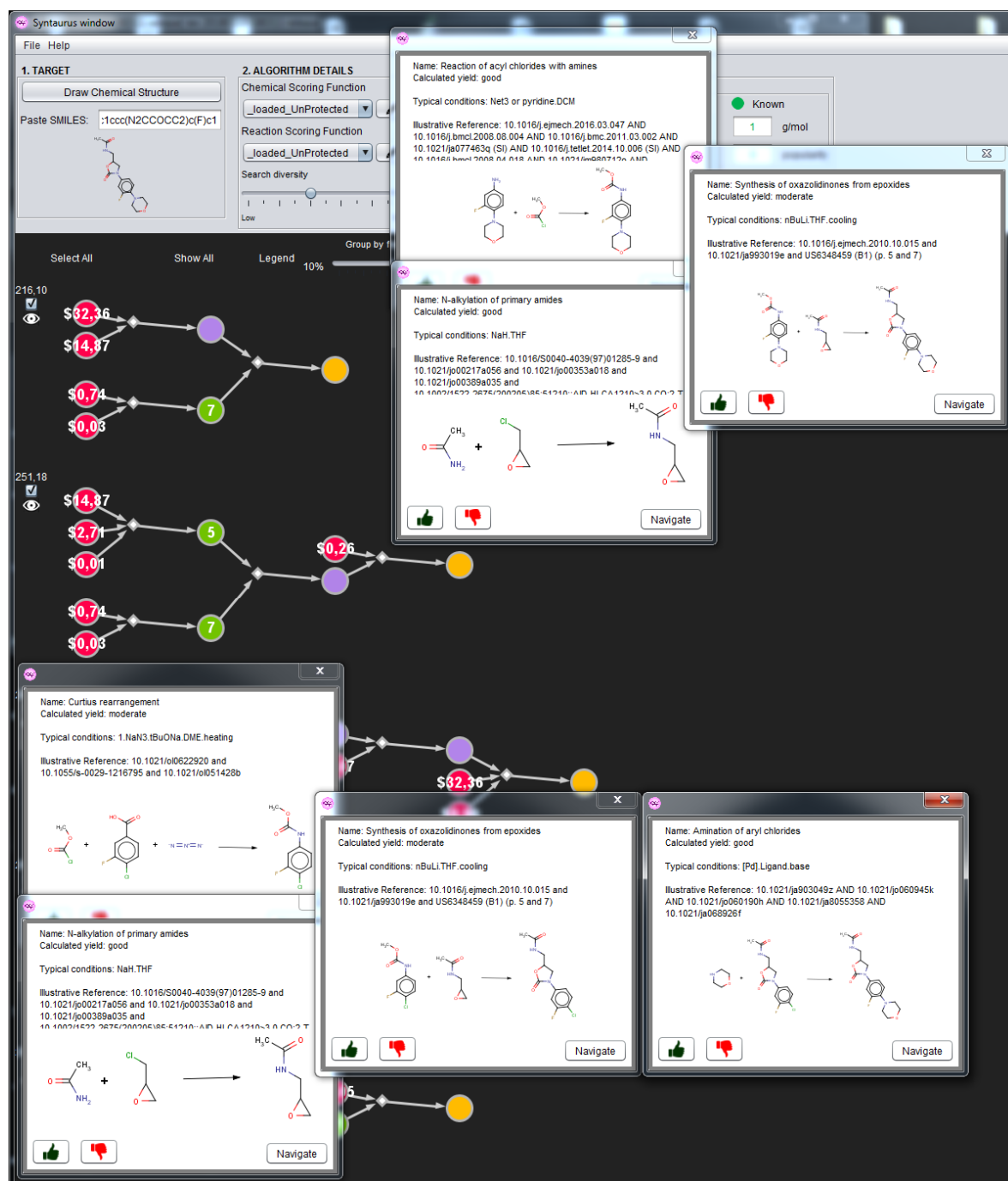


Figure S7. Raw output of Chematica's top-scoring synthetic plans obtained for Linezolid without any bonds "preserved".

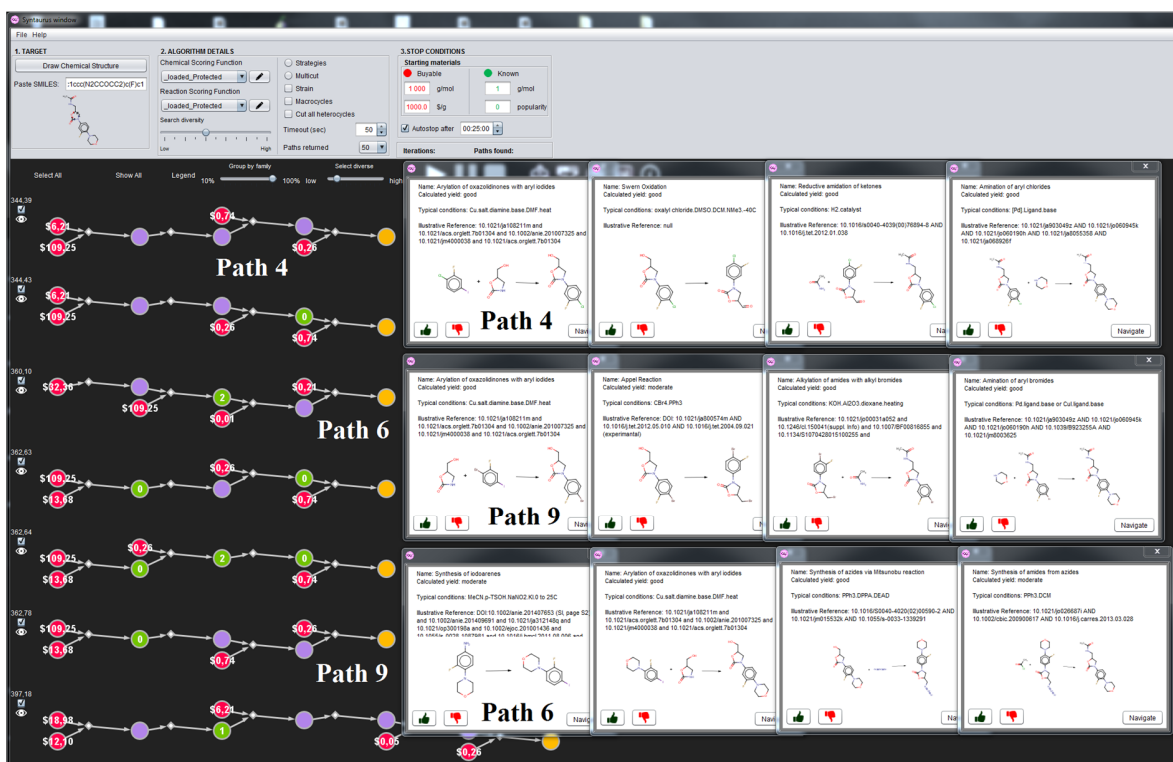
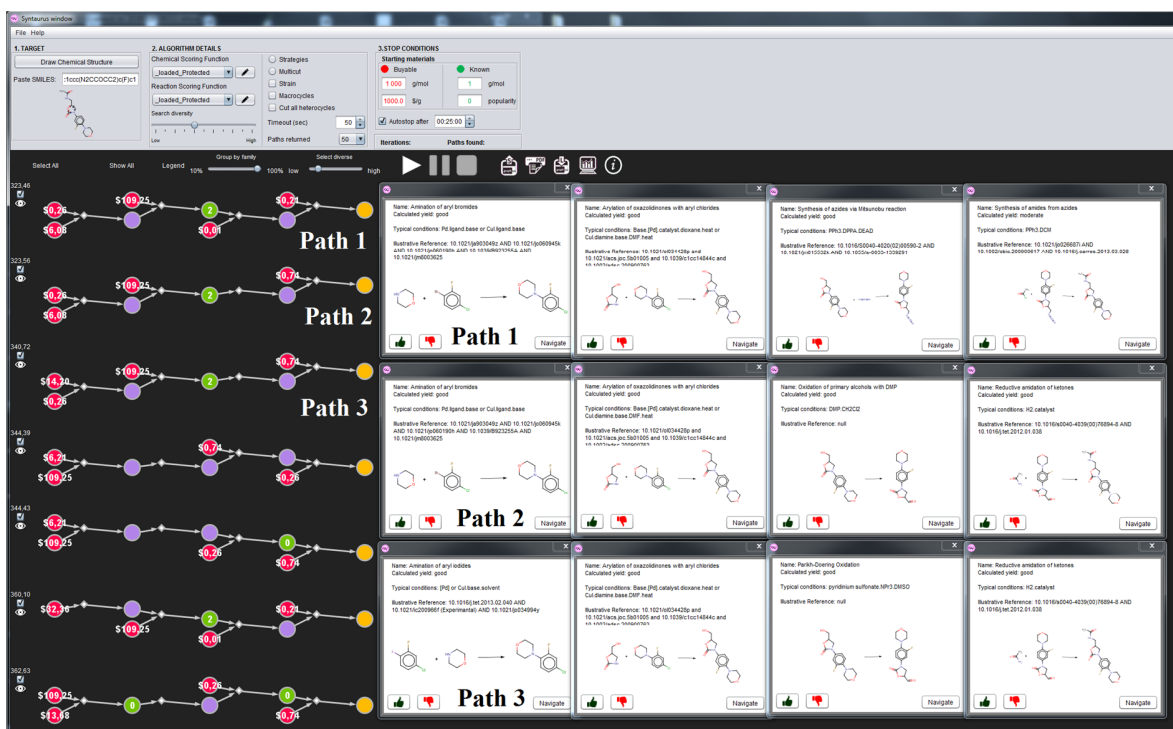


Figure S8. Raw output of Chemica's synthetic plans obtained for Linezolid when oxazolidinone ring was denoted as not-to-be-cut. Paths 5,7,8 are modifications of paths 4/8 with step order changed.

## Section S7. Screenshots showing Chematica's syntheses of Sitagliptin.

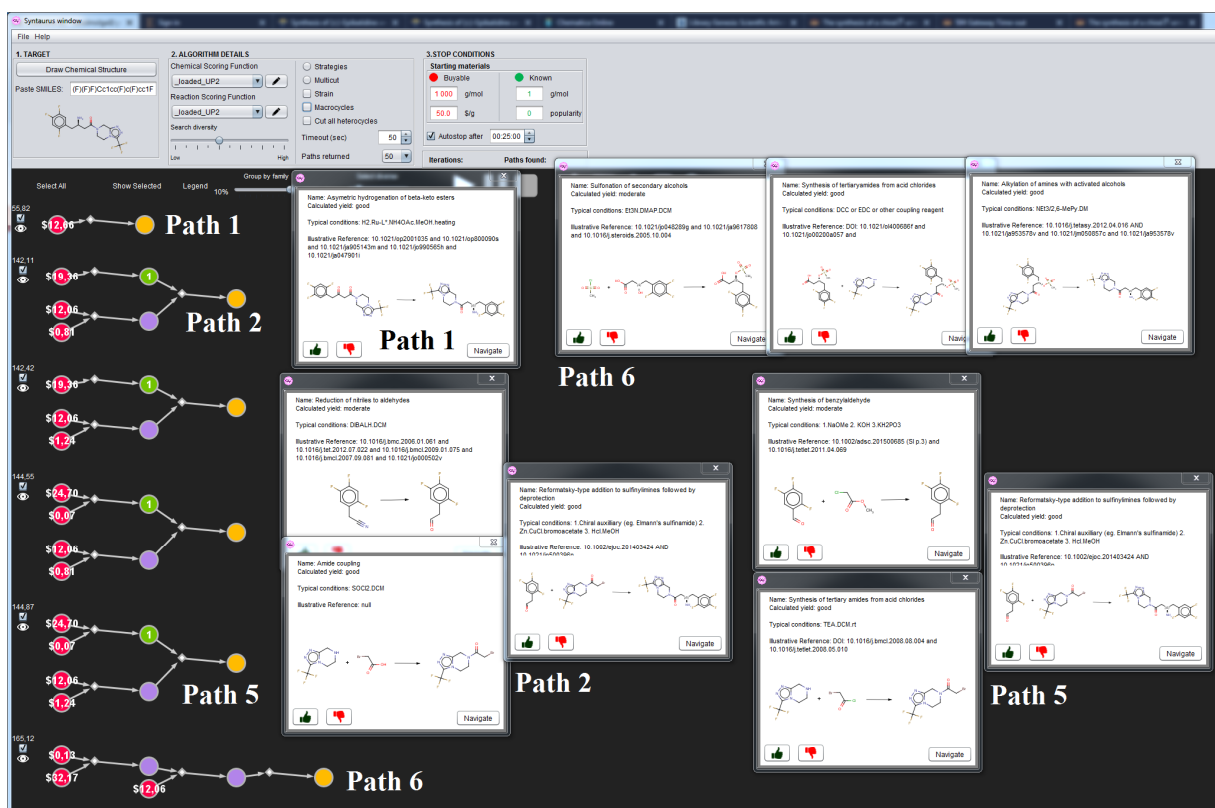


Figure S9. Raw output of Chematica's synthetic plans obtained for Sitagliptin without any bonds "preserved". Paths 3/4 are minor modifications of paths 2/5.



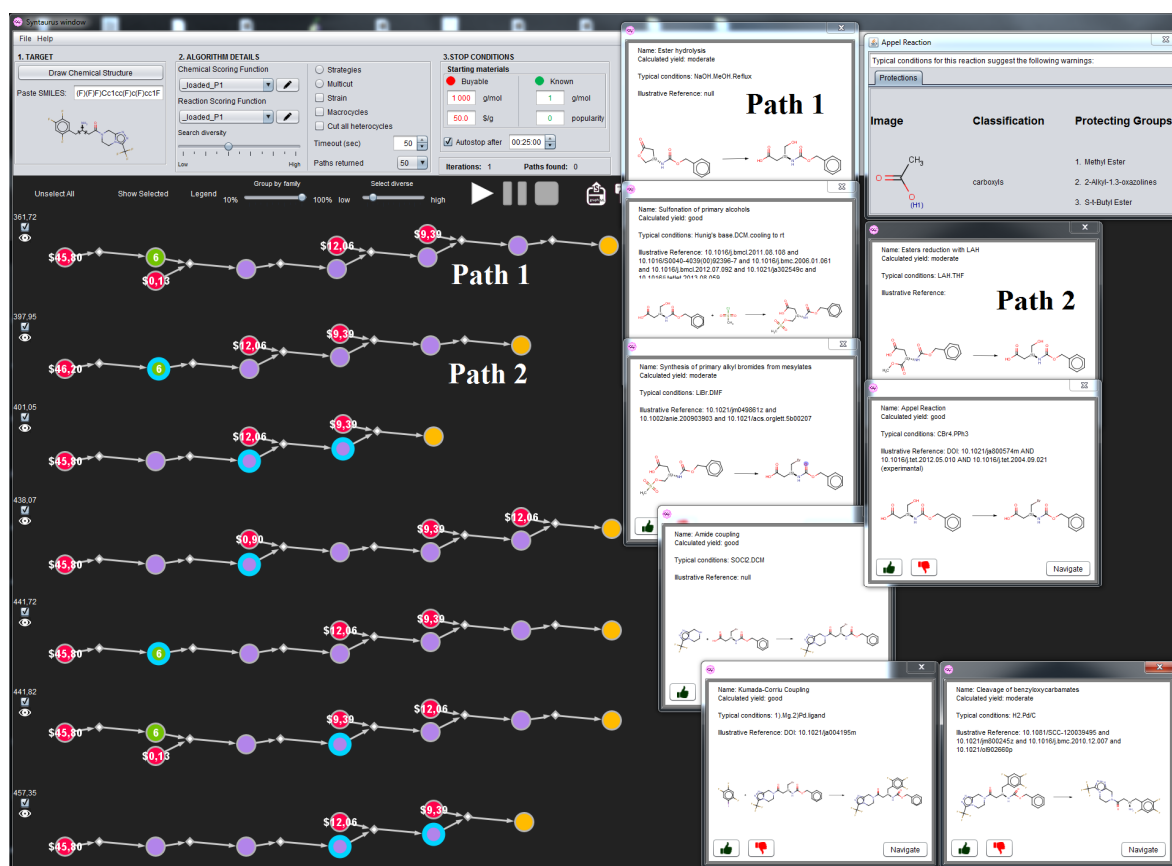
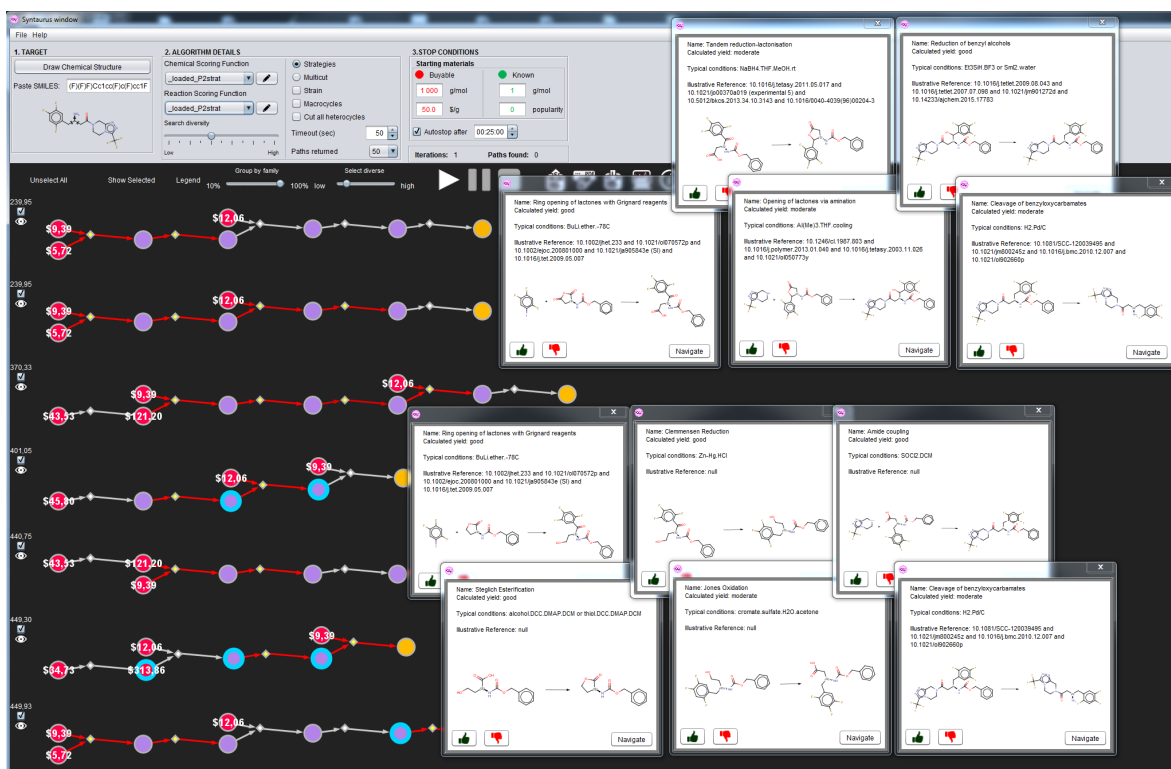


Figure S10. Raw output of Chemica's synthetic plans obtained for Sitagliptin when bonds adjacent to the stereocenter were "protected". Top-right: details of protection required for Appel reaction (Path 2).



**Figure S11. Raw output of Chematica's top-scoring synthetic plans obtained for Sitagliptin when bonds adjacent to stereocenter were protected and strategies applied.**

## Section S8. Chematica's syntheses of Panobinostat.

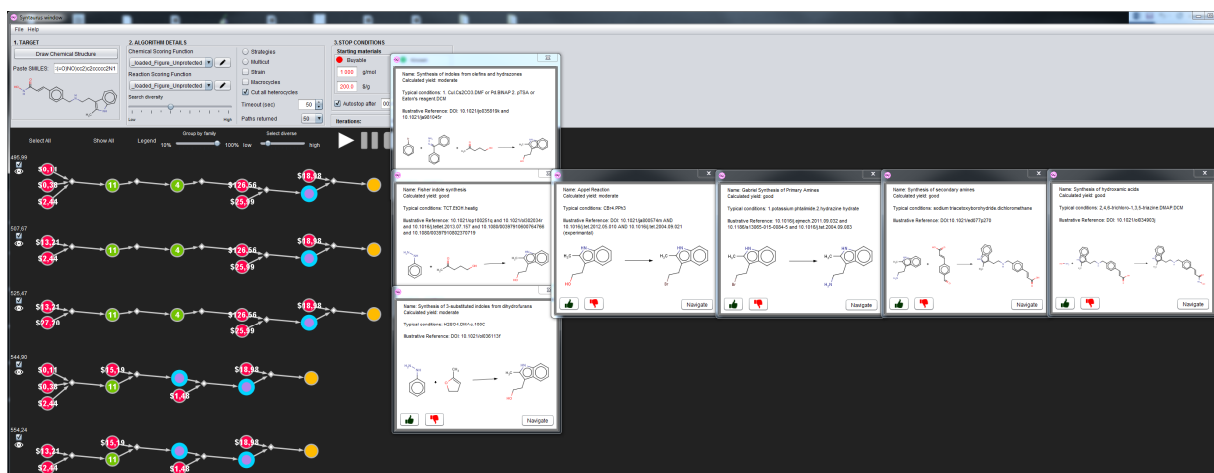


Figure S12. Raw output of Chematica's top-scoring synthetic plans for Panobinostat without "preserving" any bonds.

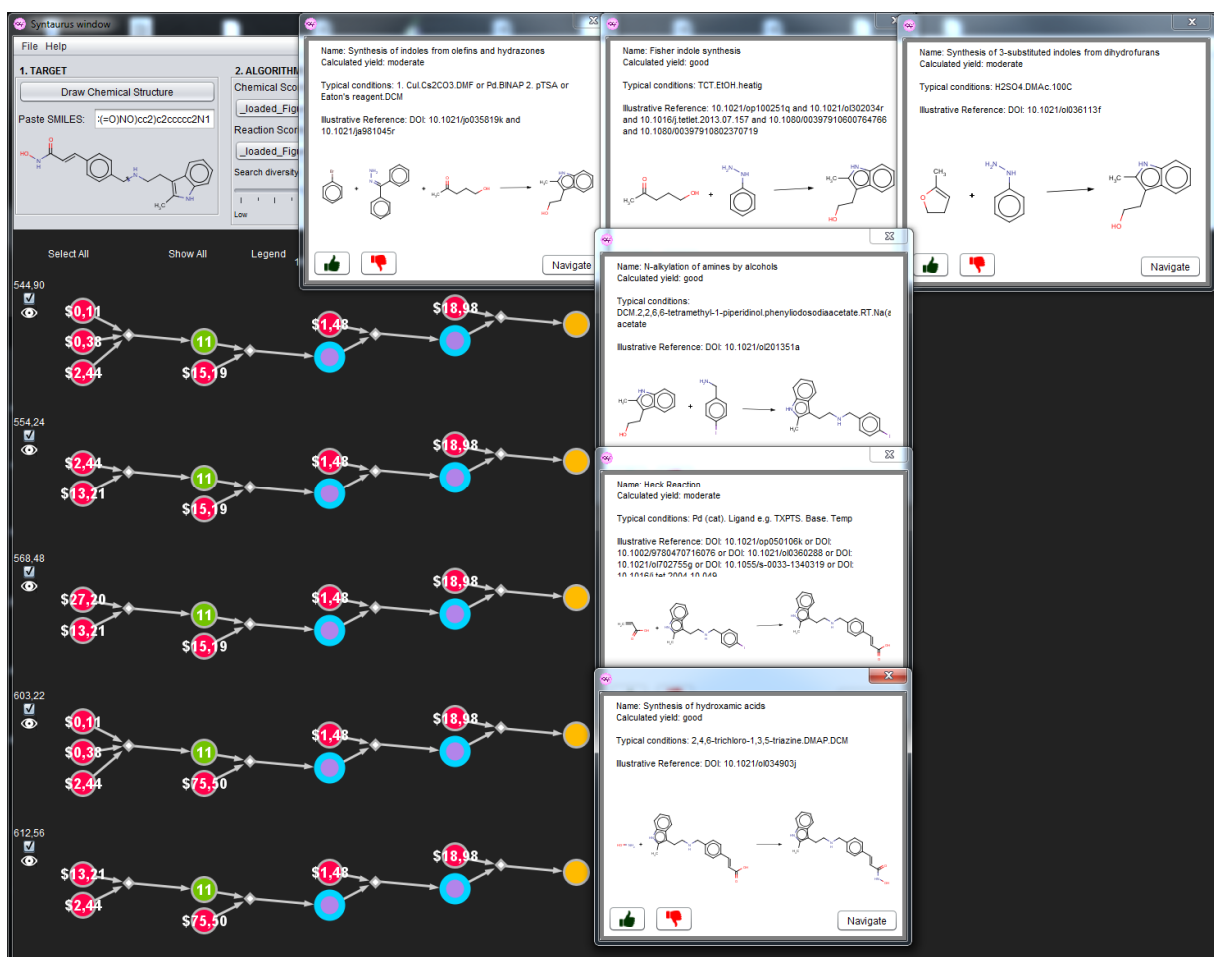
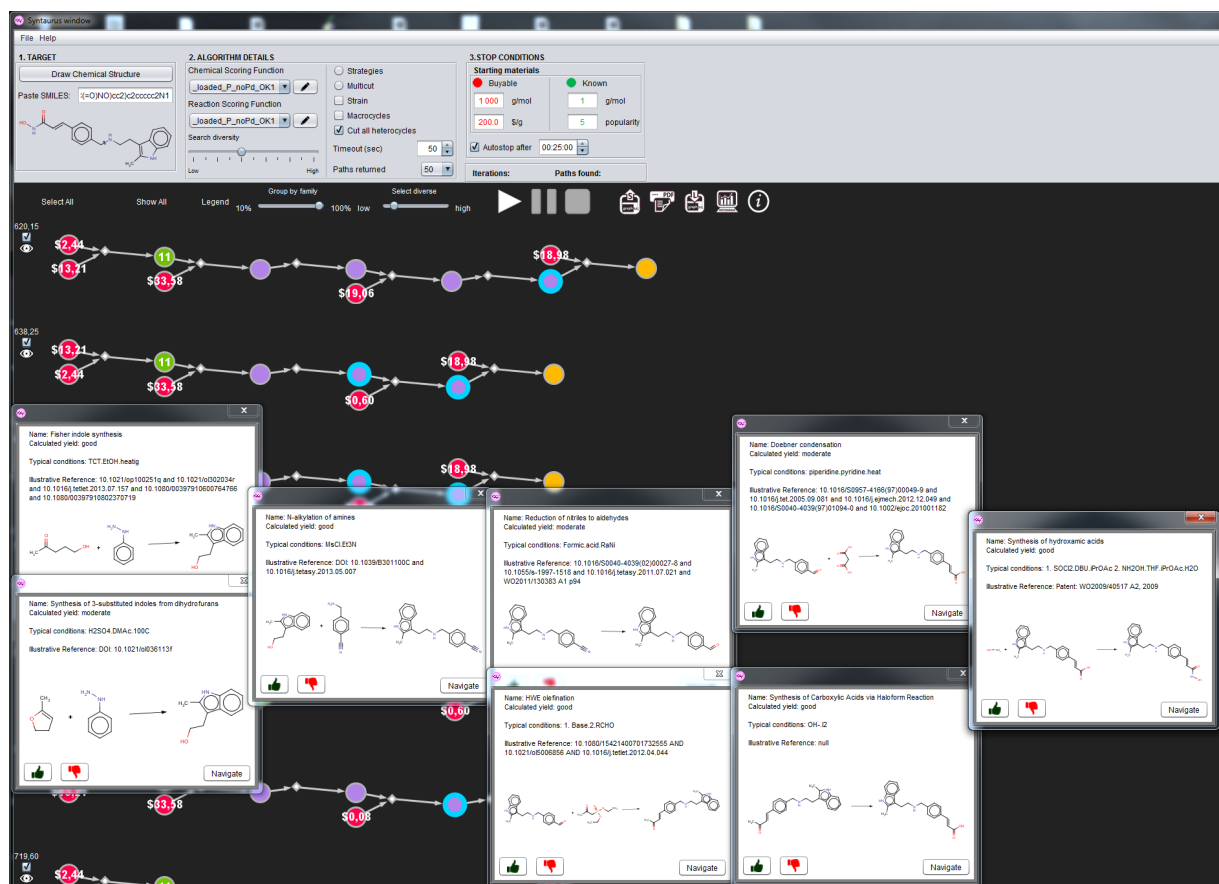


Figure S13. Raw output of Chematica's top-scoring synthetic plans for Panobinostat when one of the C-N bonds was marked as not-to-be-disconnected.



**Figure S14. Raw output of Chematica’s synthetic plans for Panobinostat when one of the C-N bonds was “preserved” and reactions requiring Pd catalysts were penalized.**

## Section S9. Details of Chematica's enantioselective syntheses of concineine

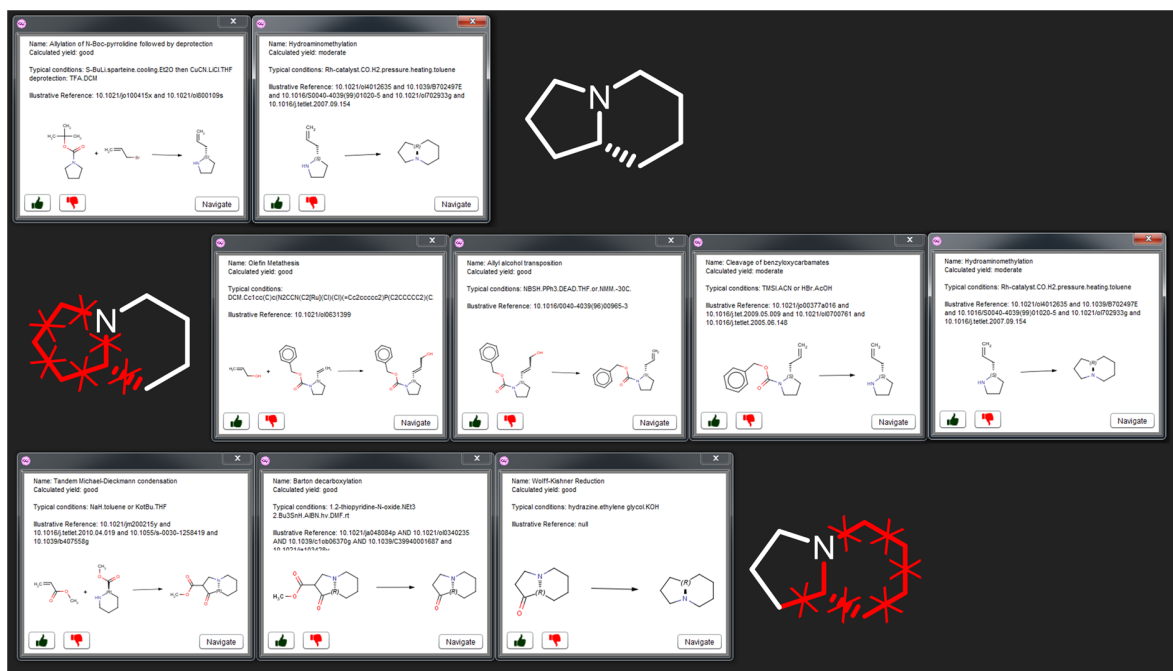


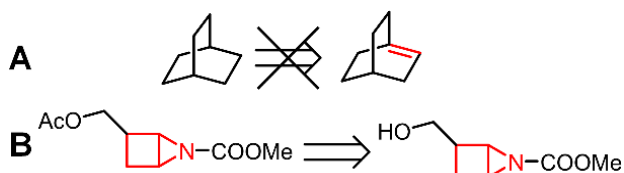
Figure S15. Details of Chematica's syntheses of concineine.

## Section S10. Outline of some important computational routines underlying Chematica.

**S10.1. General considerations.** While most factors dictating the reaction outcomes are captured by the reaction transforms which carefully delineate the scope of admissible substituents (based on reaction's mechanism and stereoelectronic requirements), there are also issues that require scrutiny beyond even vastly extended reaction "cores". These issues can be broadly sub-divided into two categories: (**C1**) that the transform leads to synthons or proceeds through a transition state that are structurally problematic (especially, excessively strained); and (**C2**) that the transform itself is so broad that enumeration of all its variants is unfeasible (e.g., enumerating all substitution patterns on all possible aromatic systems is clearly not feasible, nor is accounting for all possible substituent effects leading to different regio- or stereoisomers in pericyclic reactions such as Diels-Alder reaction or Cope rearrangement). In thinking how to circumvent such problems, it is important to emphasize the key restriction for any potential solution – namely, since in its searches for syntheses leading to non-trivial targets *Chematica* evaluates millions of synthetic possibilities, it is essential that any such solutions are very fast, taking no longer than on the order of 10 milliseconds to execute. Accordingly, it is not possible to evaluate each considered reaction on-the-fly by high-end QM or MD methods. On the other hand, it is possible to use such methods prior to actual synthetic searches to pre-calculate libraries of problematic structural motifs, or parametrize much faster heuristic or machine-learned models. In addition, our general strategy has been to avoid the use of external codes/programs (which often introduce additional problems with licensing fees, etc.) and rely on in-house solutions. As narrated briefly in this section (and also in *Angew. Chem. Int. Ed.* **55**, 5904, **2016** and *Chem* **4**, 522, **2018**, and upcoming papers mentioned in specific sub-sections below), such models – despite the imposed restrictions – perform quite well and, based on our experience with *Chematica* over the past decade, have helped to improve immensely the overall quality of synthetic pathways the program predicts.

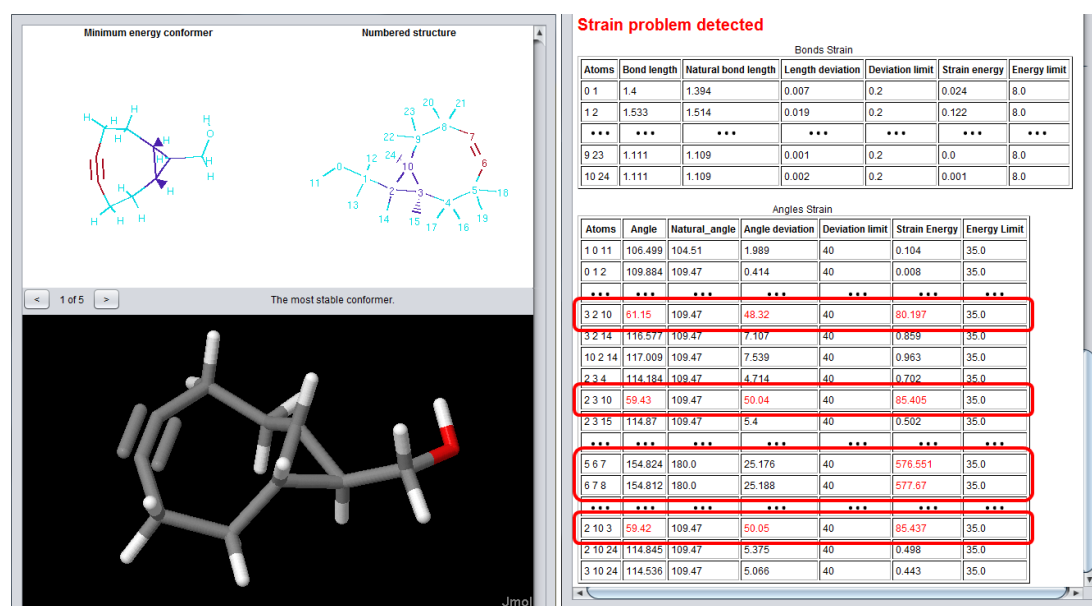
### S10.2. Problem class C1: Nonsensical or undesired structural motifs in the synthons.

Although an individual reaction transform might be "locally" fine, it can create within the synthon(s) a motif that cannot exist (e.g., a reaction introducing, in the retro direction, a double bond at a bridgehead atom of a small, bridged ring system; **Figure S16a**) or is "borderline," such that one would not like the synthesis to proceed through this motif unless it is present in the target (i.e., unless the user specifically wishes to make a molecule containing this motif, **Figure S16b**). Many of the forbidden motifs are eliminated simply based on the general chemical knowledge, e.g., the Bredt's rule for the aforementioned bridgehead atoms. For many other "suspected," motifs we performed molecular mechanics calculations at the level of UFF force-field. In



**Figure S16.** Strained motifs in retrosynthetic planning. **A**) Motif with a double bond at the bridgehead atom cannot exist and is always prohibited. **B**) In contrast, strained cyclobutane-fused aziridines can, in principle, exist but are not "welcome" synthetic intermediates – accordingly, they are permitted in *Chematica* only when the user specifies a target that contains this motif.

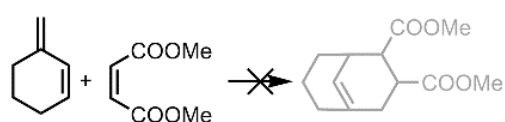
these straightforward calculations, we used the accepted non-existent motifs (recognized as such in the literature) and, for comparison, plausible motifs. We then determined the thresholds of molecular strain that separate the two classes. The results were used for three purposes – first, to create a library of several hundred motifs that are always forbidden (as in **Figure S16a**); (ii) to create a library of motifs that are forbidden unless present in the target (as in **Figure S16b**); and (iii) to use this modality after the synthetic searches are complete,” and the user can inspect a full report on any of the molecules in the syntheses generated by Chematica (**Figure S17**).



**Figure S17.** Chematica’s “Strain report” module can be used to inspect any of the intermediates present in the synthetic plans the program creates. Here, it is used to evaluate strain in bicyclo[6.1.0]nonylmethanol. Values of bond length, bond or dihedral angles that are above allowed thresholds are marked red in the table and also color-coded in the structure in the upper-left corner (light blue designate loci with no excessive strain).

**S10.3. Problem class C1: Motifs leading to impossible transition states.** In addition to structurally “faulty” synthons, there are, of course, situations in which these synthons are perfectly feasible, but the transition state, TS, between substrates and products is very high in energy and the reaction is kinetically prohibited. While full-fledged TS calculations are orders of magnitude too long to implement on on-the-fly, during Chematica’s synthetic searches, we have implemented several heuristic solutions that eliminate at least the most prevalent problems.

(i) Since all reactions in Chematica are coded based on well-defined reaction mechanisms, we can eliminate some situations in which the structure of the substrates/synthons prevents attainment of the known, proper TS conformation. One illustrative example is shown



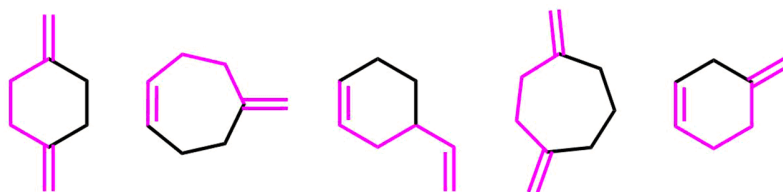
**Figure S18.** Diels-Alder reaction cannot occur when substrate is frozen in *s-trans* conformation.

in **Figure S18** whereby the *s-trans* conformation of the diene is “frozen” in the ring system and will not be able to achieve the *s-cis* conformation required for the DA's transition state. Such motifs are listed along the reaction transforms and, if detected in the synthons, a specific reaction attempt is removed from Chematica's search.

(ii) In some classes of reactions, the approximation of the TS based on the structures of the substrates is used to supplement thermodynamic considerations. The case in point here are Cope and related rearrangements in which the direction the equilibrium shifts depends on both thermodynamic considerations (electronic/substituent and ring-strain effects) as well as the kinetic factors specifying whether the thermodynamic product can be attained.

Regarding the thermodynamic component, we quantify the ability of the substituents to stabilize/destabilize double bonds by the so-called Double Bond Stabilization Energy (DBSE) concept (*J. Am. Chem. Soc.* **95**, 1179-1185, **1973**). The DBSE values for specific substituents are pre-calculated at the DFT level (M11/6-31+G(d,p)) as energy differences between this substituent present in the allylic or in the vinylic positions. The overall energetic effect, caused by changes of substituents' locations in the 1,5-diene core (as the result of the reaction), is then the sum of DBSE values. The changes in strain energy (accompanying any additional ring creation or destruction) are also pre-calculated and taken from previous studies (notably, *Eur. J. Org. Chem.* **2000**, 3117-3125, **2000**). The sum of the DBSE and ring strain energies gives the reaction enthalpy, which is then used to calculate  $\Delta G$  and the expected product/substrate ratio in the equilibrating mixture.

Regarding the kinetic component, it is aimed to qualitatively evaluate the TS structure and decide whether the thermodynamic product *can* be obtained at all. This module is based on a set of heuristics which exclude structural motifs that are not able to react in the Cope rearrangement (e.g., 1,5-dienes within ring systems, **Figure S19**).

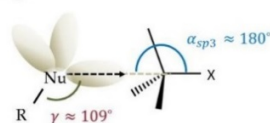


**Figure S19.** Examples of motifs forbidden as the substrates of the Cope rearrangement. The double bonds in the 1,5-diene core (pink-colored) cannot attain proper TS conformation.

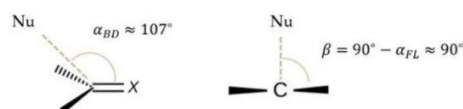


(iii) A particular case is that of intramolecular cyclizations which are essential for the synthesis of the complex natural products featuring polycyclic, bridged, or caged scaffolds. A particular synthetic transformation – say, a simple  $S_N2$  reaction – might be judged as plausible based on the endgroups of a cyclizing system, but making these groups “reach” each other to form a ring requires overcoming a high-strain TS. Looking for rapid yet relatively accurate methods to evaluate such cases, we have been combining the mechanistic knowledge of the transformations with molecular-mechanics, MM, calculations. In brief, for the most popular reaction classes we know the angles/trajectories at which the reaction partners approach one another (**Figure S20**) – we then approximate the cyclization process as proceeding along this coordinate (the calculations are usually repeated starting from different substrate conformers). For such along-the-trajectory calculations, we initially considered two freely available MM force fields (Universal Force Field (UFF) and Merck Molecular Force Field 94 (MMFF) and compared the energies they predict during some well-studied cyclizations against more precise HF/6-311+G\*\* calculations – in the end, we found that only MMFF provided decent accuracy for polycyclic/caged systems. We then used MMFF calculations to monitor molecular energy along the reaction-type-specific “coordinate of approach”. **Figure S21** illustrates the energy profiles for some cyclizations for which we were able to find publications in which these reactions were attempted and did or did not proceed (alkylation of ketone, red and green curves; from *J. Am. Chem. Soc.* **95**, 8339, **1973**), opening of epoxide (blue; *J. Mol. Catal. A Chem.*, **142**, 333, **1999**), cyclisation forming Loline’s skeleton (orange; *Nature Chemistry*, **3**, 543, **2011**), and cyclisation executed *en route* to cytosine (black; *J. Org. Chem.* **83**, 9088, **2018**). The cyclization corresponding to the highest energy (red curve) does not proceed in experiment – accordingly, the threshold for admissible cyclizations based on  $S_N2$  is set below these energy values (for the distances of reacting centers below ca. 3.5 Å; at higher separations, the “stretching” of the molecule along the reaction coordinate may be unphysical, especially for smaller ring systems). Similar calculations are performed for transforms based on other well-studied reaction mechanisms. Naturally, this is only a crude approach but, based on our experience with *Chematica* it can eliminate at least the most chemically

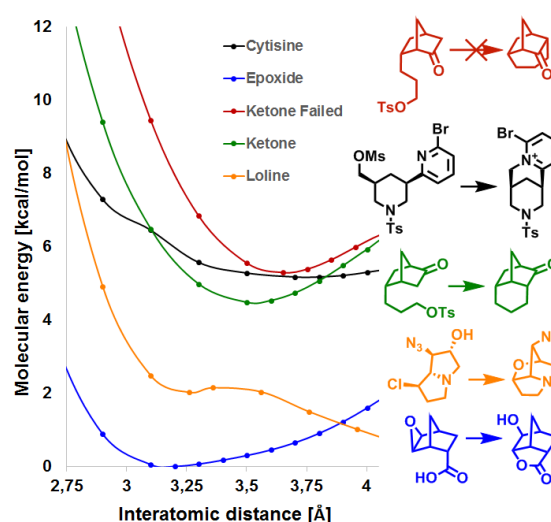
- Constraints for sp<sup>3</sup>-sp<sup>3</sup>



- sp<sup>2</sup> hybridization of central atom (Bürgi-Dunitz and Flippin-Lodge angles):



**Figure S20.** Enforcing proper trajectory of approach. For the major types of reactions, the mutual arrangement of reacting species is known, as illustrated here for a nucleophilic attack on a tetrahedral carbon, or on a sp<sup>2</sup> carbon atom (along the so-called Bürgi-Dunitz and Flippin-Lodge angles).



**Figure S21.** Examples of some possible and impossible cyclizations based on  $S_N2$  reactions.

proceed (alkylation of ketone, red and green curves; from *J. Am. Chem. Soc.* **95**, 8339, **1973**), opening of epoxide (blue; *J. Mol. Catal. A Chem.*, **142**, 333, **1999**), cyclisation forming Loline’s skeleton (orange; *Nature Chemistry*, **3**, 543, **2011**), and cyclisation executed *en route* to cytosine (black; *J. Org. Chem.* **83**, 9088, **2018**). The cyclization corresponding to the highest energy (red curve) does not proceed in experiment – accordingly, the threshold for admissible cyclizations based on  $S_N2$  is set below these energy values (for the distances of reacting centers below ca. 3.5 Å; at higher separations, the “stretching” of the molecule along the reaction coordinate may be unphysical, especially for smaller ring systems). Similar calculations are performed for transforms based on other well-studied reaction mechanisms. Naturally, this is only a crude approach but, based on our experience with *Chematica* it can eliminate at least the most chemically

“offensive” cases that a seasoned organic chemists might immediately find suspicious. We note that, for now, the codes underlying these calculations are not parallelized and still require optimization; currently, the calculations take few seconds and, consequently, are available for post-synthesis-planning scrutiny of specific reactions in *Chematica*-generated pathways (i.e., not yet on-the-fly, during synthetic searches). The results of the outlined studies are being prepared for a separate publication.

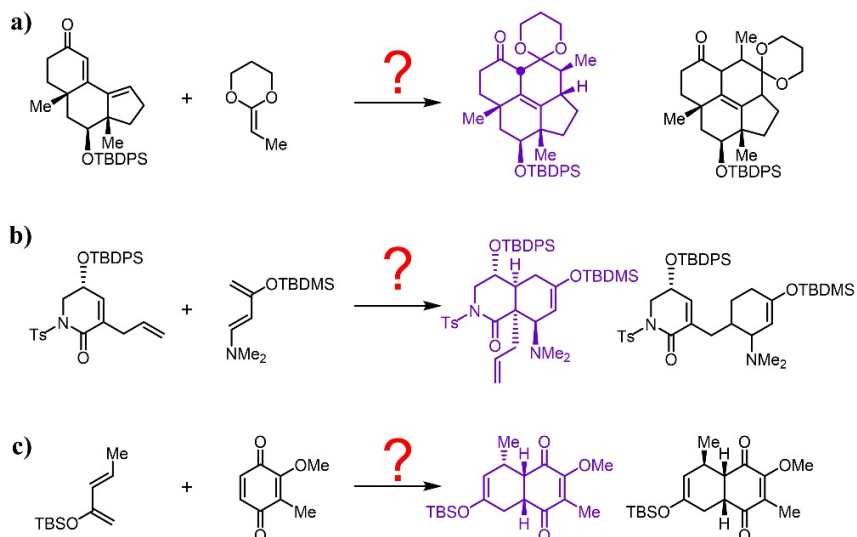
**S10.4. Problem class C2: Accounting for long-distance electronic effects – example of aromatic substitutions.** In this section, we turn our attention to reaction types for which specification of all possible arrangements and types of relevant substituents is impractical. A case in point are aromatic substitutions for which enumerating possible substituent arrangements even in the most popular aromatic systems would already require extremely large numbers of specific transforms. Instead, we encode a very general reaction transform (spanning an attacking electrophile or nucleophile plus an aromatic carbon), but determine the applicability of such a rule based on the calculations of electron populations (and related heuristics) within the aromatic systems. In the early versions of *Chematica*, the calculations were based on the Hückel method but the accuracy (assessed against published substitution patterns) was only ~75%. As we described in detail in the SI, Section 7 of our recent *Chem* paper (*Chem* **4**, 522, **2018**), other published methods were either not accurate over diverse sets of aromatic systems (e.g., those based on  $^1\text{H}$  and  $^{13}\text{C}$  NMR shifts had low predictive power, while those based on Hammett constants or the so-called electrostatic potential at nuclei, EPN, were only marginally satisfactory) or were accurate but required time-consuming QM calculations, one for each possible reaction site (e.g., the proton/electrophile affinity methods). Consequently, we developed a “hybrid” approach – one combining Hammett substituent constants, proton affinities averaged over all aromatic carbons within a specific ring type (pre-calculated at the DFT level of theory using B3LYP functional with 6-31+G\* basis set), the Hückel method (with parameters for heteroatoms taken mostly from *J. Org. Chem.* **45**, 4801-4802, **1980**), as well as some additional heuristics. This model proved rapid enough for on-the-fly calculations during retrosynthetic searches while offering accuracy around 90% (assessed against 18,000 literature examples). With more details described in the Supplementary Section **S5.1** of *Chem* **4**, 522, **2018** (and in our upcoming, more specialized publication), we narrate the model only briefly. It consists of two parts whose main subroutines are:

*Part 1. Assessing the most reactive position within each ring present.* For benzene rings, regioselectivity is determined by Hammett substituent constants with additional heuristics added to properly treat strongly donating groups. In heterocyclic rings, regioselectivity is dictated predominantly by heteroatoms, and for common ring types, we simply catalogued the literature knowledge as to which position is most active. For heteroaromatics with additional substituents, the situation is more complicated. For example, electrophilic substitutions at pyridines typically occur at the most active “meta” (“3” or “5”) positions relative to nitrogen. In pyridines bearing a strongly donating group at “meta” (“3”) position, substitution takes place not at the second “meta” (“5”) position but at the “ortho” (“6”) position relative to nitrogen. In order to include such dependencies, we supplemented these empirical rules with Hammett constants to quantitatively measure the effects of such substituents. For polycyclic aromatic hydrocarbons (PAH), detection of the most active ring and position cannot be achieved based on Hammett constants – instead, for this class of compounds,

we use the Hückel model which offers good accuracy against experimental/literature results (unlike in heterocycles, for which Hückel fails dismally).

*Part 2. Determining the most reactive ring.* If a molecule contains more than one aromatic ring, we sequentially remove less active rings. The removal procedure is itself divided into several steps. First, less active ring(s) within fused/conjugated systems are removed based on heuristics tailor-made for specific ring systems and accounting for substituent effects via the Hammett constants. In the second step, the algorithm performs pairwise comparisons of all remaining rings and – based on the heuristic rules taking into account ring type, presence of strongly donating/withdrawing groups, position of the most active site relative to a heteroatom, and more – removes the less active rings from each pair. In the third step, “activity” of the remaining rings is examined in more detail via proton affinities averaged over each ring and corrected for the presence of specific substituents. As mentioned above, these ring-averaged proton affinities (RAPA) are pre-calculated for different ring type templates (at the DFT level of theory using B3LYP functional with 6-31+G\* basis set), and finally corrected for specific substituents and substitution patterns using Hammett constants.

**S10.5. Problem class C2: Quantifying substituent effects in popular, non-aromatic reaction classes – example of Diels-Alder cycloadditions.** Whereas for aromatic substitutions, the key effects from beyond the reaction core are electronic, this is not a general situation and, typically, both electronic and steric effects need to be taken in consideration. For the vast majority of Chematica's reactions such considerations are included in the reaction transforms by carefully specifying the scope of admissible substituents. In some cases, however, enumeration of possible combinations of substituents is impractical and would, at best, result in very large numbers of reaction variants (which would slow down Chematica's operation). For the most popular transforms, for which thousands of reaction examples are available in the literature, we remedy this problem by defining a “narrow” reaction core (like in aromatic substitutions) but, instead of performing QM/MM calculations, we do Machine-Learning. As an example of this approach we focus in this section on the synthetically powerful Diels-Alder, DA, cycloadditions – in particular, on the problem how to predict their regio-, site- and diastereoselective outcomes (**Figure S22**). The brief summary based on our upcoming publication (W. Beker, E.P. Gajewska, T. Badowski, B.A. Grzybowski, *Angew. Chem. Int. Ed.* <https://doi.org/10.1002/anie.201806920>, **2018**) follows.



**Figure S22.** Examples of possible outcomes of the Diels-Alder reactions taken from some “classic” total syntheses. Products that were experimentally obtained and also correctly predicted by our Random-Forest classifiers are colored in violet. **a)** Regioselectivity in the DA reaction used by in the total synthesis of Rippertenol (Snyder *et. al.*) is dictated mainly by electronic factors. **b)** Site-selectivity in the DA reaction used by Danishefsky in the synthesis of Xestocyclamine A. Only one of the two possible dienophiles is reacting to give the violet-colored product. **c)** Diastereoselectivity of the DA reaction leading to a desired intermediate in Nicolaou’s total synthesis of Colombiasin A. Figure adapted with permission from (W. Beker, E.P. Gajewska, T. Badowski, B.A. Grzybowski, *Angew. Chem. Int. Ed.* <https://doi.org/10.1002/anie.201806920>, **2018**).

Initially, we attempted to determine such outcomes by QM methods but these calculations took hours to complete and offered, at best, accuracy of 82% (e.g., even for state-of-the-art Parr functions obtained by calculating wave functions for neutral, cationic-radical and anionic-radical species of each diene/dienophile, using B3LYP functional [open-shell for radicals] and 6-31+G\*\* basis set with diffuse functions on heavy atoms to better describe the third- and fourth-row elements). Consequently, we turned our attention to ML methods and prepared a set of several thousand of literature-reported DA precedents – about 3,000 examples in which regioselective outcomes were possible, ~1,000 with possible site-selectivity, and another ~3,000 with different possible diastereoselective outcomes. For each of these classes, we trained and tested (with five-fold cross validation) models ranging from deep neural networks to random forest classifiers. Interestingly, the choice of a specific method was of secondary importance – what mattered most is that the descriptors used to encode substituents on the diene and the dienophile carry with them real physical meaning. Indeed, we showed that when Hammett constants reflecting substituents’ propensities to donate/withdraw electrons and the so-called TSEI indices quantifying steric hindrance were used, the models achieved remarkable accuracies: 93.6% for the prediction of regioselectivity, 91.3% for site-selectivity, and 89.2% for diastereoselectivity. Notably, all other representations that do not capture the stereoelectronic effects (i.e., descriptors such as ECFP4, MACCS, or RDKit fingerprints, etc.) did much

worse in terms of accuracy and/or transferability to cases the machine has not seen during training. Full theoretical details are part of the aforementioned *Angewandte* paper (<https://doi.org/10.1002/anie.201806920>) and the reader can use our Diels-Alder predictor at <http://dielsalderapp.grzybowski.org.pl/>.

### **Section S11. Caption to Movie S1.**

**Movie S1. Design of patent-evading synthetic plans for Linezolid.** After introducing the target molecule (00:02), scoring functions (00:05-00:10) and stop conditions (00:12), a search without any additional constraints is started (00:13). After ca. 90 s, the search is stopped and the top-scoring pathways found are displayed. From these, top ones (relying on the condensation of epoxide with carbamate) are displayed in detail (00:25-00:55). Color coding: red = commercially available chemicals; green = molecules with syntheses already reported in the literature (user can further query those in Chematica's Network of Chemistry, NOC, module); violet = molecules unknown in the NOC. Desired bond set is then marked as "unbreakable" (01:00) and the search is performed again. After ca. 3 minutes, the search is stopped and the top-scoring pathways are displayed. From these, the three top-scoring ones (now utilizing commercially available hydroxymethyloxazolidinone) are scrutinized in detail (1:18-2:03). Finally, prices of commercially available starting materials (numbers over red nodes) and synthetic popularities of molecules with known syntheses (numbers over green nodes) are displayed (02:07).

Cite this: *Chem. Sci.*, 2020, 11, 6736

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 28th March 2020  
Accepted 2nd June 2020

DOI: 10.1039/d0sc01799j

rsc.li/chemical-science

# Computer-generated “synthetic contingency” plans at times of logistics and supply problems: scenarios for hydroxychloroquine and remdesivir†

Sara Szymkuć,<sup>‡a</sup> Ewa P. Gajewska,<sup>‡a</sup> Karol Molga,<sup>‡a</sup> Agnieszka Wotos,<sup>a</sup> Rafał Roszak,<sup>a</sup> Wiktor Beker,<sup>a</sup> Martyna Moskal,<sup>a</sup> Piotr Dittwald<sup>a</sup> and Bartosz A. Grzybowski<sup>‡a</sup>  \*<sup>abc</sup>

A computer program for retrosynthetic planning helps develop multiple “synthetic contingency” plans for hydroxychloroquine and also routes leading to remdesivir, both promising but yet unproven medications against COVID-19. These plans are designed to navigate, as much as possible, around known and patented routes and to commence from inexpensive and diverse starting materials, so as to ensure supply in case of anticipated market shortages of commonly used substrates. Looking beyond the current COVID-19 pandemic, development of similar contingency syntheses is advocated for other already-approved medications, in case such medications become urgently needed in mass quantities to face other public-health emergencies.

## Introduction

Faced with the eruption of the coronavirus pandemic, individual academic and clinical laboratories, funding agencies, and entire governments are intensifying efforts to develop and deploy safe and effective vaccines and/or antiviral medications. While vaccines may become available within a year or so, development and approval of a brand new drug will, most likely, require a significantly longer time, not relevant to the current exigency. Accordingly, much of the ongoing effort has been focused on drugs that are already approved and could be repurposed against COVID-19. For example, reports have been emerging in the scientific literature<sup>1,2</sup> that chloroquine (CQ) and hydroxychloroquine (HCQ) – vintage drugs to treat malaria as well as some autoimmune diseases – seem to inhibit SARS-CoV-2 infection *in vitro* by slowing down entry of viruses into the cell and by blocking their transport from early endosomes to endolysosomes,<sup>2,3</sup> causing noticeable enlargement of the former and affecting the pH levels<sup>4</sup> within the endolysosomal tract. Since HCQ is less toxic than CQ<sup>5,6</sup> and given the current scarcity of viable alternatives, the use of this drug against the COVID-19 pandemic has been sanctioned by the FDA (but only

temporarily and in clinical settings), even in the absence of comprehensive clinical data. In another example, Gilead's remdesivir – originally developed to treat hepatitis C and then tested against Ebola and Marburg disease – has shown promise<sup>7,8</sup> and recently gained the FDA's authorization for emergency use in the U.S. Still, should either HCQ, remdesivir or any other repurposed drug prove effective – and, to reiterate, this is still uncertain – the demand might soon surpass supply. Moreover, the key synthetic methods involved in their production are very often protected by patents (including some very recent ones, even for the off-label HCQ, see Fig. 1), and we cannot exclude the possibility that monetary, corporate interests might interfere with humanitarian inspirations. In addition, the failure of the worldwide logistics and supply chains that has accompanied the COVID-19 pandemic might render some key substrates temporarily unavailable, in effect delaying the execution of the proven synthetic routes and calling for alternative synthetic solutions. Anticipating such complications, we harnessed the power of Chematica<sup>9–18</sup> – an experimentally tested<sup>10,11</sup> platform for computer-assisted retrosynthesis of both known and unknown target molecules – to design syntheses of HCQ and remdesivir. We were most interested in synthetic plans that would (1) commence from various inexpensive and popular starting materials (so that the syntheses minimize the abovementioned supply problems); (2) circumvent patented methodologies whenever possible;<sup>16</sup> and (3) minimize the use of expensive methodologies and/or reagents. As described below, these analyses had different outcomes for HCQ and remdesivir. For the structurally simpler HCQ, Chematica was able to rapidly identify a large number of alternative syntheses differing in the key disconnections and meeting criteria (1)–(3). For the structurally more complex

<sup>a</sup>Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 02-224, Poland. E-mail: nanogrzybowski@gmail.com

<sup>b</sup>IBS Center for Soft and Living Matter, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

<sup>c</sup>Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0sc01799j

‡ The authors contributed equally.



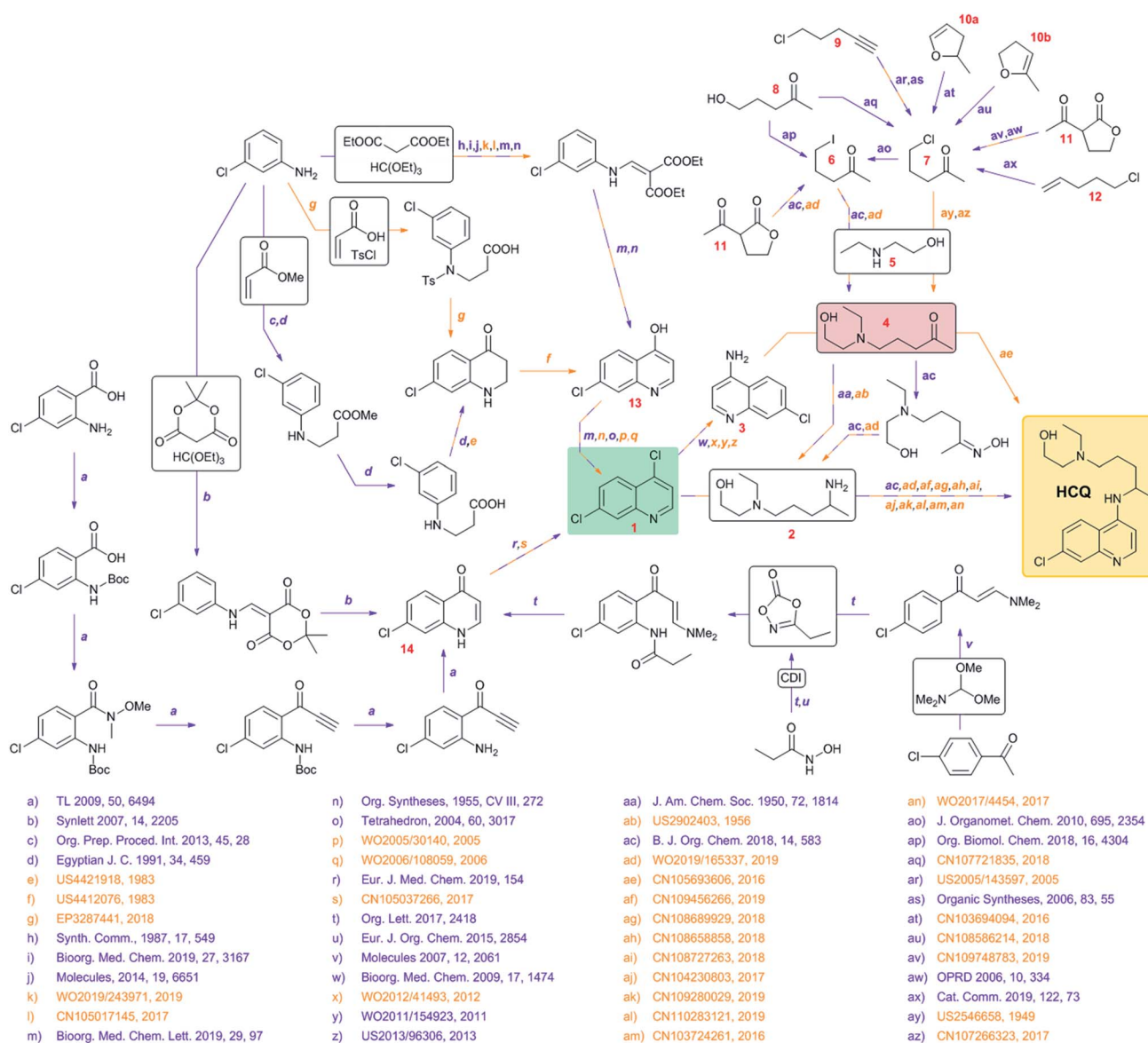


Fig. 1 Synthetic network summarizing known syntheses of hydroxychloroquine, HCQ, along with the pertinent literature (patents in orange, scientific publications in violet). The two key intermediates are highlighted by colored boxes: in terms of availability and price, 1 (green) is significantly less problematic than 4 (red). For the full literature references, see the ESI, Section S3.†

remdesivir, the program also found chemically correct routes but very similar to the current and patented approaches (with minor differences in the choices of low-cost substrates and protecting groups). These results illustrate the prowess of modern computer-assisted retrosynthesis in cases when multiple methodologies can be applied to a target of interest (here, the HCQ scenario), but also their limitations when challenged with scaffolds that can be made by only one known methodology (here, remdesivir's case). It should be noted, however, that the first of these scenarios is by far more common (see ref. 10,11 and 16 describing novel syntheses designed by Chematica and validated by experiments) and computers can be of tangible help in suggesting alternative and, as we demonstrate here, cost-effective syntheses on timescales commensurate with emergency situations such as the one presented by the

COVID-19 pandemic. In this context, we remain open to performing – on a *pro bono* basis – similar synthetic analyses for organizations considering production and unrestricted (both geographically and economically) distribution of other potential anti-COVID-19 agents, should such agents become available in the near future.

## Computational methods

Chematica is a sophisticated platform for fully automated design of pathways leading to arbitrary (*i.e.*, both known and new) targets. The software combines elements of network theory<sup>17,18</sup> with an expert knowledge-base of synthetic transformations as well as multiple reaction-evaluation routines (based on machine learning,<sup>12,13</sup> quantum mechanics,<sup>9,10</sup> and



molecular dynamics<sup>10,14</sup>) to search over vast trees of synthetic possibilities. The reaction transforms (currently, ~100 000) are expert-coded based on the underlying reaction mechanisms and are broader than any specific literature precedents (for comparison with machine extraction of rules from reaction repositories, see ref. 14). Each rule specifies the scope of admissible substituents, accounts for stereo- and regio-chemistry requirements, recognizes groups that must be protected under given reaction conditions, and identifies functionalities that are outright incompatible. The searches are guided by combinations of functions (either heuristic<sup>9,10</sup> or best-in-class AI-based<sup>13</sup>) that score both synthetic positions as well as costs of individual reactions. The pathways identified by the program terminate in either commercially available chemicals (here, more than 200 000 molecules from Sigma-Aldrich catalogs, each with the price per unit quantity; also see below for price re-scaling) or those already known in the literature (*ca.* 6 million substances, each accompanied by a measure of synthetic popularity,<sup>9,17</sup> *i.e.*, how many times a given substance was used in prior syntheses). Since the program typically identifies a large number of possible routes, the network of viable syntheses already found is queried by dynamic linear programming algorithms to select pathways with the lowest cost (propagated recursively from substrates to products with the consideration of estimated yields) and highest diversity of the proposed retrosynthetic strategies.<sup>15</sup> In setting up a particular search, the user can specify parameters influencing the economy of the solutions, notably, the upper price threshold and/or the minimal synthetic popularity of the starting materials, the relative cost of performing a reaction operation, or the desired estimated yield. The user can also eliminate certain types of transformations or unwanted reagents (*e.g.*, expensive catalysts). He/she is also able to “lock” certain bonds or fragments in the target such that they are not disconnected along the synthetic plan – as described in detail in ref. 16; this functionality is useful in navigating around patented routes. Depending on the number of imposed constraints, a typical search for a drug-like molecule takes from few to tens of minutes and within this time inspects tens to hundreds of thousands of reaction candidates. Ultimately, a user-specified number of top-scoring pathways (typically 50–100) are returned and displayed as bipartite graphs with nodes that are expandable to display molecular structures, suggested reaction conditions typical to a given reaction class, and more.<sup>9,10</sup>

## Results and discussion

The results described in the following come from various searches executed by our team over the course of two days and using three machines, each with 64 cores. Multiple searches were performed on the newest version of the program that is currently being transitioned onto the commercial Synthia™ platform owned and distributed by Sigma-Aldrich/Merck KGaA. These searches used various parameters (summarized in the ESI, Section S2†) to reflect different economic scenarios of the desired syntheses and with different types of the above-mentioned constraints. In all, the searches considered on the

order of millions of potential intermediates and synthetic plans. The common feature of the searches was the desire to offer alternatives to existing syntheses and to suggest multiple synthetic plans using diverse but inexpensive starting materials. In considering the prices of the starting materials, we realized that catalog prices from a specialty-chemicals retailer such as Sigma-Aldrich, S-A, are understandably higher than those from whole-sale producers. What is important for the validity of our analysis, however, is that the S-A prices correlate with the wholesale ones, as evidenced by the data plotted in Fig. 2 and spanning substrates of the new syntheses of HCQ we identified. We will discuss these issues in more detail along with specific routes.

### Known syntheses of hydroxychloroquine (HCQ)

To begin with, we surveyed the available literature to construct a synthetic network summarizing currently known syntheses of HCQ (Fig. 1). Somewhat remarkably, although HCQ has been off-patent for decades, a large proportion of methods involved have been patented, some quite recently, substantiating our concern of potential IP complications in case of emergency production by independent agents. These solutions hinge on the late stage attachment of the side chain performed *via* either (i) nucleophilic aromatic substitution of dichloroquinoline **1** and amine **2** or (ii) reductive amination of aminochloroquinoline **3** (itself derived from **1**) and ketone **4**, the latter being the starting material for the preparation of **2**. The two “hubs” of the network are, obviously, **1** or **4** though they are quite different from the economic and logistic points of view. The heterocyclic part of HCQ, **1**, is rather inexpensive (1.50 \$ per g from S-A, 0.26 \$ per g from Biosynth Carbosynth) and in case of supply problems, can be sourced (in 94% yield, *via* chlorination using POCl<sub>3</sub>) from hydroxychloroquinoline **13** which, in turn, can be made in ~40% yields in two steps either from 3-chloroaniline, diethyl malonate and ethyl orthoformate (respectively, 9.51 \$ per g from S-A, 0.05 \$ per g from Oakwood Chemical, OC; 0.04 \$ per g from S-A, 0.015 \$ per g from OC; 0.12 \$ per g from S-A, 0.03 \$ per g from OC) or from 3-chloroaniline, acrylic acid (or methyl acrylate) and tosyl chloride (respectively, 9.51 \$ per g from S-A, 0.05 \$ per g from OC; 1.48 \$ per g from S-A, 0.02 \$ per g from Gelest Inc.; 1.43 \$ per g from S-A, 0.03 \$ per g from Alfa Aesar; 0.02 \$ per g from S-A, 0.04 \$ per g from Alfa Aesar). Alternatively, **1** can be obtained from chloroquinolinone **14**, available *via* a similar two-step sequence starting from 3-chloroaniline, Meldrum's acid (1.75 \$ per g from S-A, 0.07 \$ per g from Aba-ChemScene) and ethyl orthoformate. Some more recent approaches for the preparation of **14** rely on different starting materials (4-chloroacetophenone or 2-amino-4-chlorobenzoic acid) but require at least four steps. Additionally, C–H activation of enamionone derived from chloroacetophenone requires an expensive bimetallic catalyst (Cp\*Co(CO)I<sub>2</sub>/AgSbF<sub>6</sub>).

In contrast, ketone **4** is not easily sourced (no prices listed on eMolecules) and is likely the production bottleneck. This intermediate can be prepared *via* alkylation of aminoalcohol **5** (0.11 \$ per g from S-A, 0.04 \$ per g from Acros Organics) with haloketones **6/7**, which in turn can be derived from





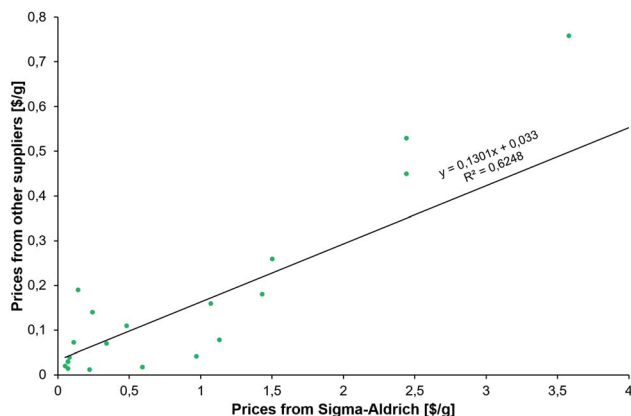


Fig. 2 Correlation between prices (scaled per gram) of the substrates for new syntheses of HCQ (see Fig. 3) from (x-axis) Sigma-Aldrich' catalog interfaced with Chematica and (y-axis) the least inexpensive options we were able to identify from larger-scale manufacturers. Not surprisingly, the latter are, on average, twelve times less expensive (median = 4.67).

hydroxyketone **8**, chloroalkyne **9**, enol ethers **10a/10b**, lactone **11** or chloroalkene **12**. These substrates, except for **8** and **11** (both available for less than 0.5 \$ per g from suppliers like Combi-Blocks or ChemScene), are relatively expensive (from 4 \$ per g to even 585 \$ per g) so these methodologies are probably unsuitable for industrial up-scaling.

### Syntheses of hydroxychloroquine (HCQ) designed automatically by Chematica

Without any search constraints, Chematica generally identified, among its top choices, many of these known solutions (or their very close analogs, differing in insignificant details). The program began to find substantially different pathways especially upon application of restrictive thresholds for the prices/popularities of the starting materials. Fig. 3 summarizes 17 routes we found most economically viable, concise, and diverse (see the ESI, Section 1† for enlarged views). In addition to routes relying on nucleophilic aromatic substitution of dichloroquinoline and reductive amination of aminochloroquinoline, the software was able to avoid these steps, replacing them with methodologies such as A3-coupling (path **15**), Cu-cat. coupling between a heteroaryl iodide and an amine (paths **2** and **3**), three-component reaction between an amine, an aldehyde and a halide under Barbier-type conditions (path **14**), or alkylation of an aromatic amine with an alkyl iodide (path **11**). Other innovative aspects of Chematica's plans are manifested in the routes to prepare the side-chain of HCQ which, as we saw before, is the major factor driving availability/cost of the overall synthesis. The machine's proposals include, for example, opening of a lactam with a Grignard reagent (paths **1** and **16**) or alkylation of a lactone followed by ring-opening to install a primary iodide functionality – which is a very convenient group for subsequent alkylation (path **13**). Other interesting approaches use the multicomponent Mannich

reaction. In pathway **2**, this reaction is combined with a subsequent Henry reaction, and in pathway **10** it follows a Curtius rearrangement. Both the Henry reaction and Curtius rearrangement are interesting alternatives to reductive amination or reduction of an oxime used for the introduction of the nitrogen atom. As already mentioned, all of these proposed routes avoid expensive catalysts and commence from inexpensive starting materials, readily available in large quantities (e.g., ethylamine at 0.018 \$ per g, 2-bromoethyl acetate at 0.22 \$ per g, 5-chloro-2-pentanone at 0.08 \$ per g, or ethanolamine at 0.012 \$ per g). Only a few of these substrates were used in previously published/patented syntheses. In Fig. 3, their prices are indicated in red font. We observe that several pathways start from materials available in ton quantities or from biofeedstocks, making these routes especially appealing for scale-up. For example, path **12** commences from hydroxyketone (which can be produced from pentoses *via* a furfural intermediate) and uses ethylene oxide to install the required hydroxyethyl side chain. Path **3** navigates the HCQ's side chain to levulinic acid, preparable on ton-scales from carbohydrates.

In addition to the above analyses, it is also prudent to consider syntheses of the dichloroquinoline starting material which, as already mentioned, is currently inexpensive but may become less so if demand surpasses supply. As shown in Fig. 4, the program rediscovered the patented approach (pathway marked as **18**) based on Meldrum's acid and also generated a similar path, **19**, starting from a commercially available dimethylamino Meldrum's acid adduct not used in the patented synthesis. Moreover, Chematica proposed three alternative routes, **20–22**, which rely on copper-catalyzed cyclization forming the quinolone ring. These pathways involve different methodologies for the preparation of the acetophenone derivative, which is commercially available but expensive (29.4 \$ per g from S-A, 11.07 \$ per g from Chemenu). Specifically, this compound can be synthesized from a methyl benzoate derivative (6.56 \$ per g from S-A, 1.16 \$ per g from PharmaBlock) either *via* Tebbe olefination followed by ketone synthesis (pathway **20**) or *via* the Grignard reaction (pathway **21**). Chematica also proposed the synthesis of the acetophenone derivative from 2-bromo-4-chlorotoluene (0.96 \$ per g from S-A, 0.35 \$ per g from BLD Pharmatech) *via* benzylic oxidation followed by the Grignard reaction and oxidation of an alcohol (pathway **22**).

We also performed searches to synthesize HCQ taking advantage of Chematica's functionality to avoid certain user-specified substructures. Here, we excluded from the searches both 8-chloro-4-chloroquinoline and 8-chloro-4-iodoquinoline scaffolds. The top-scoring routes avoiding these motifs share a common key intermediate, 7-chloro-1H-quinolin-4-one, also identified as a precursor in some of the pathways (**3–5**) leading to dichloroquinoline (see Fig. 5). However, "downstream" syntheses of this heterocycle differ in both synthetic plans. In the first pathway marked as **23**, the quinolone scaffold is assembled from a bromoacetophenone derivative and formamide *via* copper-catalyzed cyclization. These starting materials were also used to synthesize dichloroquinoline and both their



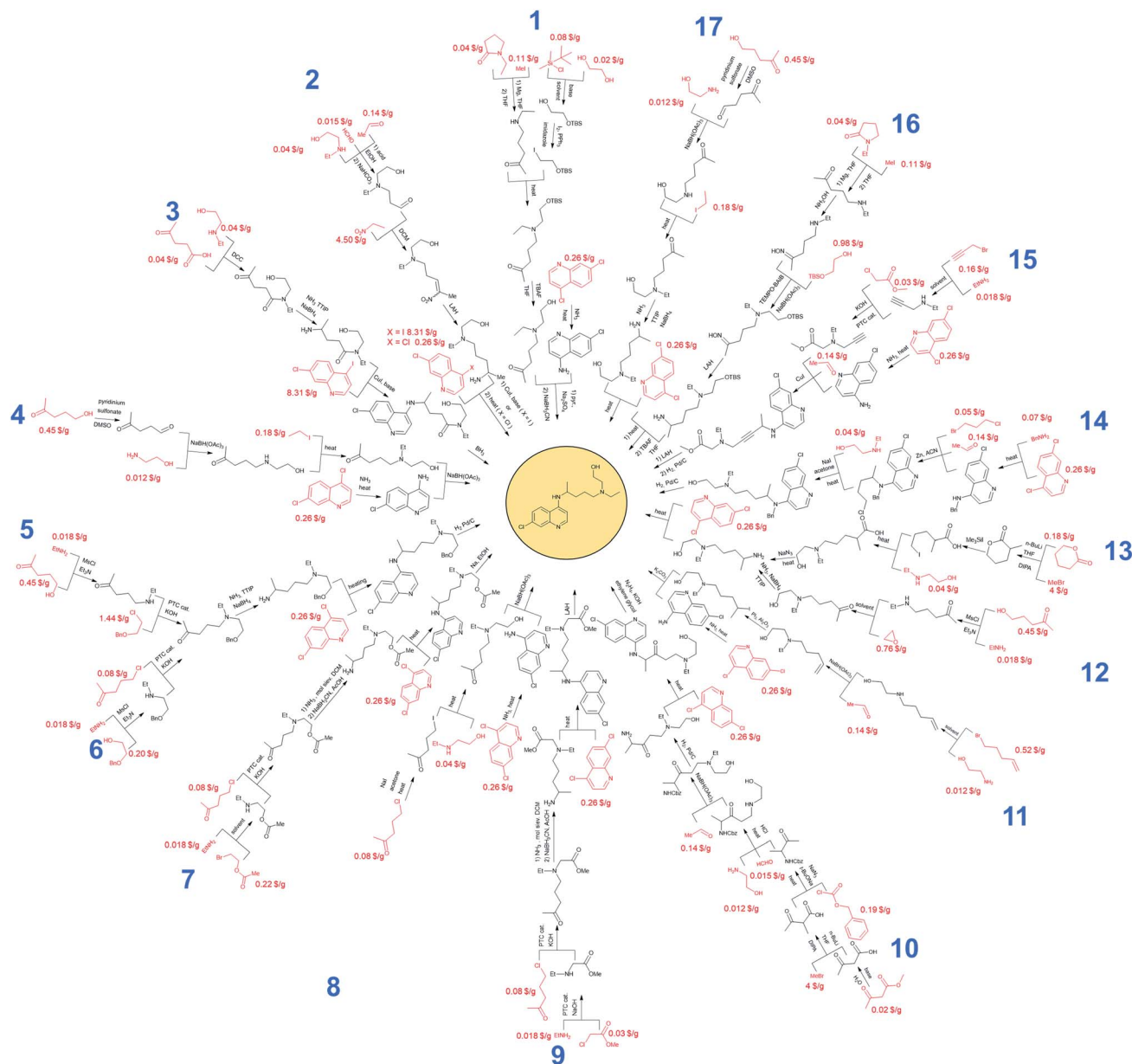


Fig. 3 Novel and economically appealing syntheses of hydroxychloroquine (HCQ) designed automatically by Chematica. Commercially available substrates and their prices (the lowest ones we were able to identify) scaled to \$ per g are colored in red.

pricing and accessibility from less expensive sources were described in the previous paragraph. The synthetic strategy for the alkyl part of hydroxychloroquine is the same as in path 16 from Fig. 3. The second synthetic route, numbered 24 in Fig. 5, uses a heterocyclic starting material, benzopyrone, which is further transformed into the quinolinone precursor in a two-step sequence. The cost of this substrate is 46.50 \$ per g at S-A and 26 \$ per g at *CombiBlocks*, making this solution less economical. The approach used to construct the alkyl part of the molecule resembles the strategy from path 5 in Fig. 3.

Finally, we comment on the conditions proposed by the software, especially on the usability of certain reagents at large scales. The reader will, no doubt, recognize many conditions used widely on industrial scales – for instance, the use of  $H_2$ /

$Pd^{19,20}$  in pathways 5, 6, 10, 14, and 15 or  $PTC^{21}$  in pathways 5, 7, 9, 15, and 23. On the other hand, the uses of reagents such as TEMPO, *n*-BuLi or LAH merit further discussion.

TEMPO, used in pathways 16 and 24, is relatively expensive, which might be a deterrent in using it on large scales as a stoichiometric reagent. However, it is widely used in industrial organic synthesis, although usually in small amounts (0.1–10% mol).<sup>22</sup> In this context, we note that the *N*-alkylation methodology (*via* a one-pot oxidation-reduction amination sequence)<sup>23</sup> proposed by the software also uses reduced amounts of TEMPO (0.2 mmol of TEMPO to 1 mmol of the substrate), suggesting some potential for scale-up. Alternatively, one could consider its replacement with 4-hydroxy derivatives (*e.g.*, 4-hydroxy-TEMPO or 4-acetamido-TEMPO)



obtained from an inexpensive (3 \$ per kg) triacetoneamine precursor.<sup>22</sup> Moreover, if the cost of the one-pot methodology still proves impractical, it could readily be replaced with a “classic” two-step approach involving oxidation of a primary alcohol followed by reductive amination. Such a strategy was, in fact, used in pathway 17 for the synthesis of 5-[ethyl(2-hydroxyethyl)amino]pentan-2-one. Therein, Chematica's suggestion for the oxidation of a primary alcohol was the sulphur trioxide/pyridine complex, an oxidant used in industry, *e.g.*, in the 190 kg scale synthesis of an HIV protease inhibitor.<sup>24</sup> For the reductive amination step, the software proposed  $\text{NaBH}(\text{OAc})_3$  (the same reductant was proposed in the one-pot procedure) for which multi-kilogram-scale syntheses were described in the literature.<sup>25,26</sup> In addition,  $\text{NaBH}(\text{OAc})_3$  is easily obtainable *in situ* from  $\text{NaBH}_4$  and  $\text{AcOH}$ .<sup>27</sup>

Regarding *n*-butyllithium – used in pathways 10 and 13 – the use of this reagent requires safety precautions, but there are examples of its usage in the pharmaceutical industry.<sup>28</sup> For instance, *n*-BuLi has been used in kilogram-scale syntheses of 2,2-dimethyl-1-(4-methylthio-5-pyrimidinyl)indane<sup>29</sup> and 1-[4-(2-chloroethoxy)phenyl]-1-(4-iodophenyl)-2-phenyl-1-butanol<sup>30</sup> as well as in the synthesis of AZD6906 performed in a flow-reactor.<sup>31</sup> In Chematica's synthetic plans, both reactions employing *n*-butyllithium are alkylations of esters and the reagent is used to generate LDA *in situ*; this approach is also used in industry owing to its cost efficiency.<sup>29</sup>

Regarding LAH – used in pathways 2, 9 and 16 – and according to the review by Magano and Dunetz on large-scale reductions,<sup>32</sup> it is the most common reagent for the reduction of acyclic esters. This assertion is supported by numerous kilogram-scale syntheses using this reducing agent (*e.g.*, Glaxo's synthesis of Sodelglitazar,<sup>33</sup> Eli Lilly's route to DPP IV Inhibitor LY2497282,<sup>34</sup> or Sanochemia's nine-step synthesis leading to (–)-galanthamine<sup>35</sup>).

### Chematica's syntheses of remdesivir

Since HCQ is a synthetically simple target, we were also interested in Chematica's performance on a more complex target such as remdesivir. In this case, the program was able to construct chemically correct pathways though the key steps in these routes – marked as A, B and C in Fig. 6 – were identical (save minor differences in the protecting groups used) to the known approach. The choice of these key retrosynthetic disconnections is perhaps not surprising given that the presence of the *C*-arylated ribose, phosphorylated alanine and pyrrolotriazine fragments leaves relatively little room for other strategies. For instance, assuming that these building blocks are kept intact during synthesis, there does not seem to be a more straightforward means to construct the cyanated *C*-aryl nucleoside than *via* addition of an appropriate organometallic reagent to a protected lactone and subsequent cyanation of the hemiacetal obtained. In light of these considerations, the major virtue of computational

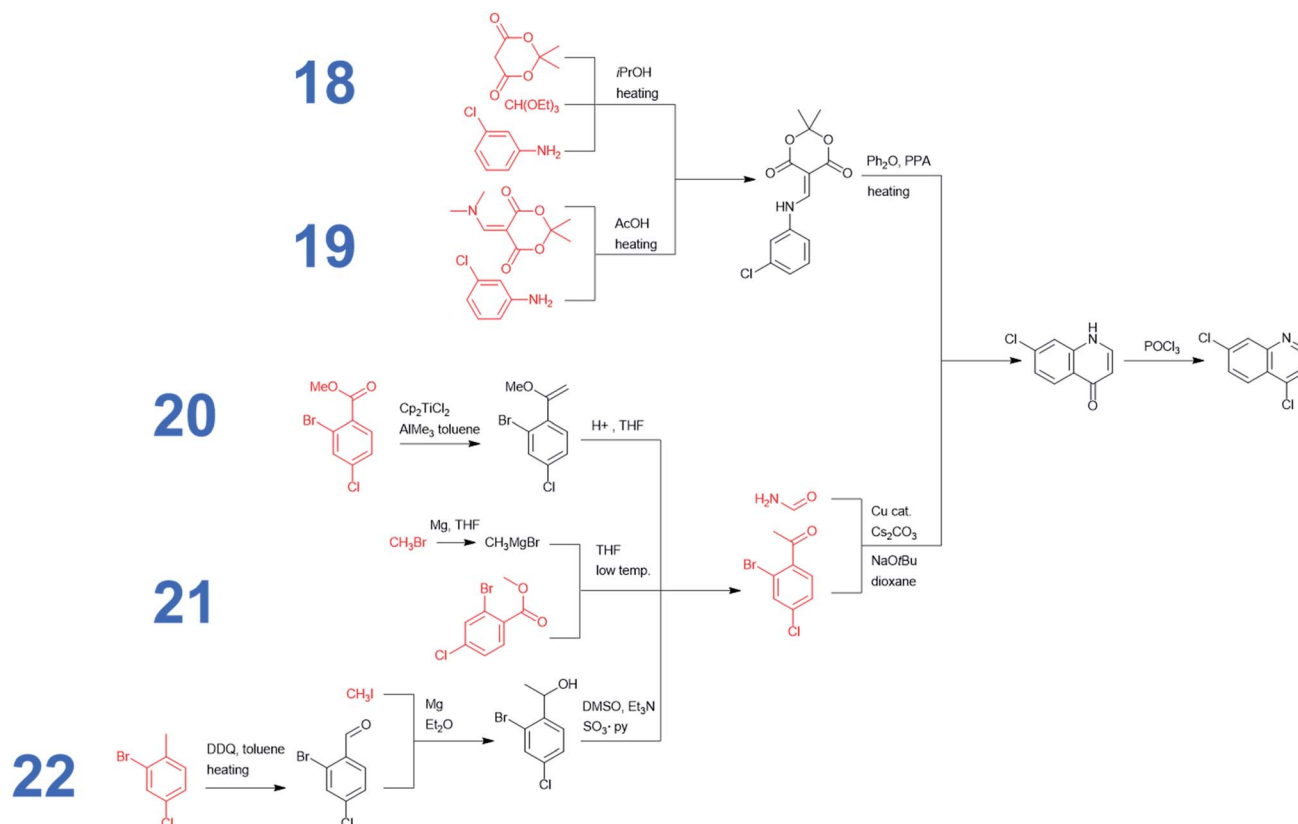
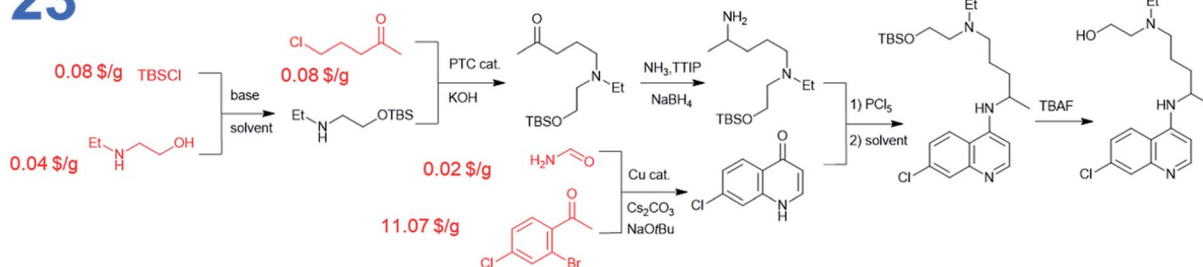


Fig. 4 Chematica-designed syntheses of the dichloroquinoline intermediate. Commercially available substrates and their prices (the lowest ones we were able to identify) scaled to \$ per g are colored in red.



23



24

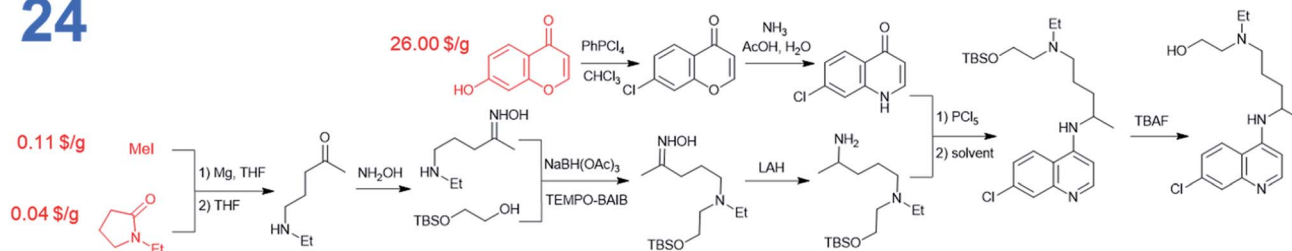


Fig. 5 Chemica-designed syntheses of HCQ constrained to avoid dichloro- or iodo, chloro-quinolone scaffolds. Commercially available substrates and their prices (the lowest ones we were able to identify) scaled to \$ per g are colored in red.

analysis is the ability of the machine to rapidly navigate the searches to inexpensive starting materials. For instance, in Chemica's synthetic plan in Fig. 6, the heterocyclic part is navigated to commercially available aminopyrrolotriazine (5.5 \$ per g from PharmaBlock), while the necessary *C*-aryl nucleoside

is constructed from ribonolactone (5.7 \$ per g from Biosynth Carbosynth). Finally, the amino acid part of remdesivir can be sourced from three different, commercially available and inexpensive (0.25–0.31 \$ per g from CoolPharm) protected (with Cbz, Boc, or Fmoc) derivatives of alanine.

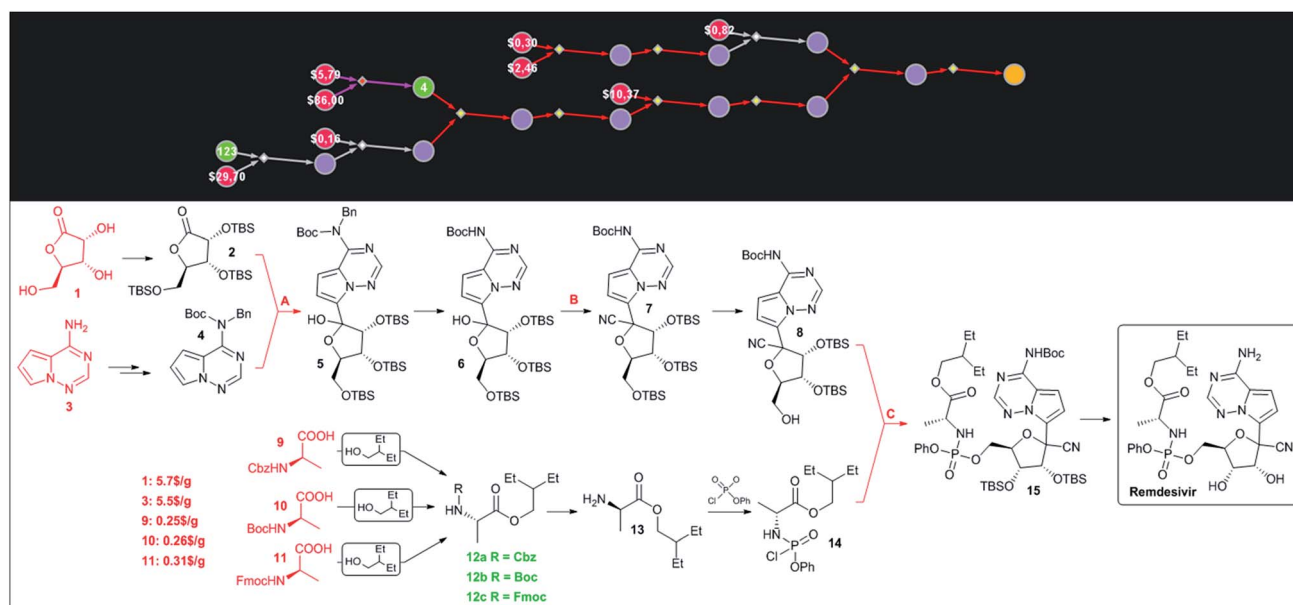


Fig. 6 Chemica-designed syntheses of remdesivir. The top panel shows the screenshot of the pathway in a graphical representation. The colors of the nodes are: yellow = target; violet = molecules not known in Chemica's literature collection; green = known molecules with numbers indicating their synthetic popularity (*i.e.*, how many times these molecules were used in prior syntheses); red = commercially available molecules with prices in \$ per g from the S-A catalog. Red- and violet-colored reaction arrows indicate reaction sequences in which the machine was strategizing over multiple steps (using, respectively, functional group interconversion and global protection/deprotection routines). Details of the pathway are shown in the bottom panel. The key steps are marked with capital letters: A – addition of an organometallic reagent to lactone, B – cyanation, and C – phosphorylation. Commercially available substrates and their prices (the lowest ones we were able to identify) scaled to \$ per g are colored in red. For Chemica's raw screenshot of these routes, see the ESI, Section S4.†



## Conclusions

In summary, we capitalized on the speed and chemical accuracy of modern computer-assisted synthetic planning to develop alternative and economical “contingency” plans for the synthesis of HCQ and remdesivir. Although these syntheses could, without doubt, also be identified by human experts alone, tracing them to inexpensive substrates while minimizing the use of previously described methodologies might be a rather tedious and time-consuming enterprise, incompatible with the COVID-19 emergency at hand. In a broader context, this exercise made us realize that the current system of chemical/pharmaceutical production is heavily reliant on efficient but few-and-far-between synthetic approaches – while this approach works at “peacetime,” it might be very vulnerable to the disruption of global supply chains of key starting materials, effectively leaving us without alternative means of production. Consequently, we advocate development of similar contingency plans for other approved drugs, in case they are needed in large quantities on short notice.

## Author contributions

K. M., E. P. G., S. S., P. D., W. B., and M. M. were the key developers of Chematica. K. M., E. P. G., and S. S. performed the synthetic analyses described in the paper. A. W. and R. R. helped with the pricing and synthetic data. B. A. G. conceived Chematica in graduate school and has directed its development ever since. All authors contributed to the writing of the manuscript.

## Conflicts of interest

The authors declare no financial interest in this work. While Chematica was originally developed and owned by B.A.G.'s Grzybowski Scientific Inventions, LLC, neither he nor the co-authors currently hold any stock in this company, which is now the property of Merck KGaA, Darmstadt, Germany. Most of the authors have, until recently, collaborated with Merck KGaA, Darmstadt, on Chematica's development within the DARPA “Make-It” award. All queries about access options to Chematica (now rebranded as Synthia™), including academic collaborations, should be directed to Dr Sarah Trice at sarah.trice@sial.com.

## Acknowledgements

This work was performed on a volunteer basis in the authors' free time (over the weekend of March 21/22, 2020). This being said, we are grateful to the U.S. DARPA who, for many years, sustained the development of Chematica under the “Make-It” award, 69461-CH-DRP #W911NF1610384. We also thank MilliporeSigma/Merck KGaA for the use of their computer resources. B. A. G. also gratefully acknowledges the personal support from the Institute for Basic Science Korea, Project Code IBS-R020-D1. A.W. gratefully acknowledges the personal

support from the National Science Center, NCN, Poland under the Symfonia Award (#2014/12/W/ST5/00592).

## References

- 1 M. Wang, R. Cao, L. Zhang, X. Yang, J. Liu, M. Xu, Z. Shi, Z. Hu, W. Zhong and G. Xiao, *Cell Res*, 2020, **30**, 269–271.
- 2 J. Liu, R. Cao, M. Xu, X. Wang, H. Zhang, H. Hu, Y. Li, Z. Hu, W. Zhong and M. Wang, *Cell Discov*, 2020, **6**, 16.
- 3 M. Mauthe, I. Orhon, C. Rocchi, X. Zhou, M. Luhr, K.-J. Hijlkema, R. P. Coppes, N. Engedal, M. Mari and F. Reggiori, *Autophagy*, 2018, **14**, 1435–1455.
- 4 M. Borkowska, M. Siek, D. V. Kolygina, Y. I. Sobolev, S. Lach, S. Kumar, Y.-K. Cho, K. Kandere-Grzybowska and B. A. Grzybowski, *Nat. Nanotechnol.*, 2020, **15**, 331–341.
- 5 H. Weniger, Review of side effects and toxicity of chloroquine, *Bull. World Health*, 1979, **79**, 906.
- 6 E. W. McChesney, *Am. J. Med.*, 1983, **75**, 11–18.
- 7 J. H. Beigel, K. M. Tomashek, L. E. Dodd, A. K. Mehta, B. S. Zingman, A. C. Kalil, E. Hohmann, H. Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R. W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T. F. Patterson, R. Paredes, D. A. Sweeney, W. R. Short, G. Touloumi, D. C. Lye, N. Ohmagari, M. Oh, G. M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M. G. Kortepeter, R. L. Atmar, C. B. Creech, J. Lundgren, A. G. Babiker, S. Pett, J. D. Neaton, T. H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak and H. C. Lane, *N. Engl. J. Med.*, 2020, DOI: 10.1056/NEJMoa2007764.
- 8 W. Yin, C. Mao, X. Luan, D.-D. Shen, Q. Shen, H. Su, X. Wang, F. Zhou, W. Zhao, M. Gao, S. Chang, Y.-C. Xie, G. Tian, H.-W. Jiang, S.-C. Tao, J. Shen, Y. Jiang, H. Jiang, Y. Xu, S. Zhang, Y. Zhang and H. E. Xu, *Science*, 2020, DOI: 10.1126/science.abc1560.
- 9 S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chemie Int. Ed.*, 2016, **55**, 5904–5937.
- 10 T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice and B. A. Grzybowski, *Chem*, 2018, **4**, 522–532.
- 11 E. P. Gajewska, S. Szymkuć, P. Dittwald, M. Startek, O. Popik, J. Młynarski and B. A. Grzybowski, *Chem*, 2020, **6**, 280–293.
- 12 W. Beker, E. P. Gajewska, T. Badowski and B. A. Grzybowski, *Angew. Chemie Int. Ed.*, 2019, **58**, 4515–4519.
- 13 T. Badowski, E. P. Gajewska, K. Molga and B. A. Grzybowski, *Angew. Chemie Int. Ed.*, 2020, **59**, 725–730.
- 14 K. Molga, E. P. Gajewska, S. Szymkuć and B. A. Grzybowski, *React. Chem. Eng.*, 2019, **4**, 1506–1521.
- 15 T. Badowski, K. Molga and B. A. Grzybowski, *Chem. Sci.*, 2019, **10**, 4640–4651.
- 16 K. Molga, P. Dittwald and B. A. Grzybowski, *Chem*, 2019, **5**, 460–473.
- 17 B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk and C. E. Wilmer, *Nat. Chem.*, 2009, **1**, 31–36.



- 18 M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski and K. J. M. Bishop, *Angew. Chemie Int. Ed.*, 2012, **51**, 7928–7932.
- 19 C. Hardouin, F. Pin, J.-F. Giffard, Y. Hervouet, J. Hublet, S. Janvier, C. Penloup, J. Picard, N. Pinault, B. Schiavi, P. Zhang, W. Zhao and X. Zhu, *Org. Process Res. Dev.*, 2019, **23**, 1932–1947.
- 20 P. Lienard, P. Gradoz, H. Greciet, S. Jegham and D. Legroux, *Org. Process Res. Dev.*, 2017, **21**, 18–22.
- 21 J. Tan and N. Yasuda, *Org. Process Res. Dev.*, 2015, **19**, 1731–1746.
- 22 R. Ciriminna and M. Pagliaro, *Org. Process Res. Dev.*, 2010, **14**, 245–251.
- 23 C. Guérin, V. Bellosta, G. Guillamot and J. Cossy, *Org. Lett.*, 2011, **13**, 3534–3537.
- 24 C. Liu, J. S. Ng, J. R. Behling, C. H. Yen, A. L. Campbell, K. S. Fuzail, E. E. Yonan and D. V. Mehrotra, *Org. Process Res. Dev.*, 1997, **1**, 45–54.
- 25 D. H. B. Ripin, S. Abele, W. Cai, T. Blumenkopf, J. M. Casavant, J. L. Doty, M. Flanagan, C. Koecher, K. W. Laue, K. McCarthy, C. Meltz, M. Munchhoff, K. Pouwer, B. Shah, J. Sun, J. Teixeira, T. Vries, D. A. Whipple and G. Wilcox, *Org. Process Res. Dev.*, 2003, **7**, 115–120.
- 26 E. J. Kiser, J. Magano, R. J. Shine and M. H. Chen, *Org. Process Res. Dev.*, 2012, **16**, 255–259.
- 27 A. F. Abdel-Magid and S. J. Mehrman, *Org. Process Res. Dev.*, 2006, **10**, 971–1031.
- 28 T. L. Rathman and W. F. Bailey, *Org. Process Res. Dev.*, 2009, **13**, 144–151.
- 29 T. J. Dietsche, D. B. Gorman, J. A. Orvik, G. A. Roth and W. R. Shiang, *Org. Process Res. Dev.*, 2000, **4**, 275–285.
- 30 D. S. Ennis, D. C. Lathbury, A. Wanders and D. Watts, *Org. Process Res. Dev.*, 1998, **2**, 287–289.
- 31 T. Gustafsson, H. Sörensen and F. Pontén, *Org. Process Res. Dev.*, 2012, **16**, 925–929.
- 32 J. Magano and J. R. Dunetz, *Org. Process Res. Dev.*, 2012, **16**, 1156–1184.
- 33 A. D. Brown, R. D. Davis, R. N. Fitzgerald, B. N. Glover, K. A. Harvey, L. A. Jones, B. Liu, D. E. Patterson and M. J. Sharp, *Org. Process Res. Dev.*, 2009, **13**, 297–302.
- 34 H. Yu, R. N. Richey, J. R. Stout, M. A. LaPack, R. Gu, V. V. Khau, S. A. Frank, J. P. Ott, R. D. Miller, M. A. Carr and T. Y. Zhang, *Org. Process Res. Dev.*, 2008, **12**, 218–225.
- 35 B. Küenburg, L. Czollner, J. Fröhlich and U. Jordis, *Org. Process Res. Dev.*, 1999, **3**, 425–431.



## Supplementary information for manuscript:

### Computer-generated “synthetic contingency” plans at times of logistics and supply problems: Scenarios for hydroxychloroquine and remdesivir.

Sara Szymkuć<sup>1+</sup>, Ewa P. Gajewska<sup>1+</sup>, Karol Molga<sup>1+</sup>, Agnieszka Wołos<sup>1</sup>, Rafał Roszak<sup>1</sup>, Wiktor Beker<sup>1</sup>, Martyna Moskal<sup>1</sup>, Piotr Dittwald<sup>1</sup> & Bartosz A. Grzybowski<sup>1,2,3\*</sup>

<sup>1</sup> Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 02-224, Poland

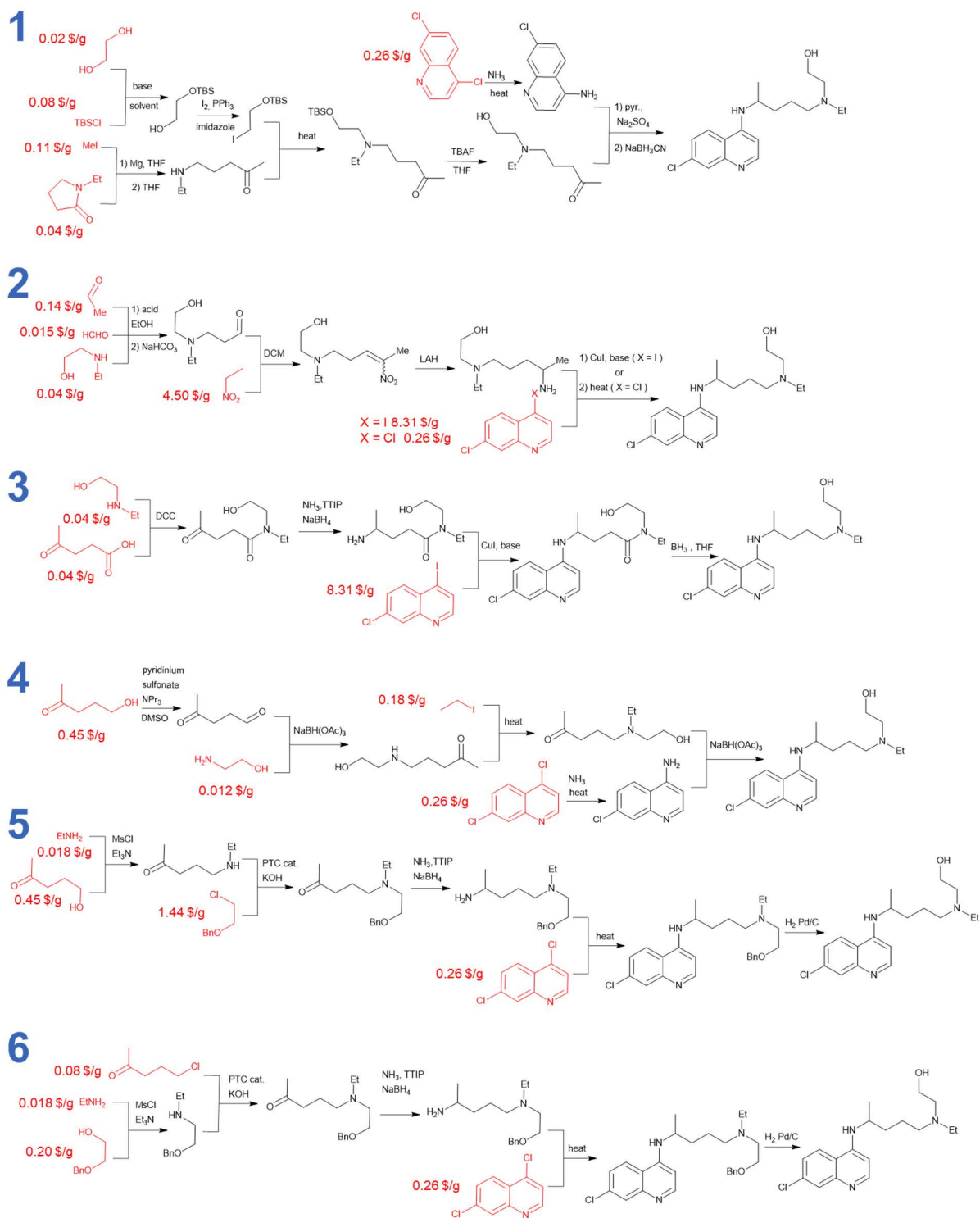
<sup>2</sup> IBS Center for Soft and Living Matter and

<sup>3</sup> Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

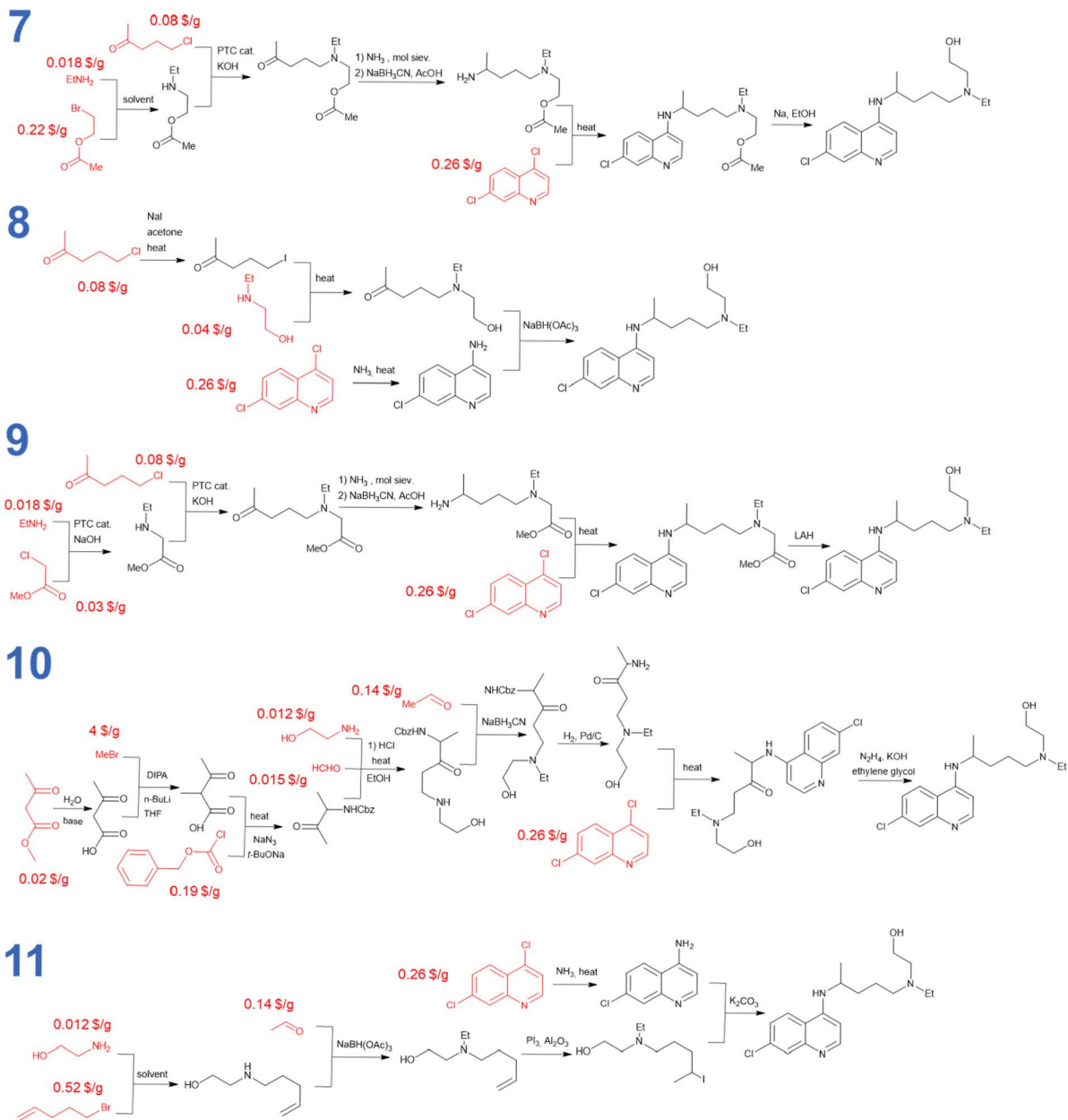
<sup>+</sup>Authors contributed equally

\*Correspondence to: [nanogrzybowski@gmail.com](mailto:nanogrzybowski@gmail.com)

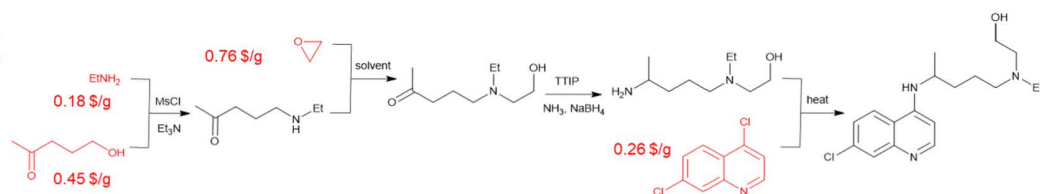
**S1. Chematica-designed pathways leading to hydroxychloroquine and presented in a form of a list.**



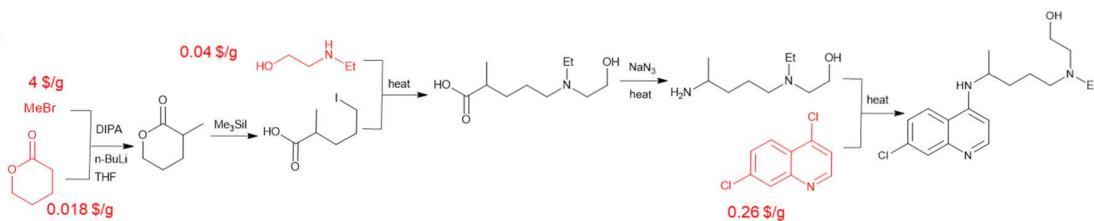




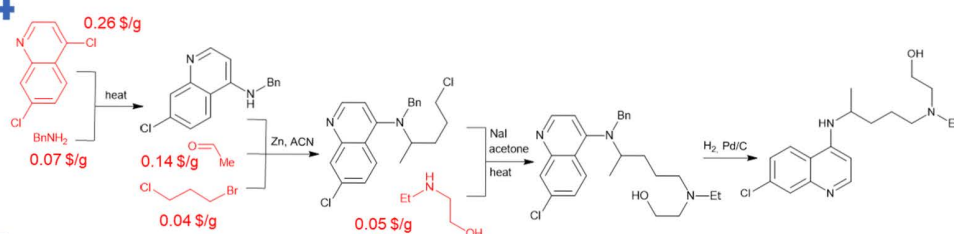
12



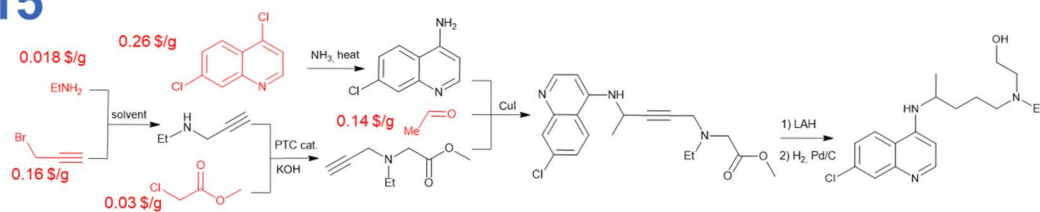
13



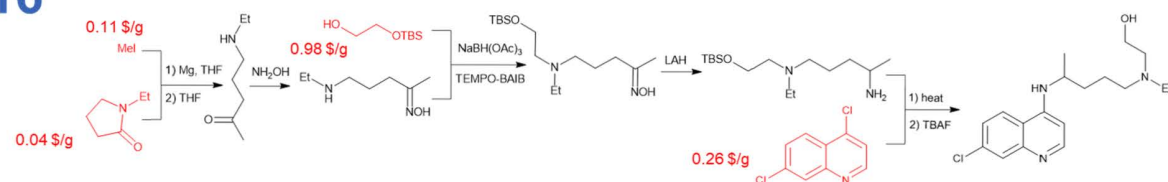
14



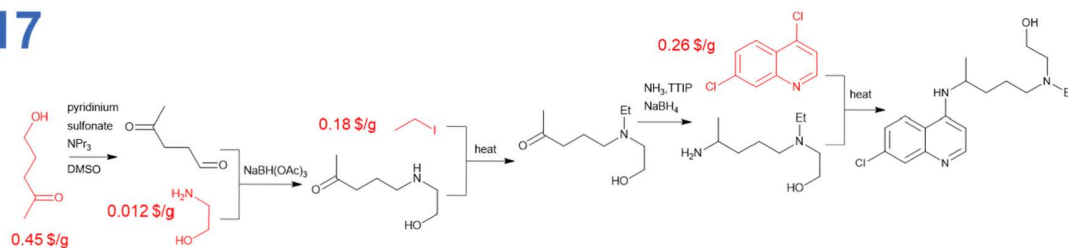
15



16



17



## S2. Details of *Chematica* searches.

The search details are structured in the following manner: (i) stop conditions (threshold prices and/or synthetic popularities), (ii) details of scoring functions, and (iii) additional parameters. Whereas most of these settings were discussed in detail in our earlier publications (see refs.S1-S4), some functionalities are new and merit brief description:

(a) Increasing the value of either “**minimum search width**” or “**max reactions per product**” increases the “breadth” of possible reactions considered at each step and, typically, yields more diverse results at the expense of longer time to complete calculations.

(b) Similarly, last five parameters for pathway 14 (cf. below) aim at increasing the diversity of results. If a certain transformation repeats in a given number of the top-scoring pathways and is identified as a “**key-step**,” it becomes disallowed after user-specified number of iterations. The key-transformation is defined as a reaction generating (in the retro direction) the highest simplification of the structure according to the specified chemical scoring function. Such reaction, in order to be identified as a “key-step,” has to be applied to an intermediate having certain percentage of the target’s mass.

(c) Four options, “**Apply\_tunnels**”, **Apply\_FGI**, **FGI\_with\_protections** and **Multirx**, are designed to strategize over multiple steps and to navigate around intermediates presenting reactive and conflicting groups. In addition, their use in combination with large penalty for protection steps (in the Reaction Scoring Function) yields pathways with complete protection/deprotection strategy (as opposed to implicit treatment of protection chemistries in the earlier versions of *Chematica*).

(c) If “**Cut all Heterocycles**” filter is “on,” all aromatic heterocycles present in a molecule have to be synthesized rather than supplied as starting materials.

(d) The “**remove\_diast**” option excludes from the search all reactions yielding mixtures of diastereoisomers. Although it does not apply here for HCQ, we usually left this option on.

All the settings used for designing synthetic routes described in the text are listed below.

**Pathways: 2 (from 8-chloro-4-iodoquinoline), 3, 5, 6, 7, 9**

**Stop points:**

Buyable molecules: 1000 g/mol, 200 \$/g

Known molecules: 1000 g/mol, 50 popularity

**Scoring Functions:**

Chemical Scoring Function: SMALLER\*\*3, SMALLER\*\*1.5

Reaction Scoring Function: 60 + 120000 \* PROTECT + 1000000 \* (CONFLICT + NON\_SELECTIVITY + FILTERS)

**Additional parameters:**

Minimum search width = 400

Max reactions per product = 60

remove\_diast = True

Macrocycles filter: ON

**Pathways: 2 (from dichloroquinoline), 4, 8, 17**

**Stop points:**

Buyable molecules: 1000 g/mol, 200 \$/g

Known molecules: 1000 g/mol, 50 popularity

**Scoring Functions:**

Chemical Scoring Function: SMALLER\*\*3, SMALLER\*\*1.5

Reaction Scoring Function:  $60 + 120000 * \text{PROTECT} + 1000000 * (\text{CONFLICT} + \text{NON\_SELECTIVITY} + \text{FILTERS}) + 50000 * \text{HIDE\_SEEK\_NAME}(['Pd'])$

**Additional parameters:**

Minimum search width = 400

Max reactions per product = 60

remove\_diast = True

Macrocycles filter: ON

**Pathways: 1,15**

**Stop points:**

Buyable molecules: 1000 g/mol, 1000 \$/gram

Known molecules: 1000 g/mol, 5 popularity

**Scoring Functions:**

Chemical Scoring Function: HOOD\*\*3, HOOD\*\*1.5

Reaction Scoring Function:  $\text{TUNNEL\_COEF} * \text{FGI\_COEF} * 60 + 10000000 * (\text{CONFLICT} + \text{NON\_SELECTIVITY} + \text{FILTERS} + \text{PROTECT})$

**Additional parameters:**

Macrocycles filter: ON

Minimum search width = 200

Max reactions per product = 30

apply\_tunnels = True

apply\_FGI = True

FGI\_with\_protections = True

multirx = True

remove\_diast = True

**Pathways: 13, 16**

**Stop points:**

Buyable molecules: 1000 g/mol, 200 \$/gram

Known molecules: 1000 g/mol, 10 popularity

**Scoring Functions:**

Chemical Scoring Function: HOOD\*\*3, HOOD\*\*1.5

Reaction Scoring Function (**for pathway 13**):  $\text{TUNNEL\_COEF} * \text{FGI\_COEF} * 60 + 10000000 * (\text{CONFLICT} + \text{NON\_SELECTIVITY} + \text{FILTERS} + \text{PROTECT}) + 10000000 * \text{HIDE\_SEEK\_NAME}(['Ozonolysis', 'Amination of aryl iodides', 'Synthesis of haloarenes via triflates']) + 10000000 * \text{HIDE\_SEEK\_SMILES}(['Fc1ccnc2cc(Cl)ccc12'])$

Reaction Scoring Function (**for pathway 16**):  $\text{TUNNEL\_COEF} * \text{FGI\_COEF} * 60 + 10000000 * (\text{CONFLICT} + \text{NON\_SELECTIVITY} + \text{FILTERS} + \text{PROTECT}) + 10000000 * \text{HIDE\_SEEK\_NAME}(['Ozonolysis', 'Amination of aryl iodides', 'Synthesis of haloarenes via triflates'])$

**Additional parameters:**

Macrocycles filter: ON  
Minimum search width = 200  
Max reactions per product = 30  
apply\_tunnels = True  
apply\_FGI = True  
FGI\_with\_protections = True  
multirx = True  
remove\_diast = True

**Pathways: 10,11,12****Stop points:**

Buyable molecules: 1000 g/mol , 50 \$/gram  
Known molecules: 1000 g/mol , 40 popularity

**Scoring Functions:**

CSF: HOOD\*\*3, HOOD\*\*1.5  
RSF: TUNNEL\_COEF \* FGI\_COEF \*60+10000000\*(CONFLICT +  
NON\_SELECTIVITY+FILTERS+PROTECT)+10000000\*HIDE\_SEEK\_NAME([  
'Ozonolysis','Amination of aryl iodides','Synthesis of haloarenes via triflates','Decarboxylative  
alkylation',' Photocatalytic','Amination of aryl tosylates','Coupling of Ammonia with Aryl  
Sulfonates','Hydroaminomethylation','Tandem Ni/Ir'])

**Additional parameters:**

Macrocycles filter: ON  
Minimum search width = 200  
Max reactions per product = 30  
apply\_tunnels = True  
apply\_FGI = True  
FGI\_with\_protections = True  
multirx = True  
remove\_diast = True

**Pathway 14****Stop points:**

Buyable molecules: 1000 g/mol, 30 \$/gram  
Known molecules: 1000 g/mol, 20 popularity

**Scoring Functions:**

CSF: HOOD\*\*3, HOOD\*\*1.5  
RSF: TUNNEL\_COEF \* FGI\_COEF \*  
120+10000000\*(CONFLICT+NON\_SELECTIVITY+FILTERS+PROTECT)+10000000\*HIDE  
\_SEEK\_NAME(['Ozonolysis','Amination of aryl iodides','Synthesis of haloarenes via  
triflates','Decarboxylative','Photocatalytic','Amination of aryl tosylates','Coupling of Ammonia  
with Aryl Sulfonates','Hydroaminomethylation','Tandem Ni/Ir'])

**Additional parameters:**

Macrocycles filter: ON  
Minimum search width = 400  
Max reactions per product = 60

apply\_tunnels = True  
apply\_FGI = True  
FGI\_with\_protections = True  
multirx = True  
remove\_diast = True  
algo\_mode = restart\_search  
key\_rx\_prod\_target\_fraction = 0.5  
restart\_search\_num\_same\_paths = 3  
restart\_search\_expanded\_spiders\_same\_paths = 800  
restart\_search\_csf = SMALLER\*\*3  
restart\_search\_csf\_fraction\_bound = 0.8

Pathways avoiding usage of dichloroquinine

### **Pathway 23**

#### **Stop points:**

Buyable molecules: 1000 g/mol, 200 \$/gram

Known molecules: 1000 g/mol , 15 popularity

#### **Scoring Functions:**

Chemical Scoring Function: HOOD\*\*3, HOOD\*\*1.5

Reaction Scoring Function: TUNNEL\_COEF \* FGI\_COEF \* 60+1000000\*(CONFLICT+NON\_SELECTIVITY+FILTERS+PROTECT+HIDE\_SEEK\_SMARTS(['[Cl,I]c1ccc2c([Cl,I])cnc2c1'])+HIDE\_SEEK\_NAME(['Synthesis of haloarenes via triflates','Coupling of Ammonia with Aryl Sulfonates','Amination of aryl tosylates','Buchwald-Hartwig type reaction','Ir/Ni','Rh-cat'])))

#### **Additional parameters:**

Macrocycles filter: ON

beam\_width = 200

max\_reactions\_per\_product = 30

apply\_tunnels = True

apply\_FGI = True

FGI\_with\_protections = True

multirx = True

remove\_diast = True

### **Pathway 24**

#### **Stop points:**

Buyable molecules: 1000 g/mol, 50 \$/gram

Known molecules: 1000 g/mol, 15 popularity

#### **Scoring Functions:**

Chemical Scoring Function: HOOD\*\*3, HOOD\*\*1.5

Reaction Scoring Function: TUNNEL\_COEF \* FGI\_COEF \* 60+1000000\*(CONFLICT+NON\_SELECTIVITY+FILTERS+PROTECT+HIDE\_SEEK\_SMARTS(['[Cl,I]c1ccc2c([Cl,I])cnc2c1'])+HIDE\_SEEK\_NAME(['Synthesis of haloarenes via triflates','Coupling of Ammonia with Aryl Sulfonates','Amination of aryl tosylates','Buchwald-Hartwig type reaction','Ir/Ni','Rh-cat','Pfizzinger Reaction'])))

#### **Additional parameters:**

Macrocycles filter: ON  
beam\_width = 200  
max\_reactions\_per\_product = 30  
apply\_tunnels = True  
apply\_FGI = True  
FGI\_with\_protections = True  
multirx = True  
remove\_diast = True  
selection\_method = yield\_rsf\_millimole\_selection  
sort\_results\_by = fully\_penalized\_cost  
selection\_penalty = 100

**Search details for the synthetic pathways leading to dichloroquinoline (the same for pathways 18-22):**

Chemical Scoring Function: HOOD\*\*1.2+50\*RINGS+50\*STEREO  
Reaction Scoring Function: TUNNEL\_COEF \* FGI\_COEF \* 60 + 10000000 \* (CONFLICT + NON\_SELECTIVITY+FILTERS+PROTECT+HIDE\_SEEK\_SMARTS(['[Cl,I]c1ccc2c([Cl,I])ccnc2c1']) + HIDE\_SEEK\_NAME(['Synthesis of haloarenes via triflates','Coupling of Ammonia with Aryl Sulfonates','Amination of aryl tosylates','Buchwald-Hartwig type reaction','Ir/Ni','Rh-cat']))  
Additional parameters:  
selection\_method = yield\_rsf\_millimole\_selection  
sort\_results\_by = fully\_penalized\_cost  
selection\_penalty = 100  
Cut all Heterocycles filter: ON  
Macrocycles filter: ON  
Buyable molecules: 1000 g/mol, 10 \$/g for pathways 1,3-5 ; 20 \$/g for pathway 2  
Known molecules: 1000 g/mol, 5 popularity

**Search details for the synthetic pathways leading to: levulinic aldehyde, 3-oxo-butanoic acid 7-chloro-quinolin-4-ylamine and N-ethyl-N-propargylamine (stop points on pathways 1,4,10,17)**

**Stop points:**

Buyable molecules: 1000 g/mol, 20 \$/gram  
Known molecules: 1000 g/mol , 60 popularity

**Scoring Functions:**

Chemical Scoring Function: HOOD\*\*3, HOOD\*\*1.5  
Reaction Scoring Function: FGI\_COEF \* 60+10000000\* (CONFLICT +NON\_SELECTIVITY+FILTERS+PROTECT)

**Additional parameters:**

beam\_width = 100  
max\_reactions\_per\_product = 30  
apply\_FGI = True  
FGI\_with\_protections = True  
multirx = True

remove\_diast = True

**Search details for the synthetic pathways leading to remdesivir:**

**Stop points:**

Buyable molecules: 1000 g/mol, 100 \$/gram

Known molecules: 1000 g/mol, 20 popularity

**Scoring Functions:**

Chemical Scoring Function: HOOD\*\*3, HOOD\*\*1.5

Reaction Scoring Function: TUNNEL\_COEF\*FGI\_COEF\*20+400000000\*PROTECT+1000000\*(CONFLICT+NON\_SELECTIVITY+FILTERS)

**Additional parameters:**

Macrocycles filter: ON

apply\_tunnels = True

beam\_width = 200

max\_reactions\_per\_product = 30

apply\_FGI = True

FGI\_with\_protections = True

multirx = True

remove\_diast = True



### S3. References cited in the main text for Figure 1

- a) T. R. Ward, B. J. Turunen, T. Haack, B. Neuenswander, W. Shadrick and G. I. Georg, *Tetrahedron Lett.*, 2009, **50**, 6494–6497.
- b) N. Al-Awadi, I. Abdelhamid, I. Abdelhamid, A. Al-Etaibi and M. Elngadi, *Synlett*, 2007, **2007**, 2205–2208.
- c) R. A. Bunce, N. R. Cain and J. G. Cooper, *Org. Prep. Proced. Int.*, 2013, **45**, 28–43.
- d) E. S. Ibrahim, M. O. A. Orabi, M. El-Badawi and M.T. Omar, *Egypt. J. Chem.*, 1991, **34**, 459–465.
- e) M. Baudouin and H. Linares, US4421918, January 15, 1982.
- f) M. Baudouin and D. Michelet, US4412076, January 15, 1982.
- g) Y. Asano, T. Kojima, O. Kurasawa, T.-T. Wong, Y. Hirata, N. Iwamura, B. Saito, Y. Tanaka, R. Arai, S. Imamura, K. Yonemori, Y. Miyamoto, S. Kitamura and O. Sano, EP3287441, April 19, 2016.
- h) H. Muñoz, J. Tamariz, H. S. Zamora, M. Lázaro and F. Labarrios, *Synth. Commun.*, 1987, **17**, 549–554.
- i) X. Simeone, M. T. Iorio, D. C. B. Siebert, S. Rehman, M. Schnürch, M. D. Mihovilovic and M. Ernst, *Bioorg. Med. Chem.*, 2019, **27**, 3167–3178.
- j) L. Forezi, N. Tolentino, A. de Souza, H. Castro, R. Montenegro, R. Dantas, M. Oliveira, F. Silva, Jr., L. Barreto, R. Burbano, B. Abraham-Vieira, R. de Oliveira, V. Ferreira, A. Cunha, F. Boechat and M. de Souza, *Molecules*, 2014, **19**, 6651–6670.
- k) S.T. Gangadharaiah, G. Sambasivam, S. Eswaran, S. Narayanan, Rk. Shandil, Rk. ; P. Kaur, V. and Potluri, WO2019243971, June 14, 2019.
- l) Z. Wang; R. Cao; X. Zhang; Z. Rong; X. Chen; X. Zhang; H. Huang; Z. Li; M. Xu; Z. Wang; J. Li and Z. Ren, CN105017145, December 12, 2012.
- m) T. G. Shruthi, S. Eswaran, P. Shivarudraiah, S. Narayanan and S. Subramanian, *Bioorg. Med. Chem. Lett.*, 2019, **29**, 97–102.
- n) C. C. Price and R. M. Roberts, *Org. Synth.*, 1955, **CVIII**, 272-274.
- o) C. Theeraladanon, M. Arisawa, A. Nishida and M. Nakagawa, *Tetrahedron*, 2004, **60**, 3017–3035.
- p) L. C. Bannen, D. S-M. Chan, J. Chen, L. E. Dalrymple, T. P. Forsyth, T. P. Huynh, V. Jammalamadaka, R. G. Khoury, J. W. Leahy, M. B. Mac, G. Mann, L. W. Mann, J. M. Nuss, J. J. Parks, C. S. Takeuchi, Y. Wang and W. Xu, WO2005030140, September 24, 2004.
- q) T. P. Forsyth; M. B. Mac; J. W. Leahy; J. M. Nuss and W. Xu, WO2006108059, April 6, 2006.

- r) T. Su, J. Zhu, R. Sun, H. Zhang, Q. Huang, X. Zhang, R. Du, L. Qiu and R. Cao, *Eur. J. Med. Chem.*, 2019, **178**, 154–167.
- s) Z. Wang, R. Cao, S. Zhang, F. Mo, F. Hu, S. Zhang, Z. Rong, X. Zhang, G. Yang; Z. Luo; S. Xia; C. Sun; R. Zhang and L. Xiong, CN105037266, December 12, 2012.
- t) P. Shi, L. Wang, K. Chen, J. Wang and J. Zhu, *Org. Lett.*, 2017, **19**, 2418–2421.
- u) V. Bizet and C. Bolm, *European J. Org. Chem.*, 2015, **2015**, 2854–2860.
- v) K. Al-Zaydi and R. Borik, *Molecules*, 2007, **12**, 2061–2079.
- w) M. V. N. de Souza, K. C. Pais, C. R. Kaiser, M. A. Peralta, M. de L. Ferreira and M. C. S. Lourenço, *Bioorg. Med. Chem.*, 2009, **17**, 1474–1480.
- x) G. Bringmann, M. Loedige, A. Kronhardt, C. Beitzinger, H. Barth and R. Benz, WO2012041493, September 28, 2011.
- y) V. Ferey, J. Mateos-Caro, R. Mondiere, P. Vayron and S. Vigne, WO2011154923, June 10, 2011.
- z) V. Ferey, J. Mateos-Caro, R. Mondiere, P. Vayron and S. Vigne, US2013096306, December 10, 2012.
- aa) A. R. Surrey and H. F. Hammer, *J. Am. Chem. Soc.*, 1950, **72**, 1814–1815.
- ab) E. F. Elslager, S. Marie-Jo, US2902403, December 3, 1956.
- ac) E. Yu, H. P. R. Mangunuru, N. S. Telang, C. J. Kong, J. Verghese, S. E. Gilliland III, S. Ahmad, R. N. Dominey and B. F. Gupton, *Beilstein J. Org. Chem.*, 2018, **14**, 583–592.
- ad) B.F. Gupton, S. Ahmad, H. P. R. Mangunuru and N. S. Telang, WO2019165337, February 25, 2019.
- ae) F. Chen; B. Hou; X. Ma; P. Gong and C. Hou, CN105693606, March 9, 2016.
- af) H. Qiang, Q. Xu and Y. Yin, CN109456266, November 12, 2018.
- ag) X. Li, G. Yan, J. Wei, P. Zhu, K. Yu, Y. Ding, W. Yu and Y. Ding, CN108689929, July 5, 2018.
- ah) Y. Zhang, K. Yu, Q. Huang, Z. Zhang, W. Yu, Y. Ding, J. Wei and P. Zhu, CN108658858, June 27, 2017.
- ai) X. Li, G. Yan, J. Wei, P. Zhu, K. Yu, Y. Ding, W. Yu and Y. Ding, CN108727263, July 5, 2018.
- aj) M. Tang, D. Gong, Z. Yang, Y. Liu, J. Yang, Z. Cai, Z. Zha, Y. Wang, CN104230803, August 28, 2014.
- ak) J. Huang, Z. Li, J. Ran, Z. Wang, Z. Wang and F. Wu, CN109280029, December 11, 2018.

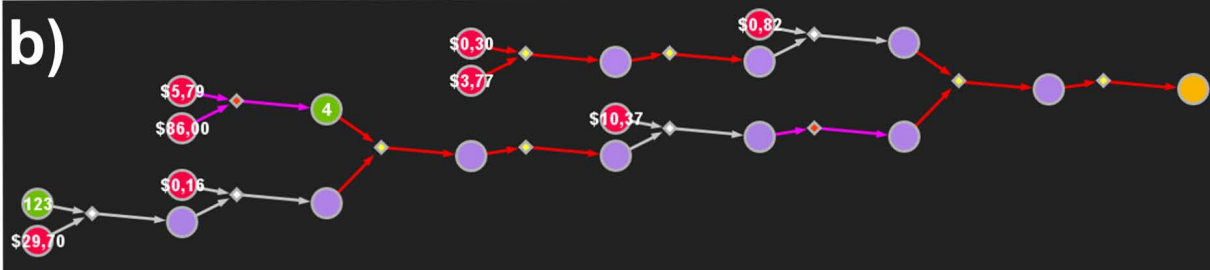
- al) J. Zhao, C. Chen, K. Yu, W. Qiu, G. Yan, W. Yu, J. Li, T. Liu, F. Liu, Z. Liu and X. Xu, CN110283121, August 6, 2019.
- am) J. Pi, Y. Ding, R. Yue, J. Wei, W. Pan and G. Xie, CN103724261, December 13, 2013.
- an) J. S. Glenn and E. A. Pham, WO2017004454, June 30, 2016.
- ao) T. K. Olszewski, C. Bomont, P. Coutrot and C. Grison, *J. Organomet. Chem.*, 2010, **695**, 2354–2358.
- ap) R. Nisal, G. P. Jose, C. Shanbhag and J. Kalia, *Org. Biomol. Chem.*, 2018, **16**, 4304–4310.
- aq) Z. Huang, CN107721835, August 13, 2016.
- ar) E. Mizushima, T. Hayashi, K. Sato and M. Tanaka, US2005143597, March 6, 2003.
- as) E. Mizushima, D.-M. Cui, D. C. Deb Nath, T. Hayashi and M. Tanaka, *Org. Synth.*, 2006, **83**, 55-60.
- at) J. Ning, X. Wu, W. Xu, H. Zhang, Z. Zhang, D. Zheng and J. Zhou, CN103694094.
- au) X. Li, CN108586214, December 1, 2018.
- av) C. Zhang, K. Yuan, L. Fan, X. Xu, D. Lu, J. Qin and P. Ju, CN109748783, January 31, 2019.
- aw) T. J. Fleck, J. J. Chen, C. V. Lu and K. J. Hanson, *Org. Process Res. Dev.*, 2006, **10**, 334–338.
- ax) X. Gao, J. Zhou and X. Peng, *Catal. Commun.*, 2019, **122**, 73–78.
- ay) A. R. Surrey, US2546658, July 23 1949.
- az) T. Deng, T. Liu, N. Peng and H. You, CN107266323, July 18, 2018.

## S4. Screenshots of *Chematica*'s syntheses of remdesivir

a)

<p>Name: Muxrx:Silylation of primary alcohols—Silylation of secondary alcohols Calculated yield: n/a</p> <p>Typical conditions: muxrx: see individual reactions</p> <p>Illustrative Reference: muxrx: see individual reactions</p> <p>Like Dislike Navigate</p>	<p>Name: Reaction of acyl chlorides with amines Calculated yield: good</p> <p>Typical conditions: NEt3 or pyridine.DCM</p> <p>Illustrative Reference: 10.1016/j.ejmech.2016.03.047 AND 10.1016/j.bmcl.2008.08.004 AND 10.1016/j.bmc.2011.03.002 AND 10.1021/jp077463a(SI) AND 10.1016/j.tetlet.2014.10.006(SI) AND 10.1016/j.bmcl.2008.04.018 AND 10.1021/jp080712a(SI)</p> <p>Like Dislike Navigate</p>	<p>Name: N-alkylation of secondary amides Calculated yield: moderate</p> <p>Typical conditions: NaH.THF</p> <p>Illustrative Reference: 10.1021/jp026391c and 10.1002/anie.201107597 (supp. info) and 10.1021/0901432a (supp. info) and 10.1016/j.tet.2011.06.026</p> <p>Like Dislike Navigate</p>
<p>Name: Furan Thiophen Pyrrole C2 RCHO Calculated yield: good</p> <p>Typical conditions: RLi or LiNR2.-78C.Et2O.then RCHO.then H+</p> <p>Illustrative Reference: 10.1055/s-1991-28395 AND 10.1016/S0040-4039(99)01876-6</p> <p>Like Dislike Navigate</p>	<p>Name: Debenzoylation Calculated yield: good</p> <p>Typical conditions: H2. Pd/C or Pd(OH)2</p> <p>Illustrative Reference: DOI: 10.1002/1521-3773(20020603)41:11:1895::AID-ANE1895&gt;3.0.CO;2-3 and 10.1021/jp400589j and 10.1021/jm8012932 (SI page S6) and 10.1021/jp070112a(SI)</p> <p>Like Dislike Navigate</p>	<p>Name: Cyanation of hemiacetals Calculated yield: good</p> <p>Typical conditions: BF3.OEt2.DCM</p> <p>Illustrative Reference: 10.1002/epoc.200300465 and 10.1055/s-2007-983882 and 10.1055/s-2007-970751 and 10.1021/ol0492647 and 10.1016/S0040-4020(96)01140-4</p> <p>Like Dislike Navigate</p>
<p>Name: Deprotection of TBS ethers Calculated yield: moderate</p> <p>Typical conditions: TBAF.THF</p> <p>Illustrative Reference: 10.1016/j.tet.2013.01.017 and 10.1016/j.tet.2004.04.042</p> <p>Like Dislike Navigate</p>	<p>Name: Reaction of phosphoryl chlorides with thiols/alcohols/phenols Calculated yield: good</p> <p>Typical conditions: DPEA.DCM</p> <p>Illustrative Reference: 10.1021/0051151f AND 10.1021/jp025510i AND 10.1016/j.orgchem.2014.08.029 AND 10.1002/adc.201000733</p> <p>Like Dislike Navigate</p>	<p>Name: Muxrx: Boc removal—Deprotection of TBS ethers Calculated yield: n/a</p> <p>Typical conditions: muxrx: see individual reactions</p> <p>Illustrative Reference: muxrx: see individual reactions</p> <p>Like Dislike Navigate</p>
<p>Name: Steglich Esterification Calculated yield: good</p> <p>Typical conditions: alcohol.DCC.DMAP.DCM or thiol.DCC.DMAP.DCM</p> <p>Illustrative Reference: 10.1002/anie.197805221</p> <p>Like Dislike Navigate</p>	<p>Name: Cleavage of benzyloxycarbamates Calculated yield: moderate</p> <p>Typical conditions: TMSI/ACN or HBr.ACOH</p> <p>Illustrative Reference: 10.1016/j.bmcl.2009.12.045 and 10.1016/j.ejmech.2013.05.017 and 10.1021/jm0701324 and 10.1016/j.bmc.2013.12.028</p> <p>Like Dislike Navigate</p>	<p>Name: Aminolysis of phosphoryl chlorides Calculated yield: good</p> <p>Typical conditions: Et3N.DCM</p> <p>Illustrative Reference: 10.1021/jp000585f AND 10.1021/jm300074y AND 10.1016/j.molstruc.2009.12.018 AND 10.1016/j.tet.2012.07.062 AND 10.1016/j.bmcl.2014.04.082 AND 10.1016/j.tet.2013.11.099</p> <p>Like Dislike Navigate</p>

b)



<p>Name: Multix:Silylation of primary alcohols—Silylation of secondary alcohols Calculated yield: n/a</p> <p>Typical conditions: multix: see individual reactions</p> <p>Illustrative Reference: multix: see individual reactions</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>	<p>Name: Reaction of acyl chlorides with amines Calculated yield: good</p> <p>Typical conditions: <math>\text{NEt}_3</math> or pyridine.DCM</p> <p>Illustrative Reference: 10.1016/j.ejmech.2016.03.047 AND 10.1016/j.bmcl.2008.08.004 AND 10.1016/j.bmc.2011.03.002 AND 10.1021/ja077483q (SI) AND 10.1016/j.tetlet.2014.10.006 (SI) AND 10.1016/j.bmcl.2008.08.004 AND 10.1016/j.bmc.2011.03.002 AND 10.1021/ja077483q (SI) AND 10.1016/j.tetlet.2014.10.006 (SI) AND 10.1016/j.bmcl.2008.08.004 AND 10.1016/j.bmc.2011.03.002 AND 10.1021/ja077483q (SI) AND 10.1016/j.tetlet.2014.10.006 (SI)</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>	<p>Name: N-alkylation of secondary amides Calculated yield: moderate</p> <p>Typical conditions: <math>\text{NaH}</math>.THF</p> <p>Illustrative Reference: 10.1021/jp026391c and 10.1002/anie.201107597 (supp. info) and 10.1021/oi901432a (supp. info) and 10.1016/j.tet.2011.08.026</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>
<p>Name: Furan Thiophen Pyrrole C2 RCHO Calculated yield: good</p> <p>Typical conditions: RLi or LNR2.-78C.Et2O.then.RCHO.then.H+</p> <p>Illustrative Reference: 10.1055/s-1991-28395 AND 10.1016/S0040-4039(99)01876-6</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>	<p>Name: Debenzylation Calculated yield: good</p> <p>Typical conditions: <math>\text{H}_2</math>.Pd/C or Pd(OH)2</p> <p>Illustrative Reference: DOI: 10.1002/1521-3773(20020603)41:111895::AID-ANIE1895&gt;3.0.CO;2-3 and 10.1021/jp400589j and 10.1021/jm8012932 (SI,page S6) and 10.1002/anie.201107597 (supp. info)</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>	<p>Name: Cyanation of hemiacetals Calculated yield: good</p> <p>Typical conditions: <math>\text{BF}_3\text{OEt}_2</math>.DCM</p> <p>Illustrative Reference: 10.1002/ajoc.200300465 and 10.1055/s-2007-983882 and 10.1055/s-2007-970751 and 10.1021/ol0492647 and 10.1016/S0040-4020(96)01140-4</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>
<p>Name: Multix:Boc removal—Deprotection of TBS ethers Calculated yield: n/a</p> <p>Typical conditions: multix: see individual reactions</p> <p>Illustrative Reference: multix: see individual reactions</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>	<p>Name: Reaction of phosphoryl chlorides with thiols/alcohols/phenols Calculated yield: good</p> <p>Typical conditions: DIPEA.DCM</p> <p>Illustrative Reference: 10.1021/od51151f AND 10.1021/jp025510i AND 10.1016/j.jrganchem.2014.08.029 AND 10.1002/adsc.201000733</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>	<p>Name: Multix:Deprotection of TBS ethers Calculated yield: n/a</p> <p>Typical conditions: multix: see individual reactions</p> <p>Illustrative Reference: multix: see individual reactions</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>
<p>Name: Steglich Esterification Calculated yield: good</p> <p>Typical conditions: alcohol.DCC.DMAP.DCM or thiol.DCC.DMAP.DCM</p> <p>Illustrative Reference: 10.1002/anie.197805221</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>	<p>Name: Boc removal Calculated yield: moderate</p> <p>Typical conditions: TFA.DCM</p> <p>Illustrative Reference: 10.1016/j.bmc.2015.01.014 and 10.1016/j.tet.2011.02.028 and 10.1016/j.tet.2015.11.027</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>	<p>Name: Aminolysis of phosphoryl chlorides Calculated yield: good</p> <p>Typical conditions: <math>\text{Et}_3\text{N}</math>.DCM</p> <p>Illustrative Reference: 10.1021/jp000585f AND 10.1021/m300074y AND 10.1016/j.molstruc.2009.12.018 AND 10.1016/j.tet.2012.07.062 AND 10.1016/j.bmcl.2014.04.082 AND 10.1016/j.tet.2013.11.099</p> <p>👍 👎 <span style="float: right;">Navigate</span></p>

c)

**Reaction 1:**  
 Name: Multix:Silylation of primary alcohols—Silylation of secondary alcohols  
 Calculated yield: n/a  
 Typical conditions: multix: see individual reactions  
 Illustrative Reference: multix: see individual reactions

**Reaction 2:**  
 Name: Reaction of acyl chlorides with amines  
 Calculated yield: good  
 Typical conditions:  $\text{NEt}_3$  or pyridine, DCM  
 Illustrative Reference: 10.1016/j.ejmech.2016.03.047 AND 10.1016/j.bmcl.2008.08.004 AND 10.1016/j.bmc.2011.03.002 AND 10.1021/ja077463q (SI) AND 10.1016/j.tetlet.2014.10.006 (SI) AND 10.1016/j.bmcl.2008.04.018 AND 10.1021/jm087172a AND 10.1021/jm087172a

**Reaction 3:**  
 Name: N-alkylation of secondary amides  
 Calculated yield: moderate  
 Typical conditions:  $\text{NaH}$ , THF  
 Illustrative Reference: 10.1021/jp026391c and 10.1002/anie.201107597 (supp. info) and 10.1021/o901432a (supp. info) and 10.1016/j.tet.2011.08.026

**Reaction 4:**  
 Name: Furan Thiophen Pyrrole C2 RCHO  
 Calculated yield: good  
 Typical conditions:  $\text{RLi}$  or  $\text{LiNR}_2$ — $\text{Et}_2\text{O}$ , then  $\text{RCHO}$ , then  $\text{H}^+$   
 Illustrative Reference: 10.1055/s-1991-28395 AND 10.1016/S0040-4039(99)01876-6

**Reaction 5:**  
 Name: Debenzylation  
 Calculated yield: good  
 Typical conditions:  $\text{H}_2$ ,  $\text{Pd/C}$  or  $\text{Pd}(\text{OH})_2$   
 Illustrative Reference: DOI: 10.1002/1521-3773(20020603)41:111895::AID-ANIE1895<3.0.CO;2-3 and 10.1021/jp400589j and 10.1021/jm8012932 (SI, page S6) and 10.1002/anie.200206034

**Reaction 6:**  
 Name: Cyanation of hemiacetals  
 Calculated yield: good  
 Typical conditions:  $\text{BF}_3 \cdot \text{OEt}_2$ , DCM  
 Illustrative Reference: 10.1002/ejoc.200300465 and 10.1055/s-2007-983882 and 10.1055/s-2007-970751 and 10.1021/ol0492647 and 10.1016/S0040-4020(96)01140-4

**Reaction 7:**  
 Name: Multix:Boc removal—Deprotection of TBS ethers  
 Calculated yield: n/a  
 Typical conditions: multix: see individual reactions  
 Illustrative Reference: multix: see individual reactions

**Reaction 8:**  
 Name: Reaction of phosphoryl chlorides with thiols/alcohols/phenols  
 Calculated yield: good  
 Typical conditions:  $\text{DIPEA}$ , DCM  
 Illustrative Reference: 10.1021/ol051151f AND 10.1021/jp025510f AND 10.1016/j.jorganchem.2014.08.029 AND 10.1002/adsc.201000733

**Reaction 9:**  
 Name: Multix:Deprotection of TBS ethers  
 Calculated yield: n/a  
 Typical conditions: multix: see individual reactions  
 Illustrative Reference: multix: see individual reactions

**Reaction 10:**  
 Name: Acid catalyzed transesterification  
 Calculated yield: moderate  
 Typical conditions:  $\text{H}^+$   
 Illustrative Reference: 10.1021/cr00020a004

**Reaction 11:**  
 Name: Deprotection of Fmoc  
 Calculated yield: moderate  
 Typical conditions: piperidine,  $\text{MeOH}$   
 Illustrative Reference: 10.1021/acs.orglett.8b03593 and 10.1055/s-0029-1219559 and 10.1016/S0957-4166(02)00212-4 and 10.1002/cnmc.201000109 and 10.1007/s00726-013-1663-1 and 10.1016/j.bmcl.2008.07.008 and 10.1021/jm087172a

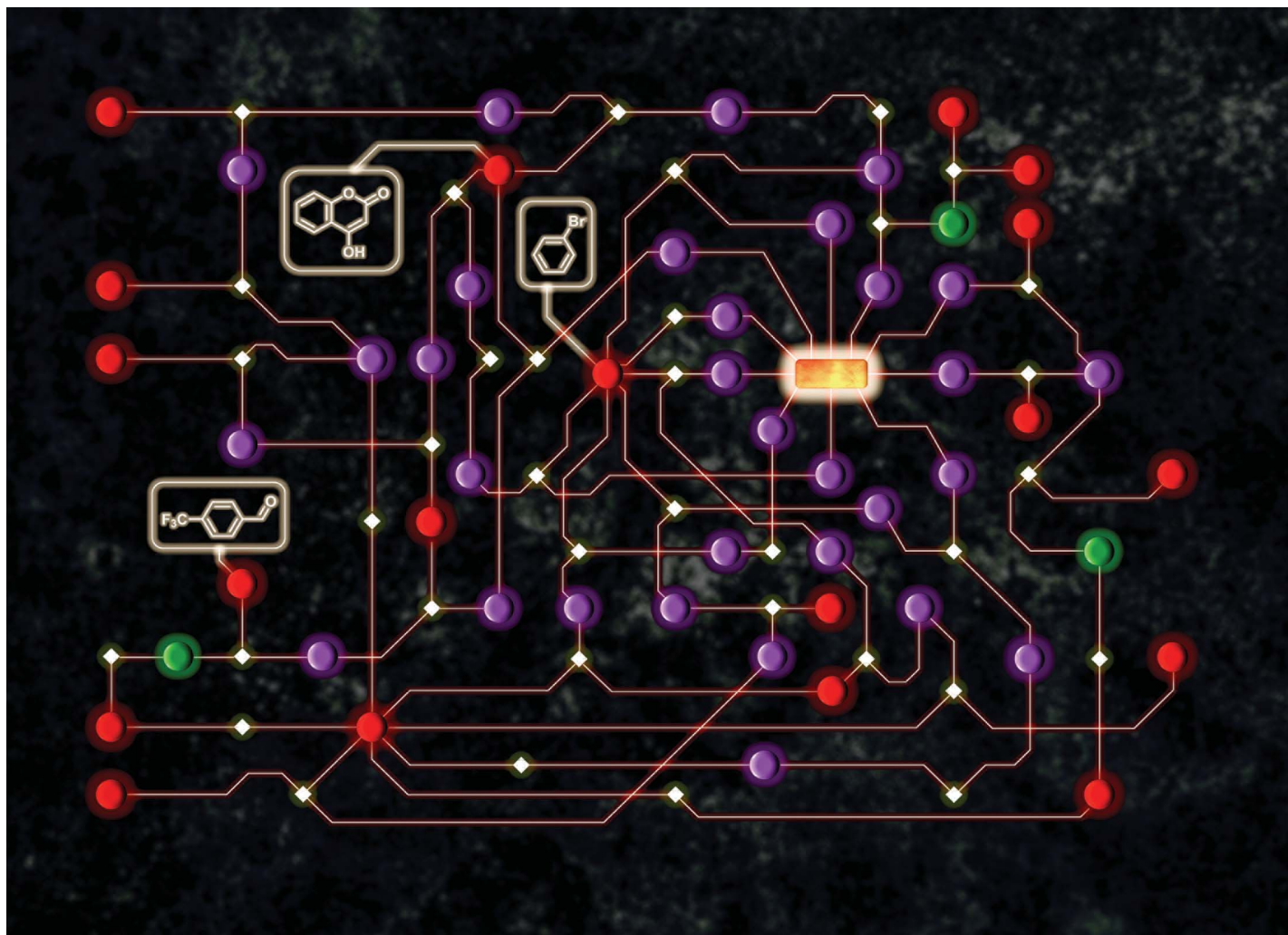
**Reaction 12:**  
 Name: Aminolysis of phosphoryl chlorides  
 Calculated yield: good  
 Typical conditions:  $\text{Et}_3\text{N}$ , DCM  
 Illustrative Reference: 10.1021/jp000585f AND 10.1021/jm300074y AND 10.1016/j.molstruc.2009.12.016 AND 10.1016/j.tet.2012.07.062 AND 10.1016/j.bmcl.2014.04.062 AND 10.1016/j.tet.2013.11.099

Chematica's syntheses of Remdesivir mimic the known approach relying on the addition of appropriate metalated heterocycle to a protected ribolactone and subsequent cyanation of the obtained hemiacetal. The side chain is constructed via sequential functionalization of phenylphosphoryl dichloride.

### Supplementary references:

- S1. S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chemie Int. Ed.*, 2016, **55**, 5904–5937.
- S2. T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice and B. A. Grzybowski, *Chem*, 2018, **4**, 522–532.
- S3. K. Molga, E. P. Gajewska, S. Szymkuć and B. A. Grzybowski, *React. Chem. Eng.*, 2019, **4**, 1506–1521.
- S4. T. Badowski, K. Molga and B. A. Grzybowski, *Chem. Sci.*, 2019, **10**, 4640–4651.





Showcasing research from Professor Bartosz Andrzej Grzybowski's laboratory, IBS Center for Soft and Living Matter and Department of Chemistry, UNIST, Ulsan, South Korea and Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland.

Computational design of syntheses leading to compound libraries or isotopically labelled targets

Whereas human chemists are trained in and accustomed to designing pathways leading to individual targets, computers can multiplex this task and design "global" synthetic plans leading to entire target libraries and/or multiple isotopomers. This study describes how network-search routines within the Chematica program can be adapted to such multi-target design while operating on one common search graph. Examples of library-wide synthetic design applied to targets of current medicinal interest illustrate how the machine skilfully constructs plans benefiting from the use of common intermediates and thus offering significant reduction of cost.

As featured in:



See Bartosz A. Grzybowski et al., *Chem. Sci.*, 2019, 10, 9219.

Cite this: *Chem. Sci.*, 2019, 10, 9219




All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 2nd June 2019  
Accepted 9th August 2019

DOI: 10.1039/c9sc02678a

rsc.li/chemical-science

# Computational design of syntheses leading to compound libraries or isotopically labelled targets†

Karol Molga,  ‡<sup>a</sup> Piotr Dittwald  ‡<sup>a</sup> and Bartosz A. Grzybowski  \*<sup>ab</sup>

Although computer programs for retrosynthetic planning have shown improved and in some cases quite satisfactory performance in designing routes leading to specific, individual targets, no algorithms capable of planning syntheses of entire target libraries – important in modern drug discovery – have yet been reported. This study describes how network-search routines underlying existing retrosynthetic programs can be adapted and extended to multi-target design operating on one common search graph, benefitting from the use of common intermediates and reducing the overall synthetic cost. Implementation in the Chematica platform illustrates the usefulness of such algorithms in the syntheses of either (i) all members of a user-defined library, or (ii) the most synthetically accessible members of this library. In the latter case, algorithms are also readily adapted to the identification of the most facile syntheses of isotopically labelled targets. These examples are industrially relevant in the context of hit-to-lead optimization and syntheses of isotopomers of various bioactive molecules.

## Introduction

Capitalizing on the advances in artificial intelligence<sup>1–4</sup> and constantly increasing computing power, recent years have brought revived interest and significant progress in the decades-old challenge of teaching computers the design of multistep organic syntheses.<sup>5–9</sup> Various platforms, differing in the underlying details of search algorithms and reaction-rule formats, have been developed<sup>10–21</sup> and one of these platforms, our own Chematica,<sup>17–21</sup> has been validated experimentally *via* successful execution of multiple routes leading to diverse, high-value, medically relevant small molecules<sup>18</sup> and, more recently, natural products.<sup>19</sup> To date, a main effort in this emerging area of chemical research has been on algorithms designing syntheses of one specified target at a time. In several practically/industrially important situations, however, it is desirable to simultaneously design routes to multiple targets. For instance, a medicinal chemist might wish to optimize an existing scaffold and place various substituents in positions of interest (*e.g.*, R<sub>1</sub>, R<sub>2</sub>, and R<sub>3</sub> in Fig. 1a). Such hit-to-lead or lead optimizations<sup>22,23</sup> encompass libraries of multiple synthetic targets, raising some pertinent questions: (1) which of the targets are most readily synthesizable? or (2) how to synthesize all of the targets while making use of some possible common

intermediates? Another situation of interest is when one wishes to design syntheses of isotopically labelled compounds that differ from the parent, non-labelled compound by a certain increment of molecular mass – this ability is important to determine drugs' pharmacokinetics,<sup>24,25</sup> to study environmental fates of pesticides,<sup>26,27</sup> to ascertain food safety,<sup>28–30</sup> or to quantify food flavourings,<sup>31–33</sup> in many cases using the so-called isotope dilution mass spectroscopy (IDMS) techniques and assays.<sup>34</sup> Because isotope labels can be placed in various positions and in various configurations within the molecule leading to isotopomers (Fig. 1b), they, again, constitute a small library of potential targets. In this case, question (1) – which of the labelled compounds offering a desired mass increase are most readily synthesizable – appears most relevant. Currently, there are no retrosynthetic algorithms with which one could address such questions for arbitrary targets. The closest analogue is our earlier work on the so-called Network of Organic Chemistry<sup>35–37</sup> (NOC), in which we used Monte-Carlo searches to select (but not design) optimal syntheses leading to multiple targets of interest.<sup>37</sup> Unfortunately, NOC is a static network comprising only published literature precedents and so its analyses are limited to already known targets and existing synthetic routes. In addition, Monte Carlo searches are computationally very intensive and typical execution times for the NOC are in days. Here, we describe significantly more efficient and general algorithms for *de novo* retrosynthetic planning (*i.e.*, planning based on general reaction rules, not existing literature precedents) producing routes to small libraries of arbitrary – that is, both known and unknown – targets, including labelled ones. In our algorithms, retrosynthetic searches for individual targets share the same search graph and can benefit from common

<sup>a</sup>Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland. E-mail: nanogrybowski@gmail.com

<sup>b</sup>IBS Center for Soft and Living Matter and Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulsan-gun, Ulsan, 689-798, South Korea

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9sc02678a

‡ These authors contributed equally.





**Fig. 1** Examples of multiple-target libraries. (a) Library of chlorcyclizine derivatives screened for the treatment of hepatitis C virus in ref. 38. (b) Isotopomers of ibuprofen. Top row has two examples of  $M + 1$  isotopomers possible with  $^{13}\text{C}$  (left) and  $^2\text{H}$  (right) labelling. Bottom row has two examples of  $M + 2$  isotopomers available via  $^{13}\text{C}$  (left) and  $^2\text{H}$  (right) labelling. Green numerals within the molecules give the total numbers of unique (note: some hydrogens and carbons are equivalent) labelled compounds for each case. For the  $M + 2$  labelling of this structurally simple target, there are already 49  $^{13}\text{C}$  and 28  $^2\text{H}$  labelling combinations. The number of isotopomers corresponds to sets without  $^2\text{H}$  labelling of the COOH group which is extremely easily exchangeable. (c) To specify multiple targets of retrosynthetic analyses, it is often convenient to use the so-called Markush structures. Here, a set of molecules is represented as a SMILES string (black) with numbered dummy atoms ([\*:1],[\*:2]) and dictionary of substituents marked in red and green for positions #1 and #2, respectively. Position [\*:1] can correspond to either an aromatic carbon or nitrogen (to yield, respectively, biphenyl and 2-phenylpyridine series), whereas position [\*:2] admits either a nitrile, CN, or F, or Cl atoms. In all, the small library thus defined will have six members shown in panel (d).

intermediates and synthetic strategies. In several cases, these searches utilize a cyclically inspected list of priority-queue-based data structures rather than a single priority-queue; this construction ensures that the overall, multi-target search is not dominated by a sub-search for any individual target. Examples of specific and in all cases viable syntheses we describe evidence that multi-target planning routines make efficient use of common intermediates, reduce the search space significantly (compared to individual, target-by-target searches), and yield complete library-wide plans within minutes. Overall, the methods we describe extend applicability of computational retrosynthetic planning to problems that are ubiquitous and important in pharmaceutical and agrochemical industries, offering savings in terms of planning time and the overall cost of synthesis of compound libraries, as well as minimization of waste (through the maximally efficient use of common

intermediates and “root” reactions in the global, library-wide reaction plans). In a broader context, the multi-target design harnesses the computer’s ability to store, analyze, and optimize large, interconnected networks of synthetic plans, which may be difficult for human chemists accustomed to planning synthetic solutions to specific targets, one target at a time.

## Computational methods

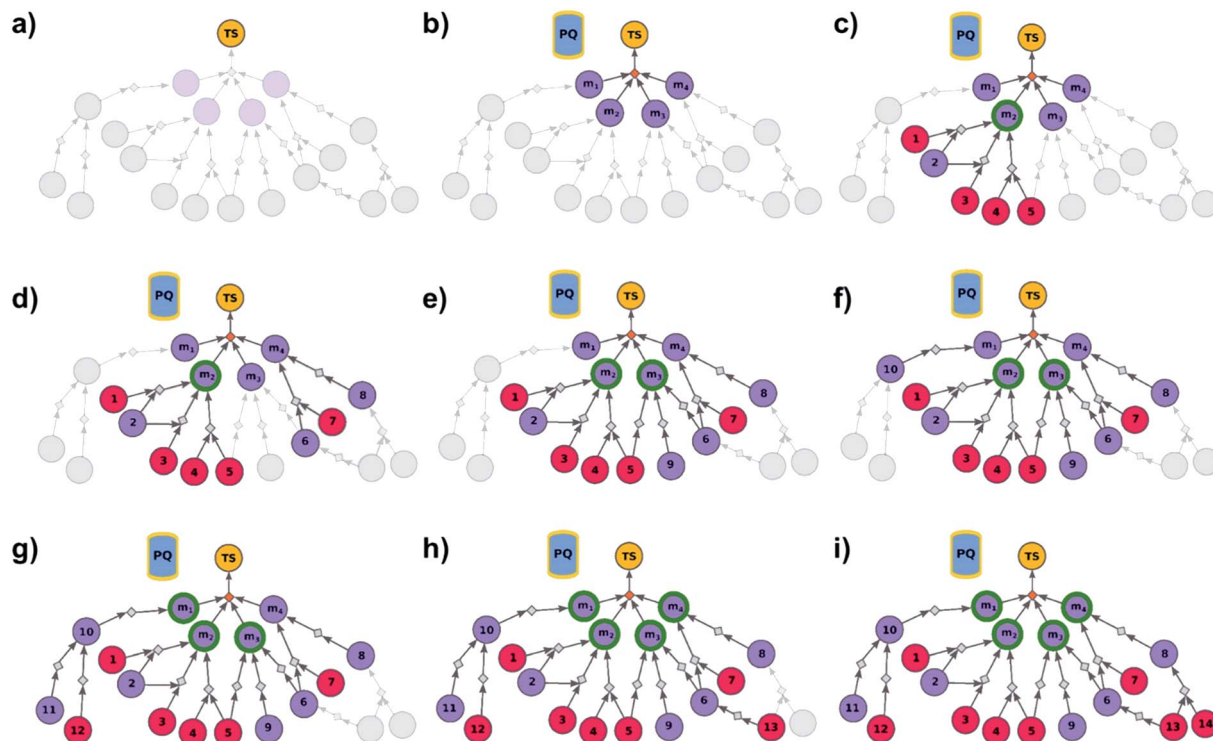
Modern retrosynthetic planners (e.g., MIT’s ASKCOS,<sup>15</sup> Waller’s MCTS,<sup>14</sup> or our Chematica<sup>17–21</sup>) differ in the origin of the underlying reaction rules (machine extracted vs. expert coded) and details of the search routines, but they all rely on the iterative expansion of parent/retron nodes into progeny/synthon nodes and on navigating (with the help of functions scoring individual reaction moves) the thus created bipartite graphs of synthetic options until simple and commercially available substrates are found. This procedure yields pathways that lead to a given target of interest and in which all chemical nodes are “synthesizable” (i.e., they are targets of at least one synthetic pathway tracing back to commercially available substrates) and the reaction nodes are “viable” (i.e., all their substrates are synthesizable). The problem we focus on in the current study is how to extend such generic routines to enable simultaneous synthesis of multiple targets (“a library”).

### Algorithm seeking the syntheses of all targets

We begin by discussing the simplest problem of identifying syntheses of all members of a library. The algorithm (Fig. 2 and pseudocode in the ESI, Section S1†) initializes by performing a dummy “multicomponent” reaction,  $r_{\text{dummy}}$ , such that the root node (i.e., the library) is “made” in one step from all molecules in the library,  $\{m_i\}$ ,  $i = 1, \dots, N$ , serving as its substrates. Chemically, this is a purely fictitious operation but algorithmically, such search-graph construction is important as it serves as an “AND” condition ensuring that any viable syntheses (cf. definition of viability above and in ref. 21) of the root node will also have to make *all* substrates for the ultimate  $r_{\text{dummy}}$  reaction – in other words, the search algorithm will not stop until viable syntheses of all  $m_i$  molecules in the library are identified. Of course, this dummy reaction is not supposed to skew the real searches for  $m_i$  in any way, and its execution cost is assigned as zero. Starting from the  $m_i$  nodes, the search graph is iteratively expanded as described before,<sup>18</sup> with the already-evaluated sets of synthons stored in a single priority-queue-based data structure (PQ), and with further synthetic navigation guided by desired scoring functions (in Chematica, the synthetic moves are guided by summing the costs of the already-performed reactions and the complexity of synthons created by each reaction;<sup>17,18</sup> in programs like ASKCOS or Waller’s MCTS, the navigation is based on the scores provided by neural networks trained on large numbers of reaction precedents<sup>14,15</sup>). The stop points for the search are known and/or commercially available molecules with prices of the latter typically taken from commercial catalogues.

As the search is allowed to progress, multiple viable syntheses are typically found, forming a solution graph from





**Fig. 2** Schematic description of the algorithm seeking the syntheses of all targets. (a) The user specifies the target set (here, TS = library of four targets), defining the root node (yellow circle) of the search graph to be explored (shaded parts represent regions of the graph not yet expanded). (b) The search graph is extended by adding four additional substance nodes (violet circles labelled  $m_1$ ,  $m_2$ ,  $m_3$ ,  $m_4$ ), linked with the root node via the same dummy reaction (orange diamond). The search algorithm utilizes a single priority-queue-based data structure, PQ, to navigate the graph expansion according to scores assigned to specific synthetic “options” encountered during the search. For this illustrative example, the algorithm first expands node  $m_2$ , having a better score than nodes  $m_1$ ,  $m_3$ , and  $m_4$ . (c). Three explored reactions (grey diamond-shaped reaction nodes) lead to possibly overlapping substrate sets – *i.e.*, nodes labelled {1, 2}, {2, 3}, and {4, 5}, respectively. Violet circular nodes denote molecules that are unknown in literature and not commercially available; red circular nodes refer to terminal, commercially available, chemicals. Here, one synthesis, *i.e.*,  $4, 5 \rightarrow m_2$ , gives viable synthesis plan for target  $m_2$  (green halo denoting that  $m_2$  is synthesizable). The search proceeds to (i) find syntheses for the remaining targets  $m_1$ ,  $m_3$ ,  $m_4$ , and (ii) find alternative synthesis plans for target  $m_2$ . (d) Search continues expanding node  $m_4$ , giving reactions with substrates {6, 7}, and {8}, then expanding node  $m_3$  (e) resulting in reactions’ substrate sets {5}, {6}, {9} (nodes 5 and 9 were already visited in previous expansions; moreover, as node 5 is terminal, path  $5 \rightarrow m_3$  is a viable synthesis for  $m_3$ ). (f) Furthermore, the region of the search graph related to the synthesis of  $m_1$  is visited by exploring node 10, and nodes 11 and 12. (g) The path  $12 \rightarrow 10 \rightarrow m_1$  is a viable synthesis, so the target  $m_1$  is now also synthesizable. (h) Then, node 6 is expanded, giving terminal node 13. Of note, 6 is a common intermediate for syntheses of targets  $m_3$ , and  $m_4$ , and these two targets have new viable syntheses ( $13 \rightarrow 6 \rightarrow m_3$ , and  $13 \rightarrow 6, 7 \rightarrow m_4$ ). Target  $m_4$  is now synthesizable, and viable synthesis plans can be retrieved for all targets from initial target set (operation performed by selection algorithm, see main text). (i) The search continues to find more synthesis options, here exploring node 8 to give reaction with substrates 13 and 14 (resulting in alternative synthesis for target  $m_4$ :  $13, 14 \rightarrow 8 \rightarrow m_4$ ).

which most economical plans can be selected (by iteratively propagating the yield-scaled costs from substrates to products and thereby assigning a realistic, monetary cost to each plan). In addition to the selection procedures detailed in our recent publication,<sup>21</sup> a unique feature of library-wide design is that we wish to promote convergent synthetic plans that make use not only of common intermediates but also of the smallest number of different synthetic methodologies – to this end, a penalty is added to any new reaction type encountered in the solution graph. Chemically, such penalization ensures that it is more economical to perform the same reaction on a, say, 2 mol scale, rather than two different types of reactions – requiring separate set-ups and likely different reagents – each on a scale of 1 mol.

We observe that if the searches for each library member were performed separately, selection would be made from each individual solution graph and no synergies in terms of common

intermediates or reaction types could, in general, be expected. Also, in terms of computational efficiency, the search on a common graph is significantly more compact than the sum of individual searches – as quantified for specific examples we discuss later in the text (*cf.* Fig. 4, 5 and Table S1†), the number of graph nodes explored before finding the first viable solution is about an order of magnitude smaller for the library-oriented, global search than for separate searches; each ran for a different library member.

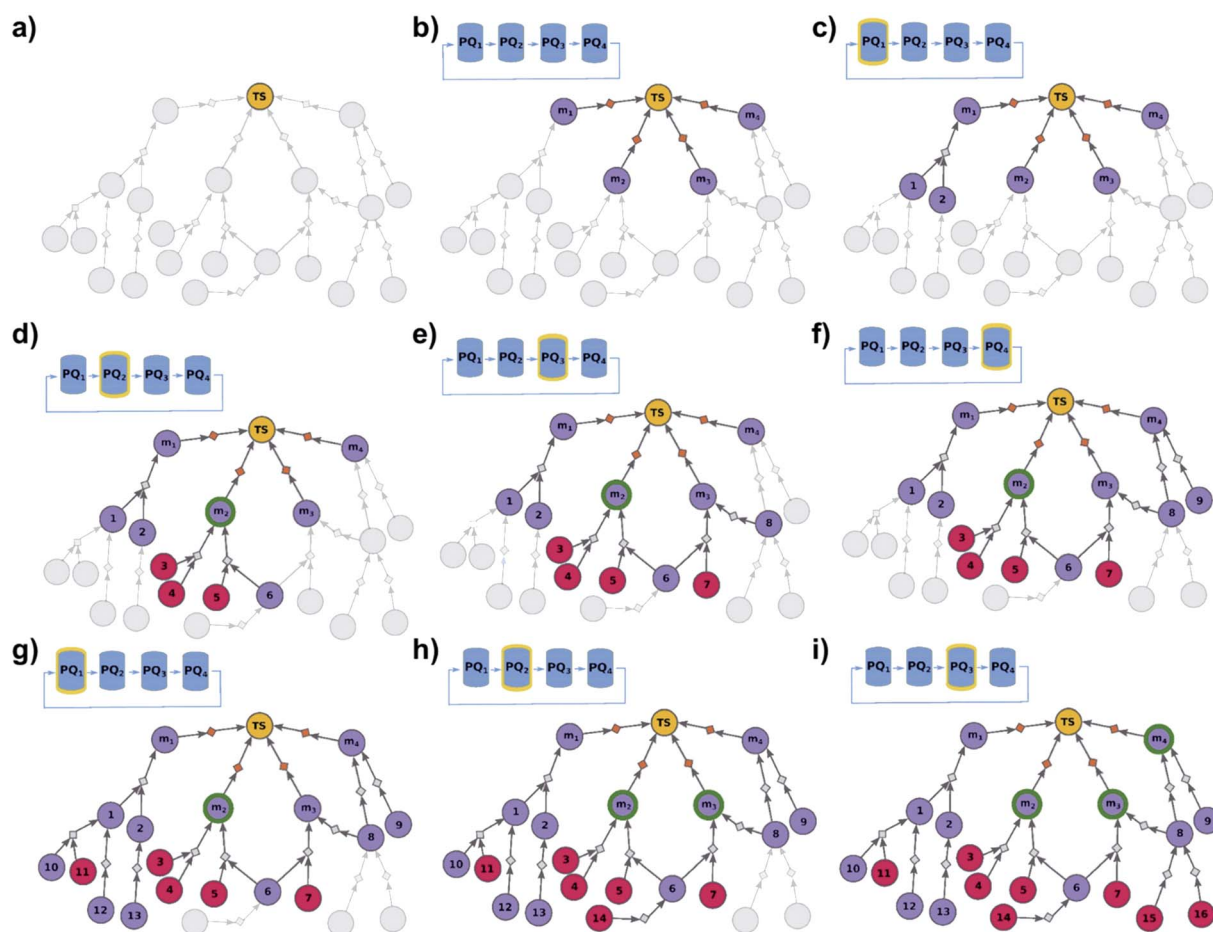
#### Algorithm seeking the easiest syntheses of some targets

In some cases, not all members of a library are equally difficult to synthesize, and one could wish to perform wet-lab execution of only those that are most readily synthesizable. Algorithmically, this task is a bit more nuanced than finding syntheses of



all targets in the library. To illustrate, let us assume that a search for the synthesis of target  $m_1$  does not find any solutions after a certain time – if this target is extremely hard to synthesize and yet the algorithm continues to find its synthesis, it might be stuck in this one search while other targets could yield solutions in shorter times. On the other hand, we do not know *a priori* if these other targets are any simpler. To overcome such problems, we implemented a multi-priority-queue algorithm that allows syntheses of different targets to be explored to

comparable levels, while sharing information about common intermediates and chemistries, and returning the best-scoring pathways taking into account the aggregated results for all targets considered. Specifically, the algorithm (Fig. 3 and pseudocode in the ESI, Section S2†) initializes not from a single “dummy” reaction (*cf.* previous section) but from  $N$  such reactions, each linking the terminal root/library node with one specific member of the library  $\{m_1, \dots, m_N\}$ . This construction serves as an “OR” condition and ensures that *any* viable



**Fig. 3** Schematic description of the algorithm seeking the syntheses of some, most-synthetically-accessible targets. (a) The root node (yellow circle) corresponds to the target set (here, TS = library of four targets). (b) The graph is extended by creating four substrates (violet circles denoted  $m_1, m_2, m_3, m_4$ ), each linked with the root node via a separate dummy reaction (orange diamonds). Additionally, four priority-queue-based data structures,  $PQ_1, PQ_2, PQ_3$ , and  $PQ_4$ , each corresponding to a separate target node, compose a priority-list, PL, inspected in circular order (first  $PQ_1$ , then  $PQ_2, PQ_3, PQ_4, PQ_1$ , etc.). PL is responsible for balanced exploration of syntheses leading to each target. (c) According to  $PQ_1$  (orange halo marks this data structure as currently inspected),  $m_1$  is expanded, identifying a reaction (grey diamond) from substrates 1 and 2 (violet circle nodes refer to unknown chemicals). (d) Then, as  $PQ_2$  is inspected, node  $m_2$  is expanded and two reactions are added to the graph (with substrate sets {3, 4} and {5, 6}, respectively; red circular nodes represent terminal, commercially available chemicals). Target  $m_2$  is now synthesizable (green halo), with a viable synthesis 3, 4  $\rightarrow$   $m_2$  already satisfying the condition of finding a synthesis of at least one member of the target library. The search continues to find alternative pathways. (e) According to  $PQ_3$ , node  $m_3$  is expanded, and reactions with substrates {6, 7} and {8} are identified (6 already appeared while searching for syntheses of  $m_2$ ). (f) Then, as  $PQ_4$  is inspected,  $m_4$  is expanded, giving reaction from node 8 (now, a common intermediate in the synthesis plans leading to  $m_3$  and  $m_4$ ) and node 9. (g) Inspection of the priority lists now returns to  $PQ_1$ , nodes 1 and 2 (previously visited while searching for the synthetic scenario of  $m_1$ ) are explored, giving new nodes 10, 11, 12, and 13. (h) As the search continues,  $PQ_2$  is inspected again, discovering viable synthesis (from terminal node 14) leading to a common intermediate of  $m_2$  and  $m_3$ , *i.e.*, node 6. Consequently,  $m_3$  becomes synthesizable, and pathways 14  $\rightarrow$  5, 6  $\rightarrow$   $m_2$  and 14  $\rightarrow$  6, 7  $\rightarrow$   $m_3$  become plausible solutions of the initial task. (i) Subsequently, the algorithm inspects  $PQ_3$ , node 8 is explored by two reactions, each starting at terminal nodes (15 and 16), and  $m_4$  becomes synthesizable (newly discovered pathways are 15  $\rightarrow$  8  $\rightarrow$   $m_3$ , 16  $\rightarrow$  8  $\rightarrow$   $m_3$ , 15  $\rightarrow$  8  $\rightarrow$   $m_4$ , and 16  $\rightarrow$  8  $\rightarrow$   $m_4$ ). All viable pathways identified during the search are retrieved and ranked according to a separate selection algorithm (see ref. 21).



pathway to the root node will also synthesize one of the  $m_i$  molecules. Moreover, to ensure that all targets  $m_1, \dots, m_N$  are being analyzed to comparable degrees, we use the priority-list, PL, rather than a single, global priority queue-based data structure, PQ. This PL (1) has length  $N$ ; (2) its  $i$ -th element,  $PL[i]$ , is a PQ initialized with a single-element set  $\{m_i\}$ ; (3) synthon sets from the PL are retrieved in circular order, *i.e.*,  $PL[1], PL[2], \dots, PL[N], PL[1], PL[2], \dots$ ; and (4) when a synthon set  $S$  is taken from  $PL[i]$ , and is further expanded into progeny synthon sets, these progenies are also inserted to  $PL[i]$ . In other words, although the search uses one common graph and still benefits from the use of common intermediates/chemistries, each target has its separate PQ which stores the synthetic options for this target and, importantly, is inspected cyclically. As solutions are being found, a selection procedure<sup>21</sup> is applied to the entire solution graph (encompassing pathways leading to different targets), to select the best *individual* syntheses (shortest, most economical, and chemically diverse routes). Unlike the “find all” modality we described earlier, for which the outcome was a global graph encompassing syntheses of all library members, the end result of the “find best” analysis is a list of individual synthetic solutions ranked in descending order of the ease of synthesis.

### Algorithm seeking the most feasible syntheses of multiple isotopomers

For this sub-problem, our aim is to find the most readily synthesizable isotopomer that increases the molecular mass by a user-specified value. The library here is a set of possible isotopomers and the search problem itself is identical to the one described in the previous section with the searches terminating in isotopically labelled (and possibly some non-labelled), commercially available starting materials. The difference lies in the way in which the target library is generated – in particular, we would like to automate the generation of all isotopomers offering the desired mass increase. The procedure to do so begins with specifying the non-labelled target molecule, a desired mass shift,  $S$  (positive or negative), and the *available\_iso* list of isotopes one wishes to use. The list should contain isotopes with only positive or only negative mass shifts, corresponding to the sign of  $S$  (*e.g.*, if  $S > 0$ ,  $^{13}\text{C}$  but not  $^{11}\text{C}$  should be used). This condition precludes generation of several nonsensical isotopomers in which, for example, a mass shift of +1 could be obtained by introducing two  $^{13}\text{C}$ s and one  $^{11}\text{C}$ . Additionally, one may wish to specify which atoms should not be labelled (*e.g.*,  $^2\text{H}$  labelled carboxylic acids, amines, or alcohols are labile, whereas  $^{13}\text{C}$  labelling should be avoided in metabolically unstable fragments such as esters or *N*-methylamines).

With such assumptions, the atoms of the target are arbitrarily ordered,  $a_1, \dots, a_M$ , and the recursive procedure (for pseudocode, see ESI, Section S3†) is applied to generate a set of desired isotopomers. To explain this procedure, let us consider methanol as the target, a hypothetical *available\_iso* = [ $^2\text{H}$ ,  $^{17}\text{O}$ , and  $^{18}\text{O}$ ], and  $S = 2$ . Let us define *labellings* ( $j, k, S$ ) as a set of labellings of atoms  $a_j, \dots, a_k$ , giving a mass shift  $S$ . Then *labellings* ( $1, M, S$ ) refer to the set of isotopomers we ultimately seek

(pending de-duplication of possible identical structures). For the specific  $\text{CH}_3\text{OH}$  target, let us order atoms  $a_1, a_2, a_3, a_4, a_5, a_6 = \text{C}, \text{H}^1, \text{H}^2, \text{H}^3, \text{O}, \text{H}^4$  (upper-right indices are used to distinguish between hydrogen atoms), and commence from  $a_1 = \text{C}$ . As *available\_iso* does not contain any isotopic label for carbon, no isotopic labelling can be applied to  $a_1$  and the set of appropriate labellings can be therefore written recursively as equation (\*) *labellings* ( $1, M, S$ ) =  $\{a_1 = ^{12}\text{C}\} \times \text{labellings}$  ( $2, M, S$ ), where  $\times$  stands for set multiplication. Moving to the second atom,  $a_2 = \text{H}^1$ , it can be labelled deuterium (mass shift of +1) or left unlabelled (no mass shift), and so we have (\*\*) *labellings* ( $2, M, S$ ) =  $\{a_2 = ^2\text{H}\} \times \text{labellings}$  ( $3, M, S - 1$ )  $\cup$   $\{a_2 = ^1\text{H}\} \times \text{labellings}$  ( $3, M, S$ ). By combining (\*) and (\*\*) we obtain *labellings* ( $1, M, S$ ) =  $\{a_1 = ^{12}\text{C}, a_2 = ^2\text{H}\} \times \text{labellings}$  ( $3, M, S - 1$ )  $\cup$   $\{a_1 = ^{12}\text{C}, a_2 = ^1\text{H}\} \times \text{labellings}$  ( $3, M, S$ ). The procedure is continued until the last atom is reached ( $a_6$  for methanol) and all acceptable labelling options are returned. For the methanol example we considered, after de-duplicating the same chemical structures (*i.e.*, removing molecules with the same canonical SMILES representation), there are five viable isotopomers, written here in the SMILES notation that is used as an input to the retrosynthetic search:  $[2\text{H}]\text{CO}[2\text{H}]$ ,  $\text{C}[18\text{O}]$ ,  $[2\text{H}][17\text{O}]\text{C}$ ,  $[2\text{H}]\text{C}[17\text{O}]$ ,  $[2\text{H}]\text{C}([2\text{H}])\text{O}$ .

## Results and discussion

### Chemical examples implemented in Chematica

The algorithms detailed in the preceding sections can be implemented in various retrosynthetic platforms. Since we have been actively involved in the development of and continue to have access to Sigma-Aldrich's Chematica, we illustrate how the algorithms function in this particular environment.

As described in several of our publications on Chematica,<sup>17–21</sup> this platform is based on the knowledge-base of over 75 000 expert-coded reaction rules reflecting reaction mechanisms, delineating carefully substituent scope as well as contextual information about potential cross-reactivity conflicts, protection requirements, selectivity issues, *etc.* (for examples, see ref. 17 and ESI of ref. 18 and 20). The rules have variants applicable to synthesis of isotopically labelled compounds and are augmented by various modules based on quantum-mechanical, molecular-mechanical, machine-learning, or heuristic measures of reactions' electronic and steric requirements.<sup>17,20,39</sup> The bipartite synthetic graphs created by the application of reaction rules are navigated with the help of scoring functions assigning costs for each reaction operation performed (with additional costs added to reactions requiring, *e.g.*, protection chemistries) and evaluating structural complexity of the sets of synthon molecules produced in each step. The searches are supplemented by routine checking of the multi-step logic of syntheses – for instance, they prevent dragging highly reactive groups along multiple steps, penalize contraction of certain macrocycles, or allow overcoming local complexity maxima by the use of the so-called tactical combinations. Once feasible routes are found, they are scored according to realistic pricing models<sup>21</sup> based on the prices of commercially available starting materials (>200 000 chemicals from Sigma-Aldrich including



~1100 isotopically labelled ones) and approximate yet realistic reaction yields.<sup>40,41</sup> The chemical correctness of all these algorithms has been corroborated by successful experimental execution of a number of Chematica-planned, multistep syntheses.<sup>18,19</sup>

The multi-target design interfaced with Chematica entails specification of the target library, either by drawing all its members in a structure editor or by defining a Markush structure written in the SMILES notation<sup>42</sup> (Fig. 1c and d) or, for isotopomers, by specifying the desired mass increase and the isotopes that can be used. The results of the searches are presented to the user as a “global” graph encompassing, depending on the search modality, syntheses of all or some, most synthetically accessible targets. Each substance and reaction node can be expanded to provide, respectively, additional structural and synthetic details (see ESI, Section S5–S11†).

### Synthesis of all members of a Prozac-derived library

Let us begin with a simple example in which we seek syntheses of all members of a small library around a selective serotonin reuptake inhibitor, fluoxetine (Prozac). The library admits four different substituents (*p*-F, *p*-Cl, *p*-CF<sub>3</sub> and H) in the aryl ether part of fluoxetine and three different moieties on the *N*-terminated side chain (NHMe, NH*Et* and NHAc), corresponding to 12 compounds in total (Fig. 4a). Within *ca.* 5 min, Chematica produced tens of global plans for the syntheses of all the library members, with the top-scoring solution graph shown in Fig. 4b and with chemical details elaborated in Fig. 4c (for raw Chematica screenshots and suggested reaction conditions, see Fig. S4†). The common root of all syntheses is the Friedel–Crafts acylation of benzene with acyl chloride derived from *N*-acetyl β-aminopropionic acid. Subsequent enantioselective reduction (controlled, *e.g.*, by the Corey–Bakshi–Shibata catalyst<sup>43</sup> or Noyori's catalyst<sup>44</sup>) yields an enantioenriched secondary alcohol which is reacted with appropriate phenols under Mitsunobu conditions to give the desired *N*-acyl series **A2–D2**. The *N*-ethyl substituted compounds **A3–D3** can then be obtained in one step *via* reduction of the acetyl moieties while the preparation of *N*-methyl series **A1–D1** requires hydrolysis of acetamide and subsequent reductive amination with formaldehyde. Importantly, the entire scheme to prepare 12 different compounds requires only 18 individual reactions and takes advantages of several common intermediates including interconversions of some library members. We note that the proposed strategy comprising enantioselective reduction of appropriate β-aminoacetophenone<sup>44–46</sup> and Mitsunobu displacement with phenol<sup>47,48</sup> has already been used in several syntheses of structurally related compounds including our synthesis of hydroxyduloxetine.<sup>18</sup> We also emphasize that this “global” synthetic plan is different from the plans that Chematica produces for each target separately – for instance, if the program's task is to make des-trifluoromethyl congener of fluoxetine **A1** (Fig. 5a–c and S5†), it uses the enantioselective allylation of aldehyde and ozonolysis mimicking Bracher's approach.<sup>49</sup> The same strategy is returned as the top solution if **A3** (Fig. 5d and S6†) or **D3** (Fig. 5e and S7†) is an individual target. We note that none of

the top-three approaches found by Chematica in single-target-oriented searches for **A1** (Fig. 5a–c and S5†) – relying on either the Friedel–Crafts acylation with carbamate protected β-aminopropionic acid<sup>50</sup> or enantioselective arylation of aldehydes<sup>51,52</sup> – are desirable for the design of the entire library. This is so because these syntheses cannot take advantage of late common intermediates and require several additional reactions (*e.g.*, for the optimal individual solution to **A1**, adaptation to the all-library synthesis would entail a total of 21 distinct reactions (allylation, four displacements with phenols, four hydroborations, and twelve aminations). The software is not using the elegant three-component Mannich reaction (the one we employed previously in ref. 18 for the construction of a structurally similar scaffold of hydroxyduloxetine) because it cannot be adapted for the synthesis of the current library – in particular, the Mannich reaction cannot be performed with acetamide instead of methylamine to yield the *N*-Ac series **A2–D2**. We also note that this design example illustrates well typical gains in terms of computational efficiency: the search on a common graph to identify the first viable solution requires exploration of *ca.* 10 times less nodes compared to the case when the syntheses of the 12 targets are searched separately (*cf.* Table S1 in the ESI, Section S4†).

### Synthesis of all members of the Almorexant-derived library

In a more chemically advanced example, we consider synthesis of a library centered around Almorexant, a drug developed by Actelion and GSK for the treatment of insomnia. The library accepts four different substituents (*p*-F, *p*-CF<sub>3</sub>, *p*-tBu and 3,4-diOMe) in the phenylethyl part of Almorexant and two different *N*-substituents, corresponding to eight compounds in total (inset in Fig. 6a). Within *ca.* 30 min, Chematica identified global synthetic plans with the top-scoring solution graph shown in Fig. 6a and with chemical details elaborated in Fig. 6b (for raw Chematica screenshots and suggested reaction conditions, see the ESI, Fig. S8†). The synthesis of each library member commences with the oxidation of appropriate terminal alkenes to aldehydes. The key formation of chiral tetrahydroisoquinolines leading to four common intermediates (marked with arrows in Fig. 6a and colored orange in Fig. 6b) is accomplished *via* enantioselective Pictet–Spengler cyclisation controlled either by a chiral auxiliary or chiral catalyst.<sup>54</sup> Subsequent alkylation with the commercially available secondary benzyl bromide **A** or condensation with a derivative of mandelic acid **B** yields the target molecules.

### Synthesis of all members of a library of RANKL/RANK inhibitors

Our last example of all-library design is important in that the computer's autonomous design can be directly compared with and validated against recent experimental work. Specifically, we challenged Chematica with the synthesis of a subset of a library of RANKL/RANK inhibitors reported very recently by Yang and co-workers.<sup>53</sup> In this task the library, represented as the Markush structure in the inset of Fig. 7a (for full representation see Fig. S9†), consisted of 20 derivatives of tryptophan with



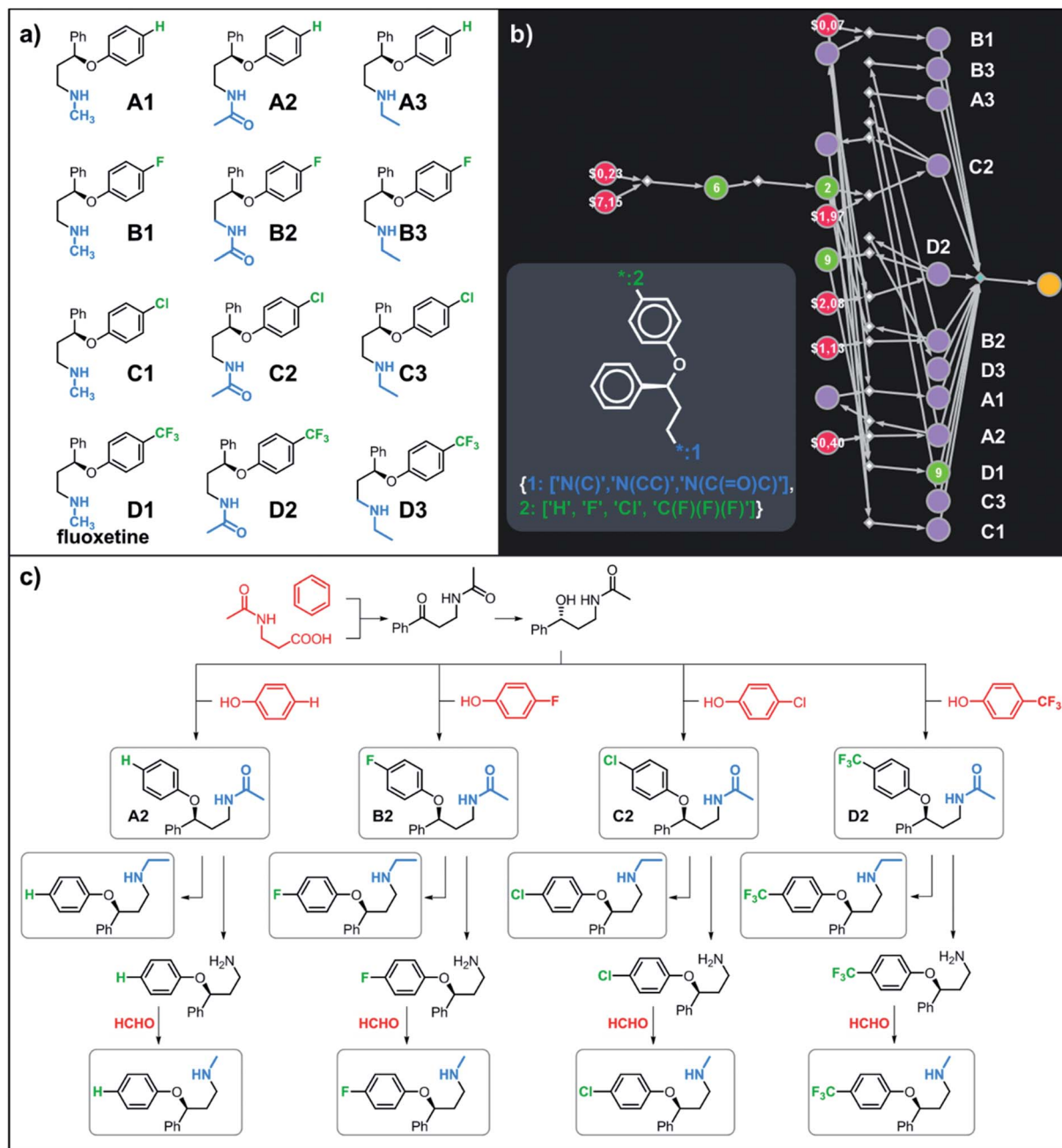


Fig. 4 Retrosynthetic search to synthesize all members of a library around fluoxetine scaffold. (a) All members of the library. (b) The corresponding Markush structure and the screenshot from Chematica showing the top-scoring solution to synthesize all of the specified targets. In this graph, yellow circular node = Markush structure; small, blue, diamond-shaped node = "dummy" reaction connecting all specific targets A1–D3 to the root node; leftmost column of 12 circular nodes = specific A1–D3 targets, mostly unknown in the NOC repository<sup>35,36</sup> (violet nodes) but one known (D1, number 9 inside of the node means that nine syntheses of this compound have been reported in the literature); red nodes = terminal, commercially available chemicals, numbers inside the nodes indicate prices per gram (from Sigma-Aldrich catalog). Note that several substances are common intermediates in multiple pathways and all 12 compounds are made in 18 steps in total. Details of synthetic plans are provided in panel (c). For raw Chematica output and further details, see Fig. S4.†

different *N*-alkyl ( $Z = \text{CH}_2$ ) or *N*-acyl fragments ( $Z = \text{C}=\text{O}$ ), different linker lengths ( $n = 1-4$ ) and substituents ( $R = \text{alkyl}$ , aryl, heteroaryl). The search was allowed to run for *ca.* 15 min and returned as the top-scoring solution the synthetic plan shown as a graph in Fig. 7a and further detailed in Fig. S10 and S11.† Within this plan, the synthesis of each library member is

accomplished in a short (three-four steps) sequence, commencing with the coupling between commercially available *N*-Boc tryptophan and 2,6-dimethylaniline. Subsequent removal of the protecting group gives a common intermediate (in Fig. 7a, the rightmost node marked with an orange arrow). Finally, coupling with appropriate carboxylic acids or primary







Fig. 5 Top-scoring solutions proposed by Chematica for the synthesis of individual members of the fluoxetine library: (a–c) target A1, (d) target A3, and (e) target D3. Node coloring scheme is as in Fig. 4. For raw Chematica output and synthetic details, see Fig. S5–S7.†

mesylates leads to *N*-acyl or *N*-alkyl target molecules, respectively. Remarkably, this approach mirrors closely the strategy used in Yang and co-workers' experiments.<sup>53</sup>

### Synthesis of the most accessible derivatives of a $\kappa$ -opioid agonist

To illustrate the synthetic design of not all but only the most accessible members of a given library, we considered derivatives of a selective  $\kappa$ -opioid agonist<sup>55</sup> ICI-199441 (Fig. 8). The ICI-199441 scaffold was decorated with three substituents in the *N*-terminated side chain, four substituents in the benzylamine part, and four combinations of halogens in the arylacetic acid part, overall corresponding to 48 distinct members of the library (Fig. 8a). The top five of the several hundreds of viable pathways



Fig. 6 Retrosynthetic search to synthesize all members of a library of analogues of Almorexant. (a) Markush structure representing proposed library (white inset, top-left), lists of substituents (bottom left) and graph representation of the top-scoring solution. Chemical details are shown in panel (b). Note that several substances (marked with arrows and colored orange in panel (b)) are common intermediates in multiple pathways while the entire library is prepared from single phenylethylamine. For raw Chematica output and further details, see Fig. S8.† Node coloring scheme is as in Fig. 4.

(identified within ~10 min) are shown in Fig. 8b–f and further detailed in Fig. S12.† In each of these plans, the molecule of interest (one node before the yellow node; compared with the scheme in Fig. 3b) can be obtained in three steps using alkylation of an appropriate secondary amine with a commercially available protected phenylglycinol, removal of the protecting group, and a sulfur catalyzed Willgerodt–Kindler (WK) reaction<sup>56</sup> yielding the desired arylacetic acid amides from acetophenones. The application of this last methodology – while





Fig. 7 Retrosynthetic search to synthesize a subset of a recently reported library of tryptophan derivatives acting as RANKL/RANK inhibitors.<sup>53</sup> (a) Markush structure representing proposed library (white inset) and graph representation of the top-scoring solution. All members of the library are prepared from one common intermediate marked with arrow in (a) and colored orange in panel (b) which details synthesis of one of the library members. For Chematica's raw output for the entire library, see Fig. S10 and S11.†

chemically correct – might, at first glance, appear counterintuitive given that such amides are usually<sup>55,57</sup> formed in reactions of acyl halides or carboxylic acids. These more conventional plans were, indeed, present in the top 100 solutions identified by the software, but the algorithm correctly gave them lower scores based on higher prices of substrates – namely, application of the WK reaction allowed for the use of cheaper acetophenone substrates rather than appropriate arylacetic acids (\$1.91 vs. \$3.23 per g of dichloroderivative and \$3.21 vs. \$5.81 per g of the 4-F-3-Cl derivative). Consequently, the diethylamino congeners were found to be more accessible than morpholino- or cyclopentylamino ones. We also note that none of the compounds substituted in the benzylamine part appeared among the most accessible targets as their synthesis requires construction of appropriate chiral aminoalcohols.

### Syntheses of isotopically labelled targets

In our last set of examples, we used the algorithm to determine which isotopically labelled compounds from a given library of isotopomers are synthetically most readily accessible.

(i) **Cinacalcet.** Fig. 9 shows five top-scoring syntheses of Amgen's cinacalcet (Sensipar/Mimpara), whose mass we wish to increase by  $S = 1$  by single labelling with either  $^{13}\text{C}$  or  $^2\text{H}$  (there



Fig. 8 Retrosynthetic analysis of ICI199441 derivatives. (a) Left portion shows the Markush structure and dictionary of substituents defining the library of 48 members; right portion shows the structure of the original ICI199441 compound. (b–f) Five top-scoring synthetic plans for library from (a). Note that in plan (d), the algorithm found a commercially available advanced intermediate (red node with price per gram = \$58.60) but continued the search until it found less expensive substrates with prices per gram below the user-specified threshold of \$50 per g. For further synthetic details, see Fig. S12.†



are 39 unique isotopomers). In the first proposed synthesis (Fig. 9a), the  $^{13}\text{C}$  isotopic label is located on the methyl group and introduced from bromomethane. In the first step, chiral-auxiliary-directed addition<sup>58,59</sup> of an organometallic reagent derived from  $^{13}\text{CH}_3\text{Br}$  yields the enantioenriched secondary amine which is then alkylated with a commercially available alcohol to give the molecule of interest. In the second plan (Fig. 9b), the  $^2\text{H}$  isotope label is introduced from  $\text{D}_2\text{O}$  during the Shapiro reaction proposed as the first step. Subsequent Rh-mediated hydroaminomethylation gives the  $^2\text{H}$ -cinacalacet also after 2 steps. In the third plan (Fig. 9c),  $^{13}\text{C}$  labelled cinacalacet is made in only one step, *via* three-component hydroaminomethylation utilizing commercially available styrene, enantioenriched secondary amine, and  $^{13}\text{C}$  carbon monoxide. We observe that such a carbonylative approach has been already used by Amgen to prepare unlabeled cinacalacet.<sup>60</sup> In the fourth plan (Fig. 9d), the labelled  $^{13}\text{C}$  atom is located at the chiral carbon and comes from the  $^{13}\text{C}$  acetic acid used in the initial Friedel–Crafts acylation of naphthalene.<sup>61</sup> Subsequent reductive amination guided by a chiral auxiliary<sup>62</sup> or a chiral catalyst<sup>63,64</sup> yields the secondary amine transformed into the target molecule following steps from the first plan. Finally, cinacalacet can be labelled with  $^2\text{H}$  located at the methyl group with a deuterium atom introduced from  $^2\text{H}$ -methyl iodide participating in direct enantioselective alkylation of naphthylacetic acid<sup>65</sup> controlled by a chiral diamine (Fig. 9e). Subsequent transformation of carboxylic acid into secondary amine *via* the Schmidt reaction<sup>66</sup> yields the amine which is subjected to the reaction with alcohol to give the target molecule. We note that the proposed approaches relying on the alkylation of chiral naphthylamine are corroborated by published syntheses<sup>67–71</sup> of unlabeled cinacalacet.

(ii) **AMG-319.** The second example in this section describes synthesis of  $M + 1$  isotopomers of Amgen's AMG-319, the

inhibitor targeted for autoimmune diseases<sup>72</sup> and head and neck squamous-cell carcinomas.<sup>73</sup> The proposed solution (Fig. 10a) commences with the Suzuki coupling of 2-pyridyl boronic acid and 2-chloropyridine. Subsequent conversion to an imine and stereoselective addition<sup>58</sup> of the organometallic reagent derived from  $^{13}\text{CH}_3\text{Br}$  yields the enantioenriched benzylic amine which is coupled with hypoxanthine in the last step. We note that this computer-designed synthetic plan resembles Amgen's original route<sup>72</sup> to unlabelled AMG-319.

(iii) **Lasmiditan.** In the third example, the algorithm is used to design syntheses of  $M + 1$  lasmiditan developed by Eli Lilly for the treatment of acute migraine. After specifying the admissible isotope label (here,  $^{13}\text{C}$ ), desired mass shift  $S = 1$  (corresponding to 15 isotopomers) and excluding  $^{13}\text{C}$  *N*-methylated isotopomer prone to hepatic cleavage observed previously for *N*-methyl piperazines,<sup>23,74–76</sup> the search was run for *ca.* 10 min and returned hundreds of viable synthetic plans from which the top-scoring one is shown in Fig. 10b. In the first step, the appropriate 2-chloropyridine is constructed in one step *via* addition of a Grignard reagent, obtained from the *N*-methyl-4-bromopiperidine, to 2-cyano-6-chloropyridine. The isotope label is introduced from  $^{13}\text{CO}_2$  used for the formation of carboxylic acid from the organolithium reagent derived from trifluorobenzene.<sup>77</sup> The following steps resemble the original route to lasmiditan.<sup>78</sup> Subsequent amination of 2-chloropyridine and reaction with labelled benzoic acid yield the molecule of interest in a four-step sequence.

(iv) **Roluperidone.** In the fourth example, Chematica designs syntheses of  $M + 1$ ,  $^{13}\text{C}$  labelled congener of roluperidone developed by Minerva Neurosciences for the treatment of schizophrenia.<sup>79</sup> The top-scoring plan returned after *ca.* 10 min is shown in Fig. 10c. The proposed short sequence commences with the *N*-alkylation of hydroxymethylpiperidine with the appropriate chloroacetophenone. Subsequent alkylation of 2-



Fig. 9 Syntheses of cinacalcet singly-labelled with either  $^{13}\text{C}$  (a, c and d) or  $^2\text{H}$  (b and e). Several viable routes were identified within 10 min; five top-scoring routes are shown. For further synthetic details, see Fig. S13.†



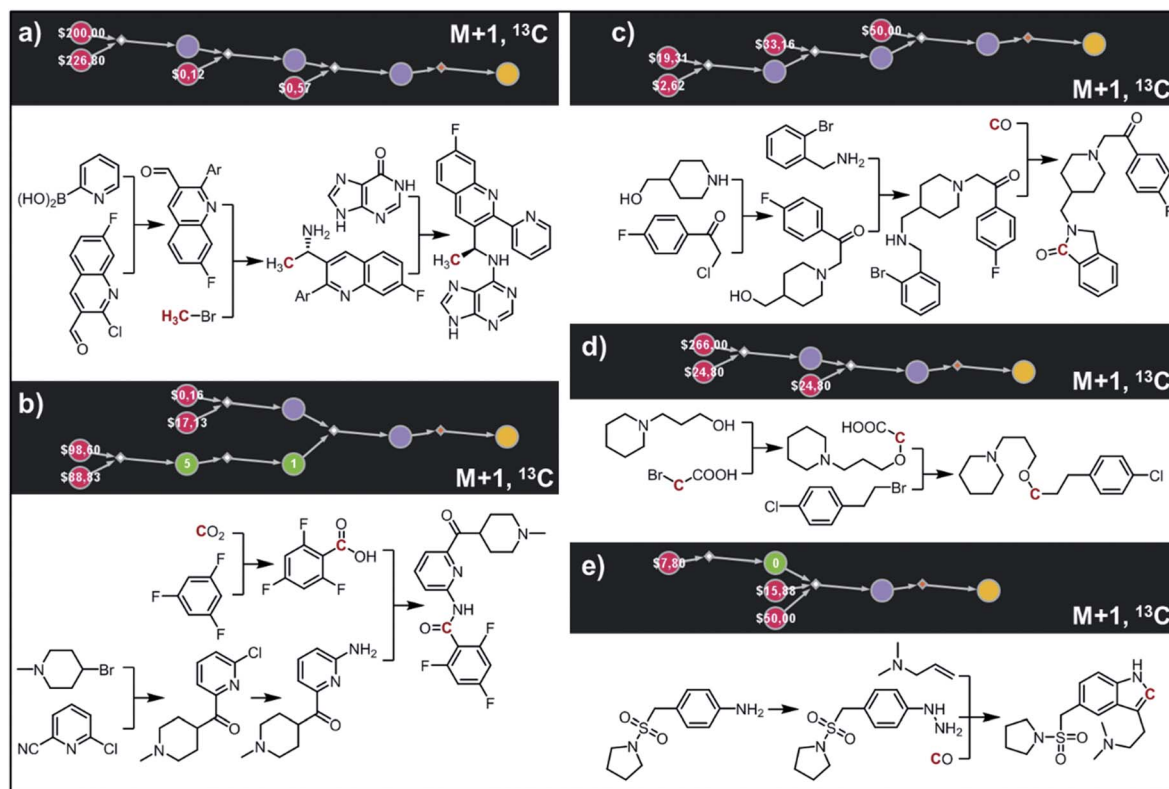


Fig. 10 Retrosynthetic design of isotopically labeled drug molecules. In all cases,  $^{13}\text{C}$  labeling was allowed and the desired mass shift was  $S = +1$ . Top-scoring and in all cases chemically viable solutions obtained for (a) a library of 21 isotopomers of AMG-319; (b) a library of 14 isotopomers of lasmiditan; (c) a library of 18 isotopomers of roluperidone; (d) a library of 13 isotopomers of pitolisant; (e) a library of 13 isotopomers of almotriptan. For further details of the pathways, see Fig. S14–S18.†

bromobenzylamine with the remaining primary alcohol yields the substrate amenable to intramolecular carbonylative amidation.<sup>80</sup> The  $^{13}\text{C}$  label is introduced in this step and sourced from  $^{13}\text{CO}$ , completing the synthesis in just three steps. The proposed plan resembles Mitsubishi's synthesis of roluperidone utilizing proposed hydroxymethylpiperidine and phenacyl bromide as building blocks and participating in  $\text{S}_{\text{N}}2$  alkylations of benzolactams and secondary amines;<sup>81</sup> moreover, the proposed carbonylative  $N$ -alkylation has been used in Astellas' synthesis of  $N$ -benzyl benzolactams.<sup>82</sup>

(v) **Pitolisant.** The fifth example illustrates efficient preparation of  $M + 1$ ,  $^{13}\text{C}$  labelled pitolisant (Wakix), developed by Bioprojet for the treatment of hypersomnia.<sup>83,84</sup> The search performed for *ca.* 10 min returned multiple solutions from which the top-scoring one is shown in Fig. 10d. The entire sequence requires only two steps and sources the  $^{13}\text{C}$  atom from labelled bromoacetic acid used in the first step to  $N$ -alkylate<sup>85</sup> hydroxypropylpiperidine. The obtained alkoxyacid is then used in MacMillan's decarboxylative coupling<sup>86</sup> with commercially available phenethyl bromide to give the target molecule. The proposed plan resembles Bioprojet's one-step solution<sup>87</sup> which also used hydroxypropylpiperidine alkylated with the appropriate alkyl bromide.

(vi) **Almotriptan.** In the sixth and last example, we aim to design routes to labelled almotriptan developed by Almirall for the treatment of severe migraine headache. After specifying the

plausible isotopes ( $^{13}\text{C}$ ) and mass shift  $S = 1$  and precluding the  $N$ -Me labelled isotopomers prone to hepatic cleavage,<sup>23,88–90</sup> the search was run for  $\sim 10$  min and returned as its top-scoring solution pathway shown in Fig. 10e. Somewhat counterintuitively, the isotope label in the most accessible isotopomer is located inside the indole ring and comes from  $^{13}\text{CO}$ . The proposed synthetic plan starts from the commercially available aniline transformed into an appropriate hydrazine. Subsequent Rh-catalyzed tandem hydroformylation/indolisation<sup>91</sup> builds the central ring system, introduces the isotope label and attaches the dimethylaminoethyl side chain and yields the target molecule in a two-step sequence. Similar, elegant carbonylative tandem indolisation was already used in Sheldon's one-pot synthesis of unlabeled melatonin.<sup>92</sup>

## Conclusions

In summary, we described how the search routines over large graphs of retrosynthetic scenarios can be adapted to find all or some members of target compound libraries. For the find-all variant, the “global” synthetic plans can benefit from the use of common intermediates and can be significantly different from optimal solutions found for each target separately – that is to say, the global solutions might be counterintuitive for human planners accustomed to optimizing specific synthetic routes rather than interconnected networks of such routes. Another



application we consider quite useful is the synthesis of isotopomers. Here, the catalogs of isotopically labelled starting materials are significantly less populous than those of unlabelled building blocks, and many blocks that are considered “basic” are not available in labelled forms. Consequently, synthetic design is often non-intuitive and it is not straightforward to predict which of the potential isotopomers would be the easiest one to make – a question our algorithms can handle rapidly and efficiently. Overall, this work extends the scope of computer-assisted synthetic planning to new problems that are common and important in pharmaceutical and agrochemical industries.

## Conflict of interest

While Chematica was originally developed and owned by B. A. G.'s Grzybowski Scientific Inventions LLC, neither he nor the co-authors no longer hold any stock in this company, which is now a property of Merck KGaA, Darmstadt, Germany. The authors continue to collaborate with Merck within the DARPA “Make-It” award. All queries about access options to Chematica (now rebranded as Synthia™), including academic collaborations, should be directed to Dr Sarah Trice at sarah.trice@sial.com.

## Acknowledgements

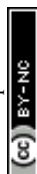
K. M., P. D. and B. A. G. thank the U.S. DARPA for generous support under the “Make-It” Award, 69461-CH-DRP #W911NF1610384. B. A. G. also gratefully acknowledges personal support from the Institute for Basic Science Korea, Project Code IBS-R020-D1. P. D. thanks Dr Tomasz Badowski for helpful insights.

## References

- 1 N. Jones, *Nature*, 2014, **505**, 146–148.
- 2 M. I. Jordan and T. M. Mitchell, *Science*, 2015, **349**, 255–260.
- 3 D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan and D. Hassabis, *Science*, 2018, **362**, 1140–1144.
- 4 I. Sutskever, O. Vinyals and Q. V. Le, *Advances in Neural Information Processing Systems*, 2014, vol. 27.
- 5 E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.
- 6 H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer and J. E. Searleman, *Science*, 1977, **197**, 1041–1049.
- 7 S. Hanessian, J. Franco and B. Larouche, *Pure Appl. Chem.*, 1990, **62**, 1887–1910.
- 8 J. B. Hendrickson, *J. Am. Chem. Soc.*, 1977, **99**, 5439–5450.
- 9 I. Ugi, J. Bauer, K. Bley, A. Dengler, A. Dietz, E. Fontain, B. Gruber, R. Herges, M. Knauer, K. Reitsam and N. Stein, *Angew. Chem., Int. Ed. Engl.*, 1993, **32**, 201–227.
- 10 O. Ravitz, *Drug Discovery Today: Technol.*, 2013, **10**, e443–e449.
- 11 M. A. Kayala and P. Baldi, *J. Chem. Inf. Model.*, 2012, **52**, 2526–2540.
- 12 J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725–732.
- 13 A. Bøgevig, H.-J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Löw, C. Oppawsky, T. Rein and H. Saller, *Org. Process Res. Dev.*, 2015, **19**, 357–368.
- 14 M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
- 15 C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281–1289.
- 16 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103–1113.
- 17 S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2016, **55**, 5904–5937.
- 18 T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Touchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice and B. A. Grzybowski, *Chem*, 2018, **4**, 522–532.
- 19 B. A. Grzybowski, *Abstr. Pap. Am. Chem. Soc.*, 2018, **256**, 5.
- 20 K. Molga, P. Dittwald and B. A. Grzybowski, *Chem*, 2019, **5**, 460–473.
- 21 T. Badowski, K. Molga and B. A. Grzybowski, *Chem. Sci.*, 2019, **10**, 4640–4651.
- 22 B. Faller and L. Urban, *Hit and Lead Profiling*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2009.
- 23 F. Z. Dörwald, *Lead Optimization for Medicinal Chemists*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2012.
- 24 A. E. Mutlib, *Chem. Res. Toxicol.*, 2008, **21**, 1672–1689.
- 25 C. J. Unkefer and R. A. Martinez, *Drug Test. Anal.*, 2012, **4**, 303–307.
- 26 C. Planas, A. Puig, J. Rivera and J. Caixach, *J. Chromatogr. A*, 2006, **1131**, 242–252.
- 27 M. B. Woudneh, M. Sekela, T. Tuominen and M. Gledhill, *J. Chromatogr. A*, 2006, **1133**, 293–299.
- 28 F. Al-Taher, K. Banaszewski, L. Jackson, J. Zweigenbaum, D. Ryu and J. Cappozzo, *J. Agric. Food Chem.*, 2013, **61**, 2378–2384.
- 29 S. Abu-El-Haj, M. J. Bogusz, Z. Ibrahim, H. Hassan and M. Al Tufail, *Food Control*, 2007, **18**, 81–90.
- 30 S. Chan, M.-F. Kong, Y.-C. Wong, S.-K. Wong and D. W. M. Sin, *J. Agric. Food Chem.*, 2007, **55**, 3339–3345.
- 31 J. Lin, D. H. Welti, F. A. Vera, L. B. Fay and I. Blank, *J. Agric. Food Chem.*, 1999, **47**, 2813–2821.
- 32 M. S. Allen, M. J. Lacey and S. Boyd, *J. Agric. Food Chem.*, 1994, **42**, 1734–1738.
- 33 V. Aubry, P. X. Etiévant, C. Giniès and R. Henry, *J. Agric. Food Chem.*, 1997, **45**, 2120–2123.
- 34 M. Berglund, in *Handbook of Stable Isotope Analytical Techniques*, ed. P. A. de Groot, Elsevier, Introduction to Isotope Dilution Mass Spectrometry (IDMS), 2004, pp. 820–834.



- 35 M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2005, **44**, 7263–7269.
- 36 B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk and C. E. Wilmer, *Nat. Chem.*, 2009, **1**, 31–36.
- 37 M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski and K. J. M. Bishop, *Angew. Chem., Int. Ed.*, 2012, **51**, 7928–7932.
- 38 S. He, J. Xiao, A. E. Dulcey, B. Lin, A. Rolt, Z. Hu, X. Hu, A. Q. Wang, X. Xu, N. Southall, M. Ferrer, W. Zheng, T. J. Liang and J. J. Marugan, *J. Med. Chem.*, 2016, **59**, 841–853.
- 39 W. Beker, E. P. Gajewska, T. Badowski and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2019, **58**, 4515–4519.
- 40 G. Skoraczynski, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski and A. Gambin, *Sci. Rep.*, 2017, **7**, 3582.
- 41 F. S. Emami, A. Vahid, E. K. Wylie, S. Szymkuć, P. Dittwald, K. Molga and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2015, **54**, 10797–10801.
- 42 <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>, accessed May 27 2019.
- 43 E. J. Corey, R. K. Bakshi and S. Shibata, *J. Am. Chem. Soc.*, 1987, **109**, 5551–5553.
- 44 T. Ohkuma, D. Ishii, H. Takeno and R. Noyori, *J. Am. Chem. Soc.*, 2000, **122**, 6510–6511.
- 45 N. A. Cortez, G. Aguirre, M. Parra-Hake and R. Somanathan, *Tetrahedron: Asymmetry*, 2013, **24**, 1297–1302.
- 46 J. Wang, D. Liu, Y. Liu and W. Zhang, *Org. Biomol. Chem.*, 2013, **11**, 3855–3861.
- 47 R. K. Rej, T. Das, S. Hazra and S. Nanda, *Tetrahedron: Asymmetry*, 2013, **24**, 913–918.
- 48 D. Liu, W. Gao, C. Wang and X. Zhang, *Angew. Chem., Int. Ed.*, 2005, **44**, 1687–1689.
- 49 F. Bracher and T. Litz, *Bioorg. Med. Chem.*, 1996, **4**, 877–880.
- 50 R. Tang, J. Zhu and Y. Luo, *Synth. Commun.*, 2006, **36**, 421–427.
- 51 J. G. Kim and P. J. Walsh, *Angew. Chem., Int. Ed.*, 2006, **45**, 4175–4178.
- 52 C. Nottingham, R. Benson, H. Müller-Bunz and P. J. Guiry, *J. Org. Chem.*, 2015, **80**, 10163–10176.
- 53 M. Jiang, L. Peng, K. Yang, T. Wang, X. Yan, T. Jiang, J. Xu, J. Qi, H. Zhou, N. Qian, Q. Zhou, B. Chen, X. Xu, L. Deng and C. Yang, *J. Med. Chem.*, 2019, **62**, 5370–5381.
- 54 J. Stöckigt, A. P. Antonchick, F. Wu and H. Waldmann, *Angew. Chem., Int. Ed.*, 2011, **50**, 8538–8564.
- 55 G. F. Costello, B. G. Main, J. J. Barlow, J. A. Carroll and J. S. Shaw, *Eur. J. Pharmacol.*, 1988, **151**, 475–478.
- 56 E. V. Brown, *Synthesis*, 1975, **1975**, 358–375.
- 57 N. Tsuritani, K. Yamada, N. Yoshikawa and M. Shibasaki, *Chem. Lett.*, 2002, **31**, 276–277.
- 58 J. A. Ellman, T. D. Owens and T. P. Tang, *Acc. Chem. Res.*, 2002, **35**, 984–995.
- 59 T. Kohara, Y. Hashimoto and K. Saigo, *Tetrahedron*, 1999, **55**, 6453–6464.
- 60 O. Thiel, C. Bernard, R. Larsen, M. J. Martinelli and M. T. Raza, WO/2009/002427, June 19, 2008.
- 61 M. Suceveanu, M. Raicopol, R. Enache, A. Finarua and S. I. Rosca, *Lett. Org. Chem.*, 2011, **8**, 690–695.
- 62 Ó. Pablo, D. Guijarro, G. Kovács, A. Lledós, G. Ujaque and M. Yus, *Chem.–Eur. J.*, 2012, **18**, 1969–1983.
- 63 C. R. Graves, K. A. Scheidt and S. T. Nguyen, *Org. Lett.*, 2006, **8**, 1229–1232.
- 64 Y. Gao, F. Yang, D. Pu, R. D. Laishram, R. Fan, G. Shen, X. Zhang, J. Chen and B. Fan, *Eur. J. Inorg. Chem.*, 2018, **2018**, 6274–6279.
- 65 C. E. Stivala and A. Zakarian, *J. Am. Chem. Soc.*, 2011, **133**, 11936–11939.
- 66 E. Brenna, M. Crotti, F. G. Gatti, A. Manfredi, D. Monti, F. Parmeggiani, S. Santangelo and D. Zampieri, *ChemCatChem*, 2014, **6**, 2425–2431.
- 67 G. B. Shinde, N. C. Niphade, S. P. Deshmukh, R. B. Toche and V. T. Mathad, *Org. Process Res. Dev.*, 2011, **15**, 455–461.
- 68 R. N. Kankan, D. R. Rao and D. R. Birari, WO/2010/100429, March 4, 2010.
- 69 R. Vlasakova and J. Hajicek, WO/2013/075679, November 21, 2012.
- 70 T. Szekeres, J. Repasi, A. Szabo, M. Benito Velez and B. Mangion, WO/2008/068625, June 8, 2007.
- 71 M. Xu, Y. Huang and M. Zhang, US2015/080608, November 28, 2014.
- 72 T. D. Cushing, X. Hao, Y. Shin, K. Andrews, M. Brown, M. Cardozo, Y. Chen, J. Duquette, B. Fisher, F. Gonzalez-Lopez de Turiso, X. He, K. R. Henne, Y.-L. Hu, R. Hungate, M. G. Johnson, R. C. Kelly, B. Lucas, J. D. McCarter, L. R. McGee, J. C. Medina, T. San Miguel, D. Mohn, V. Pattaropong, L. H. Pettus, A. Reichelt, R. M. Rzaa, J. Seganish, A. S. Tasker, R. C. Wahl, S. Wannberg, D. A. Whittington, J. Whoriskey, G. Yu, L. Zalameda, D. Zhang and D. P. Metz, *J. Med. Chem.*, 2015, **58**, 480–511.
- 73 <https://clinicaltrials.gov/ct2/show/NCT02540928>, accessed May 27 2019.
- 74 R. Hyland, E. G. H. Roe, B. C. Jones and D. A. Smith, *Br. J. Clin. Pharmacol.*, 2008, **51**, 239–248.
- 75 J. Wójcikowski, *Eur. Neuropsychopharmacol.*, 2004, **14**, 199–208.
- 76 N. Nebot, S. Crettol, F. D'Esposito, B. Tattam, D. E. Hibbs and M. Murray, *Br. J. Pharmacol.*, 2010, **161**, 1059–1069.
- 77 M. Schlosser, L. Guio and F. Leroux, *J. Am. Chem. Soc.*, 2001, **123**, 3822–3823.
- 78 D. Zhang, M.-J. Blanco, B.-P. Ying, D. Kohlman, S. X. Liang, F. Victor, Q. Chen, J. Krushinski, S. A. Filla, K. J. Hudziak, B. M. Mathes, M. P. Cohen, D. Zacherl, D. L. G. Nelson, D. B. Wainscott, S. E. Nutter, W. H. Gough, J. M. Schaus and Y.-C. Xu, *Bioorg. Med. Chem. Lett.*, 2015, **25**, 4337–4341.
- 79 <https://clinicaltrials.gov/ct2/results?term=MIN-101>, accessed May 27, 2019.
- 80 M. Mori, K. Chiba and Y. Ban, *J. Org. Chem.*, 1978, **43**, 1684–1687.
- 81 H. Yamabe, M. Okuyama, A. Nakao, M. Ooizumi and K.-I. Saito, US2003212094, February 26, 2001.
- 82 S. Yoshimura, N. Kawano, T. Kawano, D. Sasuga, T. Koike, H. Watanabe, H. Fukudome, N. Shiraiishi, R. Munakata, H. Hoshii and K. Mihara, EP2298747, July 2, 2009.



- 83 J.-C. Schwartz, *Br. J. Pharmacol.*, 2011, **163**, 713–721.
- 84 Y. Y. Syed, *Drugs*, 2016, **76**, 1313–1318.
- 85 G. Zhao, J. Wu and W.-M. Dai, *Synlett*, 2012, **23**, 2845–2849.
- 86 C. P. Johnston, R. T. Smith, S. Allmendinger and D. W. C. MacMillan, *Nature*, 2016, **536**, 322–325.
- 87 J. Sallares, I. Petschen, X. Camps, W. Schunack, H. Stark and M. Capet, WO/2007/006708, July 5, 2006.
- 88 K. Liu, F. Li, J. Lu, S. Liu, K. Dorko, W. Xie and X. Ma, *Drug Metab. Dispos.*, 2014, **42**, 863–866.
- 89 R. Dixon and A. Warrander, *Cephalalgia*, 1997, **17**, 15–20.
- 90 P. Anzenbacher and U. M. Zanger, *Metabolism of Drugs and Other Xenobiotics*, ed. P. Anzenbacher and U. M. Zanger, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2012.
- 91 A. M. Schmidt and P. Eilbracht, *J. Org. Chem.*, 2005, **70**, 5528–5535.
- 92 G. Verspui, G. Elbertse, F. A. Sheldon, M. A. P. J. Hacking and R. A. Sheldon, *Chem. Commun.*, 2000, 1363–1364.



**Supplementary Materials** for Manuscript entitled: “*Computational design of syntheses leading to compound libraries or isotopically labelled targets.*” by Karol Molga<sup>1‡</sup>, Piotr Dittwald<sup>1‡</sup>, Bartosz A. Grzybowski<sup>1,2\*</sup>

<sup>1</sup> Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland

<sup>2</sup> IBS Center for Soft and Living Matter and Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

‡ Authors contributed equally

\*Correspondence to: [nanogrzybowski@gmail.com](mailto:nanogrzybowski@gmail.com)

## CONTENTS:

**Section S1.** Pseudocode of the algorithm seeking the syntheses of all targets.

**Section S2.** Pseudocode of the algorithm seeking the easiest syntheses of some targets.

**Section S3.** Pseudocode of the algorithm generating isotopomers with a desired mass shift.

**Section S4.** Comparison of Chematica’s searches to make all members of a library vs. consecutive searches for individual targets.

**Section S5.** Details of Chematica’s retrosynthetic analyses performed for fluoxetine derivatives.

**Section S6.** Details of Chematica’s retrosynthetic analyses performed for fluoxetine derivatives in individual, target-by-target searches.

**Section S7.** Details of Chematica’s retrosynthetic analyses performed for Almorexant derivatives.

**Section S8.** Details of Chematica’s retrosynthetic analyses performed for tryptophan derivatives.

**Section S9.** Details of Chematica’s retrosynthetic analyses performed for the ICI199441 derivatives.

**Section S10.** Details of Chematica’s retrosynthetic analyses performed for <sup>13</sup>C/<sup>2</sup>H labeled Cinacalcet.

**Section S11.** Details of Chematica’s retrosynthetic analyses performed for various M+1 <sup>13</sup>C labelled drug molecules.



## Section S1. Pseudocode of the algorithm seeking the syntheses of all targets.

---

```
1: function GENERATEDUMMYREACTIONAND(TS)
  ▷ TS: targets set (library) composed of  $\{target_1, target_2, \dots, target_N\}$ 
  ▷ function generates and returns dummy reaction where TS is a product,
  and  $target_1, \dots, target_N$  are substrates
  ▷ therefore, to find retrosynthetic path for TS, search algorithm needs to
  find viable retrosynthetic scenarios for  $target_1$  AND ... AND  $target_N$ 
2:   r ← newReaction()
3:   r.addProduct(TS)
4:   for target in TS do
5:     r.addSubstrate(target)
   return {r}

6: procedure SEARCHFORLIBRARYAND(TS)
  ▷ TS: targets set (library) composed of  $\{target_1, target_2, \dots, target_N\}$ 
7:   put(PQ, node({TS}))
8:   dummyReactionNotYetGenerated ← True
9:   initializeSearchGraph(TS)
10:  while True do
11:    substratesNode ← pop(PQ)
12:    for substrate in getSubstrates(substratesNode) do
13:      if dummyReactionNotYetGenerated then
14:        progenySet ← generateDummyReactionAND(substrate)
15:        dummyReactionNotYetGenerated ← False
16:      else
17:        progenySet ← generateRetrosynthesisSteps(substrate)
18:        for progeny in progenySet do
19:          addToSearchGraph(substrate, progeny)
20:          newSubstratesNode ← node((getSubstrates(substratesNode) –
  {substrate}) ∪ progeny)
21:          if notEndOfSynthesis(newSubstratesNode) then
22:            put(PQ, newSubstratesNode)
```

---

**Figure S1.** The algorithm seeking the syntheses of all targets is the extension of existing routines for retrosynthesis search. Here, we present the pseudo-code for such a search procedure

(*searchForLibraryAND*, lines 6-22), that is run for target set, *TS*, being a user-defined library of targets  $\{target_1, \dots, target_N\}$ . The algorithm puts a node for the target set  $\{TS\}$  into priority-queue-based data structure, *PQ*, analogous to the one used for single-target search. Further, the algorithm initializes *dummyReactionNotYetGenerated* as *True*, and search graph with a single chemical node representing  $\{TS\}$  (lines 7-9). Then, the while-loop begins (the loop might be terminated, e.g., when the user decides to stop the search, if satisfactory pathways are found, or after a defined number of iterations are performed – here, loop termination is not explicitly considered for code-brevity reasons. In the first iteration of the loop, (*dummyReactionNotYetGenerated* is *True*), the algorithm calls *generateDummyReactionAND* (line 14), returning *progenySet* composed of the “multicomponent” dummy reaction  $target_1, \dots, target_N \rightarrow TS$  (lines 1-5), which is further added to the search graph (line 19). Subsequently, *newSubstratesNode = node(\{target\_1, \dots, target\_N\})* is put to *PQ* (line 22), which prioritizes the constituent nodes according to the user-provided scoring functions (see main text). As variable *dummyReactionNotYetGenerated* has been set to *False* (line 19), and will not be changed anymore, in all the subsequent iterations of the while-loop, the *progenySet* is computed as a collection of viable retrosynthetic steps (*generateRetrosynthesisSteps*, line 21). In Chematica, the related computations are based on expert-coded reaction rules, and include detection of possible cross-reactivity conflicts, protections, non-selectivity issues, etc. (see main text for references). As new retrosynthetic steps are generated, the search graph is expanded (line 19), and new synthetic options are added to *PQ* (line 22). The separate selection algorithm (see main text) is applied to retrieve a diverse set of viable retrosynthetic solutions for the requested library of targets. Please note that in the presented pseudo-code, implementation-specific optimizations such as code parallelization or search-graph representation are not considered.

## Section S2. Pseudocode of the algorithm seeking the easiest syntheses of some targets.

---

```
1: function GENERATEDUMMYREACTIONOR(TS)
  ▷ TS: targets set (library) composed of {target1, target2, ..., targetN}
2:   progenySet ← ∅
3:   for target in TS do
4:     r ← newReaction()
5:     r.addProduct(TS)
6:     r.addSubstrate(target)
7:     progenySet.update(r)
   return progenySet

8: function EXTENDIFNEEDED(PL, substratesToBeAnalyzed)
  ▷ PL: priority list; substratesToBeAnalyzed: list with substrate nodes
9:   while extensionNeeded do
10:    if allQueuesEmpty(PL) then raise Exception('PL empty')
11:    if notEmpty(PL[PL.recentlyVisited]) then
12:      substratesNode = pop(PL[PL.recentlyVisited])
13:      substratesNode.indexTakenFrom = PL.recentlyVisited
14:      substratesToBeAnalyzed.extend(substratesNode)
15:      PL.recentlyVisited ← (PL.recentlyVisited + 1) % PL.size

16: procedure SEARCHFORLIBRARYOR(TS)
  ▷ TS: targets set (library) composed of {target1, target2, ..., targetN}
  ▷ PQ: already initialized priority-queue based data structure used in single-
  target search
17:   substratesToBeAnalyzed ← [node(TS)]
18:   dummyReactionNotYetGenerated ← True
19:   initializeSearchGraph(TS)
20:   PL ← initializePL(PQ, N)
21:   while True do
22:     substratesNode ← getNew(substratesToBeAnalyzed)
23:     for substrate in getSubstrates(substratesNode) do
24:       if dummyReactionNotYetGenerated then
25:         progenySet ← generateDummyReactionOR(substrate)
26:         dummyReactionNotYetGenerated ← False
27:         index ← 0
28:         for progeny in progenySet do
29:           addToSearchGraph(substrate, progeny)
30:           putToPL(PL[index], node((getSubstrates(substratesNode) –
31:             {substrate}) ∪ progeny))
           index += 1
32:         else
33:           progenySet ← generateRetrosynthesisSteps(substrate)
34:           for progeny in progenySet do
35:             addToSearchGraph(substrate, progeny)
36:             newSubstratesNode ← node((getSubstrates(substratesNode) –
37:               {substrate}) ∪ progeny)
           if notEndOfSynthesis(newSubstratesNode) then
38:             putToPL(PL[substratesNode.indexTakenFrom], newSubstratesNode)
39:           extendIfNeeded(PL, substratesToBeAnalyzed)
```

---

**Figure S2.** The algorithm seeking the easiest syntheses of some targets extends routines used for single-target retrosynthesis search. As presented in the pseudo-code, the search procedure (*searchForLibraryOR*, lines 16-39) is run for the target set,  $\{TS\}$ , i.e., a user-defined library of  $N$  targets  $\{target_1, \dots, target_N\}$ . The algorithm first initializes *substratesToBeAnalyzed* as a list containing the only node for set  $\{TS\}$ , *dummyReactionNotYetGenerated* as *True*, and search graph with single chemical node representing  $\{TS\}$  (lines 17-19). Then, the priority list *PL* is initialized (line 20) as a list of  $N$  copies of queue-based data structures, *PQ*, used in the standard single-target search. Finally, the while-loop begins (loop termination, e.g., by the user, is not explicitly considered for code-brevity reasons). As the first iteration of the loop begins, *substrate* variable is set to *node*( $\{TS\}$ ). The algorithm calls function *generateDummyReactionOR* (line 25), returning *progenySet* composed of  $N$  dummy reactions:  $target_1 \rightarrow TS, \dots, target_N \rightarrow TS$  (lines 1-7), which are further added to the search graph (line 29). Moreover, progenies *node*( $\{target_1\}$ ), ..., *node*( $\{target_N\}$ ) are added to the subsequent elements of *PL* (line 30), allowing for comparable exploration of retrosynthetic options for each element of the initial target library. In all subsequent iterations of the while-loop, the *progenySet* is computed as a set of viable single retrosynthesis steps (line 33). In Chematica, these computations are based on expert-coded rules, and include detection of possible cross-reactivity conflicts, protections, non-selectivity issues, etc. (see main text for references). As the substrates are iteratively queried by retrosynthetic-step generator, the search graph is expanded (line 35), and *newSubstrateNode* variables corresponding to already discovered new synthetic options are added to the same element of *PL* as *substrateNode* was taken from (line 36). As algorithm advances, the list *substratesToBeAnalyzed* is extended (line 39) by taking elements from *PL*, which is inspected in a circular order (*extendIfNeeded*, line 8-15). Please note that in the presented pseudo-code, implementation-specific optimizations such as code parallelization or search-graph representation are not considered.

### Section S3. Pseudocode of the algorithm generating isotopomers with a desired mass shift.

---

```
1: function GENERATELABELLINGS(mol, i, Scurr, S)
  ▷ mol: considered molecule
  ▷ i: index of currently analyzed atom
  ▷ Scurr: current mass shift, i.e. after processing atoms 1, ..., (i - 1)
  ▷ S: desired mass shift of the whole molecule
2:   if i > mol.NumberOfAtoms() then //all atoms already processed
3:     if Scurr = S then return canonicSmiles(mol)
       return NULL //wrong mass shift
4:   a ← mol.getAtom(i)
5:   for iso ∈ allowedIsotopes(a) do
6:     if iso.massShift() + Scurr > S then continue
7:     mol2 ← mol.copy()
8:     setIsotope(mol2, i, iso) //set i-th atom as iso
9:     yield generateLabellings(mol, i + 1, iso.massShift() + Scurr, S)

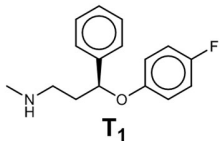
10: function GETISOTOPOMERS(mol, S)
  ▷ mol: considered molecule
  ▷ S: desired mass shift of the whole molecule
11:   V ← ∅
12:   for l ∈ generateLabellings(mol, 1, 0, S) do
13:     if l = NULL then continue
14:     V.add(l)
15:   return V
```

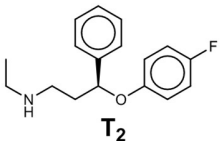
---

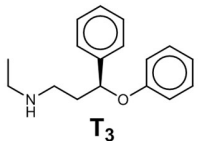
**Figure S3.** The algorithm generates plausible isotopomers (GetIsotopomers, lines 10-15) by recursively-defined analysis of possible atomic labellings (GenerateLabellings, lines 1-9). The input molecule is processed atom-by-atom, adding combinations of isotopic variants to these atoms as long as the user-defined mass shift *S* is not exceeded (lines 5-9). When all atoms are analyzed (and some of them are isotopically labelled), the currently considered isotopomer is returned if its total mass shift equals *S*, otherwise it is rejected (line 3). As the generated isotopomers are added to the final set of results (line 14), the duplicated (i.e., having the same canonical SMILES representation) entries are considered only once.

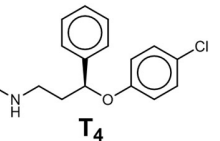
**Section S4. Comparison of the searches to make all members of a library vs. consecutive searches for individual targets.**

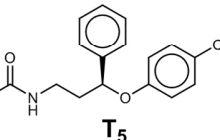
	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	T <sub>4</sub>	T <sub>5</sub>	T <sub>6</sub>	T <sub>7</sub>	T <sub>8</sub>	T <sub>9</sub>	T <sub>10</sub>	T <sub>11</sub>	T <sub>12</sub>	$\sum T_{1-12}$	Find-All	$\frac{\sum T_{1-12}}{\text{Find-All}}$
<b>run1</b>															
All nodes	3703	3653	3572	1314	1095	1924	5616	3138	1816	6577	3818	3584	39810	4228	9.4
Chemical	1553	1534	1494	569	490	802	2219	1315	764	2604	1550	1463	16357	1653	9.9
Expanded	76	70	74	20	16	32	92	71	36	111	77	80	755	72	10.5
Time (sec.)	20	18	21	12	22	11	61	19	10	68	26	26	314	55	5.7
Expanded/time	3.80	3.89	3.52	1.67	0.73	2.91	1.51	3.74	3.60	1.63	2.96	3.08	2.40	1.31	1.8
<b>run2</b>															
All nodes	4471	3721	3722	1627	1095	2486	5129	3634	1633	6569	3818	3816	41721	4228	9.9
Chemical	1872	1558	1547	699	490	1035	2038	1509	695	2594	1550	1566	17153	1653	10.4
Expanded	85	71	76	27	16	45	86	78	30	111	77	81	783	72	10.9
Time (sec.)	23	19	19	12	21	12	59	20	9	67	25	27	313	57	5.5
Expanded/time	3.70	3.74	4.00	2.25	0.76	3.75	1.46	3.90	3.33	1.66	3.08	3	2.50	1.26	2.0
<b>run3</b>															
All nodes	4319	3556	3297	1438	1095	2240	5802	4183	1629	6577	3690	3844	41670	4228	9.9
Chemical	1816	1499	1381	621	490	946	2287	1721	681	2604	1497	1579	17122	1653	10.4
Expanded	82	68	69	24	16	36	95	88	33	111	73	82	777	72	10.8
Time (sec.)	22	19	18	12	21	10	62	30	9	68	25	26	322	56	5.8
Expanded/time	3.73	3.58	3.83	2.00	0.76	3.60	1.53	2.93	3.67	1.63	2.92	3.15	2.41	1.29	1.9

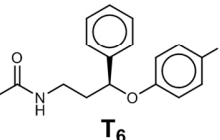
  
**T<sub>1</sub>**

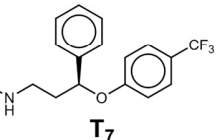
  
**T<sub>2</sub>**

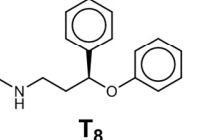
  
**T<sub>3</sub>**

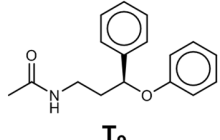
  
**T<sub>4</sub>**

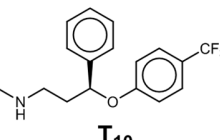
  
**T<sub>5</sub>**

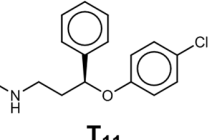
  
**T<sub>6</sub>**

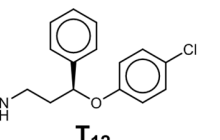
  
**T<sub>7</sub>**

  
**T<sub>8</sub>**

  
**T<sub>9</sub>**

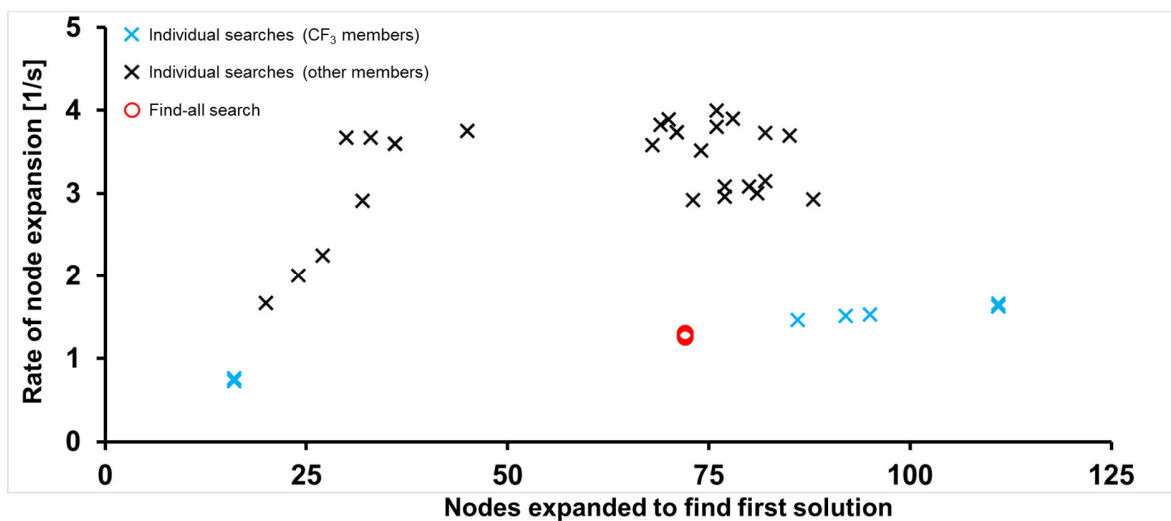
  
**T<sub>10</sub>**

  
**T<sub>11</sub>**

  
**T<sub>12</sub>**

**Table S1:** Individual measures of graph size and time elapsed for 12 single-target searches and the corresponding find-all algorithm. The searches are for the Prozac example from the main-text **Figures 4 and 5** (Markush structure C([\*:1])C[C@H](Oc1ccc([\*:2])cc1)c1ccccc1 with *substitutions\_dct* = {1: ['N(C)', 'N(CC)', 'N(C(=O)C)'], 2: ['H', 'F', 'Cl', 'C(F)(F)(F)']}). The individual targets are shown under the Table.

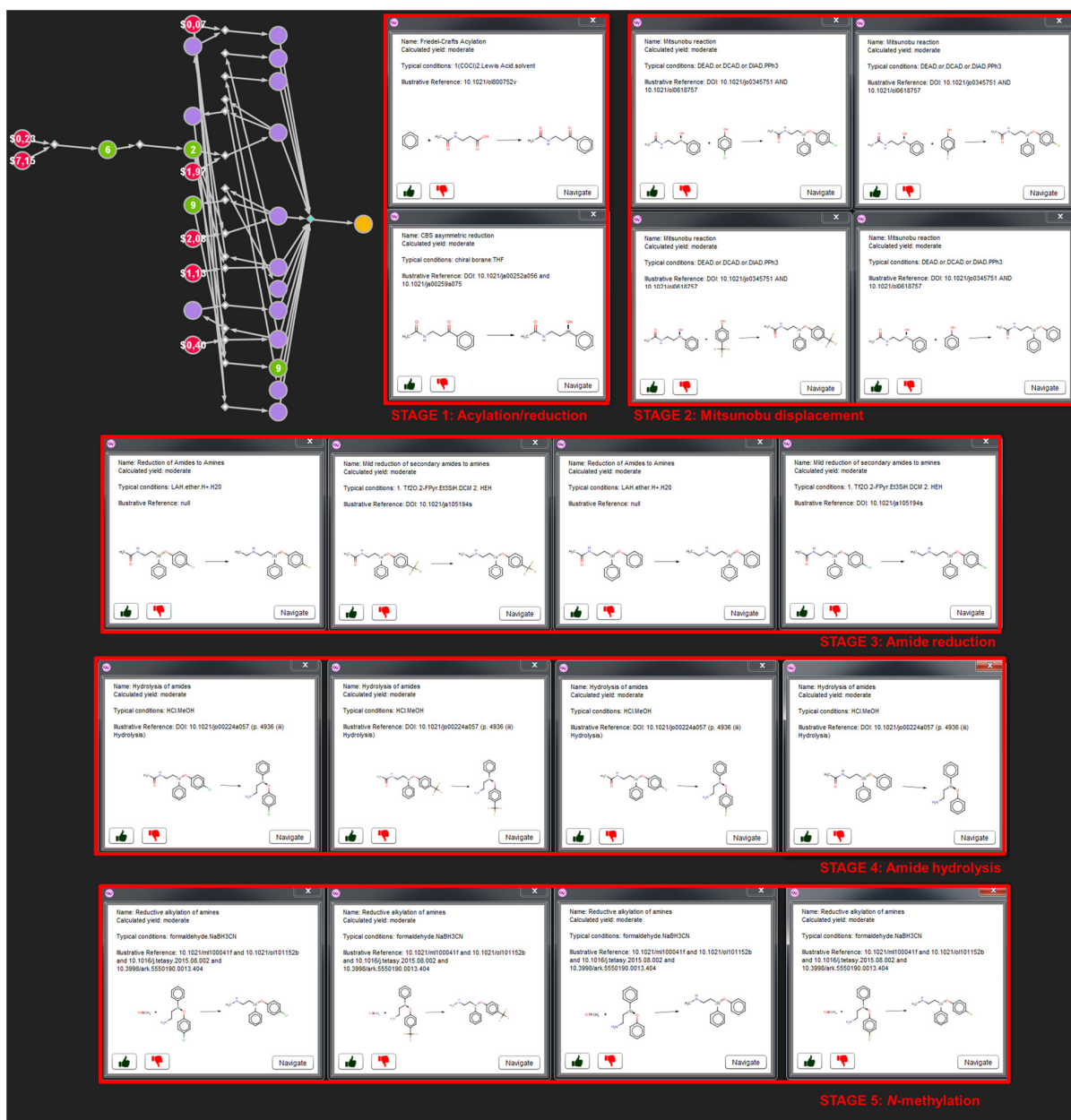
All searches were performed on a machine with 64-processor threads clocked @2.2-3.6 GHz each. For each run, as soon as the target molecule (or all target molecules in case of the find-all algorithm) became synthesizable, the timings and the search graphs were saved. Then, the following parameters were computed for each saved graph: number of all nodes, number of chemical-substance nodes (i.e., circular nodes representing chemicals), number of chemical nodes that were expanded (i.e., retrosynthetic options for related chemicals were already computed and added to the graph). An additional parameter – ratio of expanded nodes to time – was added to estimate search efficiency. The measurements were repeated three times (run1, run2, run3). The find-all algorithm performed around 5 times faster and required around 10 times fewer nodes than 12 consecutive, single-target searches. Interestingly, we observe that find-all search needs less expanded nodes than certain individual searches (e.g., for targets T<sub>1</sub>, T<sub>7</sub> or T<sub>12</sub>), which might reflect ‘synergy’ between targets in the find-all mode (i.e., retrosynthetic steps explored for synthesis of one individual target might be utilized in synthetic pathways of other targets). Of note, the searches with larger number of expanded nodes tend to have higher node expansion ratio plateauing around 3-4 nodes/second. However, this is not surprising, as we generally anticipate fewer but more complex molecules (e.g. late intermediates) to be analyzed at the beginning of the search vs. larger numbers of simpler molecules to be analyzed later in the search. Additionally, we note that searches performed for trifluoromethylated targets (T<sub>5</sub>, T<sub>7</sub> and T<sub>10</sub>) have lower ratios of node expansion (**Chart S1**) compared with other library members. This can be explained by our observation that execution of retrosynthetic steps involving highly symmetric groups, such as CF<sub>3</sub>, is relatively more computationally demanding, possibly due to multiple transformation-to-molecule matchings.



**Chart S1.** Rates of node expansion (i.e., number of expanded nodes per unit time) during single target and find-all searches. Blue crosses denote observed node expansion rates for trifluoromethylated library members T<sub>5</sub>, T<sub>7</sub> and T<sub>10</sub>. Node expansion rates for other library members and find-all search are denoted by black crosses and red circles, respectively.

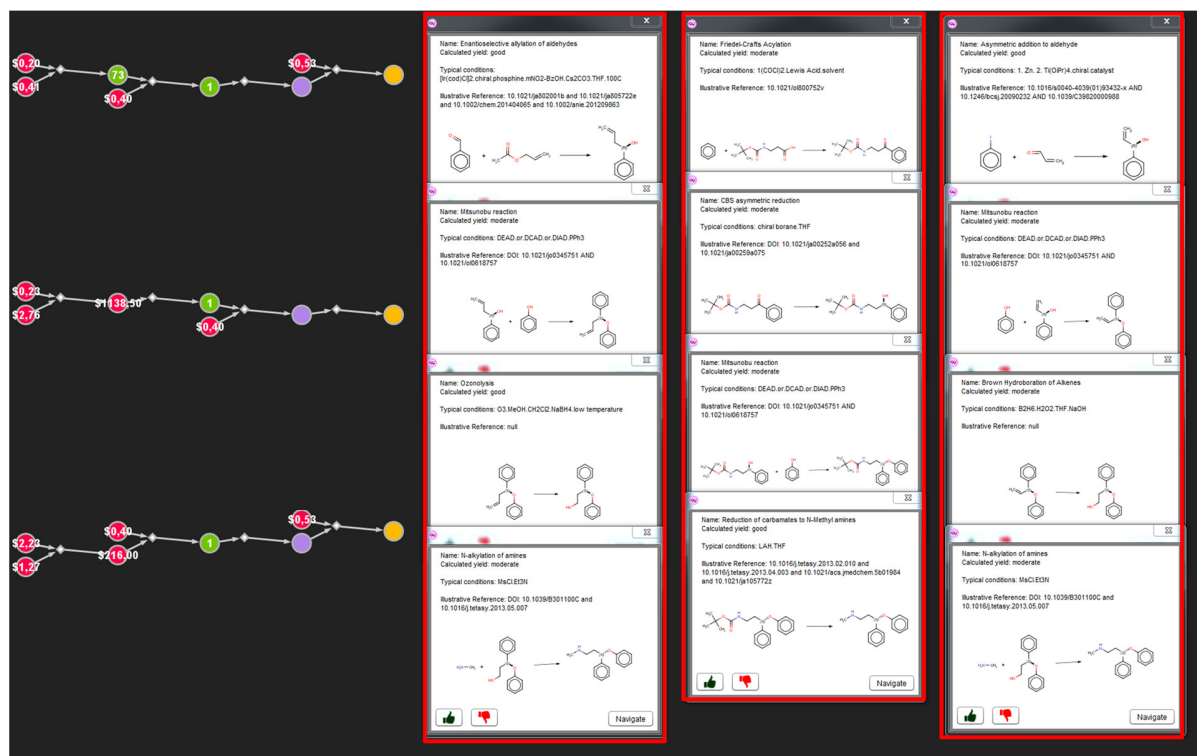


## Section S5. Details of Chematica's retrosynthetic analyses performed for fluoxetine derivatives.

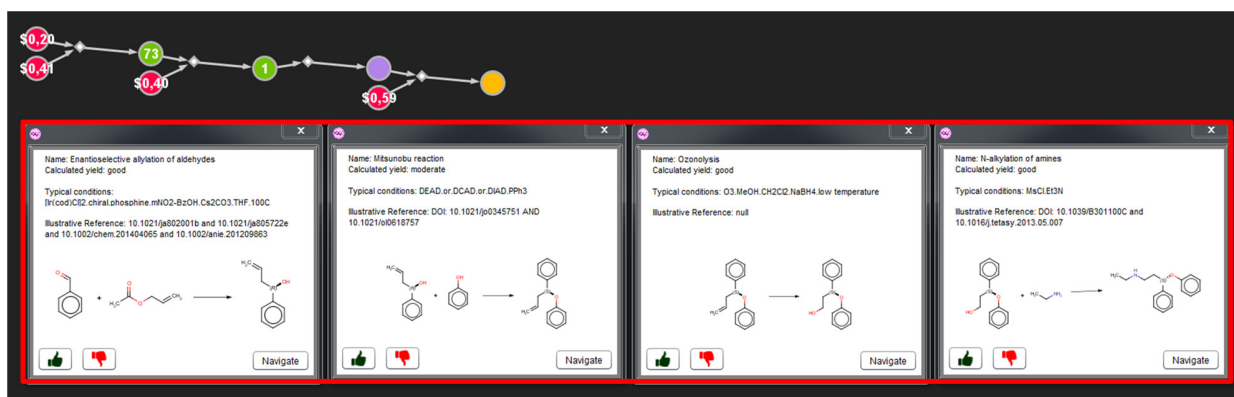


**Figure S4.** Details of Chematica's synthetic plan for the library of fluoxetine derivatives; the figure complements main-text **Figure 4**.

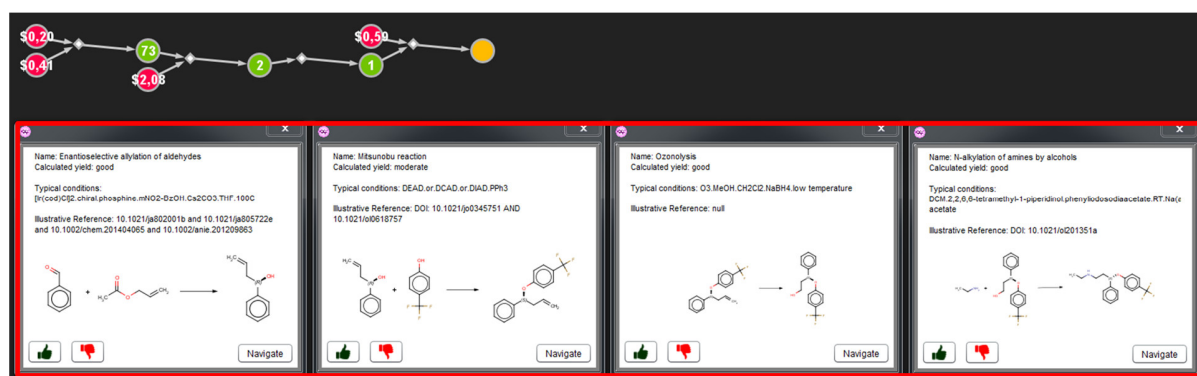
**Section S6. Details of Chematica's retrosynthetic analyses performed for fluoxetine derivatives in individual, target-by-target searches.**



**Figure S5.** Details of the top-scoring pathways identified by Chematica for the A1 member of the library of fluoxetine derivatives (cf. main-text **Figure 5a-c**).

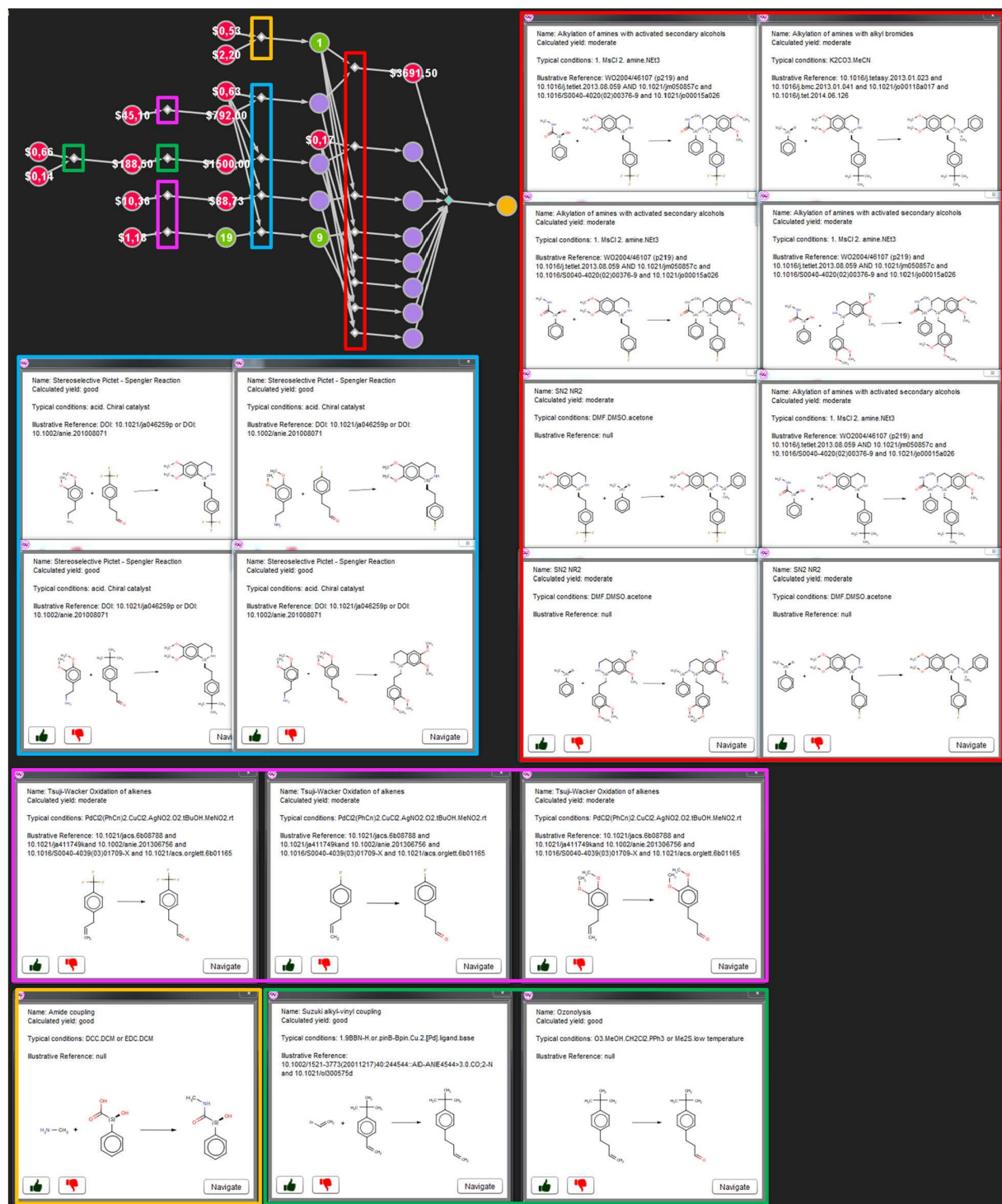


**Figure S6.** Details of the top-scoring pathway identified by Chematica for the **A3** member of the library of fluoxetine derivatives (cf. main-text **Figure 5d**).



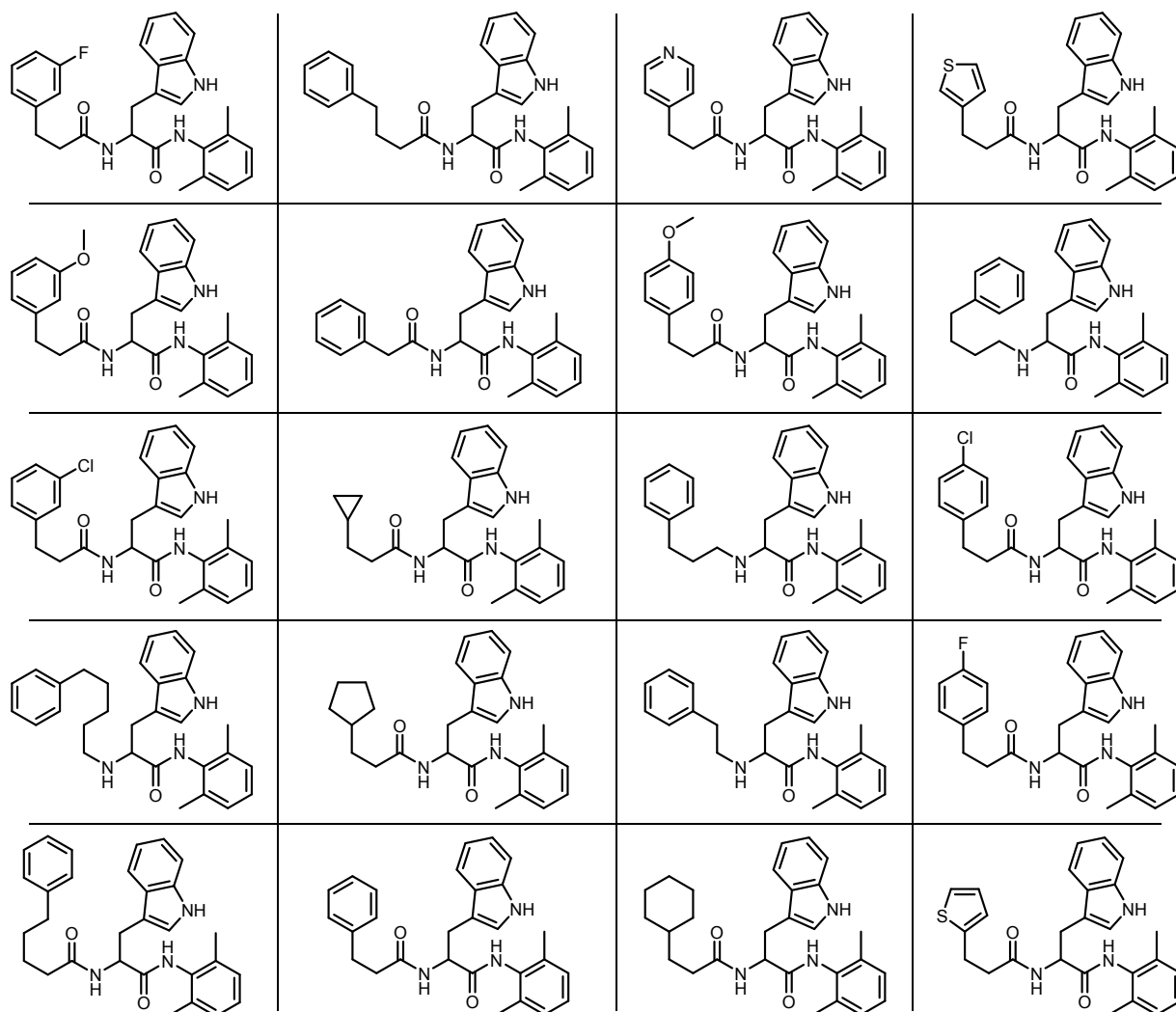
**Figure S7.** Details of the top-scoring pathway identified by Chematica for the **D3** member of the library of fluoxetine derivatives (cf. main-text **Figure 5e**).

## Section S7. Details of Chematica's retrosynthetic analyses performed for Almorexant derivatives.

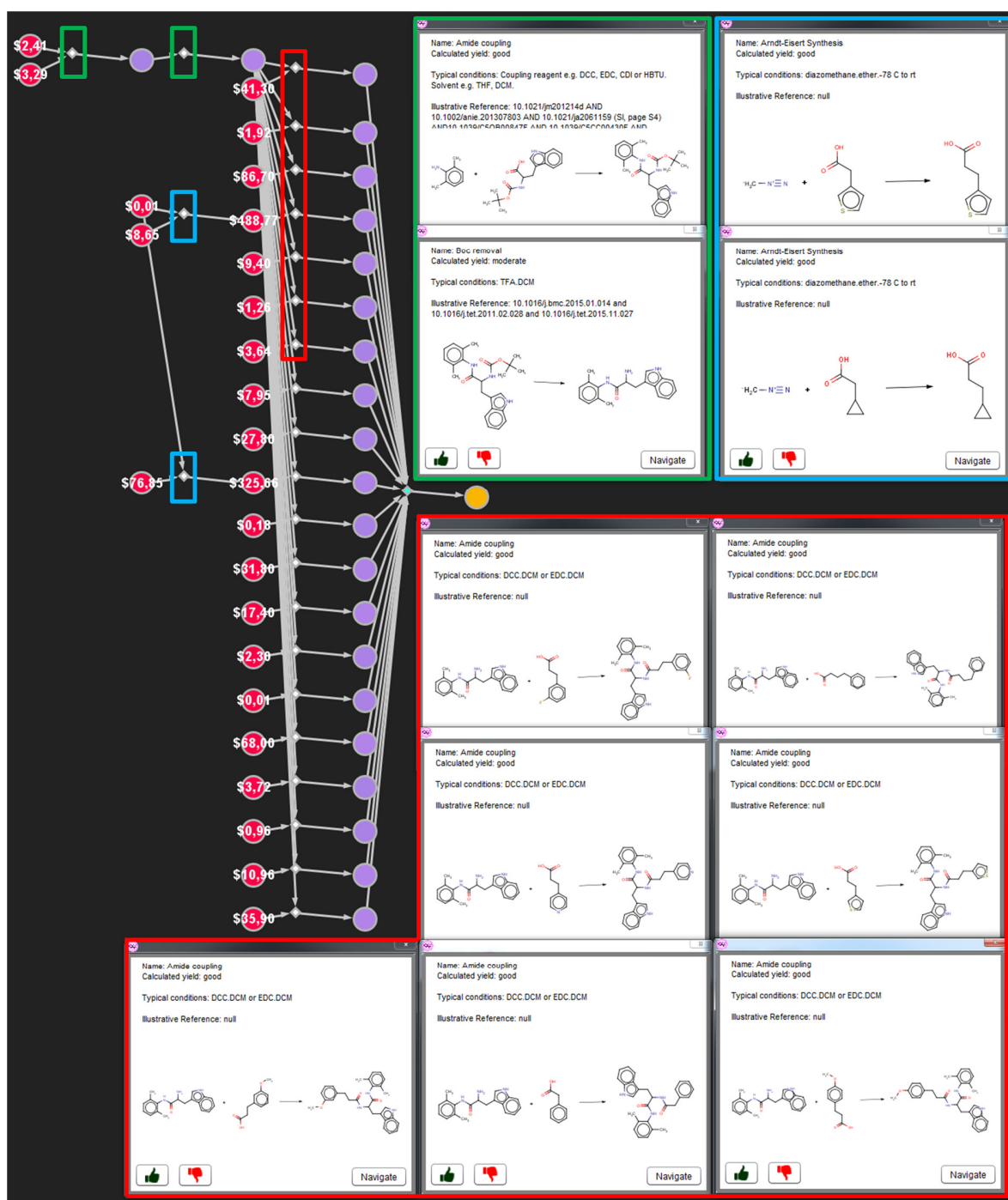


**Figure S8.** Details of Chematica's synthetic plan for the library of Almorexant derivatives; the figure complements main-text **Figure 6**.

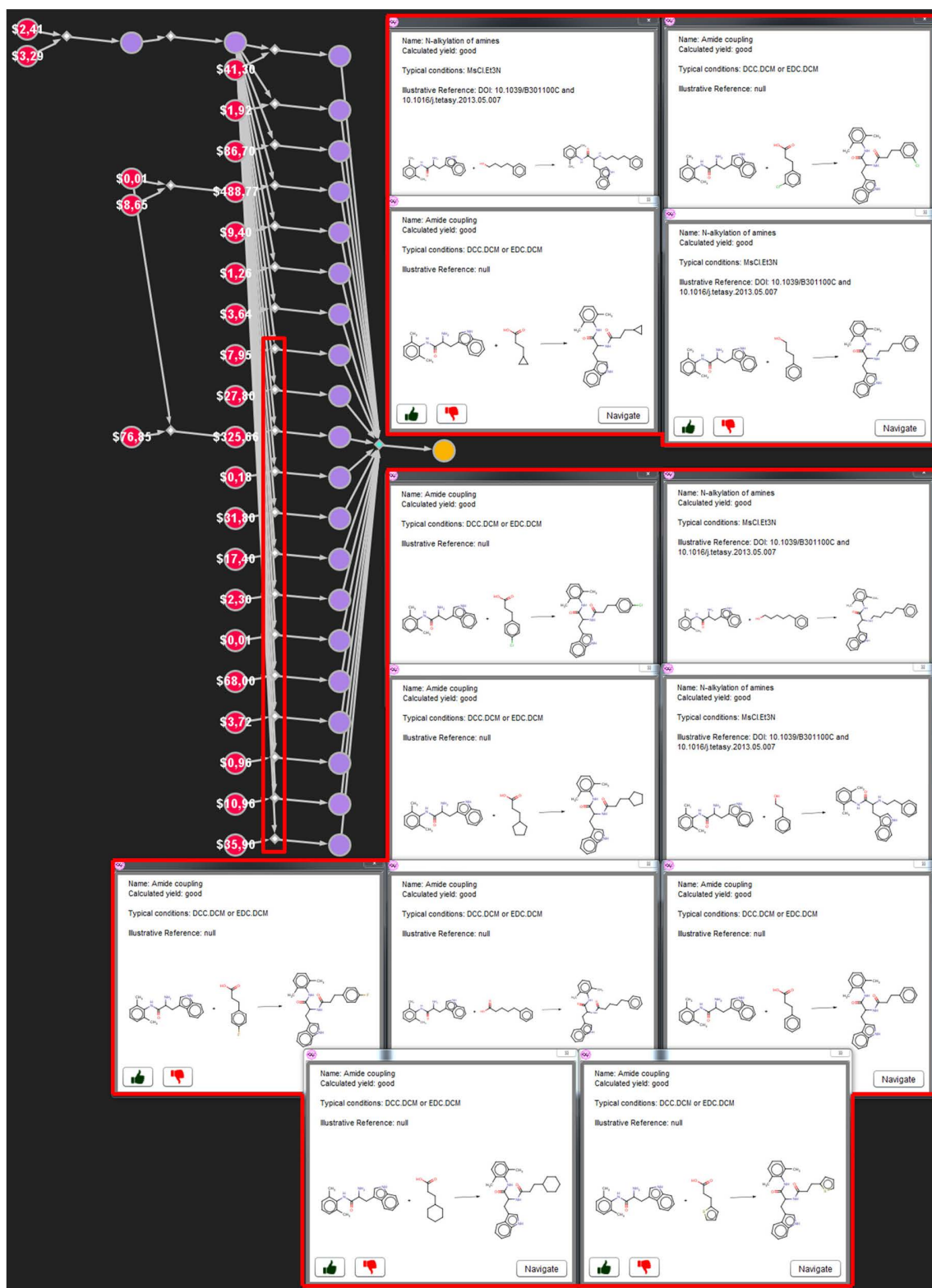
**Section S8. Details of Chemica's retrosynthetic analyses performed for tryptophan derivatives.**



**Figure S9.** Components of library of tryptophan derivatives; the figure complements main-text **Figure 7**.

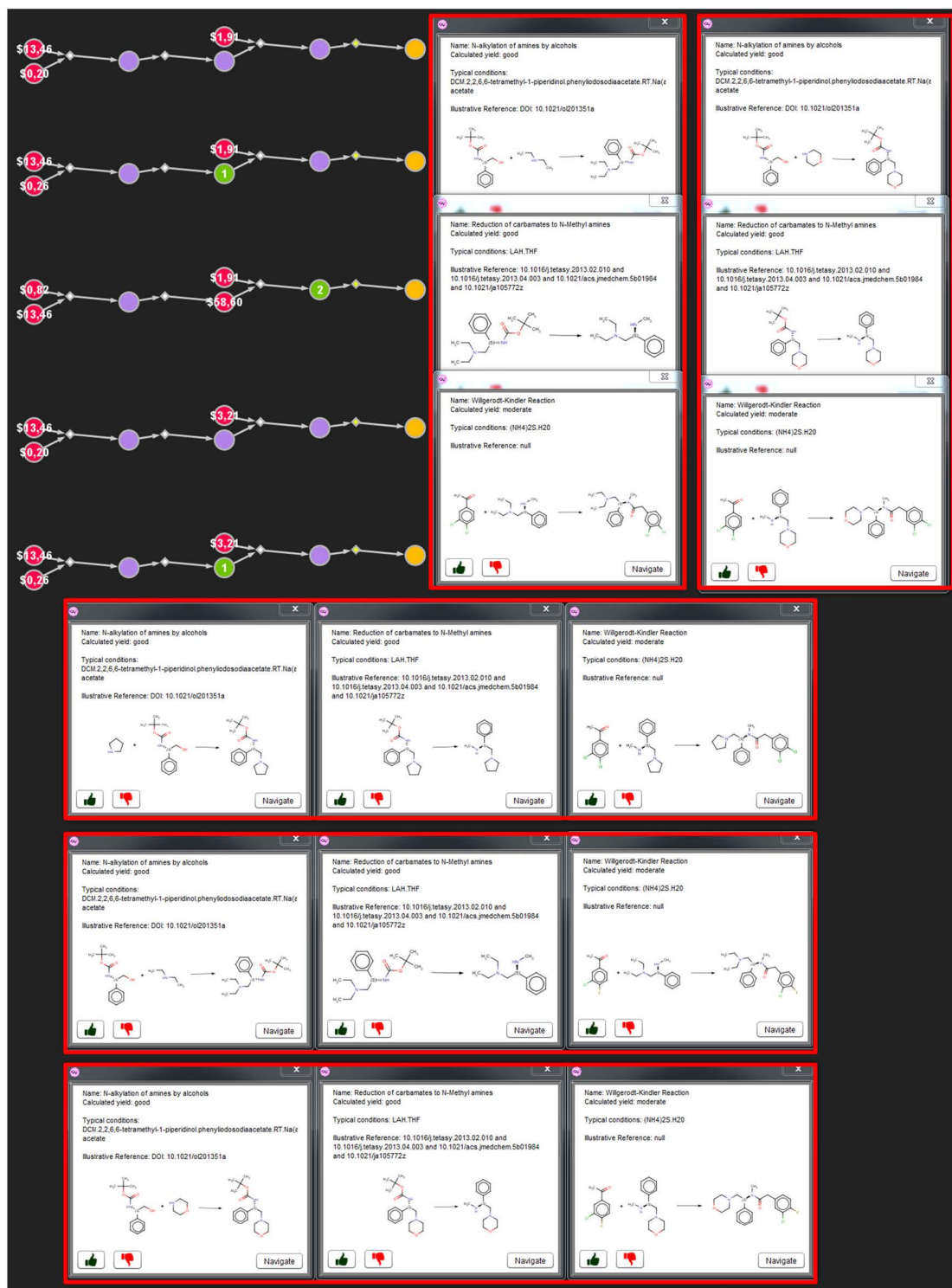


**Figure S10.** Details of Chematica's synthetic plan for the library of tryptophan derivatives; the figure complements main-text **Figure 7**.



**Figure S11.** Details of Chematica's synthetic plan for the library of tryptophan derivatives; the figure complements main-text **Figure 7**.

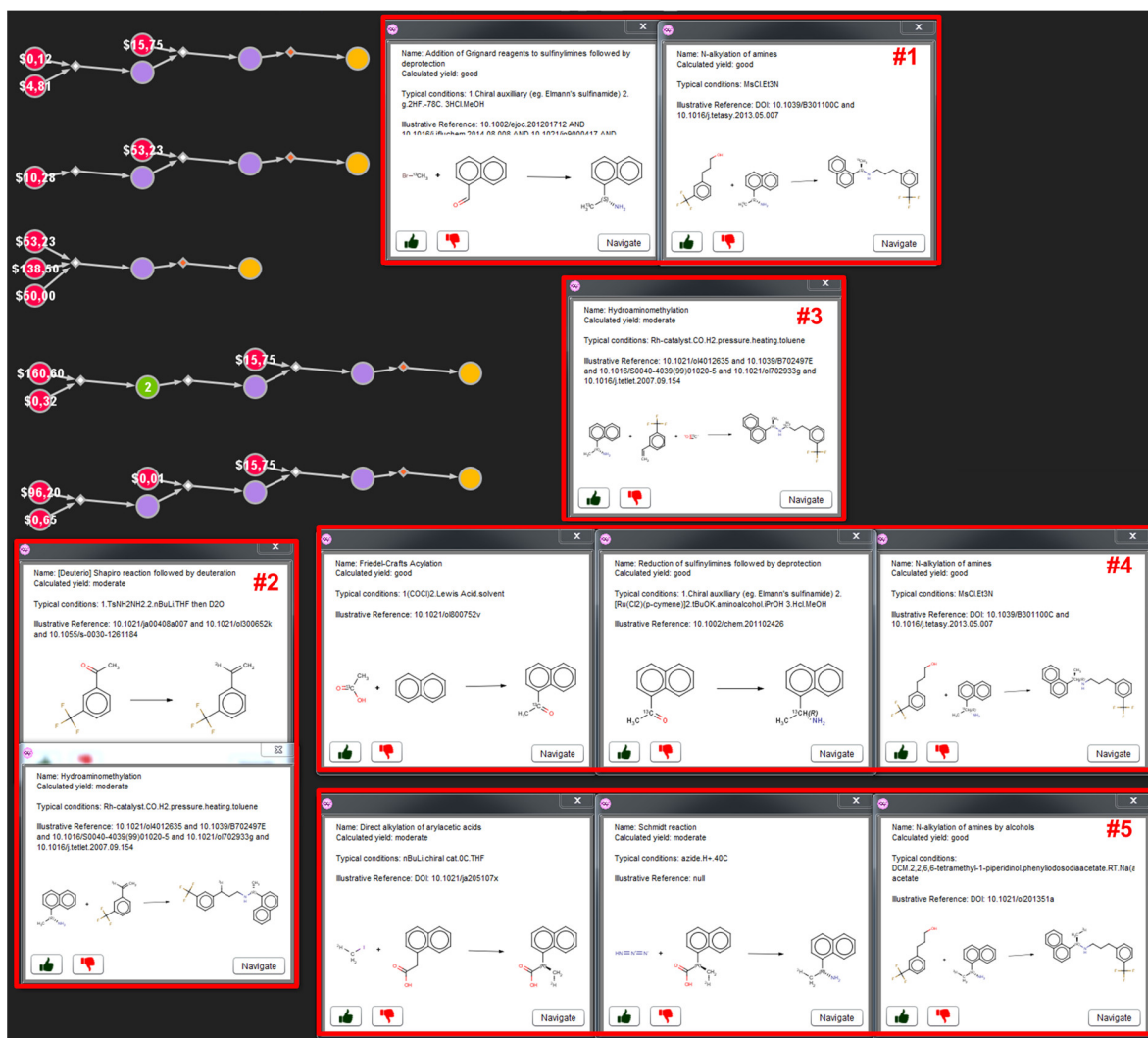
**Section S9. Details of Chematica's retrosynthetic analyses performed for the ICI199441 derivatives.**



**Figure S12.** Details of the five top-scoring pathways proposed by Chematica's for the synthesis of the most accessible members of the ICI199441 library discussed in main-text **Figure 8**.

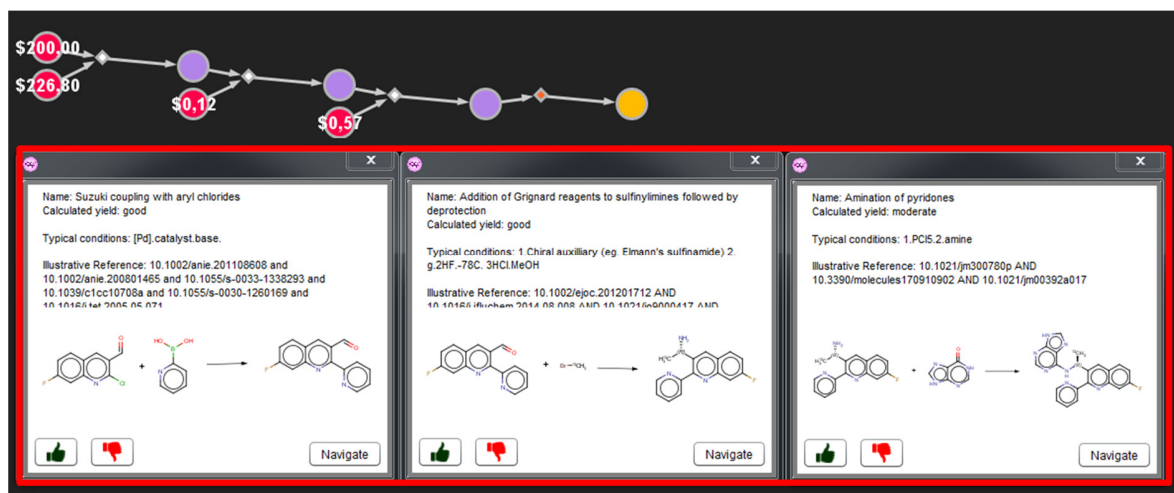


**Section S10. Details of Chematica's retrosynthetic analyses performed for  $^{13}\text{C}/^2\text{H}$  labeled Cinacalcet.**

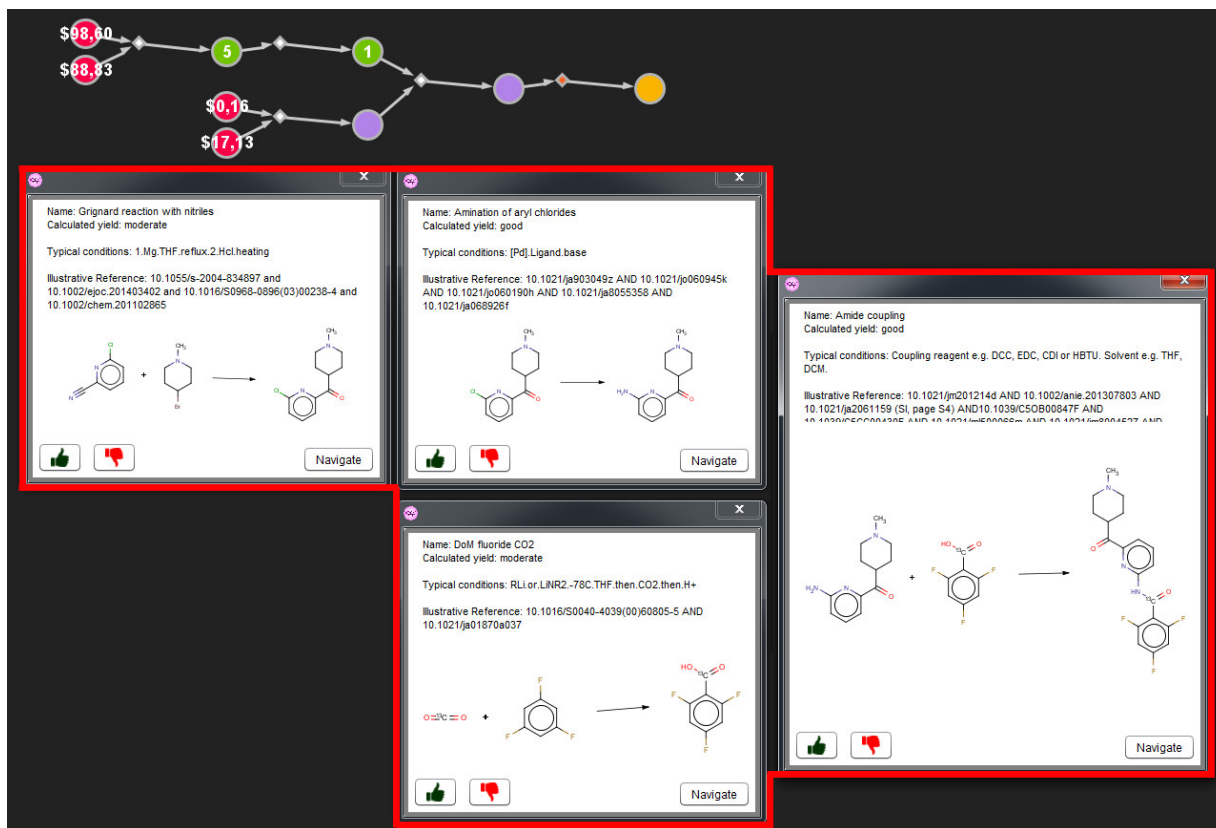


**Figure S13.** Details of the five top-scoring pathways identified by Chematica searching for the most accessible  $^{13}\text{C}/^2\text{H}$  labeled M+1 isotopomers of Cinacalcet (cf. main-text Figure 9).

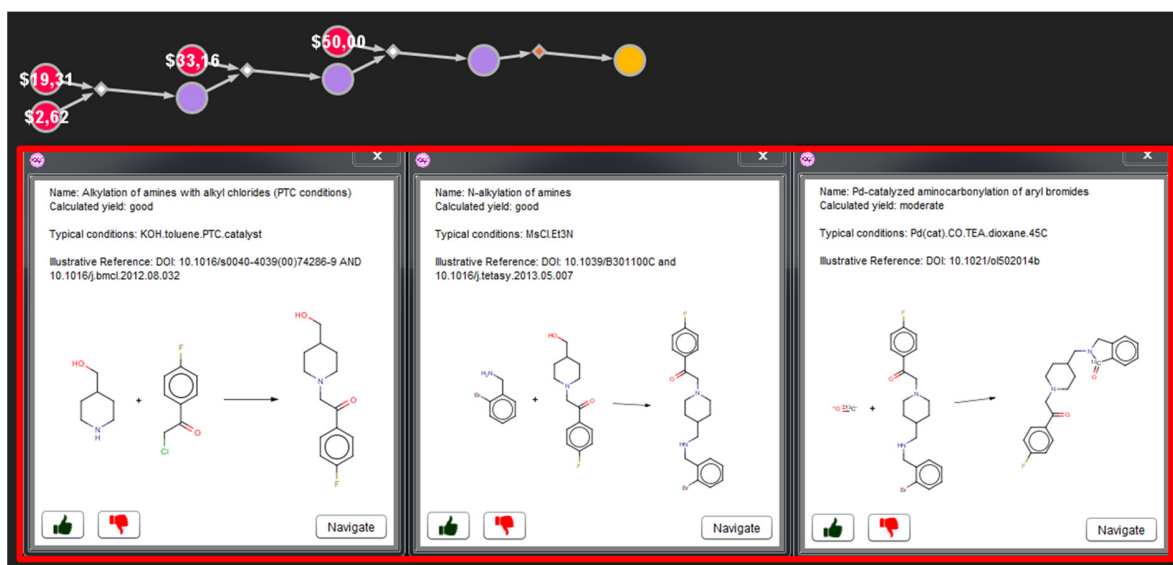
**Section S11. Details of Chematica's retrosynthetic analyses performed for various M+1 <sup>13</sup>C labelled drug molecules.**



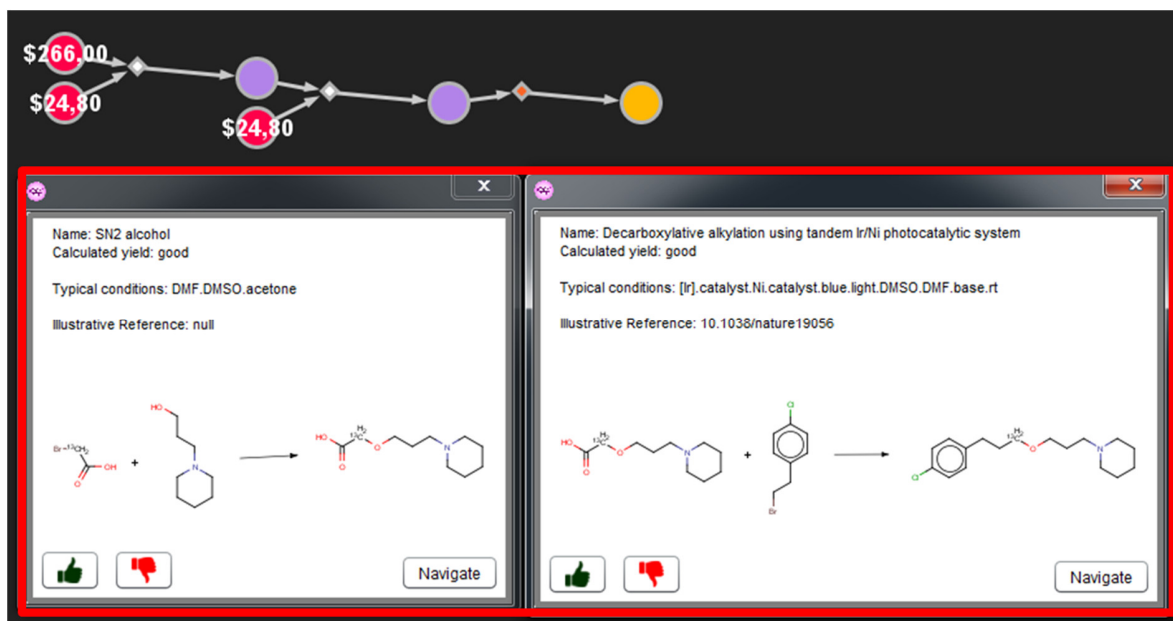
**Figure S14.** Details of the top-scoring pathway proposed by Chematica for the most accessible <sup>13</sup>C labeled M+1 isotomer of AMG-319 (cf. main-text **Figure 10a**).



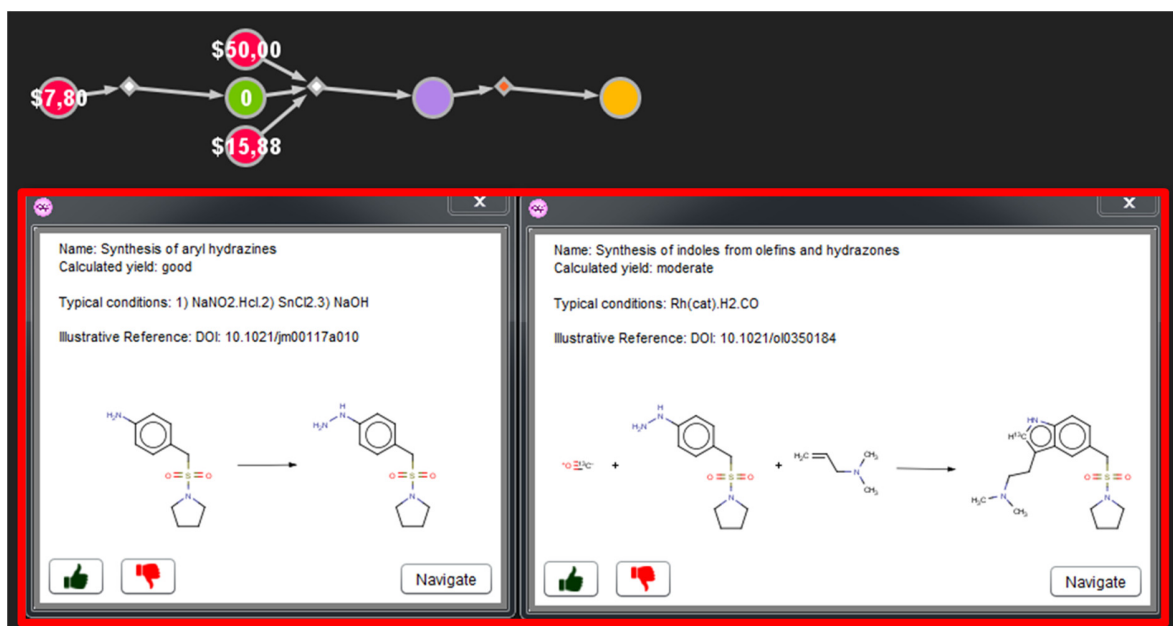
**Figure S15.** Details of the top-scoring pathway proposed by Chematica for the most accessible <sup>13</sup>C labeled M+1 isotomer of Lasmiditan (cf. main-text **Figure 10b**).



**Figure S16.** Details of the top-scoring pathway proposed by Chematica for the most accessible  $^{13}\text{C}$  labeled M+1 isotopomer of Roluperidone (cf. main-text **Figure 10c**).



**Figure S17.** Details of the top-scoring pathway proposed by Chematica for the most accessible  $^{13}\text{C}$  labeled M+1 isotopomer of Pitolisant (cf. main-text **Figure 10d**).



**Figure S18.** Details of the top-scoring pathway proposed by Chematica for the most accessible <sup>13</sup>C labeled M+1 isotopomer of Almotriptan (cf. main-text **Figure 10e**).

# A computer algorithm to discover iterative sequences of organic reactions

Karol Molga<sup>1,2</sup>, Sara Szymkuć<sup>1,2</sup>, Patrycja Gołębiowska<sup>1</sup>, Oskar Popik<sup>1</sup>, Piotr Dittwald<sup>1</sup>,  
Martyna Moskal<sup>1,2</sup>, Rafal Roszak<sup>1,2</sup>, Jacek Mlynarski<sup>1</sup>✉ and Bartosz A. Grzybowski<sup>1,2,3,4</sup>✉

DOSTĘP OGRANICZONY

<sup>1</sup>Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland. <sup>2</sup>Allchemy, Inc., Highland, IN, USA. <sup>3</sup>IBS Center for Soft and Living Matter, UNIST, Ulsan, South Korea. <sup>4</sup>Department of Chemistry, UNIST, Ulsan, South Korea. ✉e-mail: [jacek.mlynarski@gmail.com](mailto:jacek.mlynarski@gmail.com); [nanogrzybowski@gmail.com](mailto:nanogrzybowski@gmail.com)

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY



DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY



DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY

DOSTĘP OGRANICZONY



DOSTĘP OGRANICZONY

## Supplementary information

---

# A computer algorithm to discover iterative sequences of organic reactions

---

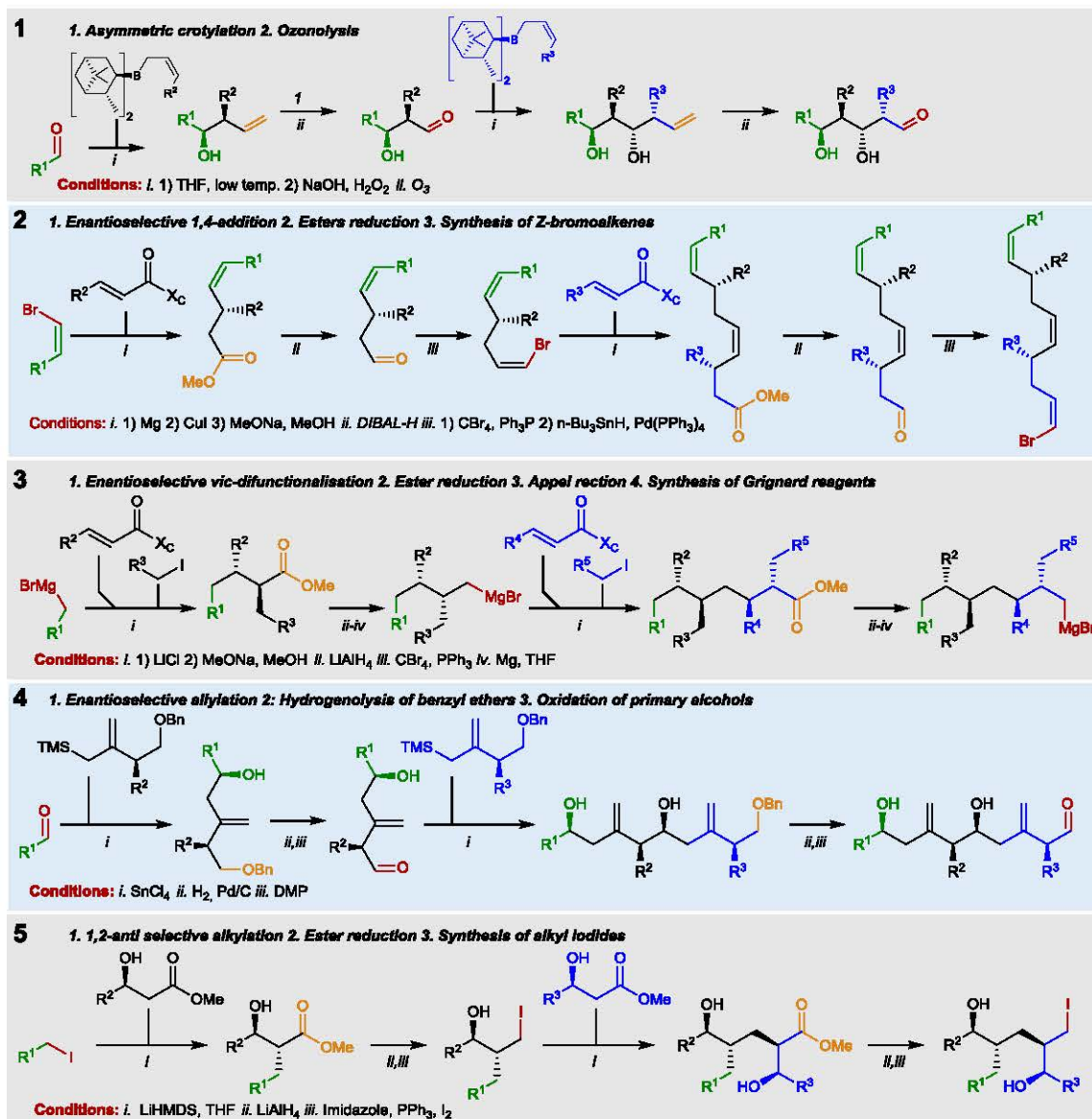
In the format provided by the authors and unedited

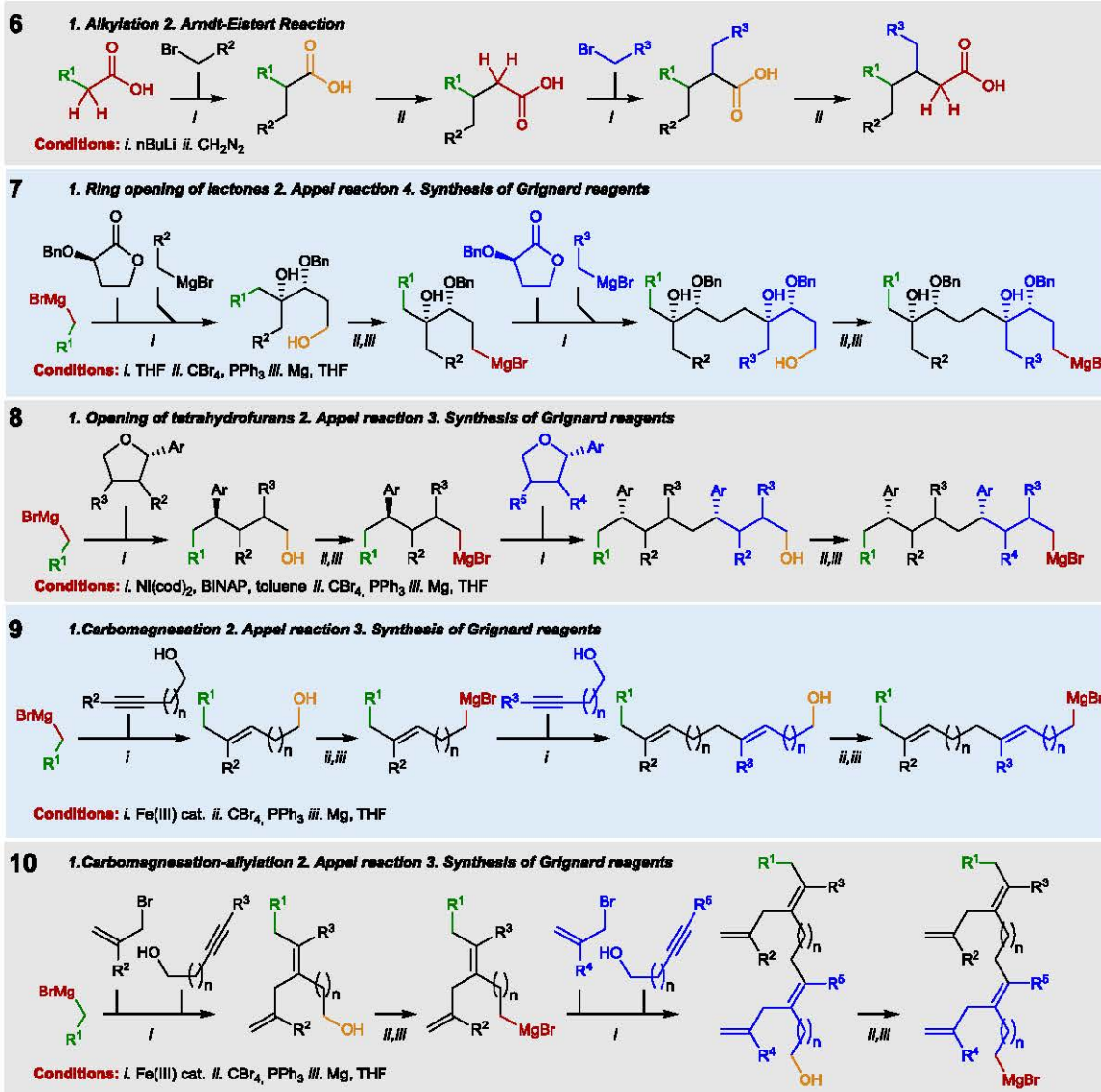
**Supplementary Information** for Manuscript titled „*Computer algorithm discovers iterative sequences of organic reactions*” by K. Molga, S. Szymkuć, P. Gołębiowska, O. Popik, P. Dittwald, M. Moskal, R. Roszak, J. Mlynarski\* & B.A. Grzybowski<sup>1\*</sup>

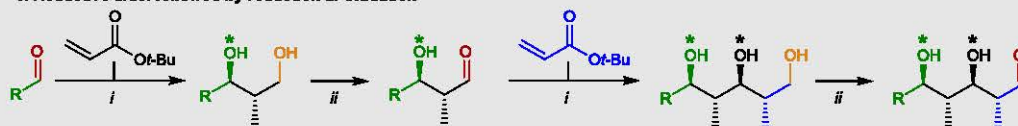
## CONTENTS

Section S1. Additional examples of unprecedented iterative sequences discovered by the “basic” algorithm from Figure 2a.....	2
Section S2. Examples of iterative sequences found by the “basic” algorithm from Figure 2a and analogous to – but not identical with – iterations already described in the literature. ....	6
Section S3. Additional examples of unprecedented iterative sequences found by the “advanced” algorithm from Figure 2b,c.....	7
Section S4. Examples of previously known iterative sequences rediscovered by the algorithm. ....	18
Section S5. General experimental procedures.....	20
Section S6. Iterative synthesis of 1,5,n polyols .....	21
Section S7. Iterative synthesis of 1,4,n polyols .....	29
Section S8. Determination of absolute configuration of newly formed stereogenic centers (Mosher ester analysis).....	36
Section S9. Iterative synthesis of monhexocin’s fragment .....	44
Section S10. Literature precedents of heterocycle-forming reactions. ....	54
Section S11. Spectroscopic data.....	55
Section S12. User Manual for Allchemy’s “Iterator” module.....	87
Section S13. Selection of conditions for iterative sequences. ....	93
Section S14. Pseudocode for the algorithm to identify iterative sequences.....	96
Section S15. References .....	101

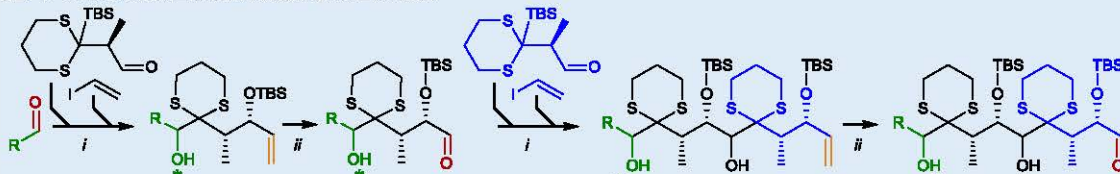
Section S1. Additional examples of unprecedented iterative sequences discovered by the “basic” algorithm from Figure 2a.



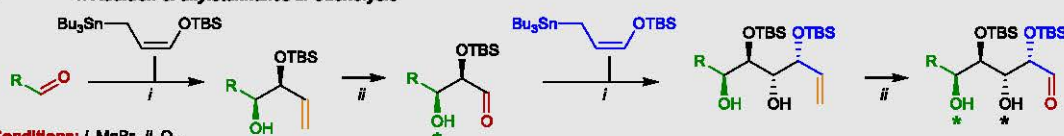


**11 1. Reductive aldol followed by reduction 2. Oxidation**

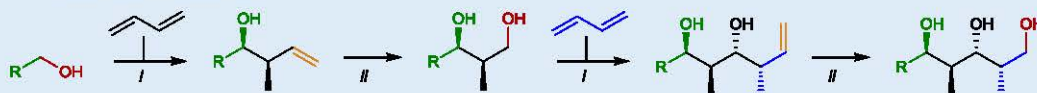
Conditions: i. 1) chiral borane, Et<sub>2</sub>O, 0 °C 2) RCHO -78 °C 3) LiAlH<sub>4</sub> // DMP

**12 1. Tricomponent ARC-II type coupling 2. Ozonolysis**

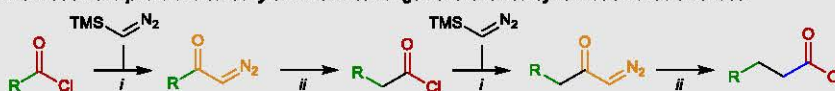
Conditions: i. t-BuLi then HMPA // O<sub>3</sub>

**13 1. Addition of allylstannanes 2. Ozonolysis**

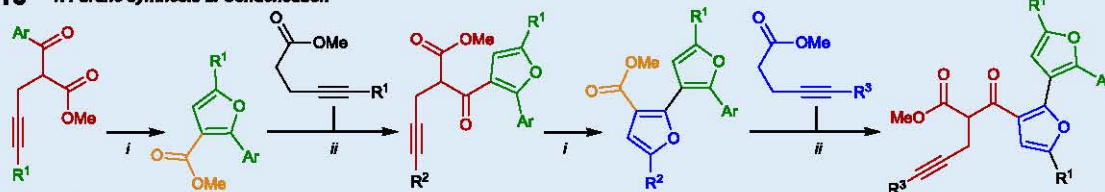
Conditions: i. MgBr<sub>2</sub> // O<sub>3</sub>

**14 1. Crotylation 2. Ozonolysis**

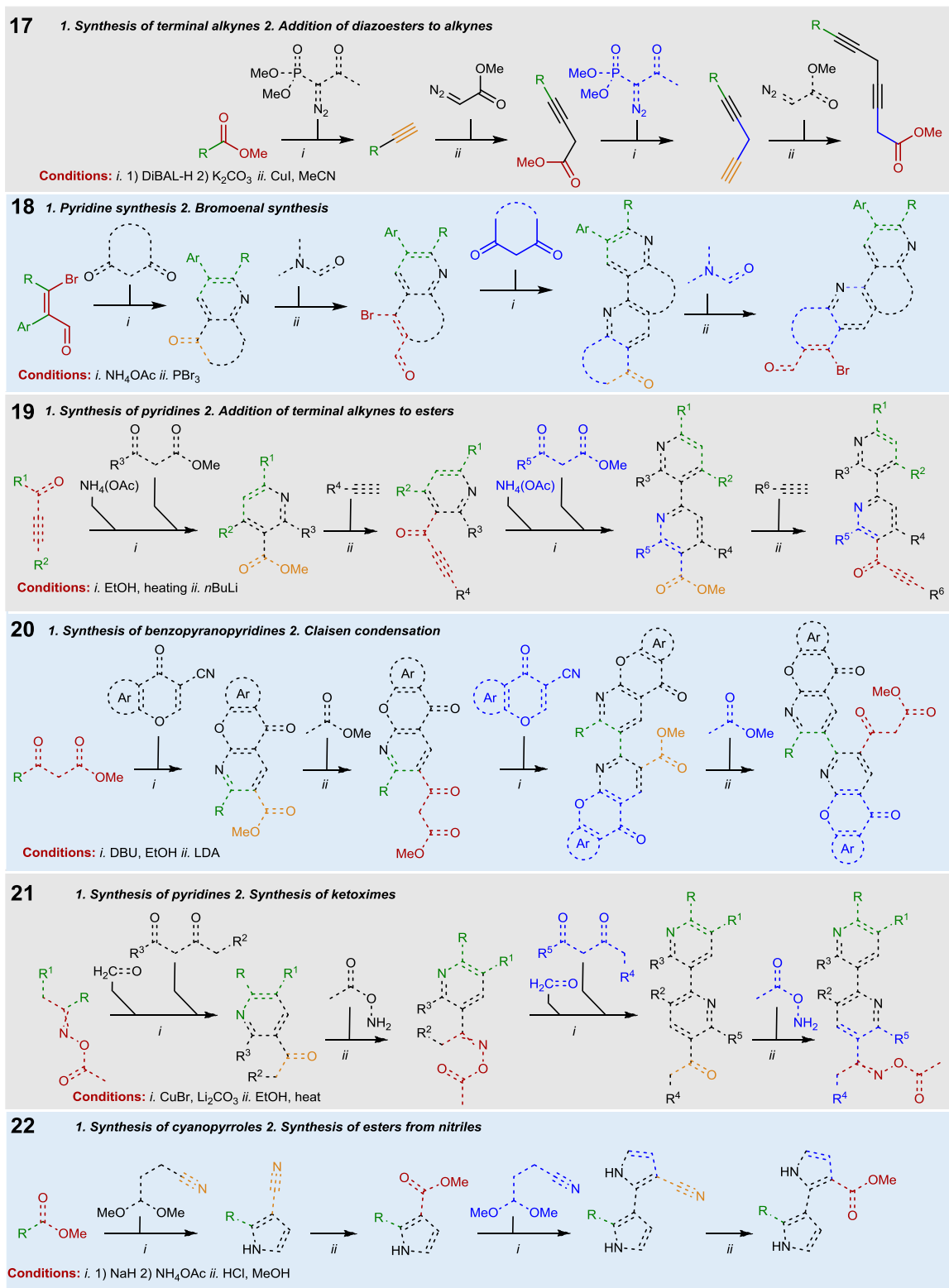
Conditions: i. [Ru]-complex, chiral phosphine, chiral acid // O<sub>3</sub> then NaBH<sub>4</sub>

**15 1. Formation of alpha-diazo carbonyls 2. Wolff rearrangement followed by formation of acid halides**

Conditions: i. Et<sub>3</sub>N, THF, MeCN // ii. 1) Ag(I) cat., hv 2) oxalyl chloride

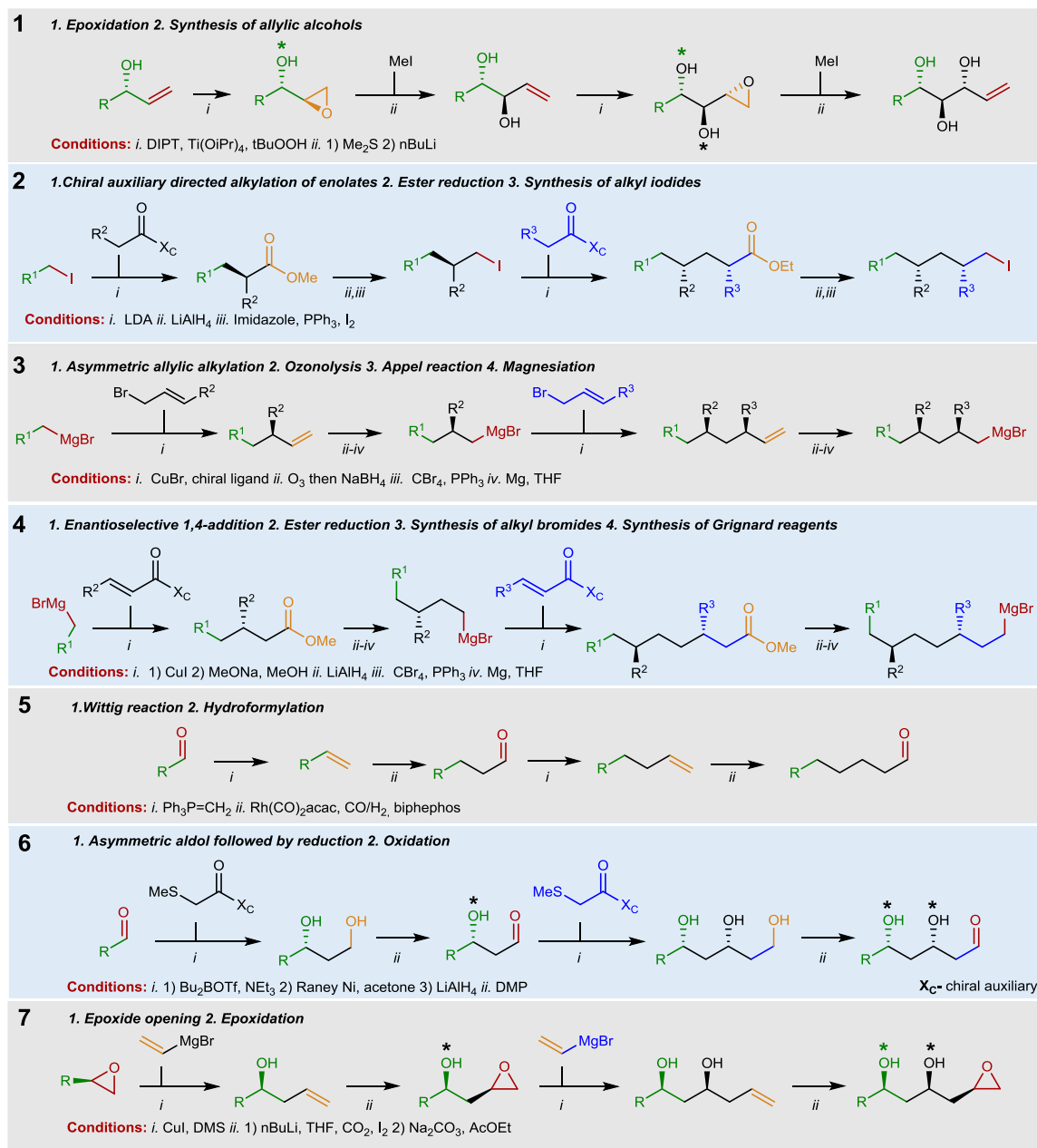
**16 1. Furane synthesis 2. Condensation**

Conditions: i. Bi(OTf)<sub>3</sub> // LDA



**Figure S1.** Examples of new iterative sequences discovered by the “basic” algorithm from main-text Figure 2a.

**Section S2. Examples of iterative sequences found by the “basic” algorithm from Figure 2a and analogous to – but not identical with – iterations already described in the literature.**

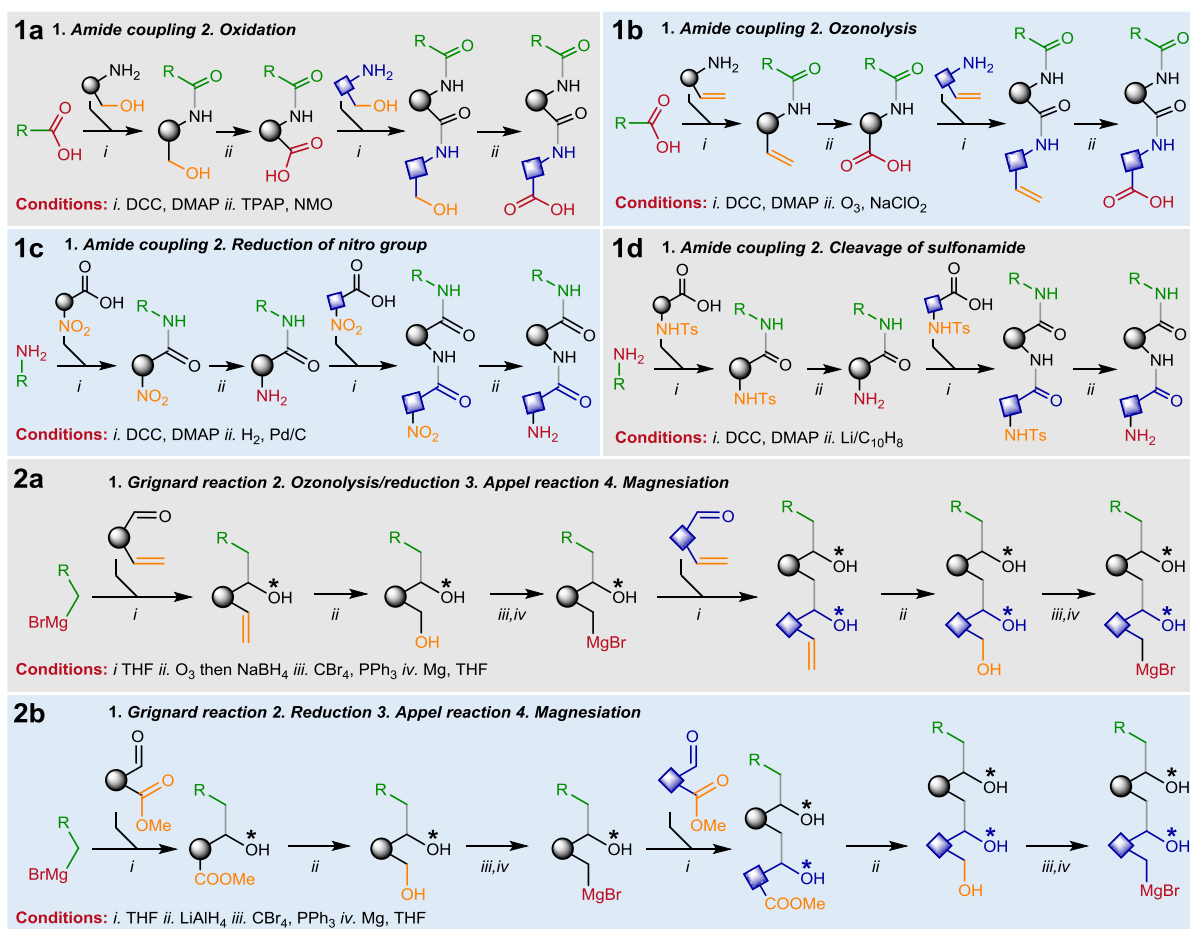


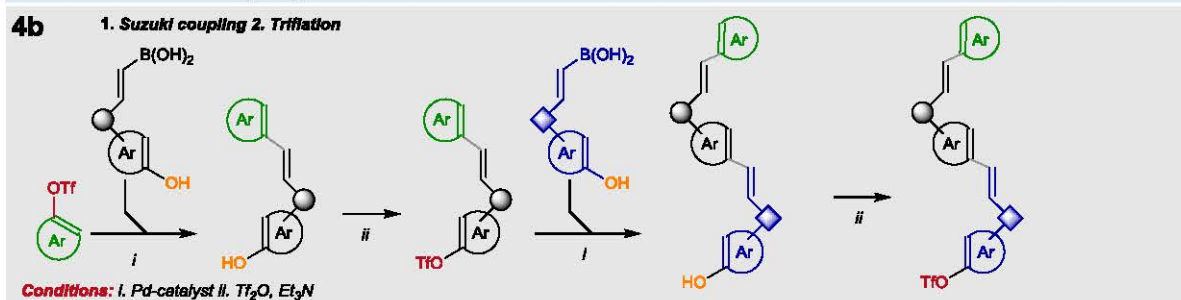
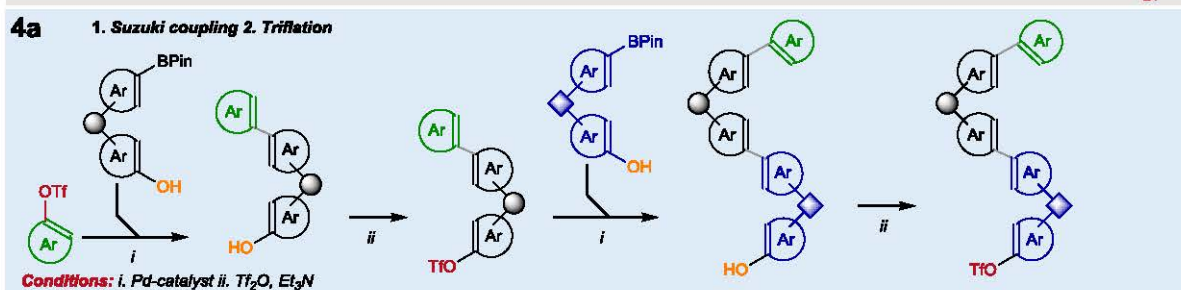
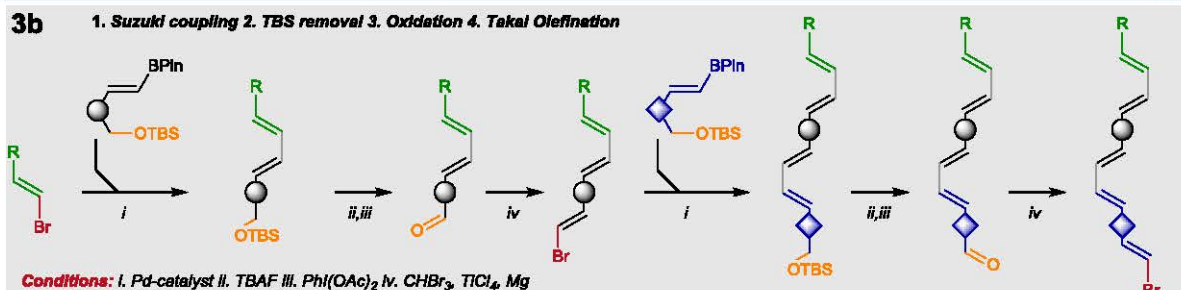
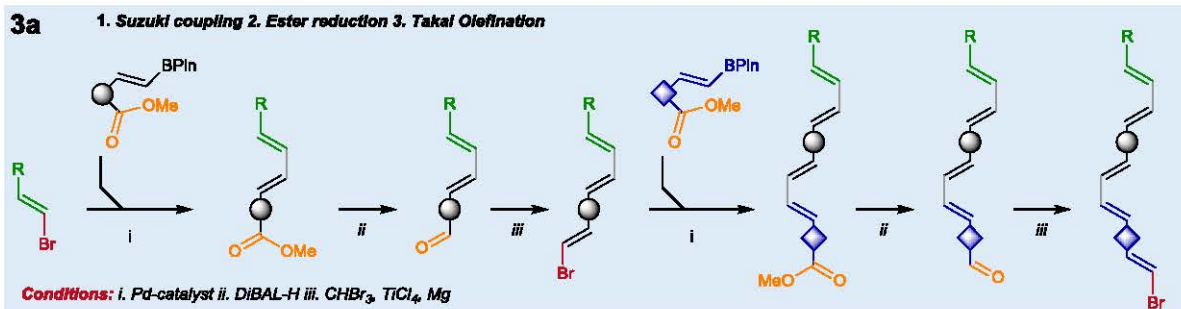
**Figure S2.** Examples of iterative sequences found by the “basic” algorithm from main-text Figure 2a and analogous to – but not identical with – iterations already described in the literature.

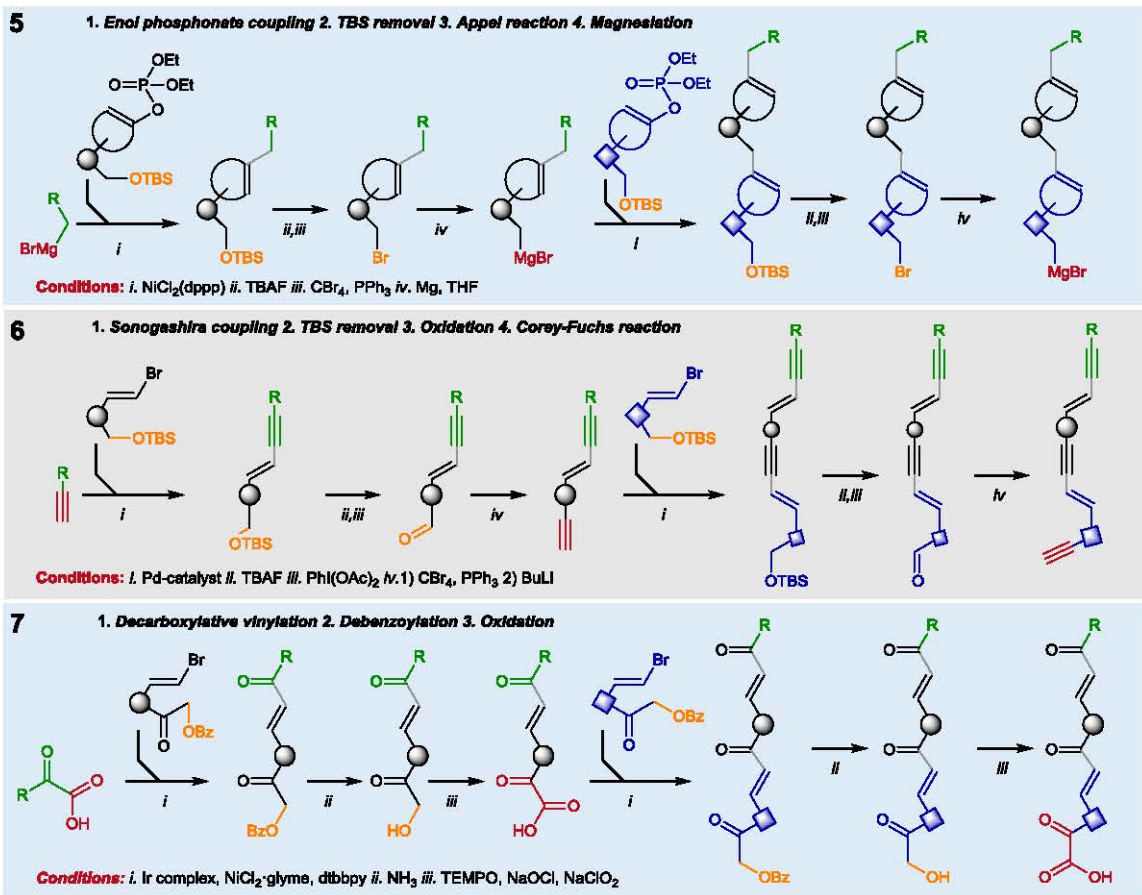


## Section S3. Additional examples of unprecedented iterative sequences found by the “advanced” algorithm from Figure 2b,c.

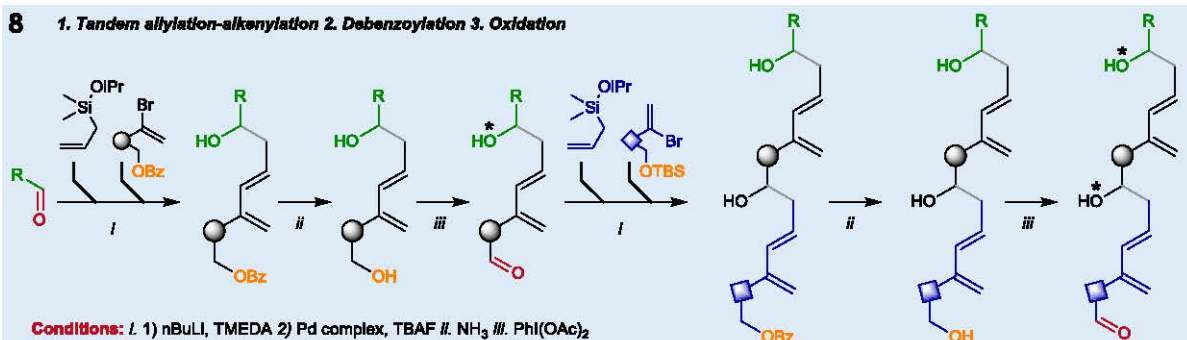
Note: In sequences #1 and #2 below, the final structures (peptides and *1,n*-polyols) can also be prepared by known iterations. The iterations found by the algorithm are shown here because they allow for the use of alternative reagents and/or conditions. For example, amines and carboxylic acids used in known (also rediscovered by our algorithm) iterative amide couplings are commonly regenerated either under acidic (from  $-\text{NH}^t\text{Boc}/-\text{COO}^t\text{Bu}$  groups) or basic (from  $-\text{NH}^t\text{Fmoc}/-\text{COOEt}$  groups) conditions. Sequences 1A-D shown below enable regeneration of necessary functional groups under reductive (amine from sulfonamide or nitro groups) or oxidative (carboxylic acid from alcohol or alkene) conditions thus enabling coupling with substrates possessing both acid- and base-labile fragments.



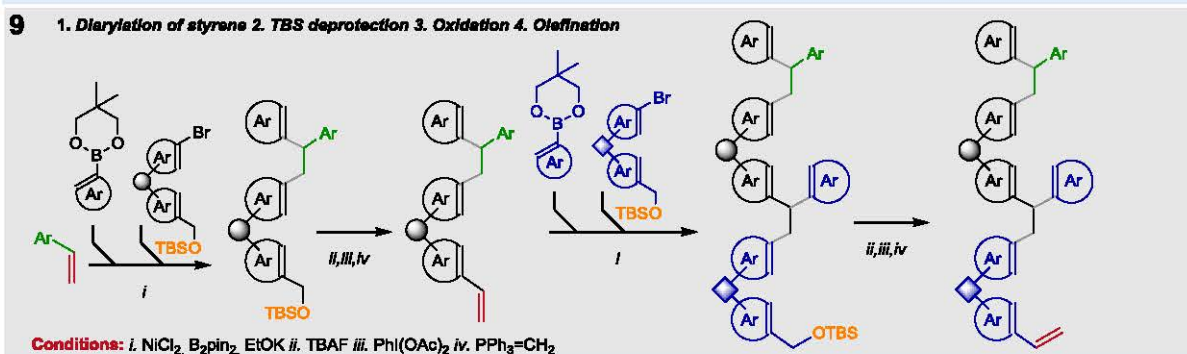




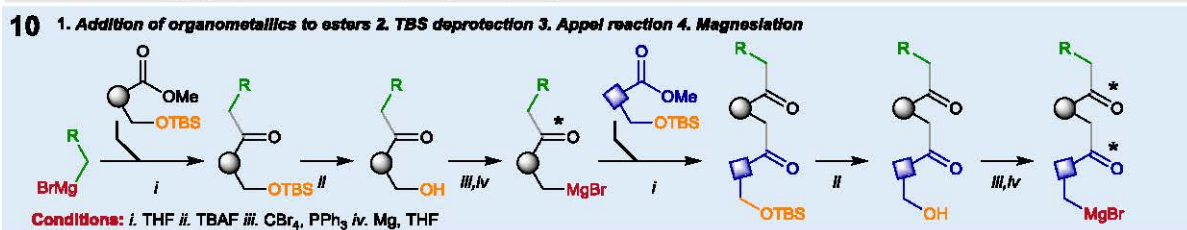
**8** 1. Tandem allylation-alkenylation 2. Debenzoylation 3. Oxidation



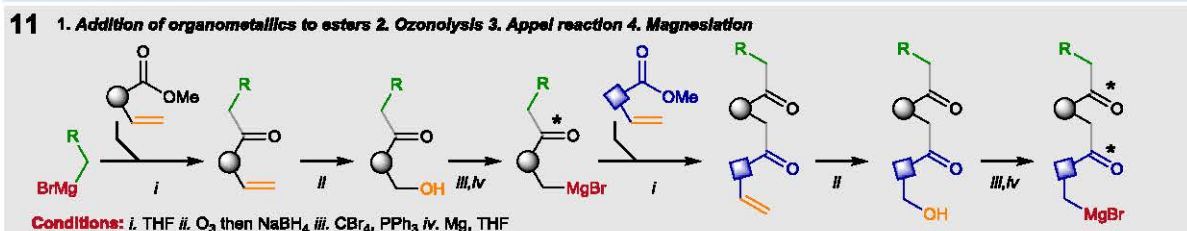
**9** 1. Diarylation of styrene 2. TBS deprotection 3. Oxidation 4. Olefination



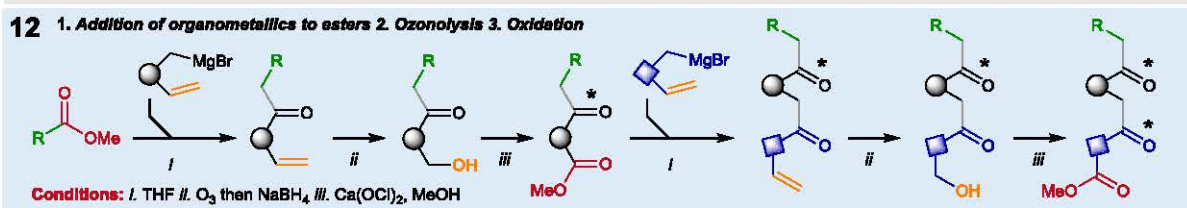
**10** 1. Addition of organometallics to esters 2. TBS deprotection 3. Appel reaction 4. Magnesium

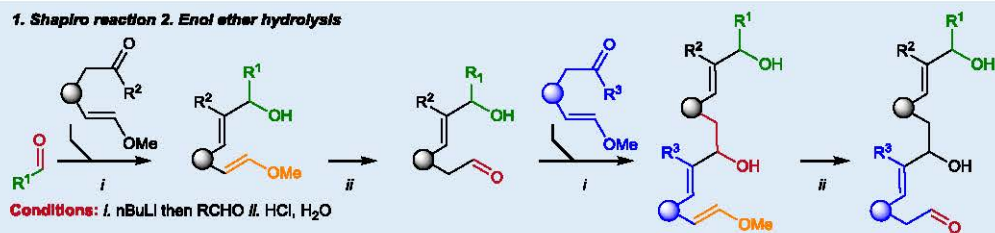
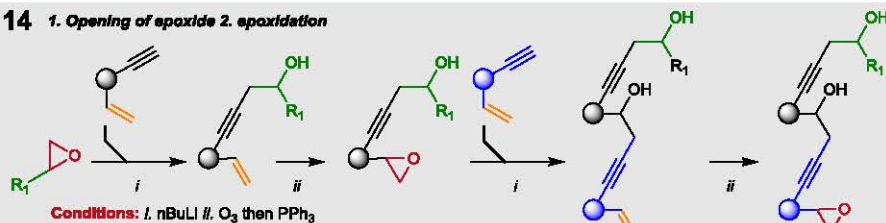
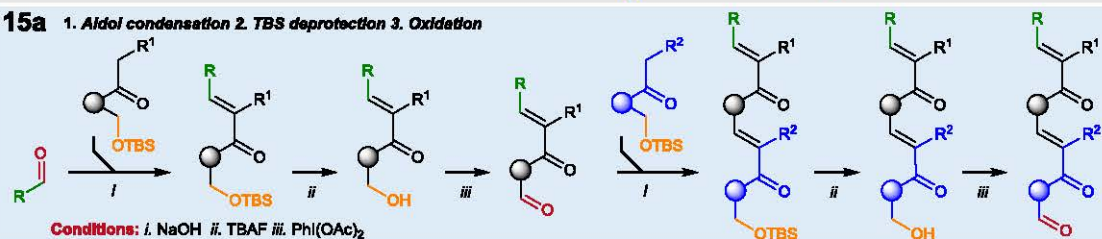
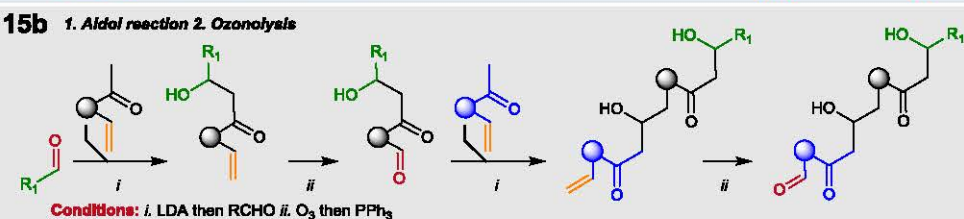
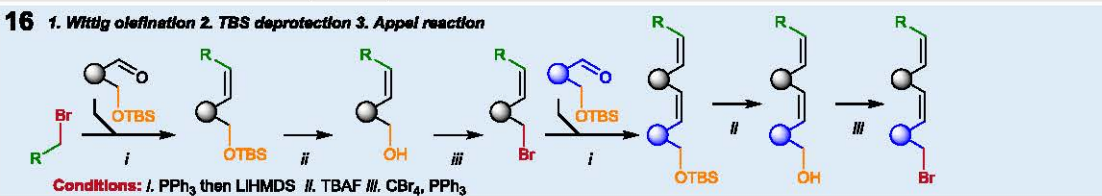
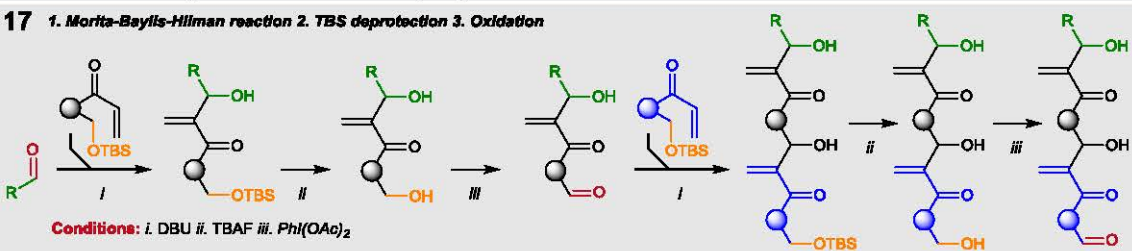


**11** 1. Addition of organometallics to esters 2. Ozonolysis 3. Appel reaction 4. Magnesium

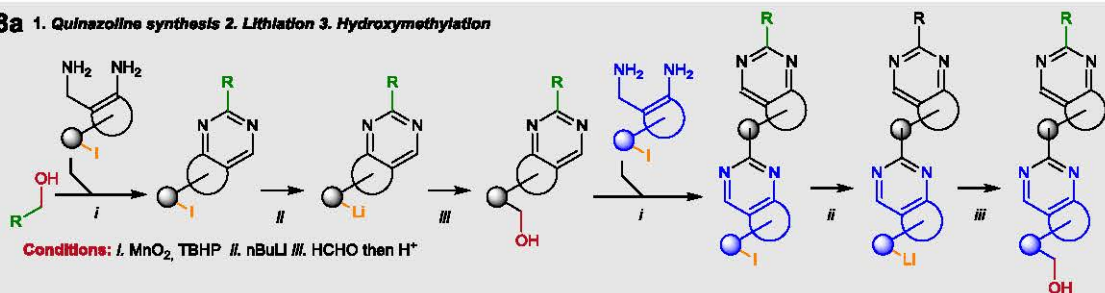


**12** 1. Addition of organometallics to esters 2. Ozonolysis 3. Oxidation

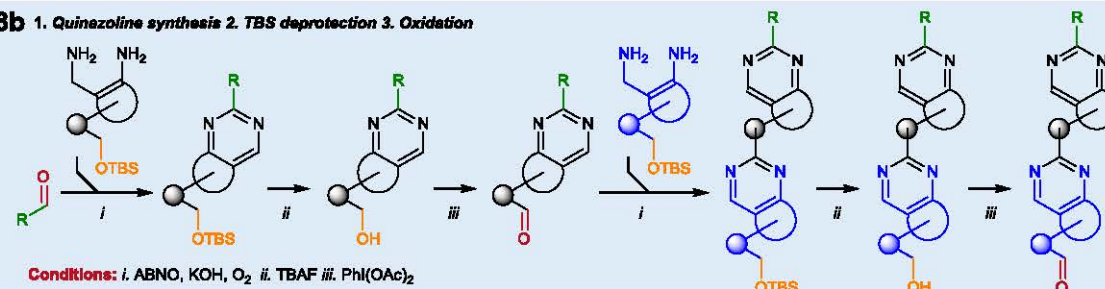


**13** 1. Shapiro reaction 2. Enol ether hydrolysis**14** 1. Opening of epoxide 2. epoxidation**15a** 1. Aldol condensation 2. TBS deprotection 3. Oxidation**15b** 1. Aldol reaction 2. Ozonolysis**16** 1. Wittig olefination 2. TBS deprotection 3. Appel reaction**17** 1. Morita-Baylis-Hillman reaction 2. TBS deprotection 3. Oxidation

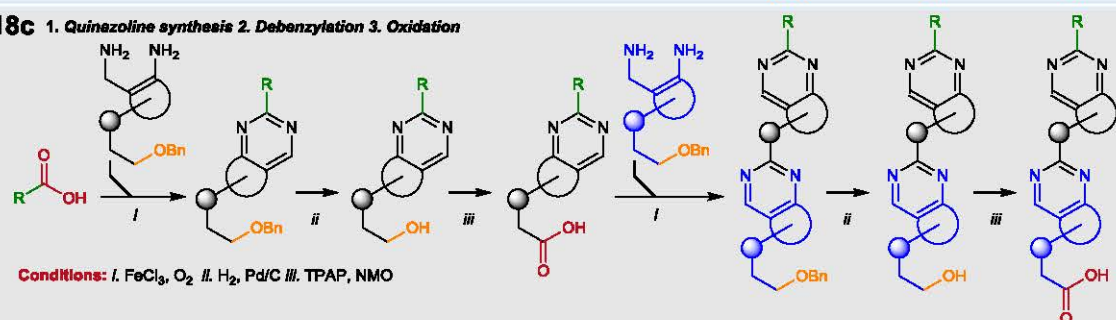
**18a** 1. Quinazoline synthesis 2. Lithiation 3. Hydroxymethylation



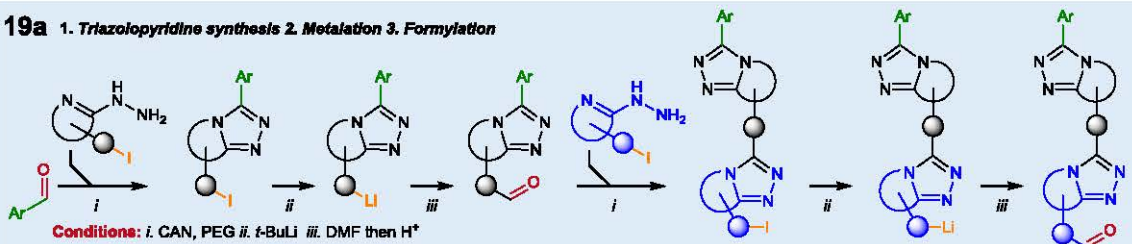
**18b** 1. Quinazoline synthesis 2. TBS deprotection 3. Oxidation



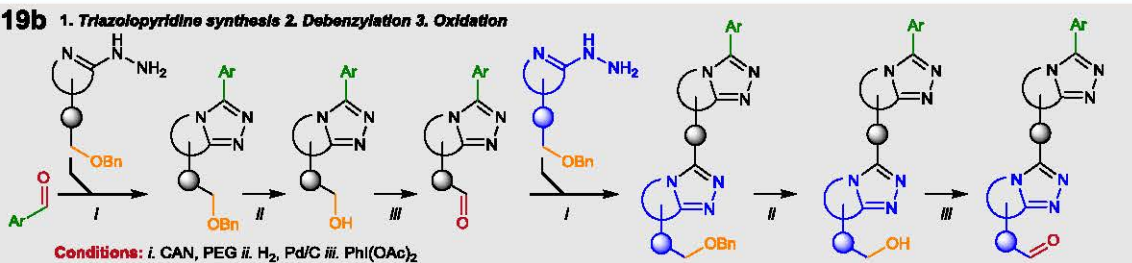
**18c** 1. Quinazoline synthesis 2. Debenzylation 3. Oxidation



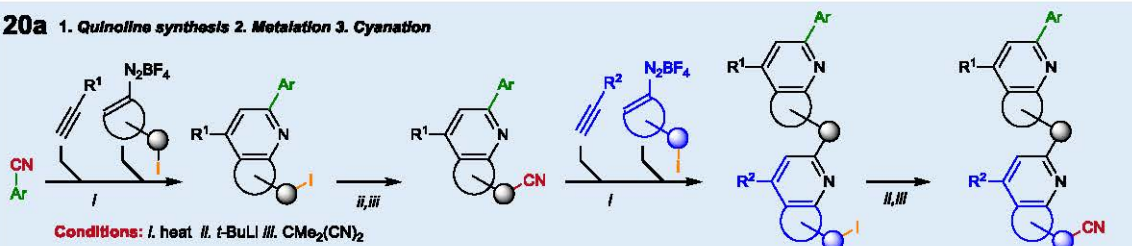
**19a** 1. Triazolopyridine synthesis 2. Metalation 3. Formylation



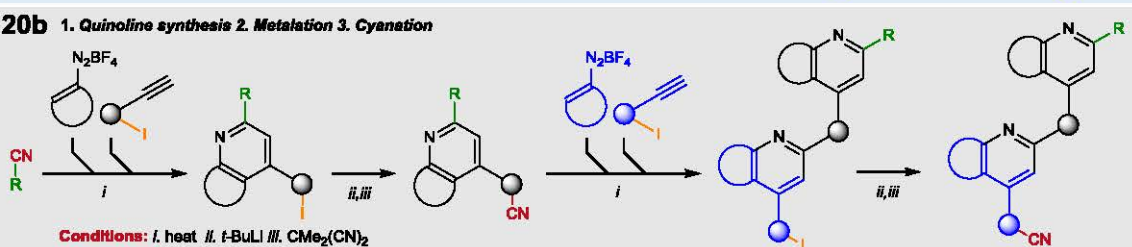
**19b** 1. Triazolopyridine synthesis 2. Debenzylation 3. Oxidation

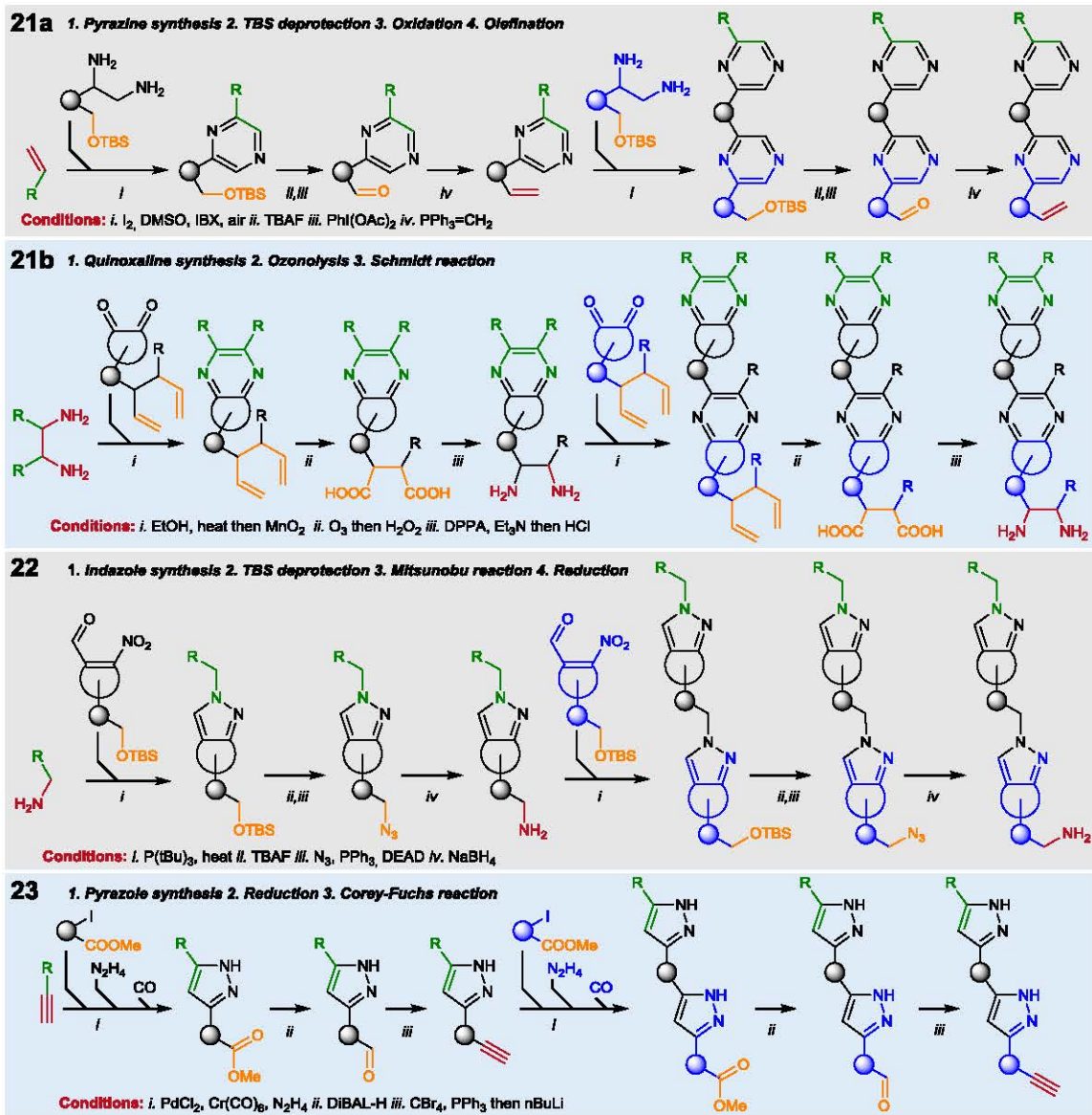


**20a** 1. Quinoline synthesis 2. Metalation 3. Cyanation

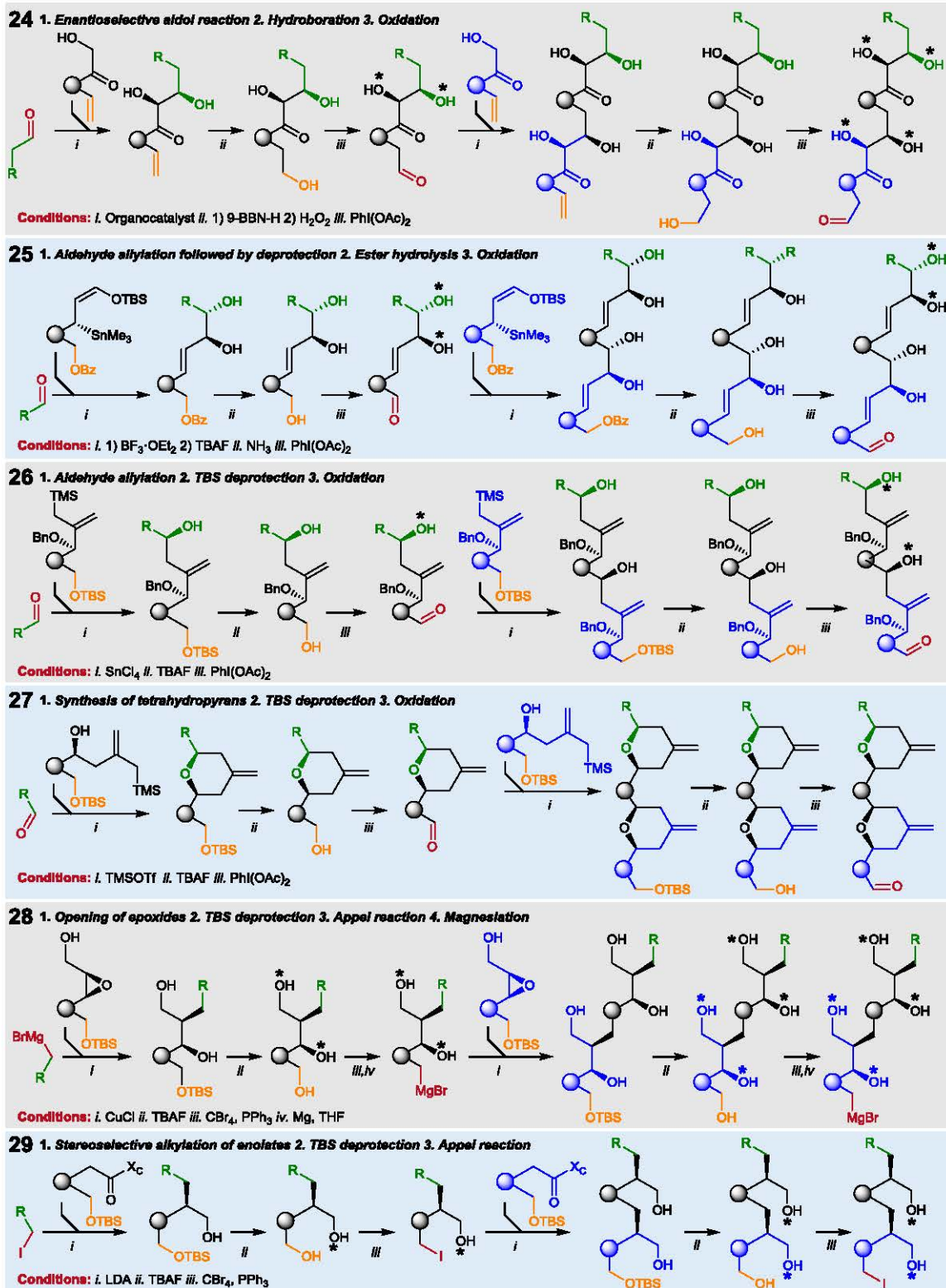


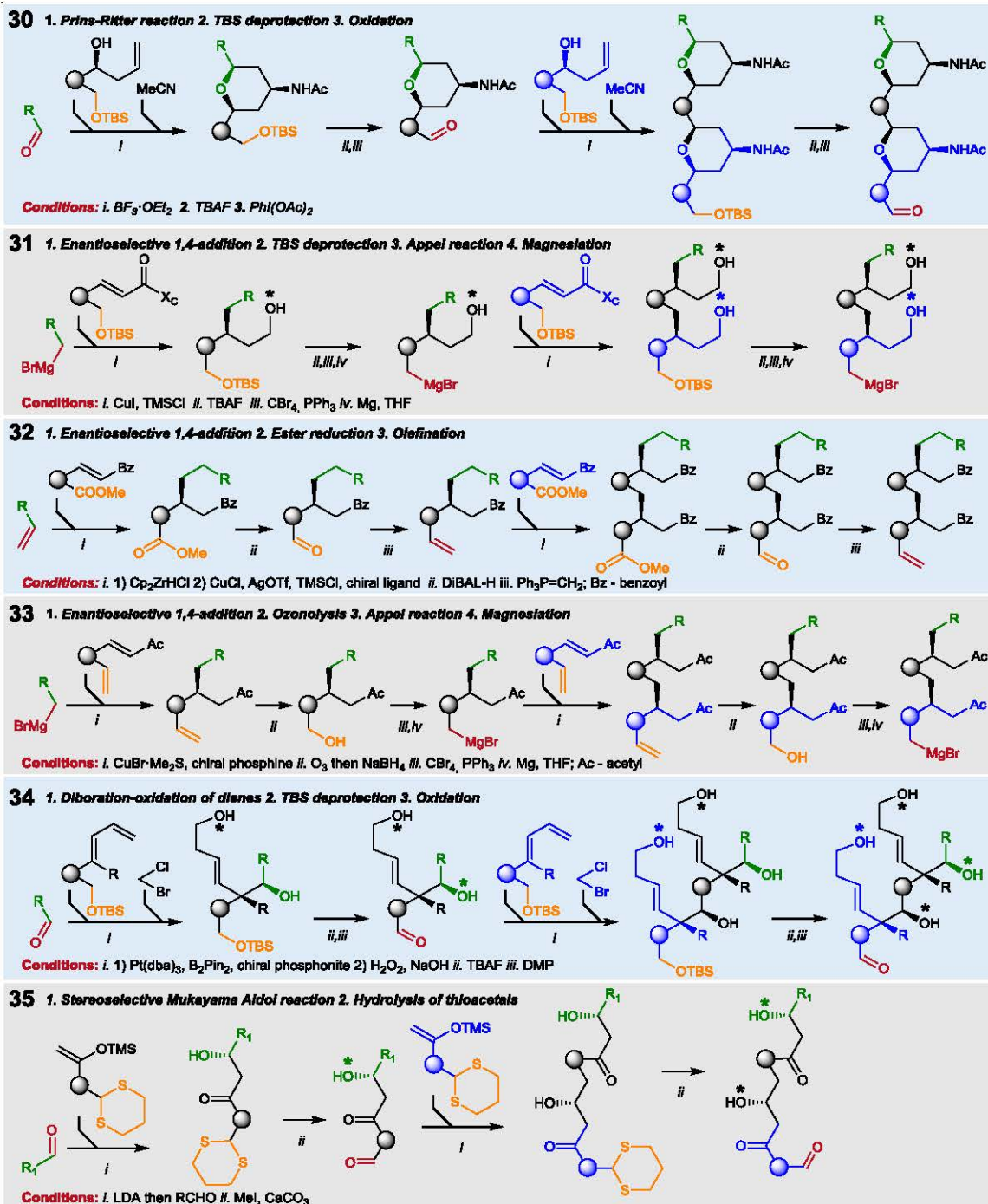
**20b** 1. Quinoline synthesis 2. Metalation 3. Cyanation

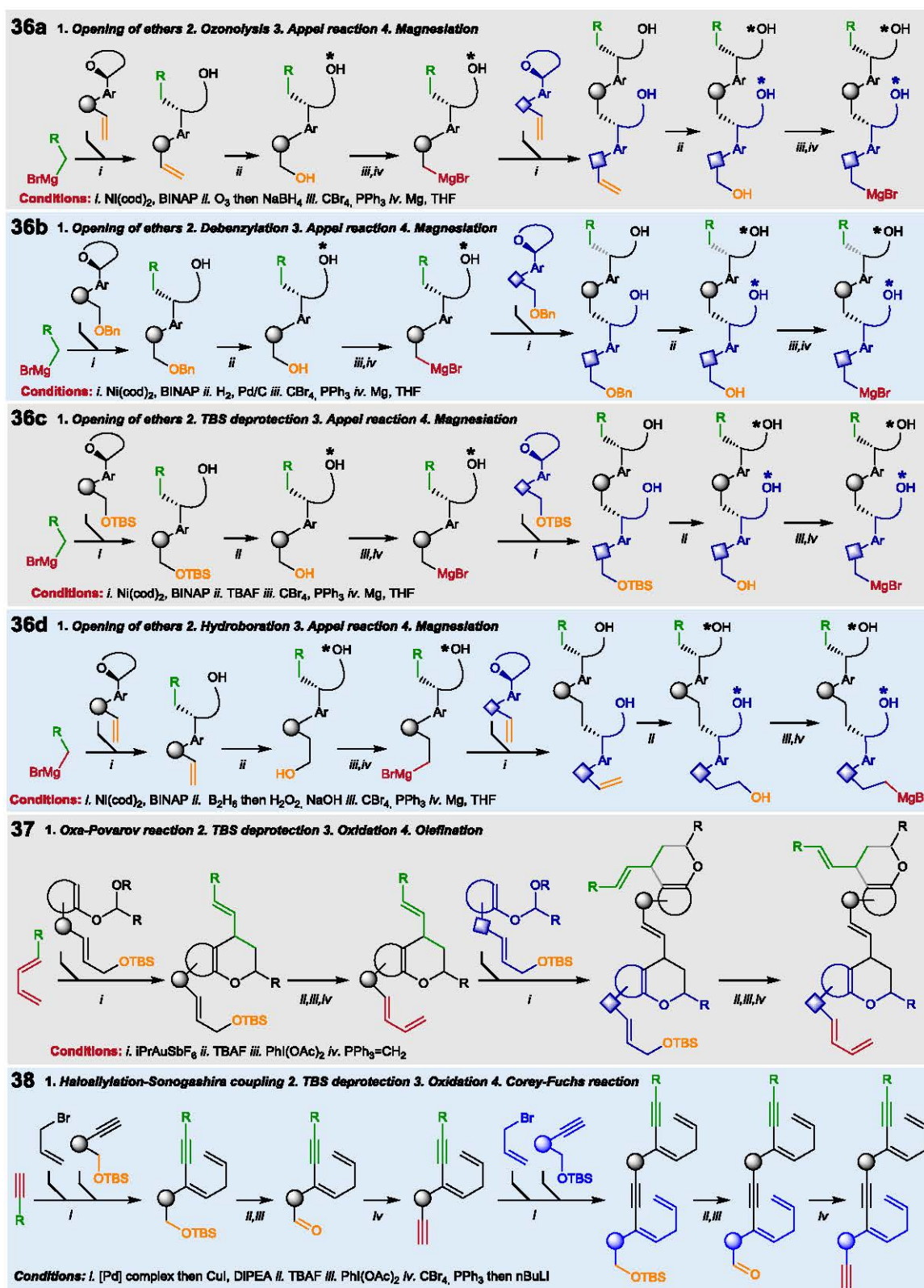






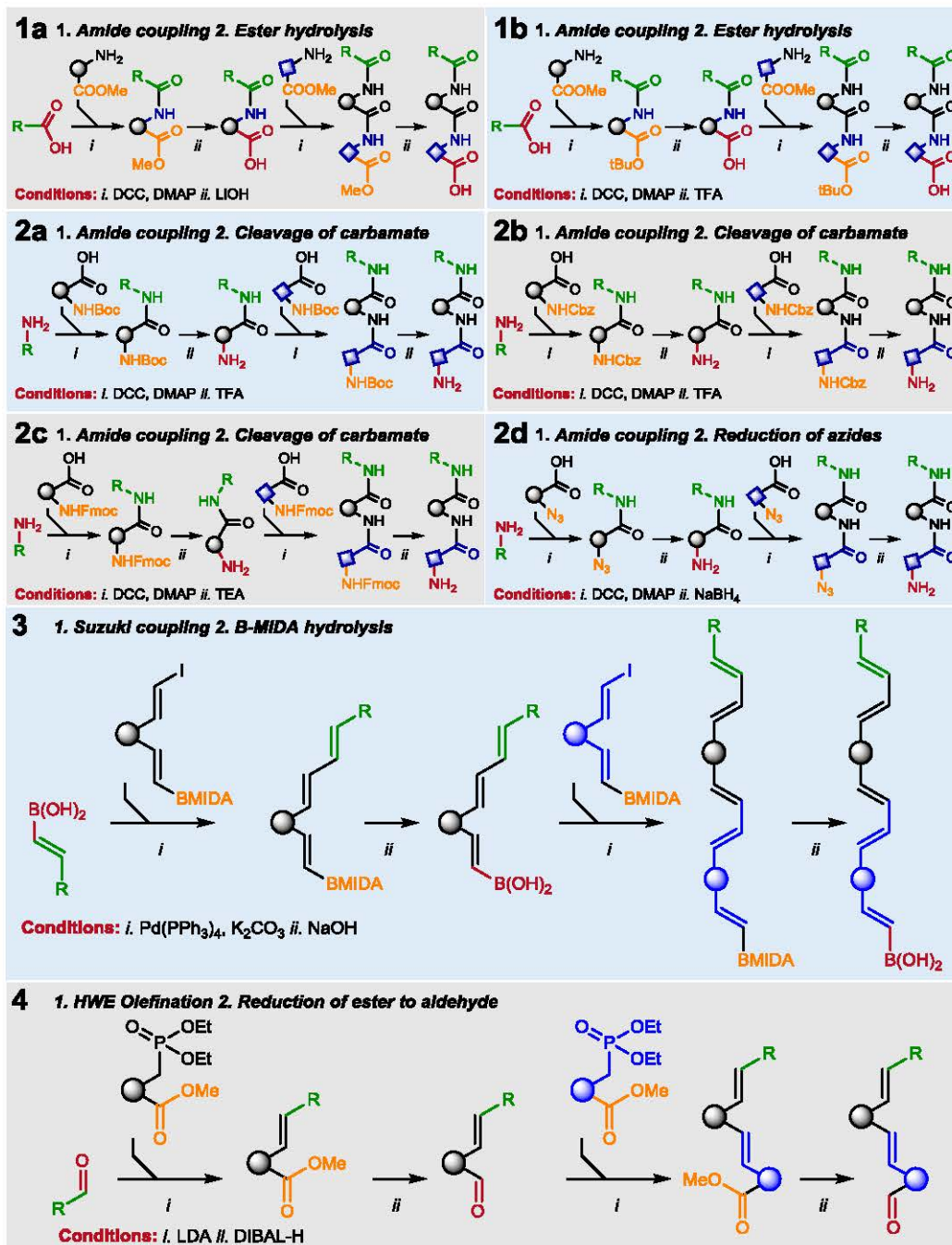


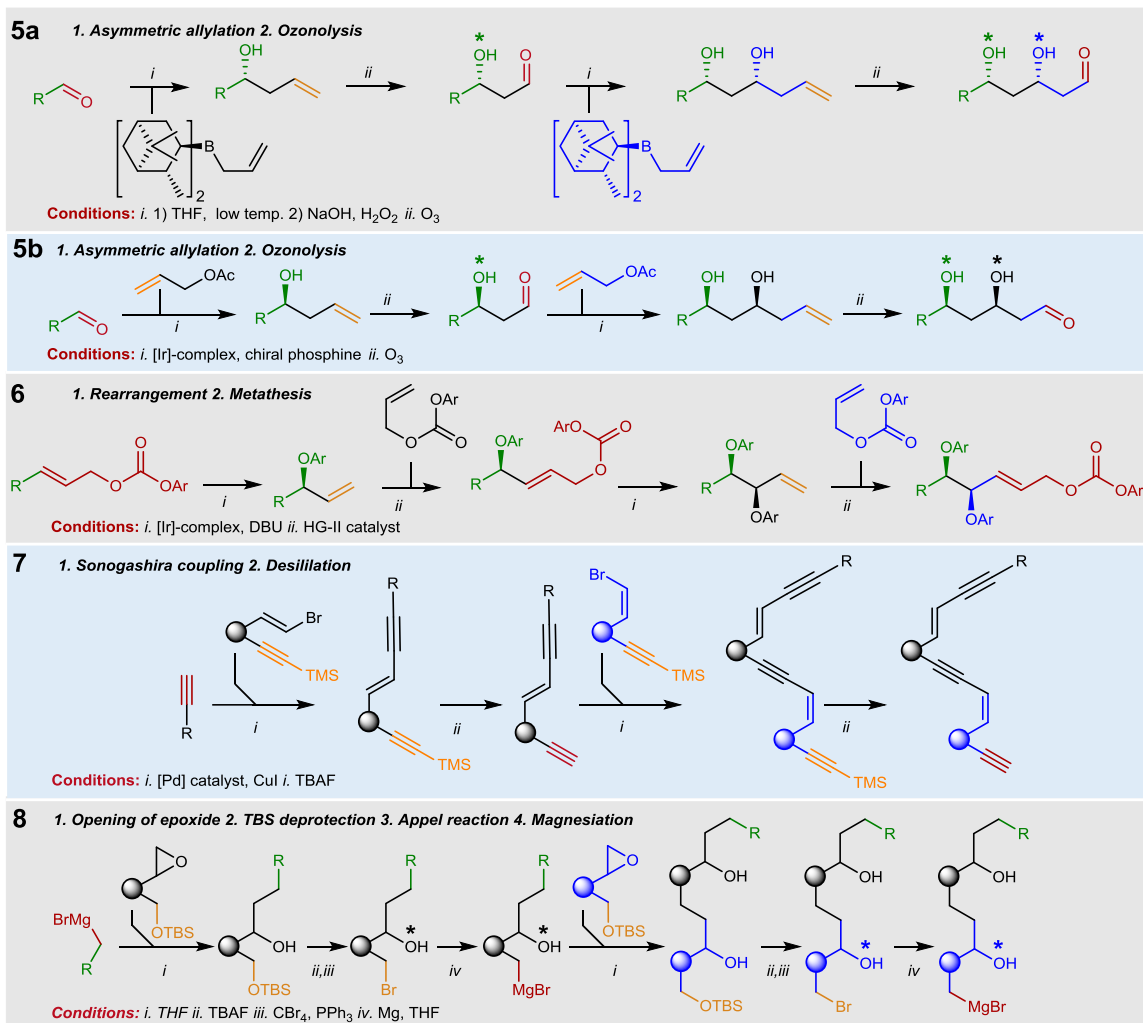




**Figure S3.** Additional examples of new iterative sequences found by the “advanced” algorithm from main-text Figure 2b,c.

Section S4. Examples of previously known iterative sequences rediscovered by the algorithm.



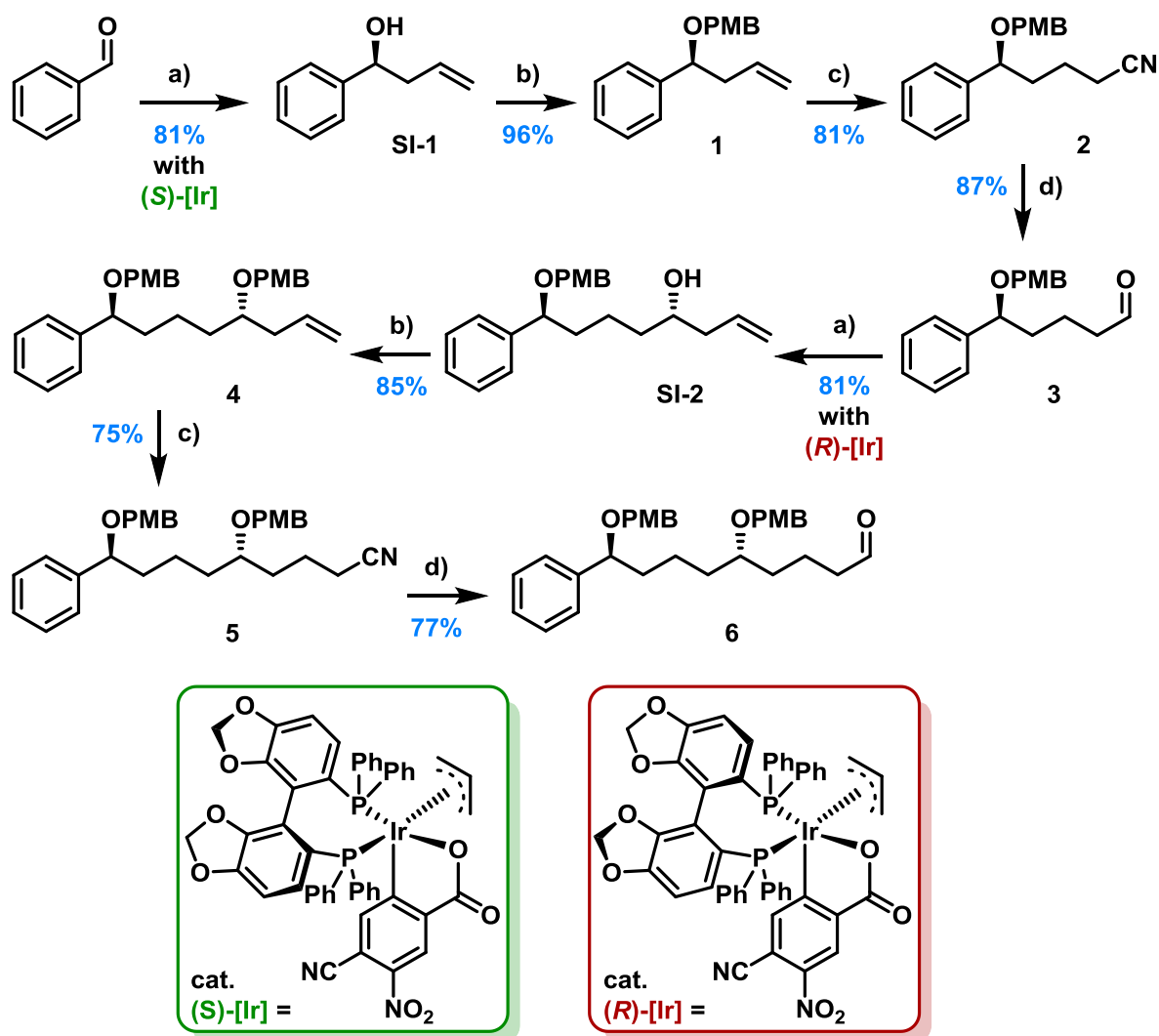


**Figure S4.** Examples of previously known iterative sequences rediscovered by the algorithm described in the main text.

## Section S5. General experimental procedures.

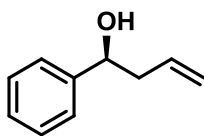
All starting materials and reagents were obtained from commercial sources and, unless otherwise noted, were used as received. All solvents used were freshly distilled prior to use.  $^1\text{H}$  NMR spectra were recorded at 400, 500 or 600 MHz and  $^{13}\text{C}$  NMR spectra were recorded at 100, 125 or 150 MHz with complete proton decoupling. Chemical shifts are given in  $\delta$  relative to the residual signals of the deuterated solvents. High-resolution mass spectra were acquired using electron ionization (EI) or electrospray ionization (ESI) modes with a time-of-flight detector. Infrared (IR) spectra were recorded on a Fourier transform infrared (FT-IR) spectrometer as a thin film on a NaCl plate (film). HPLC analysis were performed on a HPLC system equipped with chiral stationary phase columns with an UV detector. Optical rotations were measured at room temperature with a polarimeter. TLC was performed with aluminum plates coated with 60 F254 silica gel. Plates were visualized with UV light (254 nm) and by treatment with ethanolic *p*-anisaldehyde with sulfuric and glacial acetic acid followed by heating, aqueous cerium(IV) sulfate solution with molybdic and sulfuric acid followed by heating, or aqueous potassium permanganate with sodium hydroxide and potassium carbonate solution followed by heating. Reaction products were purified by flash chromatography using silica gel 60 (230-400 mesh).

## Section S6. Iterative synthesis of 1,5,*n* polyols



Reagents and conditions: (a) allyl acetate, Krische's Ir Catalyst ((*S*)-Ir or (*R*)-Ir), Cs<sub>2</sub>CO<sub>3</sub>, *i*-PrOH, THF, 100°C, 16-18 h; (b) PMBCl, NaH, TBAI, DMF, 0°C to rt, 17-20 h; (c) Zn(CN)<sub>2</sub>, NiCl<sub>2</sub>·6H<sub>2</sub>O, dppp, Zn, DMAP, H<sub>2</sub>O, MeCN, 80°C, 22-24 h; (d) DIBAL-H, DCM, -78°C to rt, 1.5-2 h.

**Scheme S1.** Iterative synthesis of 1,5,*n* polyols *via* asymmetric allylation.



**(S)-1-phenylbut-3-en-1-ol (SI-1).** Prepared *via* adaptation of procedure from <sup>R1</sup>.

An oven-dried vacuum Schlenk tube equipped with a magnetic stir bar was charged with Cs<sub>2</sub>CO<sub>3</sub> (0.078 g, 0.24 mmol, 6 mol%) and Krische Ir Catalyst ((*S*)-SEGPHOS, 4-cyano-3-nitrobenzoate ligated) (0.207 g, 0.20 mmol, 5 mol%). The reaction vessel was placed under an atmosphere of argon, and anhydrous THF (20 mL), benzaldehyde (0.407 mL, 4.00 mmol), allyl acetate (0.863 mL, 8.00 mmol, 2 equiv) and 2-propanol (0.612 mL, 8.00 mmol, 2 equiv) were added by syringe. The reaction vessel was sealed and the reaction mixture was stirred at 100 °C and was monitored by TLC. After 16 h, the mixture was allowed to reach rt and was concentrated *in vacuo*. The residue was purified by flash column chromatography (hexane/ethyl acetate 9:1) to give **SI-1** (0.480 g, 81%, 97% *ee* by HPLC analysis) as a slightly yellow liquid.

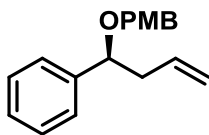
**HPLC** (Chiralcel OD-H, hexane/*i*-PrOH 95:5, flow rate 0.5 mL/min, λ 220 nm): *t*<sub>R(minor)</sub> = 18.2 min (*R*), *t*<sub>R(major)</sub> = 19.4 min (*S*); (lit.<sup>R2</sup> *t*<sub>R</sub> = 20.7 min (*R*), *t*<sub>R</sub> = 22.1 min (*S*));

[α]<sup>27</sup><sub>D</sub> -63.0 (c 1.19, CHCl<sub>3</sub>) (lit.<sup>R3</sup> [α]<sup>25</sup><sub>D</sub> -63.2 (c 0.95, CHCl<sub>3</sub>) (for 97% *ee*));

<sup>1</sup>H NMR (500 MHz, CDCl<sub>3</sub>) δ 7.38 – 7.31 (m, 4H), 7.29 – 7.25 (m, 1H), 5.80 (dddd, *J* = 17.0, 10.3, 7.6, 6.6 Hz, 1H), 5.20 – 5.10 (m, 2H), 4.73 (dd, *J* = 7.6, 5.2 Hz, 1H), 2.58 – 2.44 (m, 2H), 2.00 (brs, 1H);

<sup>13</sup>C NMR (125 MHz, CDCl<sub>3</sub>) δ 144.0, 134.6, 128.5, 127.7, 125.9, 118.5, 73.4, 44.0;

**IR** (film, DCM) 3541, 3378, 3075, 3030, 2979, 2930, 1641, 1493 cm<sup>-1</sup>.



**(S)-1-methoxy-4-(((1-phenylbut-3-en-1-yl)oxy)methyl)benzene (1).**

NaH (60 % dispersion in mineral oil) (0.132 g, 3.31 mmol, 2 equiv) was added to a flask containing **SI-1** (0.245 g, 1.65 mmol) and TBAI (0.061 g, 0.17 mmol, 10 mol%) in anhydrous DMF (3.30 mL) cooled to 0 °C. Reaction mixture was flushed with argon and stirred for 30 min at 0 °C. Then, 4-methoxybenzyl chloride (0.448 mL, 3.31 mmol, 2 equiv) was added dropwise



and the mixture was stirred at rt. The reaction was monitored by TLC and after 18 h the mixture was quenched by addition of sat.  $\text{NH}_4\text{Cl}$  solution. The mixture was extracted with ethyl acetate. Combined organic phases were dried over  $\text{Na}_2\text{SO}_4$  and concentrated *in vacuo*. Crude product was purified by flash column chromatography (hexane/ethyl acetate 98:2) to give **1** (0.425 g, 96%) as a slightly yellow liquid.

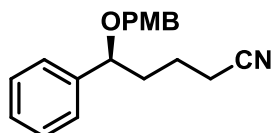
$[\alpha]^{28}_{\text{D}} -70.7$  (c 1.09,  $\text{C}_6\text{H}_6$ ) (lit.<sup>R4</sup>  $[\alpha]^{20}_{\text{D}} = +67.3$  (c 0.95,  $\text{C}_6\text{H}_6$ ) (for (*R*)-enantiomer);

$^1\text{H NMR}$  (400 MHz,  $\text{CDCl}_3$ )  $\delta$  7.40 – 7.28 (m, 5H), 7.25 – 7.21 (m, 2H), 6.90 – 6.85 (m, 2H), 5.78 (ddt,  $J = 17.1, 10.2, 6.9$  Hz, 1H), 5.09 – 4.96 (m, 2H), 4.41 (d,  $J = 11.5$  Hz, 1H), 4.35 (dd,  $J = 7.6, 5.9$  Hz, 1H), 4.22 (d,  $J = 11.4$  Hz, 1H), 3.81 (s, 3H), 2.68 – 2.57 (m, 1H), 2.43 (dddt,  $J = 14.2, 7.2, 5.9, 1.3$  Hz, 1H);

$^{13}\text{C NMR}$  (100 MHz,  $\text{CDCl}_3$ )  $\delta$  159.3, 142.2, 135.1, 130.8, 129.5, 128.5, 127.8, 127.1, 116.9, 113.9, 81.0, 70.2, 55.4, 42.8;

**HRMS** (ESI)  $m/z$ :  $[\text{M}+\text{Na}]^+$  Calcd for  $\text{C}_{18}\text{H}_{20}\text{O}_2\text{Na}$  291.1361; Found 291.1356;

**IR** (film) 3067, 3030, 3003, 2934, 2906, 2861, 2837, 1612, 1513, 1455  $\text{cm}^{-1}$ .



**(S)-5-((4-methoxybenzyl)oxy)-5-phenylpentanenitrile (2)**. Prepared *via* adaptation of procedure from <sup>R5</sup>.

An oven-dried vacuum Schlenk tube equipped with a magnetic stir bar was placed under an atmosphere of argon and charged with  $\text{NiCl}_2 \cdot 6\text{H}_2\text{O}$  (0.018 g, 0.08 mmol, 5 mol%), dppp (0.037 g, 0.09 mmol, 6 mol%), zinc powder (0.098 g, 1.50 mmol, 1 equiv),  $\text{Zn}(\text{CN})_2$  (0.106 g, 0.90 mmol, 0.6 equiv), DMAP (0.183 g, 1.50 mmol, 1 equiv), anhydrous  $\text{CH}_3\text{CN}$  (7.50 mL), compound **1** (0.403 g, 1.50 mmol) and water (0.054 mL, 3.00 mmol, 2 equiv). The reaction vessel was sealed and the reaction mixture was stirred at 80 °C and was monitored by TLC. After 24 h, the mixture was allowed to reach rt. Next, it was filtered through a short pad of silica gel and washed with ethyl acetate. The solvent was concentrated *in vacuo* and the residue was purified by flash column chromatography (hexane/ethyl acetate 85:15) to give **2** (0.357 g, 81%) as a colorless oil.

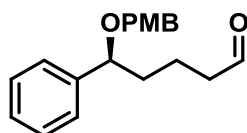
$[\alpha]^{19}_{\text{D}}$  -91.0 (c 1.29,  $\text{CHCl}_3$ );

$^1\text{H NMR}$  (400 MHz,  $\text{CDCl}_3$ )  $\delta$  7.43 – 7.36 (m, 2H), 7.35 – 7.30 (m, 3H), 7.24 – 7.19 (m, 2H), 6.92 – 6.86 (m, 2H), 4.42 (d,  $J = 11.4$  Hz, 1H), 4.32 (dd,  $J = 8.1, 4.3$  Hz, 1H), 4.18 (d,  $J = 11.4$  Hz, 1H), 3.82 (s, 3H), 2.37 – 2.22 (m, 2H), 1.99 – 1.88 (m, 1H), 1.87 – 1.75 (m, 2H), 1.73 – 1.61 (m, 1H);

$^{13}\text{C NMR}$  (100 MHz,  $\text{CDCl}_3$ )  $\delta$  159.4, 141.9, 130.4, 129.5, 128.7, 128.0, 126.7, 119.7, 114.0, 79.9, 70.2, 55.4, 37.2, 22.1, 17.1;

**HRMS** (ESI)  $m/z$ :  $[\text{M}+\text{Na}]^+$  Calcd for  $\text{C}_{19}\text{H}_{21}\text{NO}_2\text{Na}$  318.1470; Found 318.1475;

**IR** (film, DCM) 3061, 3030, 3003, 2935, 2865, 2838, 2245, 1612, 1585, 1513, 1454  $\text{cm}^{-1}$ .



**(S)-5-((4-methoxybenzyl)oxy)-5-phenylpentanal (3).**

To a stirred solution of **2** (0.325 g, 1.10 mmol) in anhydrous DCM (11.0 mL) cooled to  $-78$  °C was added DIBAL-H solution (1.0 M in DCM) (1.320 mL, 1.32 mmol, 1.2 equiv) in a drop-wise manner. The mixture was stirred at  $-78$  °C and was monitored by TLC. After 1.5 h, the reaction was quenched by addition of sat.  $\text{Na}_2\text{SO}_4$  solution (0.143 mL) and the mixture was allowed to slowly reach rt. The mixture was diluted with DCM and sat. potassium sodium tartrate solution was added. After the layers were separated, the aqueous phase was extracted with DCM. Combined organic phases were washed with 50% potassium sodium tartrate solution and then brine, dried over  $\text{Na}_2\text{SO}_4$  and concentrated *in vacuo*. Crude product was purified by flash column chromatography (hexane/ethyl acetate 85:15) to give **3** (0.284 g, 87%) as a colorless oil.

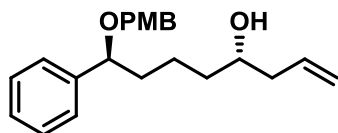
$[\alpha]^{19}_{\text{D}}$  -79.9 (c 1.42,  $\text{CHCl}_3$ );

$^1\text{H NMR}$  (400 MHz,  $\text{CDCl}_3$ )  $\delta$  9.71 (t,  $J = 1.7$  Hz, 1H), 7.41 – 7.28 (m, 5H), 7.24 – 7.18 (m, 2H), 6.92 – 6.85 (m, 2H), 4.40 (d,  $J = 11.3$  Hz, 1H), 4.30 (dd,  $J = 7.8, 5.0$  Hz, 1H), 4.18 (d,  $J = 11.3$  Hz, 1H), 3.81 (s, 3H), 2.48 – 2.31 (m, 2H), 1.92 – 1.75 (m, 2H), 1.74 – 1.50 (m, 2H);

$^{13}\text{C NMR}$  (100 MHz,  $\text{CDCl}_3$ )  $\delta$  202.5, 159.3, 142.4, 130.7, 129.5, 128.6, 127.8, 126.9, 113.9, 80.7, 70.2, 55.4, 43.8, 37.8, 18.8;

**HRMS** (ESI)  $m/z$ :  $[M+Na]^+$  Calcd for  $C_{19}H_{22}O_3Na$  321.1467; Found 321.1461;

**IR** (film, DCM) 3061, 3029, 3002, 2935, 2863, 2837, 2721, 1722, 1612, 1585, 1513  $cm^{-1}$ .



**(4S,8S)-8-((4-methoxybenzyl)oxy)-8-phenyloct-1-en-4-ol (SI-2)**. Prepared *via* adaptation of procedure from <sup>R1</sup>.

An oven-dried vacuum Schlenk tube equipped with a magnetic stir bar was charged with  $CS_2CO_3$  (0.030 g, 0.09 mmol, 12 mol%) and Krische Ir Catalyst ((*R*)-SEGPHOS, 4-cyano-3-nitrobenzoate ligated) (0.081 g, 0.08 mmol, 10 mol%). The reaction vessel was placed under an atmosphere of argon, and anhydrous THF (3.90 mL), **3** (0.233 g, 0.78 mmol), allyl acetate (0.168 mL, 1.56 mmol, 2 equiv) and 2-propanol (0.119 mL, 1.56 mmol, 2 equiv) were added by syringe. The reaction vessel was sealed and the reaction mixture was stirred at 100 °C and was monitored by TLC. After 18 h, the mixture was allowed to reach rt and was concentrated *in vacuo*. The residue was purified by flash column chromatography (hexane/ethyl acetate 4:1) to give **SI-2** (0.215 g, 81%) as a colorless oil.

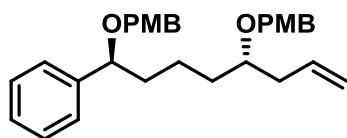
$[\alpha]^{19}_D$  -68.4 (c 1.08,  $CHCl_3$ );

**<sup>1</sup>H NMR** (500 MHz,  $CDCl_3$ )  $\delta$  7.39 – 7.27 (m, 5H), 7.24 – 7.20 (m, 2H), 6.89 – 6.84 (m, 2H), 5.85 – 5.73 (m, 1H), 5.15 – 5.05 (m, 2H), 4.39 (d,  $J = 11.4$  Hz, 1H), 4.28 (dd,  $J = 7.9, 5.4$  Hz, 1H), 4.17 (d,  $J = 11.4$  Hz, 1H), 3.80 (s, 3H), 3.64 – 3.58 (m, 1H), 2.28 – 2.21 (m, 1H), 2.14 – 2.04 (m, 1H), 1.93 – 1.82 (m, 1H), 1.69 – 1.58 (m, 2H), 1.53 – 1.37 (m, 3H);

**<sup>13</sup>C NMR** (125 MHz,  $CDCl_3$ )  $\delta$  159.3, 142.8, 135.0, 130.8, 129.6, 128.6, 127.7, 126.9, 118.2, 113.9, 81.2, 70.6, 70.2, 55.4, 42.0, 38.4, 36.7, 22.3;

**HRMS** (ESI)  $m/z$ :  $[M+Na]^+$  Calcd for  $C_{22}H_{28}O_3Na$  363.1936; Found 363.1944;

**IR** (film, DCM) 3436, 3066, 3029, 3001, 2933, 2862, 1640, 1612, 1586, 1513  $cm^{-1}$ .



**4,4'-((((1*S*,5*S*)-1-phenyloct-7-ene-1,5-diyl)bis(oxy))bis(methylene))bis(methoxybenzene) (4).**

NaH (60 % dispersion in mineral oil) (0.040 g, 1.00 mmol, 2 equiv) was added to a flask containing **SI-2** (0.170 g, 0.50 mmol) and TBAI (0.018 g, 0.05 mmol, 10 mol%) in anhydrous DMF (1.00 mL) cooled to 0 °C. Reaction mixture was flushed with argon and stirred for 30 min at 0 °C. Then, 4-methoxybenzyl chloride (0.136 mL, 1.00 mmol, 2 equiv) was added dropwise and the mixture was stirred at rt. The reaction was monitored by TLC and after 17 h the mixture was quenched by addition of sat. NH<sub>4</sub>Cl solution. The mixture was extracted with ethyl acetate. Combined organic phases were dried over Na<sub>2</sub>SO<sub>4</sub> and concentrated *in vacuo*. Crude product was purified by flash column chromatography (hexane/ethyl acetate 95:5) to give **4** (0.195 g, 85%) as a colorless oil.

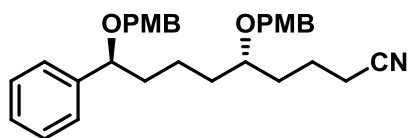
$[\alpha]_D^{19}$  -54.0 (c 1.17, CHCl<sub>3</sub>);

<sup>1</sup>H NMR (400 MHz, CDCl<sub>3</sub>) δ 7.40 – 7.27 (m, 5H), 7.24 – 7.19 (m, 4H), 6.90 – 6.82 (m, 4H), 5.81 (ddt, *J* = 17.3, 10.3, 7.1 Hz, 1H), 5.09 – 5.04 (m, 1H), 5.04 – 5.01 (m, 1H), 4.47 (d, *J* = 11.2 Hz, 1H), 4.40 (d, *J* = 4.2 Hz, 1H), 4.37 (d, *J* = 4.2 Hz, 1H), 4.27 (dd, *J* = 7.8, 5.5 Hz, 1H), 4.18 (d, *J* = 11.2 Hz, 1H), 3.80 (s, 6H), 3.42 – 3.33 (m, 1H), 2.34 – 2.19 (m, 2H), 1.90 – 1.79 (m, 1H), 1.68 – 1.55 (m, 1H), 1.54 – 1.35 (m, 4H);

<sup>13</sup>C NMR (100 MHz, CDCl<sub>3</sub>) δ 159.2, 159.2, 142.9, 135.2, 131.2, 131.0, 129.5, 129.4, 128.5, 127.6, 127.0, 116.9, 113.9, 113.9, 81.3, 78.2, 70.7, 70.2, 55.4, 38.5, 38.5, 33.9, 22.0;

HRMS (ESI) *m/z*: [M+Na]<sup>+</sup> Calcd for C<sub>30</sub>H<sub>36</sub>O<sub>4</sub>Na 483.2511; Found 483.2495;

IR (film, DCM) 3064, 3030, 3001, 2935, 2861, 2837, 1639, 1612, 1586, 1513 cm<sup>-1</sup>.



**(5R,9S)-5,9-bis((4-methoxybenzyl)oxy)-9-phenylnonanenitrile (5).** Prepared *via* adaptation of procedure from <sup>R5</sup>.

An oven-dried vacuum Schlenk tube equipped with a magnetic stir bar was placed under an atmosphere of argon and charged with NiCl<sub>2</sub> · 6H<sub>2</sub>O (0.004 g, 0.02 mmol, 5 mol%), dppp (0.009 g, 0.02 mmol, 6 mol%), zinc powder (0.023 g, 0.35 mmol, 1 equiv), Zn(CN)<sub>2</sub> (0.025 g, 0.21 mmol, 0.6 equiv), DMAP (0.043 g, 0.35 mmol, 1 equiv), anhydrous CH<sub>3</sub>CN (1.75 mL), compound **4** (0.161 g, 0.35 mmol) and water (0.013 mL, 0.70 mmol, 2 equiv). The reaction vessel was sealed and the reaction mixture was stirred at 80 °C and was monitored by TLC. After 22 h, the mixture was allowed to reach rt. Next, it was filtered through a short pad of silica gel and washed with ethyl acetate. The solvent was concentrated *in vacuo* and the residue was purified by flash column chromatography (hexane/ethyl acetate 85:15) to give **5** (0.128 g, 75%) as a colorless oil.

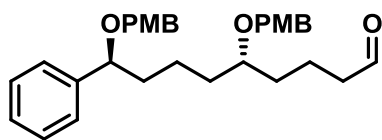
[α]<sub>D</sub><sup>20</sup> -26.4 (c 0.98, CHCl<sub>3</sub>);

<sup>1</sup>H NMR (400 MHz, CDCl<sub>3</sub>) δ 7.41 – 7.29 (m, 5H), 7.24 – 7.18 (m, 4H), 6.89 – 6.84 (m, 4H), 4.44 – 4.37 (m, 2H), 4.34 (d, *J* = 11.2 Hz, 1H), 4.28 (dd, *J* = 7.7, 5.5 Hz, 1H), 4.18 (d, *J* = 11.2 Hz, 1H), 3.80 (s, 3H), 3.80 (s, 3H), 3.38 – 3.31 (m, 1H), 2.33 – 2.22 (m, 2H), 1.86 (dddd, *J* = 13.4, 9.6, 7.7, 4.7 Hz, 1H), 1.73 – 1.40 (m, 8H), 1.39 – 1.30 (m, 1H);

<sup>13</sup>C NMR (100 MHz, CDCl<sub>3</sub>) δ 159.3, 159.3, 142.8, 130.8, 130.8, 129.5, 129.5, 128.6, 127.7, 126.9, 119.8, 114.0, 113.9, 81.1, 77.4, 70.7, 70.2, 55.4, 38.5, 33.6, 32.8, 21.8, 21.5, 17.3;

HRMS (ESI) *m/z*: [M+Na]<sup>+</sup> Calcd for C<sub>31</sub>H<sub>37</sub>NO<sub>4</sub>Na 510.2620; Found 510.2613;

IR (film, DCM) 3060, 3030, 3001, 2936, 2863, 2837, 2244, 1612, 1585, 1513, 1455 cm<sup>-1</sup>.



**(5S,9S)-5,9-bis((4-methoxybenzyl)oxy)-9-phenylnonanal (6)**

To a stirred solution of **5** (0.029 g, 0.06 mmol) in anhydrous DCM (0.60 mL) cooled to  $-78\text{ }^{\circ}\text{C}$  was added DIBAL-H solution (1.0 M in DCM) (0.072 mL, 0.07 mmol, 1.2 equiv) in a drop-wise manner. The mixture was stirred at  $-78\text{ }^{\circ}\text{C}$  and was monitored by TLC. After 2 h, the reaction was quenched by addition of sat.  $\text{Na}_2\text{SO}_4$  solution (0.008 mL) and the mixture was allowed to slowly reach rt. The mixture was diluted with DCM and sat. potassium sodium tartrate solution was added. After the layers were separated, the aqueous phase was extracted with DCM. Combined organic phases were washed with 50% potassium sodium tartrate solution and then brine, dried over  $\text{Na}_2\text{SO}_4$  and concentrated *in vacuo*. Crude product was purified by flash column chromatography (hexane/ethyl acetate 4:1) to give **6** (0.023 g, 77%) as a slightly yellow oil.

$[\alpha]_{\text{D}}^{20}$   $-33.9$  (c 1.59,  $\text{CHCl}_3$ );

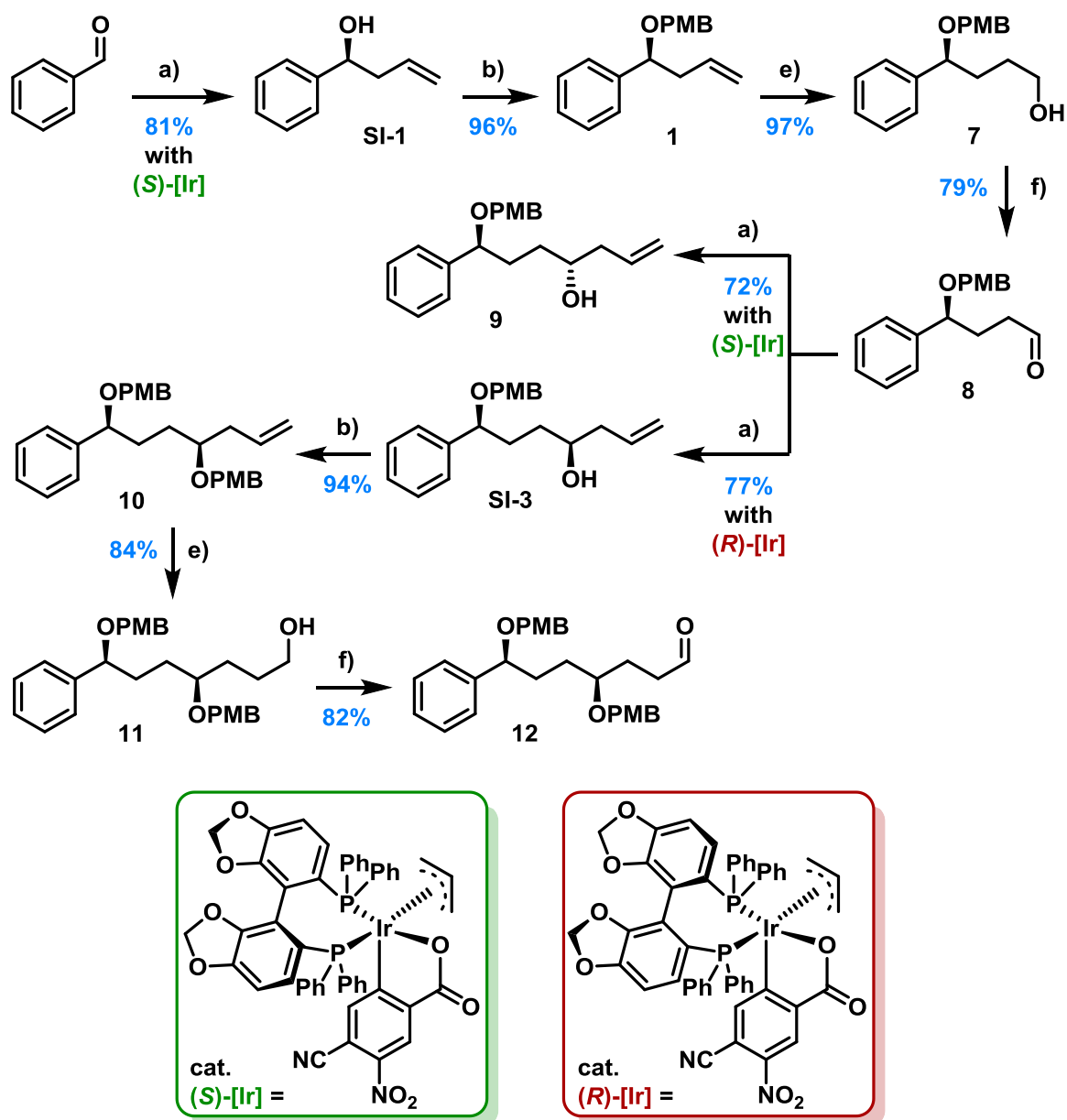
$^1\text{H NMR}$  (400 MHz,  $\text{CDCl}_3$ )  $\delta$  9.72 (t,  $J = 1.8$  Hz, 1H), 7.40 – 7.29 (m, 5H), 7.24 – 7.17 (m, 4H), 6.89 – 6.79 (m, 4H), 4.42 – 4.34 (m, 3H), 4.27 (dd,  $J = 7.7, 5.6$  Hz, 1H), 4.17 (d,  $J = 11.4$  Hz, 1H), 3.80 (s, 3H), 3.79 (s, 3H), 3.32 (p,  $J = 5.6$  Hz, 1H), 2.42 – 2.35 (m, 2H), 1.91 – 1.80 (m, 1H), 1.75 – 1.58 (m, 4H), 1.53 – 1.41 (m, 4H), 1.38 – 1.30 (m, 1H);

$^{13}\text{C NMR}$  (100 MHz,  $\text{CDCl}_3$ )  $\delta$  202.6, 159.3, 142.9, 131.1, 130.9, 129.5, 129.5, 128.6, 127.6, 127.0, 113.9, 81.2, 78.2, 70.7, 70.2, 55.4, 55.4, 44.0, 38.6, 33.7, 33.4, 21.9, 18.2;

**HRMS** (ESI)  $m/z$ :  $[\text{M}+\text{Na}]^+$  Calcd for  $\text{C}_{31}\text{H}_{38}\text{O}_5\text{Na}$  513.2617; Found 513.2627;

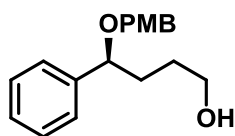
**IR** (film, DCM) 3061, 3030, 3000, 2935, 2862, 2720, 1723, 1612, 1586, 1513  $\text{cm}^{-1}$ .

## Section S7. Iterative synthesis of 1,4,n polyols



Reagents and conditions: (a) allyl acetate, Krische's Ir Catalyst ((S)-Ir or (R)-Ir), Cs<sub>2</sub>CO<sub>3</sub>, *i*-PrOH, THF, 100°C, 16-18 h; (b) PMBCl, NaH, TBAI, DMF, 0°C to rt, 17-20 h; (e) 1) 9-BBN, THF, reflux, 1-2 h, 2) NaOH (aq) (0.5 M), H<sub>2</sub>O<sub>2</sub> (30%), EtOH, 50°C, 1 h; (f) PCC, MS 4 Å, DCM, rt, 1.5 h.

**Scheme S2.** Iterative synthesis of 1,4,n polyols *via* asymmetric allylation.



**(S)-4-((4-methoxybenzyl)oxy)-4-phenylbutan-1-ol (7).**

To a solution of **1** (0.309 g, 1.15 mmol) in anhydrous THF (4.60 mL) was added drop-wise 9-borabicyclo[3.3.1]nonane solution (0.5 M in THF) (6.90 mL, 3.45 mmol, 3 equiv) over 10 min and the mixture was heated under reflux for 2 h. After cooling, the reaction mixture was treated with ethanol (3.00 mL), 2 M NaOH solution (1.50 mL) and H<sub>2</sub>O<sub>2</sub> solution (30% (w/w) in H<sub>2</sub>O) (1.50 mL) and the mixture was stirred at 50 °C for 1 h. The reaction was monitored by TLC and upon completion sat. K<sub>2</sub>CO<sub>3</sub> solution and Et<sub>2</sub>O were added. The aqueous phase was extracted with Et<sub>2</sub>O, combined organic phases were dried over Na<sub>2</sub>SO<sub>4</sub> and concentrated *in vacuo*. Crude product was purified by flash column chromatography (hexane/ethyl acetate 7:3) to give **7** (0.319 g, 97%) as a slightly yellow oil.

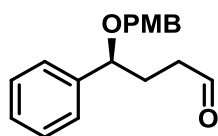
[ $\alpha$ ]<sup>23</sup><sub>D</sub> -82.2 (c 1.08, CHCl<sub>3</sub>);

<sup>1</sup>H NMR (400 MHz, CDCl<sub>3</sub>)  $\delta$  7.40 – 7.28 (m, 5H), 7.24 – 7.19 (m, 2H), 6.90 – 6.85 (m, 2H), 4.40 (d, *J* = 11.3 Hz, 1H), 4.33 (dd, *J* = 8.1, 4.7 Hz, 1H), 4.20 (d, *J* = 11.3 Hz, 1H), 3.80 (s, 3H), 3.61 (t, *J* = 6.2 Hz, 2H), 2.03 – 1.82 (m, 2H), 1.80 – 1.66 (m, 2H), 1.64 – 1.55 (m, 1H);

<sup>13</sup>C NMR (100 MHz, CDCl<sub>3</sub>)  $\delta$  159.4, 142.5, 130.4, 129.7, 128.6, 127.7, 126.9, 114.0, 81.2, 70.3, 62.9, 55.4, 35.2, 29.5;

HRMS (ESI) *m/z*: [M+Na]<sup>+</sup> Calcd for C<sub>18</sub>H<sub>22</sub>O<sub>3</sub>Na 309.1467; Found 309.1471;

IR (film, DCM) 3399, 3061, 3030, 3001, 2935, 2865, 1612, 1585, 1513 cm<sup>-1</sup>.



**(S)-4-((4-methoxybenzyl)oxy)-4-phenylbutanal (8).**

To a solution of **7** (0.315 g, 1.10 mmol) in anhydrous DCM (5.50 mL) were added molecular sieves 4 Å (beads) (0.550 g) and pyridinium chlorochromate (0.474 g, 2.20 mmol, 2 equiv) and the mixture was stirred at rt. The reaction was monitored by TLC and after 1.5 h Et<sub>2</sub>O (5.50 mL) was added. After 30 min of intensive stirring, the mixture was filtered through a Celite



plug, which was thoroughly washed with Et<sub>2</sub>O. The filtrate was washed with water and brine, dried over Na<sub>2</sub>SO<sub>4</sub> and concentrated *in vacuo*. Crude product was purified by flash column chromatography (hexane/ethyl acetate 9:1) to give **8** (0.246 g, 79%) as a colorless oil.

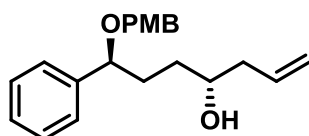
$[\alpha]^{20}_{\text{D}}$  -91.1 (c 1.31, CHCl<sub>3</sub>);

<sup>1</sup>H NMR (400 MHz, CDCl<sub>3</sub>) δ 9.73 (t, *J* = 1.7 Hz, 1H), 7.42 – 7.29 (m, 5H), 7.23 – 7.18 (m, 2H), 6.92 – 6.81 (m, 2H), 4.40 (d, *J* = 11.3 Hz, 1H), 4.35 (dd, *J* = 8.3, 4.8 Hz, 1H), 4.18 (d, *J* = 11.3 Hz, 1H), 3.81 (s, 3H), 2.58 – 2.42 (m, 2H), 2.13 (ddt, *J* = 14.2, 8.3, 7.1 Hz, 1H), 2.01 (dddd, *J* = 14.2, 7.5, 6.8, 4.8 Hz, 1H);

<sup>13</sup>C NMR (100 MHz, CDCl<sub>3</sub>) δ 202.2, 159.3, 141.8, 130.4, 129.5, 128.7, 127.9, 126.8, 113.9, 80.0, 70.3, 55.4, 40.6, 31.1;

HRMS (ESI) *m/z*: [M-H]<sup>-</sup> Calcd for C<sub>18</sub>H<sub>19</sub>O<sub>3</sub> 283.1334; Found 283.1329;

IR (film, DCM) 3061, 3030, 3003, 2934, 2836, 2725, 1721, 1611, 1585, 1512 cm<sup>-1</sup>.



**(4R,7S)-7-((4-methoxybenzyl)oxy)-7-phenylhept-1-en-4-ol (9)**. Prepared *via* adaptation of procedure from <sup>R1</sup>.

An oven-dried vacuum Schlenk tube equipped with a magnetic stir bar was charged with Cs<sub>2</sub>CO<sub>3</sub> (0.007 g, 0.020 mmol, 12 mol%) and Krische Ir Catalyst ((*S*)-SEGPHOS, 4-cyano-3-nitrobenzoate ligated) (0.019 g, 0.018 mmol, 10 mol%). The reaction vessel was placed under an atmosphere of argon, and anhydrous THF (0.90 mL), **8** (0.051 g, 0.180 mmol), allyl acetate (0.039 mL, 0.360 mmol, 2 equiv) and 2-propanol (0.028 mL, 0.360 mmol, 2 equiv) were added by syringe. The reaction vessel was sealed and the reaction mixture was stirred at 100 °C and was monitored by TLC. After 16 h, the mixture was allowed to reach rt and was concentrated *in vacuo*. The residue was purified by flash column chromatography (hexane/ethyl acetate 4:1) to give **9** (0.042 g, 72%) as a slightly yellow oil.

$[\alpha]^{24}_{\text{D}}$  -64.4 (c 1.28, CHCl<sub>3</sub>);

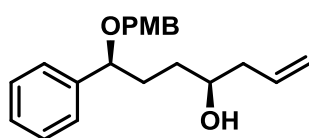
<sup>1</sup>H NMR (500 MHz, CDCl<sub>3</sub>) δ 7.40 – 7.28 (m, 5H), 7.24 – 7.19 (m, 2H), 6.91 – 6.82 (m, 2H), 5.80 (dddd, *J* = 20.3, 9.7, 7.8, 6.7 Hz, 1H), 5.12 – 5.10 (m, 1H), 5.10 – 5.08 (m, 1H), 4.40 (d, *J*

= 11.3 Hz, 1H), 4.33 (dd,  $J = 7.8, 5.0$  Hz, 1H), 4.20 (d,  $J = 11.3$  Hz, 1H), 3.81 (s, 3H), 3.68 – 3.61 (m, 1H), 2.29 – 2.22 (m, 1H), 2.19 – 2.11 (m, 1H), 2.03 (s, 1H), 1.93 – 1.79 (m, 2H), 1.73 – 1.62 (m, 1H), 1.47 – 1.36 (m, 1H);

$^{13}\text{C NMR}$  (125 MHz,  $\text{CDCl}_3$ )  $\delta$  159.3, 142.6, 135.1, 130.6, 129.6, 129.6, 128.6, 127.7, 126.9, 117.9, 113.9, 81.4, 70.7, 70.3, 55.4, 42.0, 34.6, 33.4;

**HRMS** (ESI)  $m/z$ :  $[\text{M}+\text{Na}]^+$  Calcd for  $\text{C}_{21}\text{H}_{26}\text{O}_3\text{Na}$  349.1780; Found 349.1781;

**IR** (film, DCM) 3442, 3079, 2932, 2862, 1641, 1613, 1512  $\text{cm}^{-1}$ .



**(4S,7S)-7-((4-methoxybenzyl)oxy)-7-phenylhept-1-en-4-ol (SI-3)**. Prepared *via* adaptation of procedure from <sup>R1</sup>.

An oven-dried vacuum Schlenk tube equipped with a magnetic stir bar was charged with  $\text{Cs}_2\text{CO}_3$  (0.033 g, 0.10 mmol, 12 mol%) and Krische Ir Catalyst (*(R)*-SEGPHOS, 4-cyano-3-nitrobenzoate ligated) (0.088 g, 0.09 mmol, 10 mol%). The reaction vessel was placed under an atmosphere of argon, and anhydrous THF (4.25 mL), **8** (0.242 g, 0.85 mmol), allyl acetate (0.183 mL, 1.70 mmol, 2 equiv) and 2-propanol (0.130 mL, 1.70 mmol, 2 equiv) were added by syringe. The reaction vessel was sealed and the reaction mixture was stirred at 100 °C and was monitored by TLC. After 16 h, the mixture was allowed to reach rt and was concentrated *in vacuo*. The residue was purified by flash column chromatography (hexane/ethyl acetate 4:1) to give **SI-3** (0.214 g, 77%) as a slightly yellow oil.

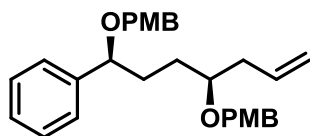
$[\alpha]^{23}_{\text{D}} -79.7$  (c 1.31,  $\text{CHCl}_3$ );

$^1\text{H NMR}$  (400 MHz,  $\text{CDCl}_3$ )  $\delta$  7.39 – 7.28 (m, 5H), 7.23 – 7.19 (m, 2H), 6.91 – 6.83 (m, 2H), 5.86 – 5.73 (m, 1H), 5.12 – 5.10 (m, 1H), 5.09 – 5.06 (m, 1H), 4.41 (d,  $J = 11.4$  Hz, 1H), 4.31 (dd,  $J = 8.3, 4.9$  Hz, 1H), 4.19 (d,  $J = 11.4$  Hz, 1H), 3.81 (s, 3H), 3.65 – 3.55 (m, 1H), 2.27 – 2.09 (m, 3H), 2.01 – 1.89 (m, 1H), 1.85 – 1.72 (m, 1H), 1.60 – 1.48 (m, 2H);

$^{13}\text{C NMR}$  (100 MHz,  $\text{CDCl}_3$ )  $\delta$  159.3, 142.6, 135.1, 130.5, 129.6, 128.6, 127.7, 126.9, 117.9, 113.9, 81.1, 70.8, 70.2, 55.4, 42.1, 34.9, 33.4;

**HRMS** (ESI)  $m/z$ :  $[\text{M}+\text{Na}]^+$  Calcd for  $\text{C}_{21}\text{H}_{26}\text{O}_3\text{Na}$  349.1780; Found 349.1782;

IR (film, DCM) 3423, 3072, 3030, 3003, 2932, 2863, 1640, 1612, 1586, 1513  $\text{cm}^{-1}$ .



**4,4'-((((1S,4S)-1-phenylhept-6-ene-1,4-diyl)bis(oxy))bis(methylene))bis(methoxybenzene) (10).**

NaH (60 % dispersion in mineral oil) (0.052 g, 1.30 mmol, 2 equiv) was added to a flask containing **SI-3** (0.212 g, 0.65 mmol) and TBAI (0.024 g, 0.07 mmol, 10 mol%) in anhydrous DMF (1.30 mL) cooled to 0 °C. Reaction mixture was flushed with argon and stirred for 30 min at 0 °C. Then, 4-methoxybenzyl chloride (0.176 mL, 1.30 mmol, 2 equiv) was added dropwise and the mixture was stirred at rt. The reaction was monitored by TLC and after 20 h the mixture was quenched by addition of sat.  $\text{NH}_4\text{Cl}$  solution. The mixture was extracted with ethyl acetate. Combined organic phases were dried over  $\text{Na}_2\text{SO}_4$  and concentrated *in vacuo*. Crude product was purified by flash column chromatography (hexane/ethyl acetate 95:5) to give **10** (0.272 g, 94%) as a colorless oil.

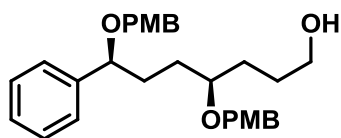
$[\alpha]^{23}_{\text{D}} -58.1$  (c 1.05,  $\text{CHCl}_3$ );

$^1\text{H NMR}$  (400 MHz,  $\text{CDCl}_3$ )  $\delta$  7.39 – 7.27 (m, 5H), 7.25 – 7.19 (m, 4H), 6.89 – 6.83 (m, 4H), 5.79 (ddt,  $J = 17.2, 10.2, 7.1$  Hz, 1H), 5.09 – 5.05 (m, 1H), 5.05 – 4.98 (m, 1H), 4.45 – 4.33 (m, 3H), 4.24 (dd,  $J = 7.6, 5.4$  Hz, 1H), 4.16 (d,  $J = 11.5$  Hz, 1H), 3.81 (s, 3H), 3.80 (s, 3H), 3.44 – 3.32 (m, 1H), 2.31 – 2.21 (m, 2H), 2.01 – 1.89 (m, 1H), 1.78 – 1.61 (m, 2H), 1.52 – 1.39 (m, 1H);

$^{13}\text{C NMR}$  (100 MHz,  $\text{CDCl}_3$ )  $\delta$  159.3, 159.2, 142.8, 135.1, 131.2, 130.9, 129.5, 129.4, 128.5, 127.6, 127.0, 116.9, 113.9, 80.9, 77.8, 70.6, 70.1, 55.4, 38.3, 33.9, 29.9;

**HRMS** (ESI)  $m/z$ :  $[\text{M}+\text{Na}]^+$  Calcd for  $\text{C}_{29}\text{H}_{34}\text{O}_4\text{Na}$  469.2355; Found 469.2357;

IR (film, DCM) 3063, 3030, 3001, 2934, 2861, 2836, 1612, 1585, 1513  $\text{cm}^{-1}$ .



**(4R,7S)-4,7-bis((4-methoxybenzyl)oxy)-7-phenylheptan-1-ol (11)**

To a solution of **10** (0.277 g, 0.62 mmol) in anhydrous THF (2.86 mL) was added drop-wise 9-borabicyclo[3.3.1]nonane solution (0.5 M in THF) (3.72 mL, 1.86 mmol, 3 equiv) over 10 min and the mixture was heated under reflux for 1 h. After cooling the reaction mixture was treated with ethanol (1.90 mL), 2 M NaOH solution (0.95 mL) and H<sub>2</sub>O<sub>2</sub> solution (30% (w/w) in H<sub>2</sub>O) (0.95 mL) and the mixture was stirred at 50 °C for 1 h. The reaction was monitored by TLC and upon completion sat. K<sub>2</sub>CO<sub>3</sub> solution and Et<sub>2</sub>O were added. The aqueous phase was extracted with Et<sub>2</sub>O, combined organic phases were dried over Na<sub>2</sub>SO<sub>4</sub> and concentrated *in vacuo*. Crude product was purified by flash column chromatography (hexane/ethyl acetate 7:3) to give **11** (0.243 g, 84%) as a colorless oil.

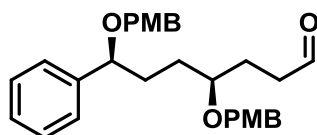
[ $\alpha$ ]<sup>19</sup><sub>D</sub> -36.7 (c 1.21, CHCl<sub>3</sub>);

<sup>1</sup>H NMR (400 MHz, CDCl<sub>3</sub>)  $\delta$  7.39 – 7.35 (m, 2H), 7.33 – 7.28 (m, 3H), 7.23 – 7.19 (m, 4H), 6.90 – 6.83 (m, 4H), 4.42 – 4.33 (m, 3H), 4.26 (dd, *J* = 7.7, 5.1 Hz, 1H), 4.17 (d, *J* = 11.4 Hz, 1H), 3.81 (s, 3H), 3.80 (s, 3H), 3.60 – 3.54 (m, 2H), 3.42 – 3.34 (m, 1H), 1.98 – 1.85 (m, 2H), 1.76 – 1.67 (m, 2H), 1.61 – 1.42 (m, 5H);

<sup>13</sup>C NMR (100 MHz, CDCl<sub>3</sub>)  $\delta$  159.3, 159.3, 142.7, 130.9, 130.8, 129.5, 128.6, 127.7, 127.0, 113.9, 113.9, 81.1, 78.2, 70.5, 70.1, 63.1, 55.4, 55.4, 33.8, 30.3, 29.6, 28.7;

HRMS (ESI) *m/z*: [M+Na]<sup>+</sup> Calcd for C<sub>29</sub>H<sub>36</sub>O<sub>5</sub>Na 487.2460; Found 487.2455;

IR (film, DCM) 3413, 3060, 3030, 3000, 2936, 2863, 1612, 1586, 1513 cm<sup>-1</sup>.



**(4S,7S)-4,7-bis((4-methoxybenzyl)oxy)-7-phenylheptanal (12)**

To a solution of **11** (0.107 g, 0.23 mmol) in anhydrous DCM (1.15 mL) were added molecular sieves 4 Å (beads) (0.115 g) and pyridinium chlorochromate (0.099 g, 0.46 mmol, 2 equiv) and the mixture was stirred at rt. The reaction was monitored by TLC and after 1.5 h Et<sub>2</sub>O (1.15

mL) was added. After 30 min of intensive stirring the mixture was filtered through a Celite plug, which was thoroughly washed with Et<sub>2</sub>O. The filtrate was washed with water and brine, dried over Na<sub>2</sub>SO<sub>4</sub> and concentrated *in vacuo*. Crude product was purified by flash column chromatography (hexane/ethyl acetate 85:15) to give **12** (0.087 g, 82%) as a colorless oil.

$[\alpha]_{\text{D}}^{19}$  -31.2 (c 1.12, CHCl<sub>3</sub>);

**<sup>1</sup>H NMR** (400 MHz, CDCl<sub>3</sub>) δ 9.70 (t, *J* = 1.7 Hz, 1H), 7.41 – 7.28 (m, 5H), 7.24 – 7.16 (m, 4H), 6.91 – 6.82 (m, 4H), 4.42 – 4.34 (m, 2H), 4.30 (d, *J* = 11.3 Hz, 1H), 4.26 (dd, *J* = 7.8, 5.1 Hz, 1H), 4.17 (d, *J* = 11.3 Hz, 1H), 3.81 (s, 3H), 3.80 (s, 3H), 3.41 – 3.30 (m, 1H), 2.42 (td, *J* = 7.2, 1.7 Hz, 2H), 1.95 – 1.80 (m, 2H), 1.78 – 1.64 (m, 3H), 1.49 – 1.37 (m, 1H);

**<sup>13</sup>C NMR** (100 MHz, CDCl<sub>3</sub>) δ 202.5, 159.3, 159.3, 142.6, 130.8, 129.5, 128.6, 127.7, 127.0, 113.9, 113.9, 80.9, 77.2, 70.5, 70.1, 55.4, 55.4, 40.0, 33.7, 29.7, 26.3;

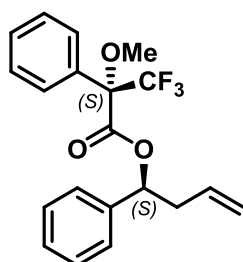
**HRMS** (ESI) *m/z*: [M+Na]<sup>+</sup> Calcd for C<sub>29</sub>H<sub>34</sub>O<sub>5</sub>Na 485.2304; Found 485.2295;

**IR** (film, DCM) 3030, 3001, 2935, 2862, 2837, 1722, 1612, 1585, 1513 cm<sup>-1</sup>.

## Section S8. Determination of absolute configuration of newly formed stereogenic centers (Mosher ester analysis)

**General procedure GP-1.** Procedure adapted from <sup>R6</sup>.

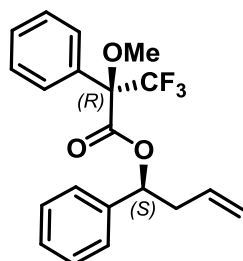
Alcohol (1 equiv) was transferred to a screw-capped 4 mL glass vial and anhydrous DCM ( $c = 0.07$  M) was added followed by addition of anhydrous pyridine (3.1 equiv) and (*R*)- or (*S*)-MTPA-Cl (1.9 equiv). The vial was sealed and the mixture was stirred at rt for 12 h. Upon completion, the reaction mixture was diluted with Et<sub>2</sub>O and water was added. The aqueous phase was extracted with Et<sub>2</sub>O. Combined organic phases were dried over Na<sub>2</sub>SO<sub>4</sub> and concentrated *in vacuo*. Crude product was purified by flash column chromatography.



**(*S*)-1-phenylbut-3-en-1-yl (*S*)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-4).**

According to general procedure **GP-1**, the reaction was performed with compound **SI-1** (10 mg, 0.07 mmol). After purification by flash column chromatography (hexane/ethyl acetate 20:1) Mosher ester **SI-4** was obtained (21 mg, 85%) as a colorless oil.

<sup>1</sup>H NMR (500 MHz, CDCl<sub>3</sub>)  $\delta$  7.44 – 7.40 (m, 3H), 7.38 – 7.30 (m, 7H), 6.03 (dd,  $J = 8.1, 5.6$  Hz, 1H), 5.61 (ddt,  $J = 17.2, 10.3, 7.0$  Hz, 1H), 5.05 – 4.99 (m, 2H), 3.45 (q,  $J = 1.2$  Hz, 3H), 2.70 (dddd,  $J = 14.4, 8.3, 7.2, 1.2$  Hz, 1H), 2.58 (dddd,  $J = 14.4, 6.9, 5.6, 1.2$  Hz, 1H).



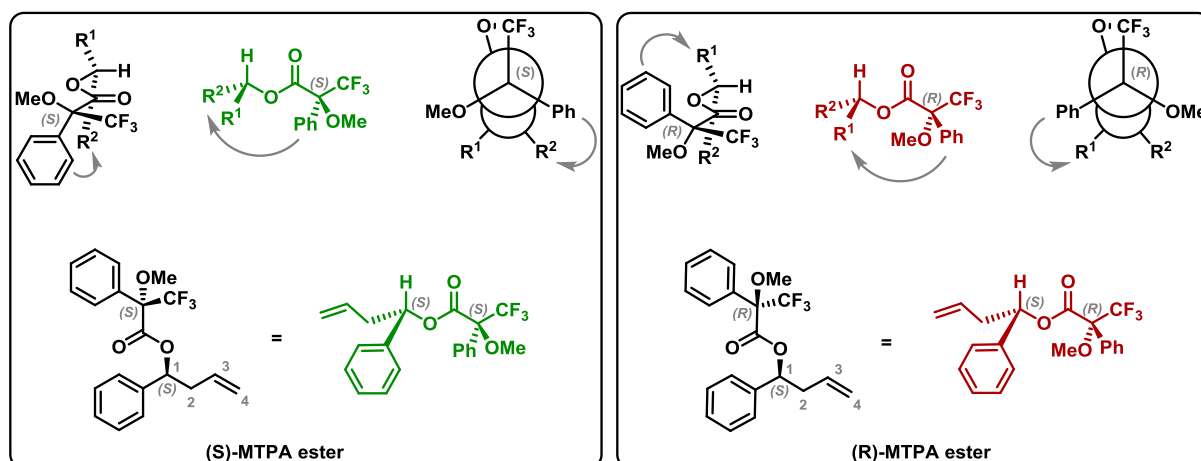
**(*S*)-1-phenylbut-3-en-1-yl (*R*)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-5).**

According to general procedure **GP-1**, the reaction was performed with compound **SI-1** (10 mg, 0.07 mmol). After purification by flash column chromatography (hexane/ethyl acetate 20:1) Mosher ester **SI-5** was obtained (23 mg, 94%) as a colorless oil.

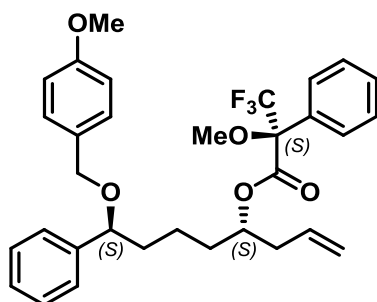
$^1\text{H}$  NMR (500 MHz,  $\text{CDCl}_3$ )  $\delta$  7.41 – 7.34 (m, 3H), 7.33 – 7.28 (m, 5H), 7.24 – 7.21 (m, 2H), 5.97 (dd,  $J = 8.4, 5.4$  Hz, 1H), 5.74 (ddt,  $J = 17.1, 10.3, 6.9$  Hz, 1H), 5.16 – 5.08 (m, 2H), 3.54 (q,  $J = 1.3$  Hz, 3H), 2.74 (dddt,  $J = 14.6, 8.4, 7.4, 1.3$  Hz, 1H), 2.61 (dddt,  $J = 14.6, 6.7, 5.4, 1.3$  Hz, 1H).

**Table S1.**  $\Delta\delta$  ( $=\delta_S - \delta_R$ ) data for the (*S*)-MTPA Mosher ester **SI-4** and (*R*)-MTPA Mosher ester **SI-5**

	$\delta$ ( <i>S</i> )-ester <b>SI-4</b> (ppm)	$\delta$ ( <i>R</i> )-ester <b>SI-5</b> (ppm)	$\Delta\delta^{\text{SR}}$ ( $=\delta_S - \delta_R$ )	
			ppm	Hz (500 MHz)
H-1	6.03	5.97	0.06	30
H-2 <sub>b</sub>	2.58	2.61	-0.03	-15
H-2 <sub>a</sub>	2.70	2.74	-0.04	-20
H-4 <sub>a</sub> & H-4 <sub>b</sub>	5.01	5.12	-0.11	-55
H-3	5.61	5.74	-0.13	-65



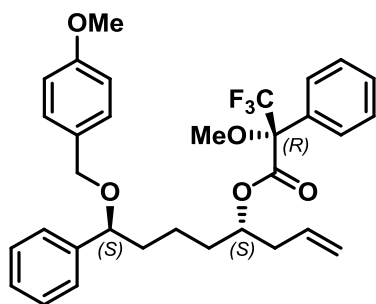
**Figure S5.** Conformational analysis of each of the diastereoisomeric MTPA esters of **SI-4** and **SI-5**. Gray arrow indicates the phenyl group shielding effect.



**(4*S*,8*S*)-8-((4-methoxybenzyl)oxy)-8-phenyloct-1-en-4-yl (S)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-6).**

According to general procedure **GP-1**, the reaction was performed with compound **SI-2** (14 mg, 0.04 mmol). After purification by flash column chromatography (hexane/ethyl acetate 95:5) Mosher ester **SI-6** was obtained (20 mg, 87%) as a colorless oil.

**<sup>1</sup>H NMR** (600 MHz, CDCl<sub>3</sub>) δ 7.55 – 7.48 (m, 2H), 7.40 – 7.32 (m, 5H), 7.31 – 7.26 (m, 3H), 7.22 – 7.17 (m, 2H), 6.90 – 6.84 (m, 2H), 5.72 (ddt, *J* = 19.3, 9.6, 7.0 Hz, 1H), 5.12 (p, *J* = 6.4 Hz, 1H), 5.10 – 5.06 (m, 2H), 4.36 (d, *J* = 11.4 Hz, 1H), 4.17 (dd, *J* = 7.9, 5.6 Hz, 1H), 4.14 (d, *J* = 11.4 Hz, 1H), 3.80 (s, 3H), 3.54 – 3.46 (m, 3H), 2.37 (td, *J* = 6.4, 5.9, 1.5 Hz, 2H), 1.80 (dddd, *J* = 13.2, 10.2, 7.8, 5.0 Hz, 1H), 1.57 – 1.51 (m, 3H), 1.38 – 1.29 (m, 1H), 1.23 – 1.16 (m, 1H).



**(4*S*,8*S*)-8-((4-methoxybenzyl)oxy)-8-phenyloct-1-en-4-yl (R)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-7).**

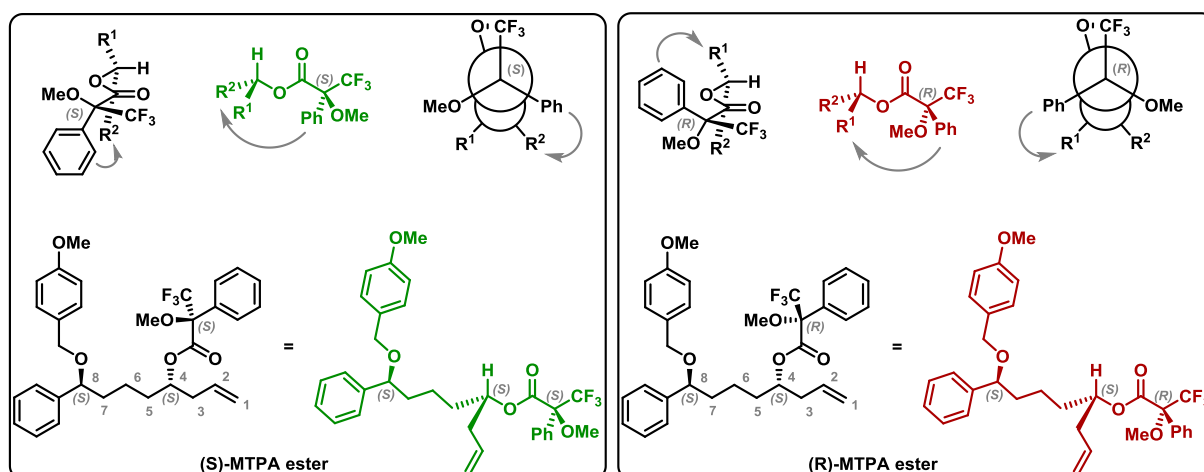
According to general procedure **GP-1**, the reaction was performed with compound **SI-2** (14 mg, 0.04 mmol). After purification by flash column chromatography (hexane/ethyl acetate 95:5) Mosher ester **SI-7** was obtained (19 mg, 83%) as a colorless oil.

**<sup>1</sup>H NMR** (600 MHz, CDCl<sub>3</sub>) δ 7.54 – 7.48 (m, 2H), 7.39 – 7.33 (m, 5H), 7.32 – 7.28 (m, 3H), 7.23 – 7.16 (m, 2H), 6.89 – 6.83 (m, 2H), 5.58 (ddt, *J* = 17.3, 10.5, 7.1 Hz, 1H), 5.09 (p, *J* = 5.9 Hz, 1H), 5.01 – 4.94 (m, 2H), 4.38 (d, *J* = 11.4 Hz, 1H), 4.24 (dd, *J* = 7.7, 5.7 Hz, 1H), 4.16 (d, *J* = 11.4 Hz, 1H), 3.80 (s, 3H), 3.47 – 3.43 (m, 3H), 2.29 (td, *J* = 6.5, 5.7, 1.3 Hz, 2H), 1.86 (dddd, *J* = 13.0, 10.2, 7.7, 5.0 Hz, 1H), 1.65 – 1.54 (m, 3H), 1.50 – 1.39 (m, 1H), 1.32 (dddd, *J* = 16.0, 12.3, 10.2, 5.9 Hz, 1H).

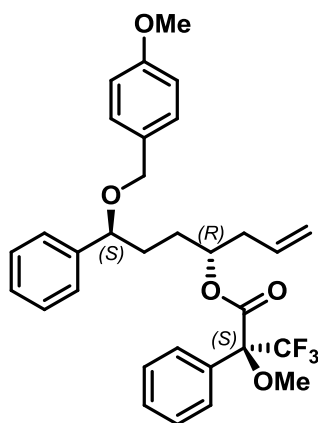


**Table S2.**  $\Delta\delta$  ( $=\delta_S - \delta_R$ ) data for the (*S*)-MTPA Mosher ester **SI-6** and (*R*)-MTPA Mosher ester **SI-7**

	$\delta$ ( <i>S</i> )-ester <b>SI-6</b> (ppm)	$\delta$ ( <i>R</i> )-ester <b>SI-7</b> (ppm)	$\Delta\delta^{SR}$ ( $=\delta_S - \delta_R$ )	
			ppm	Hz (600 MHz)
H-2	5.72	5.58	0.14	84
H-1 <sub>a</sub> &H-1 <sub>b</sub>	5.08	4.98	0.10	60
H-3 <sub>a</sub> &H-3 <sub>b</sub>	2.37	2.29	0.08	48
OMe	3.50	3.45	0.05	30
H-4	5.12	5.09	0.03	18
OMe (PMB)	3.80	3.80	0.00	0
CH <sub>2</sub> Ar (PMB)	4.36	4.38	-0.02	-12
CH <sub>2</sub> Ar (PMB)	4.14	4.16	-0.02	-12
H-7	1.80	1.86	-0.06	-36
H-8	4.17	4.24	-0.07	-42



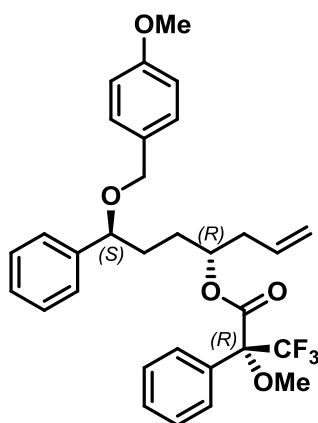
**Figure S6.** Conformational analysis of each of the diastereoisomeric MTPA esters of **SI-6** and **SI-7**. Gray arrow indicates the phenyl group shielding effect.



**(4*R*,7*S*)-7-((4-methoxybenzyl)oxy)-7-phenylhept-1-en-4-yl (S)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-8).**

According to general procedure **GP-1**, the reaction was performed with compound **9** (13 mg, 0.04 mmol). After purification by flash column chromatography (hexane/ethyl acetate 95:5) Mosher ester **SI-8** was obtained (18 mg, 83%) as a colorless oil.

<sup>1</sup>H NMR (400 MHz, CDCl<sub>3</sub>) δ 7.52 – 7.47 (m, 2H), 7.37 – 7.27 (m, 8H), 7.23 – 7.16 (m, 2H), 6.89 – 6.82 (m, 2H), 5.65 – 5.52 (m, 1H), 5.19 – 5.09 (m, 1H), 5.01 – 4.94 (m, 2H), 4.38 (d, *J* = 11.4 Hz, 1H), 4.24 (dd, *J* = 7.7, 4.9 Hz, 1H), 4.16 (d, *J* = 11.4 Hz, 1H), 3.80 (s, 3H), 3.49 (q, *J* = 1.3 Hz, 3H), 2.34 – 2.28 (m, 2H), 1.91 – 1.77 (m, 2H), 1.72 – 1.65 (m, 1H), 1.61 – 1.53 (m, 1H).



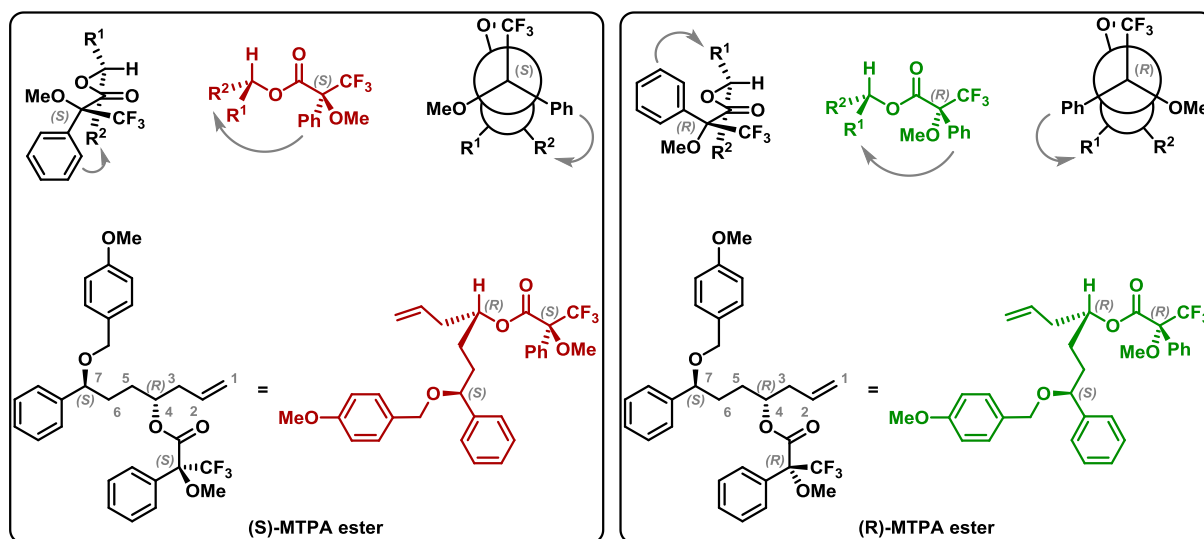
**(4*R*,7*S*)-7-((4-methoxybenzyl)oxy)-7-phenylhept-1-en-4-yl (R)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-9).**

According to general procedure **GP-1**, the reaction was performed with compound **9** (10 mg, 0.03 mmol). After purification by flash column chromatography (hexane/ethyl acetate 95:5) Mosher ester **SI-9** was obtained (14 mg, 86%) as a colorless oil.

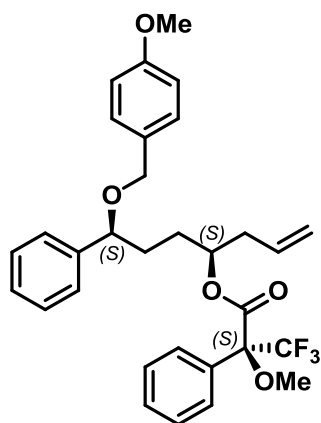
$^1\text{H}$  NMR (400 MHz,  $\text{CDCl}_3$ )  $\delta$  7.52 – 7.46 (m, 2H), 7.39 – 7.27 (m, 6H), 7.23 – 7.16 (m, 4H), 6.88 – 6.79 (m, 2H), 5.77 – 5.66 (m, 1H), 5.20 – 5.11 (m, 1H), 5.11 – 5.02 (m, 2H), 4.34 (d,  $J$  = 11.4 Hz, 1H), 4.15 (dd,  $J$  = 8.3, 4.9 Hz, 1H), 4.12 (d,  $J$  = 11.4 Hz, 1H), 3.80 (s, 3H), 3.52 (q,  $J$  = 1.3 Hz, 3H), 2.44 – 2.30 (m, 2H), 1.87 – 1.77 (m, 1H), 1.72 – 1.62 (m, 1H), 1.56 – 1.43 (m, 2H).

**Table S3.**  $\Delta\delta$  ( $=\delta_S - \delta_R$ ) data for the (*S*)-MTPA Mosher ester **SI-8** and (*R*)-MTPA Mosher ester **SI-9**

	$\delta$ ( <i>S</i> )-ester <b>SI-8</b> (ppm)	$\delta$ ( <i>R</i> )-ester <b>SI-9</b> (ppm)	$\Delta\delta^{\text{SR}}$ ( $=\delta_S - \delta_R$ )	
			ppm	Hz (400 MHz)
H-7	4.24	4.15	0.09	36
CH <sub>2</sub> Ar (PMB)	4.38	4.34	0.04	16
CH <sub>2</sub> Ar (PMB)	4.16	4.12	0.04	16
OMe (PMB)	3.80	3.80	0.00	0
H-4	5.13	5.15	-0.02	-8
OMe	3.49	3.52	-0.03	-12
H-3 <sub>a</sub> &H-3 <sub>b</sub>	2.31	2.38	-0.07	-28
H-1 <sub>a</sub> &H-1 <sub>b</sub>	4.98	5.07	-0.09	-36
H-2	5.59	5.71	-0.12	-48



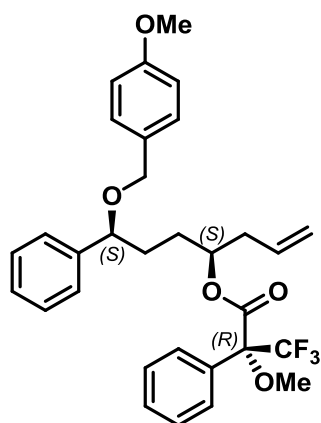
**Figure S7.** Conformational analysis of each of the diastereoisomeric MTPA esters of **SI-8** and **SI-9**. Gray arrow indicates the phenyl group shielding effect.



**(4*S*,7*S*)-7-((4-methoxybenzyl)oxy)-7-phenylhept-1-en-4-yl (S)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-10).**

According to general procedure **GP-1**, the reaction was performed with compound **SI-3** (36 mg, 0.11 mmol). After purification by flash column chromatography (hexane/ethyl acetate 95:5) Mosher ester **SI-10** was obtained (54 mg, 90%) as a colorless oil.

<sup>1</sup>H NMR (400 MHz, CDCl<sub>3</sub>) δ 7.54 – 7.48 (m, 2H), 7.39 – 7.27 (m, 6H), 7.26 – 7.22 (m, 2H), 7.21 – 7.17 (m, 2H), 6.89 – 6.84 (m, 2H), 5.77 – 5.64 (m, 1H), 5.21 – 5.12 (m, 1H), 5.11 – 5.04 (m, 2H), 4.37 (d, *J* = 11.5 Hz, 1H), 4.22 (dd, *J* = 8.0, 4.8 Hz, 1H), 4.13 (d, *J* = 11.4 Hz, 1H), 3.81 (s, 3H), 3.53 (q, *J* = 1.3 Hz, 3H), 2.44 – 2.29 (m, 2H), 1.80 – 1.67 (m, 2H), 1.64 – 1.45 (m, 2H).



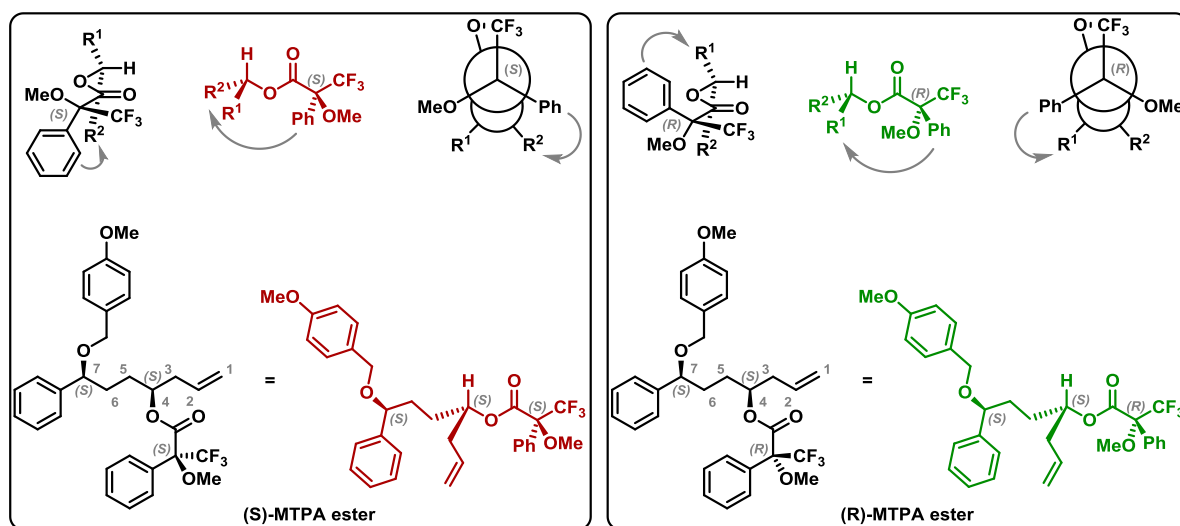
**(4*S*,7*S*)-7-((4-methoxybenzyl)oxy)-7-phenylhept-1-en-4-yl (R)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-11).**

According to general procedure **GP-1**, the reaction was performed with compound **SI-3** (29 mg, 0.09 mmol). After purification by flash column chromatography (hexane/ethyl acetate 95:5) Mosher ester **SI-11** was obtained (45 mg, 92%) as a colorless oil.

$^1\text{H NMR}$  (400 MHz,  $\text{CDCl}_3$ )  $\delta$  7.54 – 7.48 (m, 2H), 7.39 – 7.27 (m, 8H), 7.23 – 7.19 (m, 2H), 6.90 – 6.84 (m, 2H), 5.64 – 5.52 (m, 1H), 5.18 – 5.10 (m, 1H), 5.02 – 4.92 (m, 2H), 4.39 (d,  $J = 11.4$  Hz, 1H), 4.29 (dd,  $J = 8.1, 4.5$  Hz, 1H), 4.16 (d,  $J = 11.4$  Hz, 1H), 3.81 (s, 3H), 3.49 (q,  $J = 1.3$  Hz, 3H), 2.38 – 2.24 (m, 2H), 1.92 – 1.75 (m, 2H), 1.71 – 1.56 (m, 2H).

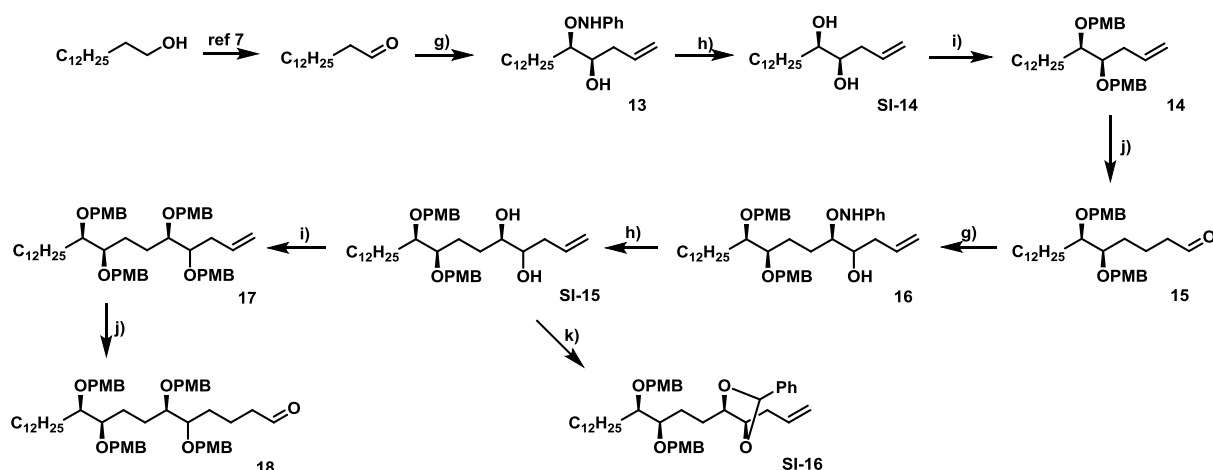
**Table S4.**  $\Delta\delta$  ( $=\delta_S - \delta_R$ ) data for the (*S*)-MTPA Mosher ester **SI-10** and (*R*)-MTPA Mosher ester **SI-11**.

	$\delta$ ( <i>S</i> )-ester SI-10 (ppm)	$\delta$ ( <i>R</i> )-ester SI-11 (ppm)	$\Delta\delta^{\text{SR}}$ ( $=\delta_S - \delta_R$ )	
			ppm	Hz (400 MHz)
H-2	5.70	5.58	0.12	48
H-1 <sub>a</sub> &H-1 <sub>b</sub>	5.07	4.97	0.10	40
H-3 <sub>a</sub> &H-3 <sub>b</sub>	2.37	2.30	0.07	28
OMe	3.53	3.49	0.04	16
H-4	5.16	5.15	0.01	4
OMe (PMB)	3.81	3.81	0.00	0
CH <sub>2</sub> Ar (PMB)	4.37	4.39	-0.02	-8
CH <sub>2</sub> Ar (PMB)	4.13	4.16	-0.03	-12
H-7	4.22	4.29	-0.07	-28



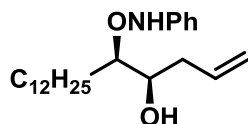
**Figure S8.** Conformational analysis of each of the diastereoisomeric MTPA esters of **SI-10** and **SI-11**. Gray arrow indicates the phenyl group shielding effect.

## Section S9. Iterative synthesis of monhexocin's fragment



### Scheme S3. Iterative synthesis towards monhexocin.

Reagents and conditions: (g) L-proline, PhNO, allyl bromide, NaI, TBAI, DMF, 0 °C, 3 h, 59-62%, (h) Zn, AcOH, THF, H<sub>2</sub>O, 60 °C, 1 h, 91-99%, (i) PMBCl, NaH, TBAI, DMF, 0 °C to rt, 17 h, 88%, (j) Rh(CO)<sub>2</sub>acac, 6-DPPon, TBAI, Ac<sub>2</sub>O, HCOOH, MS 4 Å, DMF, 80 °C, 20 h, 39-61%, (k) PhCHO, CSA, MS 4 Å, PhH, 0 °C, 1 h, 37%.



**(4R,5R)-5-((phenylamino)oxy)heptadec-1-en-4-ol (13)**. Prepared *via* adaptation of procedure from<sup>R7</sup>.

To a solution of 1-tetradecanal (1.7 g, 8 mmol, 3 equiv) in anhydrous DMF (30 mL), L-proline (921 mg, 8 mmol) was added and stirred at rt for 1 h. Then, nitrosobenzene (285 mg, 2.67 mmol, 1 equiv) was added. The endpoint of the reaction was monitored by its color change from green to orange. After 45 min, the solution was cooled to 0 °C and allylindium iodide solution (prepared by heating for 1 h at 70 °C: granular indium (916 mg, 8 mmol), NaI (1.2 g, 8 mmol), allyl bromide (1.38 mL, 16 mmol) and anhydrous DMF (10 mL)) was slowly added. The stirring was kept at 0 °C for 2 h. It was then diluted with Et<sub>2</sub>O and washed with water. The aqueous layer was extracted with Et<sub>2</sub>O. The combined organic layers were dried over Na<sub>2</sub>SO<sub>4</sub>, filtered, and concentrated *in vacuo*. Crude product was purified by column chromatography (hexane/ethyl acetate 9:1) to give two diastereoisomeric products as yellow oils with 59% yield for *syn*-configured product **13** (faster-eluting diastereoisomer) and 40% yield for *anti*-

configured product (slower-eluting diastereoisomer). The diastereomeric ratio of the products were determined by preparing Mosher esters of slower-eluting diastereoisomer (analysis below).

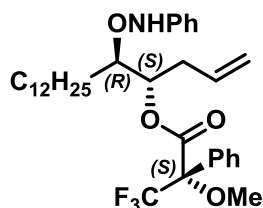
$[\alpha]_D^{21}$  19.17 (c 0.9, C<sub>6</sub>H<sub>6</sub>);

<sup>1</sup>H NMR (500 MHz, CDCl<sub>3</sub>) δ 5.91 – 5.81 (m, 1H), 5.19 – 5.13 (m, 2H), 3.53 – 3.42 (m, 2H), 2.39 – 2.33 (m, 1H), 2.27 – 2.20 (m, 1H), 2.09 – 2.03 (m, 2H), 1.54 – 1.42 (m, 3H), 1.26 (s, 19H), 0.88 (t, *J* = 6.8 Hz, 3H);

<sup>13</sup>C NMR (125 MHz, CDCl<sub>3</sub>) δ 148.3, 134.7, 129.0, 122.5, 117.8, 115.1, 85.5, 72.7, 38.2, 31.9, 29.8, 29.7, 29.6, 29.6, 29.5, 29.5, 29.3, 25.6, 22.7, 14.1;

IR (film, CH<sub>2</sub>Cl<sub>2</sub>); 3270, 3075, 2925, 2853, 1643, 1603, 1522, 1495, 1466, 1360, 1280, 1183, 1166, 1145, 1111, 1051, 1026, 997, 913, 822, 769, 741, 694 cm<sup>-1</sup>;

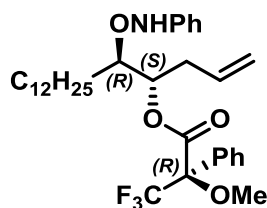
HRMS (ESI) *m/z*: [M + H]<sup>+</sup> Calcd for C<sub>23</sub>H<sub>50</sub>NO<sub>2</sub> 362.3059; Found 362.3057.



**(4*S*,5*R*)-5-((phenylamino)oxy)heptadec-1-en-4-yl (S)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-12).**

According to general procedure **GP-1**, the reaction was performed with diastereoisomer of compound **13** (20 mg, 0.055 mmol). After purification by flash column chromatography (hexane/ethyl acetate 95:5) Mosher ester **SI-12** was obtained (27 mg, 84%) as a white solid.

<sup>1</sup>H NMR (400 MHz, C<sub>6</sub>D<sub>6</sub>) δ 7.26 (d, *J* = 7.7 Hz, 2H), 6.87 – 6.71 (m, 4H), 6.68 – 6.57 (m, 4H), 6.03 – 5.89 (m, 1H), 5.10 (d, *J* = 17.1 Hz, 1H), 5.03 (d, *J* = 10.1 Hz, 1H), 4.62 (brs, 1H), 4.18 (brs, 1H), 3.70 (d, *J* = 9.4 Hz, 1H), 3.45 (s, 3H), 2.62 – 2.46 (m, 1H), 2.16 – 2.05 (m, 1H), 1.87 – 1.82 (m, 1H), 1.30 – 1.13 (m, 16H), 1.11 – 1.03 (m, 3H), 0.93 – 0.89 (m, 2H), 0.89 – 0.86 (m, 3H).



**(4*S*,5*R*)-5-((phenylamino)oxy)heptadec-1-en-4-yl (*R*)-3,3,3-trifluoro-2-methoxy-2-phenylpropanoate (SI-13).**

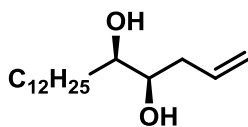
According to general procedure **GP-1**, the reaction was performed with diastereoisomer of compound **13** (20 mg, 0.055 mmol). After purification by flash column chromatography (hexane/ethyl acetate 95:5). Mosher ester **SI-13** was obtained (23 mg, 72%) as a colorless oil.

<sup>1</sup>H NMR (400 MHz, C<sub>6</sub>D<sub>6</sub>) δ 7.58 (dd, *J* = 6.9, 2.8 Hz, 1H), 7.02 (dd, *J* = 5.3, 2.0 Hz, 1H), 6.79 (d, *J* = 7.2 Hz, 1H), 6.72 (brs, 2H), 6.44 (brs, 1H), 5.98 – 5.84 (m, 1H), 5.02 (d, *J* = 17.1 Hz, 1H), 4.97 (d, *J* = 10.2 Hz, 1H), 4.90 (brs, 1H), 4.15 (brs, 1H), 3.44 (d, *J* = 10.1 Hz, 1H), 3.08 (brs, 3H), 2.46 (brs, 1H), 2.02 – 1.87 (m, 2H), 1.32 – 1.20 (m, 16H), 1.15 (brs, 3H), 0.96 (brs, 2H), 0.91 – 0.85 (m, 3H).

**Table S5.**  $\Delta\delta$  ( $=\delta_S - \delta_R$ ) data for the (*S*)-MTPA Mosher ester **SI-12** and (*R*)-MTPA Mosher ester **SI-13**

	$\delta$ ( <i>S</i> )-ester SI-12 (ppm)	$\delta$ ( <i>R</i> )-ester SI-13 (ppm)	$\Delta\delta^{SR}$ ( $=\delta_S - \delta_R$ )	
			ppm	Hz (400 MHz)
H-5	3.70	3.49	0.21	84
H-3 <sub>a</sub>	2.10	1.98	0.12	48
H-1	5.06	5.04	0.02	8
H-3 <sub>b</sub>	2.53	2.51	0.02	8
H-2	5.96	5.96	0	0
H-4	4.18	4.19	-0.01	-4
H-6 <sub>b</sub>	1.35	1.36	-0.01	-4
H-17	0.87	0.93	-0.06	-24
Alkyl chain	1.24	1.30	-0.06	-24
H-6 <sub>a</sub>	1.87	1.98	-0.11	-44
-NHPh	6.79	7.06	-0.27	-108
-NHPh	4.63	4.94	-0.31	-124
-NHPh	7.25	7.62	-0.37	-148





**(4R,5R)-heptadec-1-ene-4,5-diol (SI-14).** Prepared under conditions from<sup>R8</sup>.

Diastereoisomer **13** (550 mg, 1.52 mmol) was dissolved in a 1:1 THF/H<sub>2</sub>O mixture (22 mL). Acetic acid (33.5 mL) and Zn dust (3.78 g, 57 mmol) was added. The mixture was stirred at 60 °C for 1 h. After cooling to rt, the mixture was diluted with Et<sub>2</sub>O and filtered through a plug of silica gel, which was washed by Et<sub>2</sub>O. After evaporation of the solvents, the resulting crude mixture was redissolved in EtOAc, preadsorbed on silica gel and purified by column chromatography (hexane/ethyl acetate 4:1) to give **SI-14** (374 mg, 91%) as a white solid.

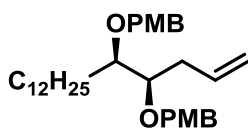
$[\alpha]^{21}_D$  7.28 (c 2.4, CHCl<sub>3</sub>);

<sup>1</sup>H NMR (500 MHz, CDCl<sub>3</sub>) δ 5.91 – 5.81 (m, 1H), 5.19 – 5.13 (m, 2H), 3.53 – 3.42 (m, 2H), 2.39 – 2.33 (m, 1H), 2.28 – 2.20 (m, 1H), 2.09 – 2.03 (m, 2H), 1.54 – 1.43 (m, 3H), 1.26 (s, 19H), 0.88 (t, *J* = 6.8 Hz, 3H);

<sup>13</sup>C NMR (125 MHz, CDCl<sub>3</sub>) δ 134.5, 118.2, 73.9, 73.2, 38.3, 33.6, 31.9, 29.7, 29.6, 29.6, 29.6, 29.3, 25.6, 22.7, 14.1;

IR (film, CH<sub>2</sub>Cl<sub>2</sub>); 3193, 2916, 2848, 1643, 1469, 1436, 1281, 1065, 989, 915, 872, 718 cm<sup>-1</sup>;

HRMS (ESI) *m/z*: [M + Na]<sup>+</sup> Calcd for C<sub>17</sub>H<sub>34</sub>O<sub>2</sub>Na 293.2408; Found 293.2416.



**4,4'-((((4R,5R)-heptadec-1-ene-4,5-diyl)bis(oxy))bis(methylene))bis(methoxybenzene) (14).**

NaH (60% dispersion in mineral oil, 213 mg, 5.32 mmol, 4 equiv) was added to a flask containing **SI-14** (360 mg, 1.33 mmol) and TBAI (25 mg, 0.067 mmol, 5 mol%) in anhydrous DMF (15 mL) cooled to 0 °C. Reaction mixture was stirred at 0 °C for 30 min. Then, PMBCl (0.72 mL, 5.32 mmol, 4 equiv) was added dropwise and the mixture was stirred at rt for 17 h. Saturated NH<sub>4</sub>Cl was added to quench the reaction followed by the extraction with EtOAc. The combined organic layers were dried over Na<sub>2</sub>SO<sub>4</sub>, filtered, and concentrated *in vacuo*. Crude

product was purified by column chromatography (hexane/ethyl acetate 95:5) to give **14** (598 mg, 88%, >99% *ee* by HPLC analysis) as a colorless oil.

**HPLC** (Chiralcel OD-H, hexane/*i*-PrOH 99.5:0.5, flow rate 1 mL/min,  $\lambda$  220 nm):  $t_R$  = 8.1 min (*S,S*),  $t_R$  = 8.4 min (*S,R*),  $t_R$  = 9.0 min (*R,R*),  $t_R$  = 13.7 min (*R,S*);

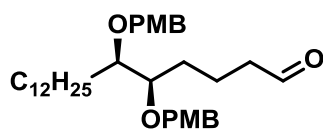
$[\alpha]_D^{20}$  -1.36 (c 2.1, CHCl<sub>3</sub>);

**<sup>1</sup>H NMR** (400 MHz, CDCl<sub>3</sub>)  $\delta$  7.28 – 7.22 (m, 4H), 6.89 – 6.84 (m, 4H), 5.90 – 5.78 (m, 1H), 5.12 – 5.01 (m, 2H), 4.58 – 4.43 (m, 4H), 3.80 (s, 6H), 3.50 (dt,  $J$  = 7.7, 4.6 Hz, 1H), 3.41 (dt,  $J$  = 8.7, 4.3 Hz, 1H), 2.44 – 2.36 (m, 1H), 2.30 – 2.19 (m, 1H), 1.63 – 1.53 (m, 1H), 1.49 – 1.34 (m, 2H), 1.26 (m, 20H), 0.92 – 0.86 (m, 3H);

**<sup>13</sup>C NMR** (100 MHz, CDCl<sub>3</sub>)  $\delta$  159.2, 135.8, 131.1, 131.0, 129.5, 129.5, 116.5, 113.7, 79.7, 79.6, 72.3, 72.1, 55.3, 34.7, 31.9, 29.9, 29.8, 29.7, 29.7, 29.7, 29.6, 29.4, 25.9, 22.7, 14.1;

**IR** (film, CH<sub>2</sub>Cl<sub>2</sub>); 3072, 2999, 2925, 2853, 2063, 1739, 1640, 1612, 1586, 1513, 1464, 1357, 1302, 1248, 1174, 1089, 1038, 912, 822 cm<sup>-1</sup>;

**HRMS** (ESI)  $m/z$ :  $[M + Na]^+$  Calcd for C<sub>33</sub>H<sub>50</sub>O<sub>4</sub>Na 533.3607; Found 533.3588.



**(5*R*,6*R*)-5,6-bis((4-methoxybenzyl)oxy)octadecanal (15).**

Mixture of catalyst Rh(CO)<sub>2</sub>acac (30 mg, 0.115 mmol, 10 mol%), ligand 6-DPPon (64 mg, 0.23 mmol, 20 mol%), additive TBAI (10.6 mg, 0.029 mmol, 2.5 mol%), 4 Å molecular sieves and anhydrous DMF (8 mL) was stirred in an ampule flushed with argon. Then **14** (585 mg, 1.15 mmol) dissolved in anhydrous DMF (8 mL) was added, stirred at rt for 5 min and Ac<sub>2</sub>O (653  $\mu$ L, 6.9 mmol, 6 equiv) and HCOOH (340  $\mu$ L, 8.97 mmol, 7.8 equiv) were successively added. The reaction mixture was stirred at 80 °C for 20 h. It was then cooled to rt and poured directly on silica gel. Column chromatography (hexane/ethyl acetate 9:1) afforded aldehyde **15** (378 mg, 61%) as a brownish oil.

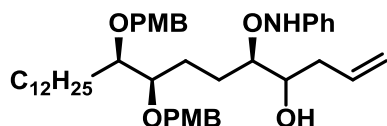
$[\alpha]_D^{20}$  13.15 (c 1.57, CHCl<sub>3</sub>);

**<sup>1</sup>H NMR** (400 MHz, C<sub>6</sub>D<sub>6</sub>) δ 9.29 (t, *J* = 1.6 Hz, 1H), 7.24 (dd, *J* = 13.1, 8.6 Hz, 4H), 6.79 (dd, *J* = 8.6, 5.3 Hz, 4H), 4.51 (dd, *J* = 11.4, 3.9 Hz, 2H), 4.41 (dd, *J* = 26.3, 11.4 Hz, 2H), 3.53 – 3.47 (m, 1H), 3.46 – 3.40 (m, 1H), 3.30 (s, 6H), 1.85 – 1.78 (m, 2H), 1.75 – 1.38 (m, 7H), 1.34 – 1.20 (m, 20H), 0.90 – 0.85 (m, 3H);

**<sup>13</sup>C NMR** (100 MHz, C<sub>6</sub>D<sub>6</sub>) δ 200.3, 159.4, 131.4, 131.2, 129.3, 129.3, 113.7, 79.5, 79.3, 72.1, 72.0, 54.4, 43.6, 32.0, 30.0, 29.9, 29.8, 29.8, 29.4, 29.3, 26.3, 22.7, 18.7, 14.0;

**IR** (film, CHCl<sub>3</sub>); 2999, 2925, 2853, 2717, 2062, 1724, 1612, 1586, 1513, 1463, 1359, 1302, 1248, 1174, 1070, 1037, 821cm<sup>-1</sup>;

**HRMS** (ESI) *m/z*: [M + Na]<sup>+</sup> Calcd for C<sub>34</sub>H<sub>52</sub>O<sub>5</sub>Na 563.3712; Found 563.3706.



**(5*R*,8*R*,9*R*)-8,9-bis((4-methoxybenzyl)oxy)-5-((phenylamino)oxy)henicos-1-en-4-ol (**16**).**

Prepared *via* adaptation of procedure from<sup>R7</sup>.

To a solution of aldehyde **15** (362 mg, 0.67 mmol, 3 equiv) in anhydrous DMF (2.5 mL), L-proline (77 mg, 0.67 mmol) was added and stirred at rt for 17 h. Then, nitrosobenzene (24 mg, 0.223 mmol, 1 equiv) was added. The endpoint of the reaction was monitored by its color change from green to orange. After 1.5 h, the solution was cooled to 0 °C and allylindium iodide solution (prepared by heating for 1 h at 70 °C: granular indium (73 mg, 0.67 mmol), NaI (95 mg, 0.67 mmol), allyl bromide (110 μL, 1.27 mmol) and anhydrous DMF (1 mL)) was slowly added. The stirring was kept at 0 °C for 2 h then at rt for 15 h. It was then diluted with Et<sub>2</sub>O and washed with water. The aqueous layer was extracted with Et<sub>2</sub>O. The combined organic layers were dried over Na<sub>2</sub>SO<sub>4</sub>, filtered, and concentrated *in vacuo*. Crude product was purified by column chromatography (hexane/acetone 85:15) to give inseparable 1:1 mixture of diastereoisomers **16** (95 mg, 62%) as yellow oil.

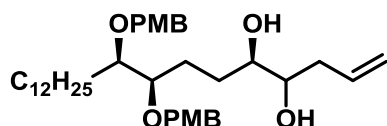
**<sup>1</sup>H NMR** (400 MHz, C<sub>6</sub>D<sub>6</sub>) δ 7.30 – 7.20 (m, 4H), 7.11 – 7.06 (m, 2H), 6.93 – 6.86 (m, 2H), 6.84 – 6.73 (m, 6H), 5.97 – 5.74 (m, 1H), 5.10 – 4.96 (m, 2H), 4.59 – 4.53 (m, 2H), 4.45 (ddd, *J* = 19.7, 11.4, 5.4 Hz, 2H), 3.99 – 3.93 (m, 1H), 3.81 – 3.73 (m, 1H), 3.62 – 3.53 (m, 2H), 3.32

– 3.25 (m, 6H), 2.59 – 2.49 (m, 1H), 2.40 – 2.30 (m, 1H), 2.29 – 2.14 (m, 2H), 2.10 – 1.91 (m, 2H), 1.86 – 1.70 (m, 2H), 1.49 – 1.37 (m, 1H), 1.35 – 1.17 (m, 20H), 0.91 – 0.82 (m, 3H);

$^{13}\text{C}$  NMR (100 MHz,  $\text{C}_6\text{D}_6$ )  $\delta$  159.4, 159.4, 149.1, 149.0, 135.5, 135.3, 131.4, 131.2, 129.4, 129.4, 129.3, 128.8, 121.8, 121.8, 116.9, 114.8, 114.7, 113.8, 113.8, 113.7, 85.8, 85.4, 79.7, 72.4, 72.2, 72.1, 71.6, 59.7, 54.4, 37.6, 32.0, 30.0, 29.8, 29.8, 29.4, 26.3, 24.4, 22.7, 14.0;

IR (film,  $\text{CH}_2\text{Cl}_2$ ); 3446, 3267, 2999, 2925, 2853, 2062, 1688, 1640, 1611, 1513, 1494, 1464, 1358, 1302, 1248, 1174, 1037, 914, 821, 762  $\text{cm}^{-1}$ ;

HRMS (ESI)  $m/z$ :  $[\text{M} + \text{Na}]^+$  Calcd for  $\text{C}_{43}\text{H}_{63}\text{NO}_6\text{Na}$  712.4553; Found 712.4542.



**(5R,8R,9R)-8,9-bis((4-methoxybenzyl)oxy)henicos-1-ene-4,5-diol (SI-15)**. Prepared under conditions from<sup>R8</sup>.

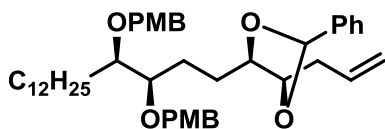
Diastereoisomers **16** (90 mg, 0.13 mmol) were dissolved in a 1:1 THF/H<sub>2</sub>O mixture (2 mL). Acetic acid (3 mL) and Zn dust (296 mg, 4.48 mmol) was added. The mixture was stirred at 60 °C for 1 h. After cooling to rt, the mixture was diluted with Et<sub>2</sub>O and filtered through a plug of silica gel, which was washed by Et<sub>2</sub>O. After evaporation of the solvents, the resulting crude mixture was redissolved in EtOAc, preadsorbed on silica gel and purified by column chromatography (hexane/ethyl acetate 3:2) to give **SI-15** (77 mg, 99%) as a colorless oil.

$^1\text{H}$  NMR (400 MHz,  $\text{CDCl}_3$ )  $\delta$  7.26 – 7.19 (m, 4H), 6.89 – 6.83 (m, 4H), 5.89 – 5.77 (m, 1H), 5.17 – 5.08 (m, 2H), 4.58 – 4.40 (m, 4H), 3.79 (s, 6H), 3.59 – 3.32 (m, 4H), 2.36 (s, 2H), 2.31 – 2.10 (m, 2H), 1.80 – 1.66 (m, 1H), 1.63 – 1.48 (m, 3H), 1.48 – 1.36 (m, 3H), 1.26 (s, 23H), 0.91 – 0.85 (m, 3H);

$^{13}\text{C}$  NMR (100 MHz,  $\text{CDCl}_3$ )  $\delta$  159.3, 159.2, 135.1, 134.7, 130.9, 130.5, 129.7, 129.7, 129.5, 118.0, 117.8, 113.8, 113.8, 79.7, 79.4, 79.4, 74.0, 73.4, 73.2, 72.3, 72.2, 55.3, 38.3, 36.2, 31.9, 29.8, 29.7, 29.7, 29.7, 29.6, 29.4, 27.9, 26.0, 22.7, 14.1;

IR (film,  $\text{CH}_2\text{Cl}_2$ ); 3425, 3073, 2998, 2925, 2853, 2063, 1881, 1708, 1640, 1612, 1586, 1514, 1465, 1358, 1302, 1249, 1210, 1173, 1065, 1038, 914, 822, 755  $\text{cm}^{-1}$ ;

**HRMS** (ESI)  $m/z$ :  $[M + Na]^+$  Calcd for  $C_{37}H_{58}O_6Na$  621.4131; Found 621.4144.



**(2*S*,4*R*,5*R*)-4-allyl-5-((3*R*,4*R*)-3,4-bis((4-methoxybenzyl)oxy)hexadecyl)-2-phenyl-1,3-dioxolane (SI-16)**. Prepared under conditions from<sup>R9</sup>.

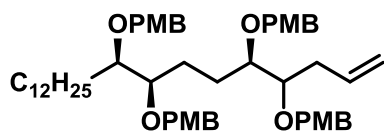
Mixture of diastereoisomers **SI-15** (20 mg, 0.033 mmol) was dissolved in PhH (0.35 mL) with camphorsulfonic acid (10 mg, 0.033 mmol) and 4 Å molecular sieves. Solution was cooled to 0 °C and PhCHO (85 µL, 0.825 mmol) was added. The mixture was stirred at 0 °C for 1 h. It was then warmed to rt and poured directly on silica gel. Column chromatography (hexane/ethyl acetate 95:5) afforded *threo/threo* diastereoisomer **SI-16** (8.5 mg, 37%) as a colorless oil. Structure was confirmed by NOE experiment.

**<sup>1</sup>H NMR** (600 MHz,  $C_6D_6$ )  $\delta$  7.35 – 7.32 (m, 2H), 7.02 (d,  $J = 8.6$  Hz, 2H), 6.97 (d,  $J = 8.6$  Hz, 2H), 6.90 – 6.88 (m, 2H), 6.84 – 6.80 (m, 1H), 6.53 (dd,  $J = 8.6, 7.8$  Hz, 4H), 5.67 – 5.58 (m, 1H), 5.50 (s, 1H), 4.83 – 4.78 (m, 1H), 4.77 – 4.74 (m, 1H), 4.32 (dd,  $J = 11.4, 4.4$  Hz, 2H), 4.21 (dd,  $J = 39.4, 11.4$  Hz, 2H), 3.69 – 3.64 (m, 1H), 3.63 – 3.59 (m, 1H), 3.33 (m, 2H), 3.03 (s, 6H), 2.10 – 2.04 (m, 1H), 1.82 – 1.77 (m, 1H), 1.73 – 1.64 (m, 1H), 1.63 – 1.47 (m, 4H), 1.41 – 1.30 (m, 3H), 1.21 – 1.11 (m, 3H), 1.06 – 0.98 (m, 20H), 0.62 (m, 3H);

**<sup>13</sup>C NMR** (150 MHz,  $C_6D_6$ )  $\delta$  159.4, 159.4, 138.7, 135.1, 131.4, 131.3, 129.5, 129.3, 128.8, 128.1, 127.9, 127.7, 127.6, 126.9, 116.5, 113.7, 103.1, 79.9, 78.8, 78.7, 78.4, 72.1, 71.8, 54.4, 54.4, 34.7, 32.0, 30.0, 29.8, 29.8, 29.8, 29.4, 26.7, 26.3, 26.2, 22.7, 14.0;

**IR** (film,  $CH_2Cl_2$ ); 3464, 3070, 2925, 2853, 1745, 1641, 1612, 1586, 1513, 1462, 1376, 1301, 1248, 1173, 1090, 1066, 1037, 916, 821, 758  $cm^{-1}$ ;

**HRMS** (ESI)  $m/z$ :  $[M + Na]^+$  Calcd for  $C_{44}H_{62}O_6Na$  709.4443; Found 709.4444.



**4,4',4'',4'''-((((5*R*,8*R*,9*R*)-henicos-1-ene-4,5,8,9-tetrayl)tetrakis(oxy))tetrakis(methylene))tetrakis(methoxybenzene) (17).**

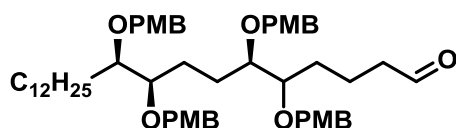
NaH (60% dispersion in mineral oil, 15 mg, 0.368 mmol, 4 equiv) was added to a flask containing **SI-14** (55 mg, 0.092 mmol) and TBAI (1.7 mg, 4.6 μmol, 5 mol%) in anhydrous DMF (1 mL) cooled to 0 °C. Reaction mixture was stirred at 0 °C for 30 min. Then, PMBCl (50 μL, 0.368 mmol, 4 equiv) was added dropwise and the mixture was stirred at rt for 17 h. Saturated NH<sub>4</sub>Cl was added to quench the reaction followed by the extraction with EtOAc. The combined organic layers were dried over Na<sub>2</sub>SO<sub>4</sub>, filtered, and concentrated *in vacuo*. Crude product was purified by column chromatography (hexane/ethyl acetate 95:5) to give **17** (68 mg, 88%) as a colorless oil.

<sup>1</sup>H NMR (400 MHz, CDCl<sub>3</sub>) δ 7.27 – 7.18 (m, 8H), 6.87 – 6.79 (m, 8H), 5.95 – 5.76 (m, 1H), 5.14 – 4.99 (m, 2H), 4.63 – 4.39 (m, 8H), 3.78 (s, 12H), 3.53 – 3.36 (m, 4H), 2.45 – 2.17 (m, 2H), 1.88 – 1.53 (m, 5H), 1.48 – 1.38 (m, 2H), 1.33 – 1.23 (m, 19H), 0.92 – 0.85 (m, 3H);

<sup>13</sup>C NMR (100 MHz, CDCl<sub>3</sub>) δ 159.1, 129.5, 129.4, 129.3, 113.7, 71.9, 55.2, 31.9, 30.7, 29.8, 29.7, 29.7, 29.4, 22.7, 14.1;

IR (film, CH<sub>2</sub>Cl<sub>2</sub>); 3070, 2999, 2925, 2853, 2062, 1881, 1612, 1586, 1513, 1464, 1356, 1301, 1248, 1173, 1087, 1037, 913, 821, 756 cm<sup>-1</sup>;

HRMS (ESI) *m/z*: [M + Na]<sup>+</sup> Calcd for C<sub>53</sub>H<sub>74</sub>O<sub>8</sub>Na 861.5281; Found 861.5273.



**(6*R*,9*R*,10*R*)-5,6,9,10-tetrakis((4-methoxybenzyl)oxy)docosanal (18).**

Mixture of catalyst Rh(CO)<sub>2</sub>acac (1.8 mg, 7.2 μmol, 10 mol%), ligand 6-DPPon (4 mg, 14.4 μmol, 20 mol%), additive TBAI (0.6 mg, 1.8 μmol, 2.5 mol%), 4 Å molecular sieves and anhydrous DMF (0.6 mL) was stirred in an ampule flushed with argon. Then **17** (60 mg, 0.072 mmol) dissolved in anhydrous DMF (0.6 mL) was added, stirred at rt for 5 min and Ac<sub>2</sub>O (38.4

$\mu\text{L}$ , 0.432 mmol, 6 equiv) and  $\text{HCOOH}$  (20  $\mu\text{L}$ , 0.562 mmol, 7.8 equiv) were successively added. The reaction mixture was stirred at 80 °C for 20 h. It was then cooled to rt and poured directly on silica gel. Column chromatography (hexane/ethyl acetate 4:1) afforded aldehyde **18** (24 mg, 39%) as a brownish oil.

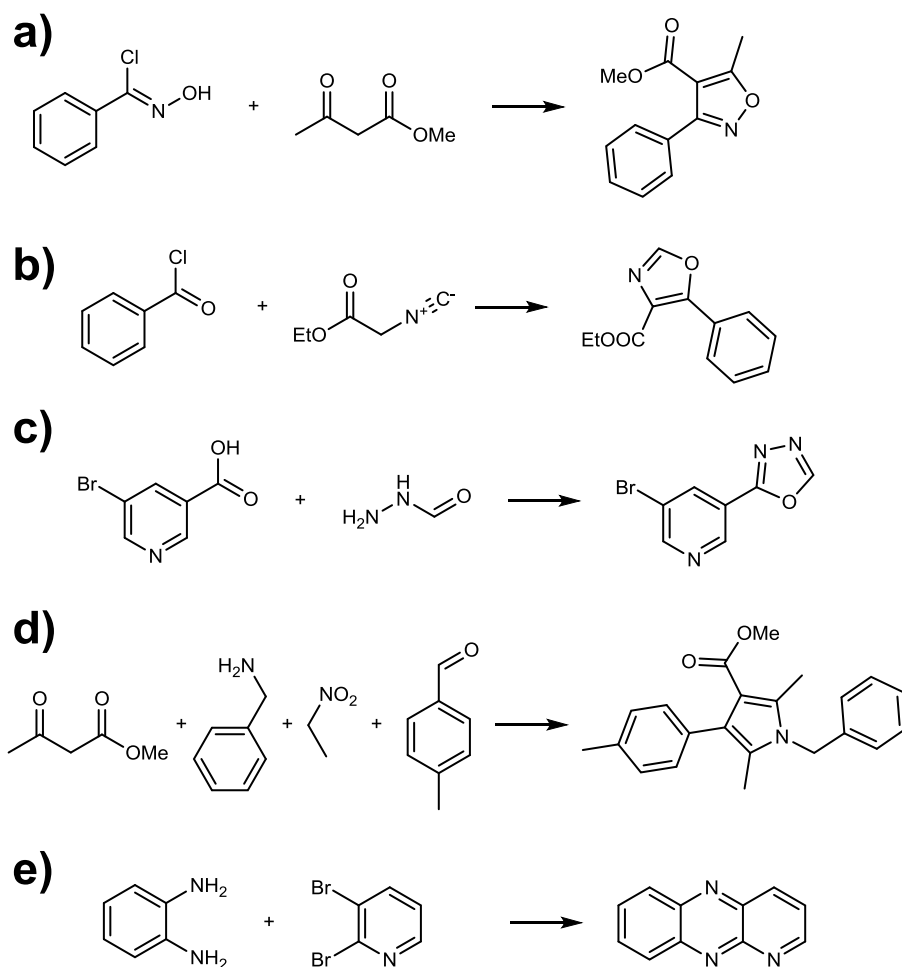
**$^1\text{H}$  NMR** (600 MHz,  $\text{C}_6\text{D}_6$ )  $\delta$  9.28 – 9.25 (m, 1H), 7.32 – 7.19 (m, 8H), 6.81 – 6.74 (m, 8H), 4.68 – 4.55 (m, 3H), 4.53 – 4.42 (m, 4H), 4.38 – 4.32 (m, 1H), 3.63 – 3.49 (m, 3H), 3.45 – 3.38 (m, 1H), 3.30 – 3.25 (m, 12H), 2.14 – 2.02 (m, 1H), 1.99 – 1.82 (m, 4H), 1.82 – 1.77 (m, 2H), 1.67 – 1.56 (m, 4H), 1.51 – 1.41 (m, 2H), 1.34 – 1.20 (m, 19H), 0.89 – 0.83 (m, 3H);

**$^{13}\text{C}$  NMR** (150 MHz,  $\text{C}_6\text{D}_6$ )  $\delta$  200.4, 200.4, 159.3, 131.5, 131.4, 131.4, 131.2, 129.3, 129.3, 129.3, 129.2, 129.2, 127.9, 113.7, 113.7, 113.7, 81.1, 80.9, 80.6, 80.5, 80.2, 80.1, 79.3, 72.2, 72.0, 71.8, 71.8, 71.7, 71.6, 54.4, 43.6, 43.5, 32.0, 31.0, 30.1, 30.0, 29.9, 29.8, 29.8, 29.8, 29.5, 29.3, 27.5, 27.3, 27.2, 26.5, 26.2, 26.1, 22.7, 18.7, 18.5, 14.0;

**IR** (film,  $\text{CH}_2\text{Cl}_2$ ); 2999, 2925, 2853, 2061, 1882, 1724, 1612, 1586, 1513, 1463, 1356, 1301, 1248, 1174, 1087, 1037, 821, 756  $\text{cm}^{-1}$ ;

**HRMS** (ESI)  $m/z$ :  $[\text{M} + \text{Na}]^+$  Calcd for  $\text{C}_{54}\text{H}_{76}\text{O}_9\text{Na}$  891.5387; Found 891.5389.

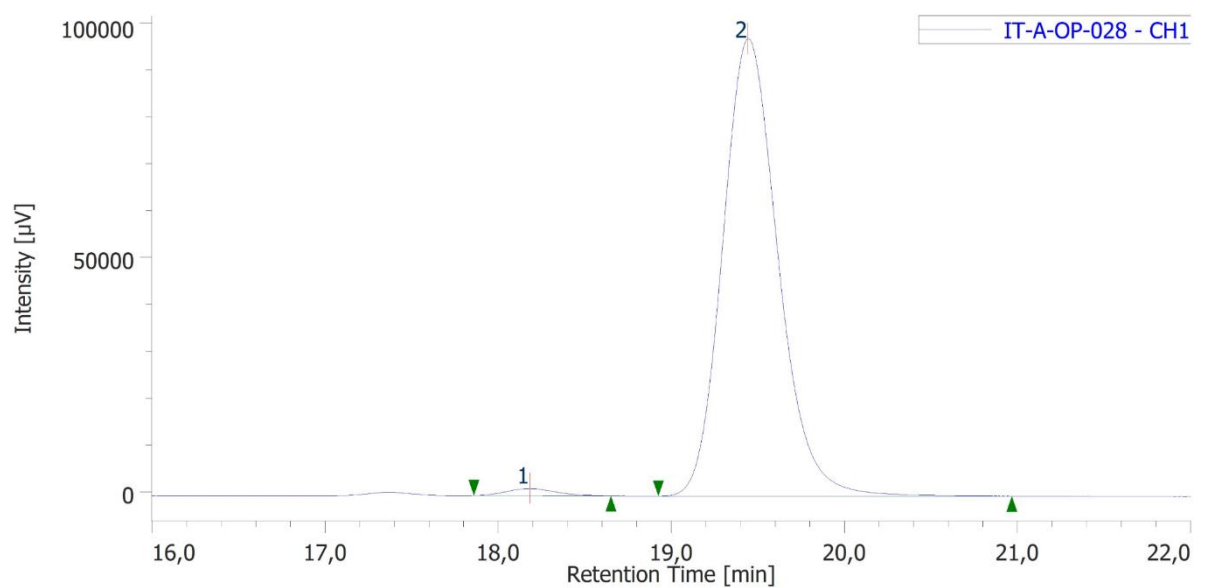
## Section S10. Literature precedents of heterocycle-forming reactions.



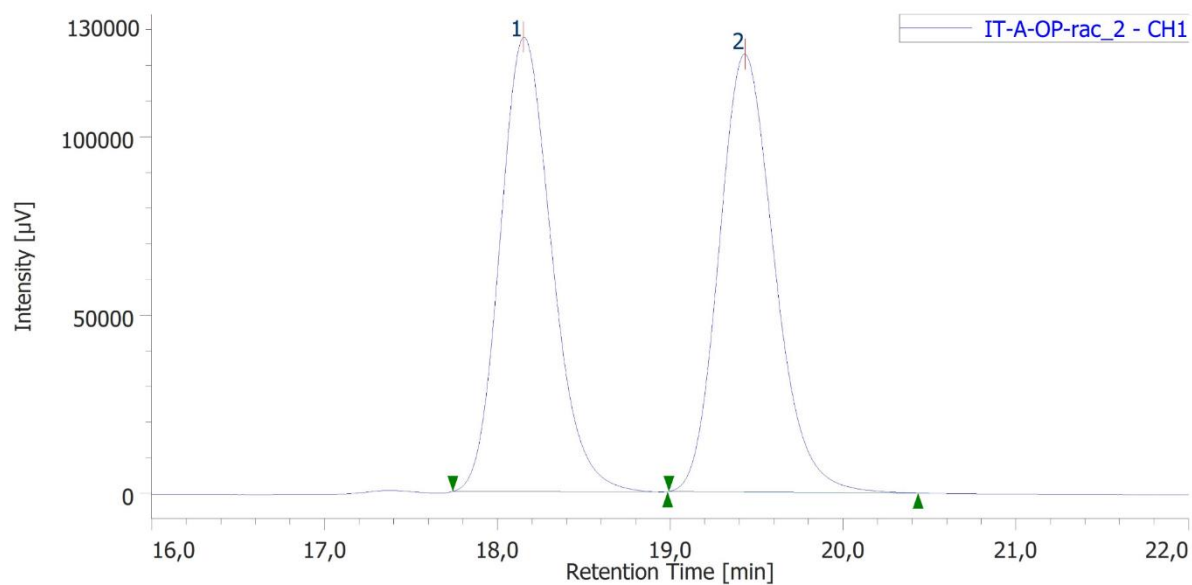
**Figure S9.** Literature precedents of heterocycle-forming reactions that are part of iterative sequences from main-text Figure 3. **a)** Synthesis of isoxazoles via condensation of imidoyl chlorides with active methylene compounds<sup>R10</sup>; **b)** Synthesis of oxazoles from isocyanides<sup>R11</sup>; **c)** Synthesis of 1,3,4-oxadiazoles from *N*-formyl hydrazine<sup>R12</sup>; **d)** Four-component synthesis of pyrroles<sup>R13</sup>; **e)** Synthesis of phenazines from dibromoarenes<sup>R14</sup>.



## Section S11. Spectroscopic data



#	Peak Name	CH	tR [min]	Area [µV·sec]	Height [µV]	Area%	Height%	Quantity	NTP	Resolution	Symmetry Factor	Warning
1	Unknown	1	18,183	29473	1519	1,343	1,534	N/A	19798	2,312	1,148	
2	Unknown	1	19,442	2165495	97501	98,657	98,466	N/A	18334	N/A	1,188	



#	Peak Name	CH	tR [min]	Area [µV·sec]	Height [µV]	Area%	Height%	Quantity	NTP	Resolution	Symmetry Factor	Warning
1	Unknown	1	18,150	2643690	127332	49,211	50,931	N/A	17984	2,297	1,194	
2	Unknown	1	19,433	2728429	122679	50,789	49,069	N/A	18026	N/A	1,174	

**Figure S10.** HPLC chromatogram of compound **SI-1** (top) and racemate of **SI-1** (bottom).

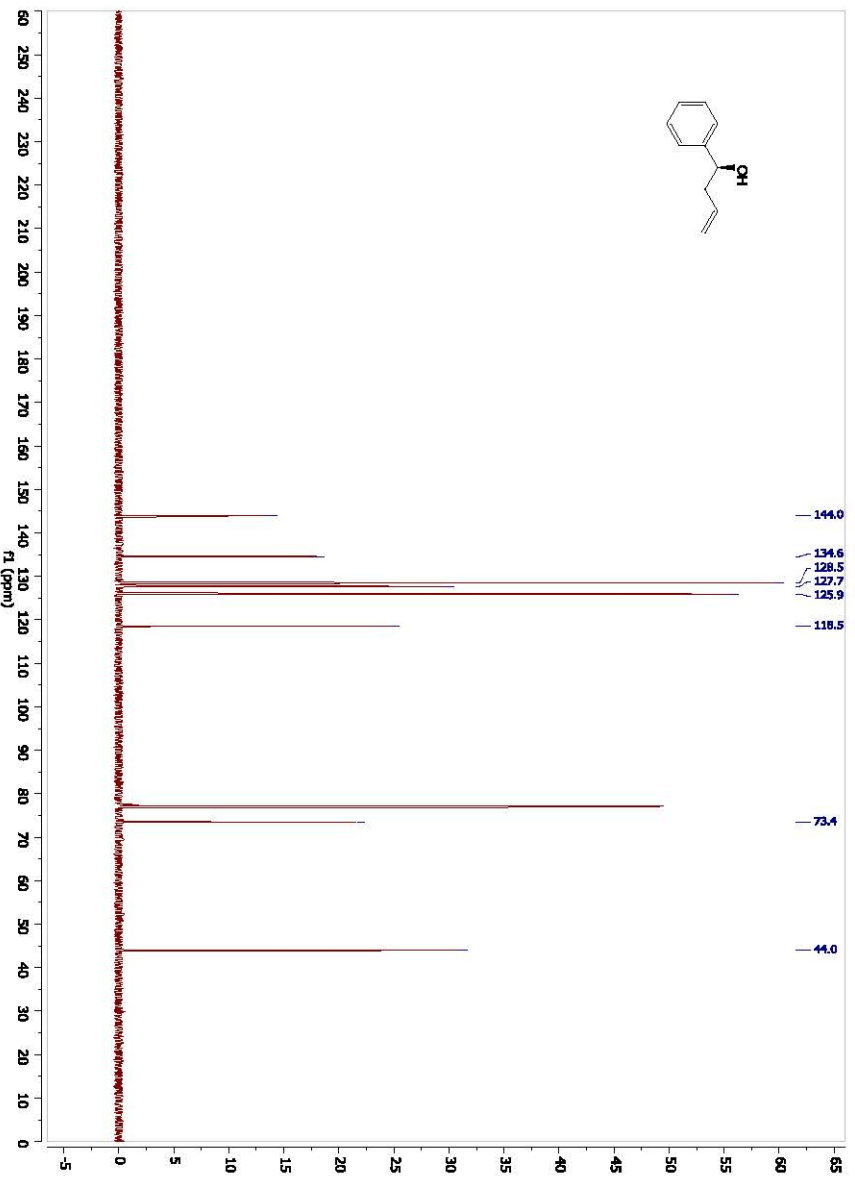
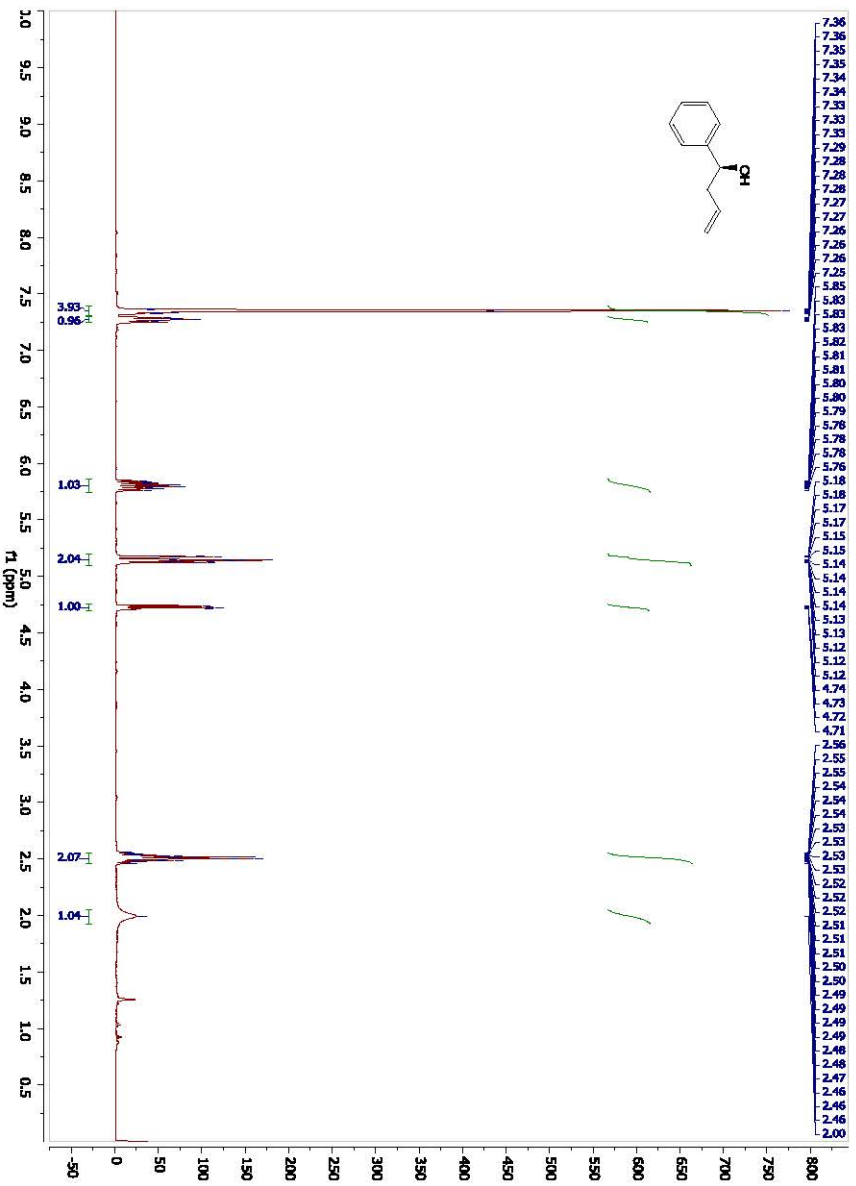


Figure S11. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound SI-1.

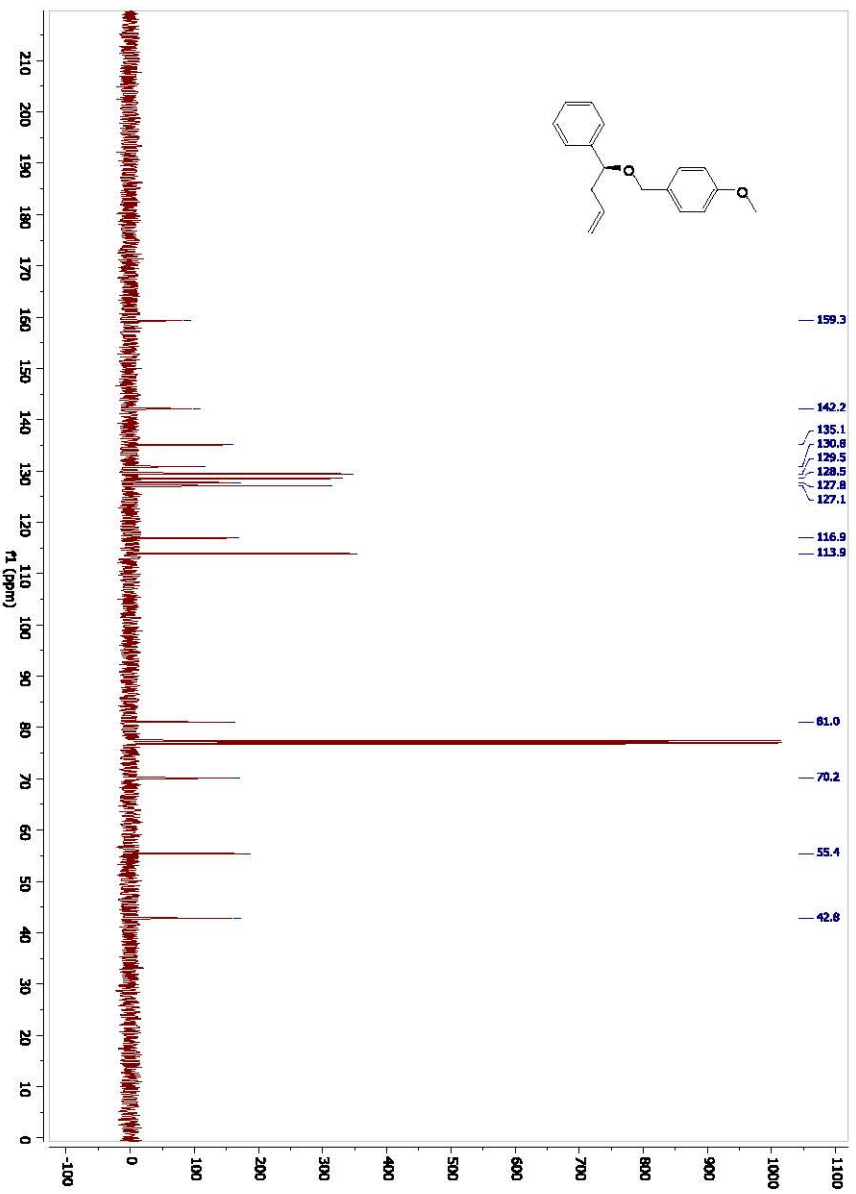
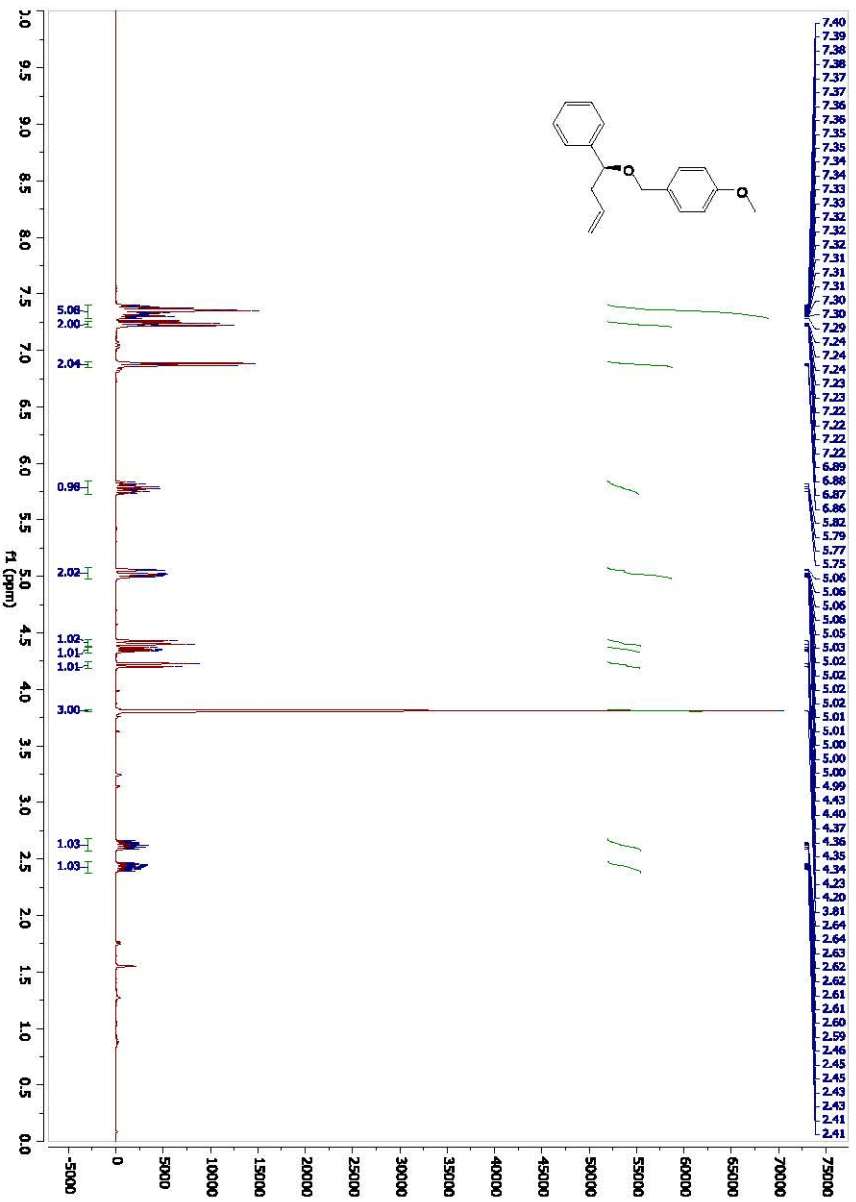
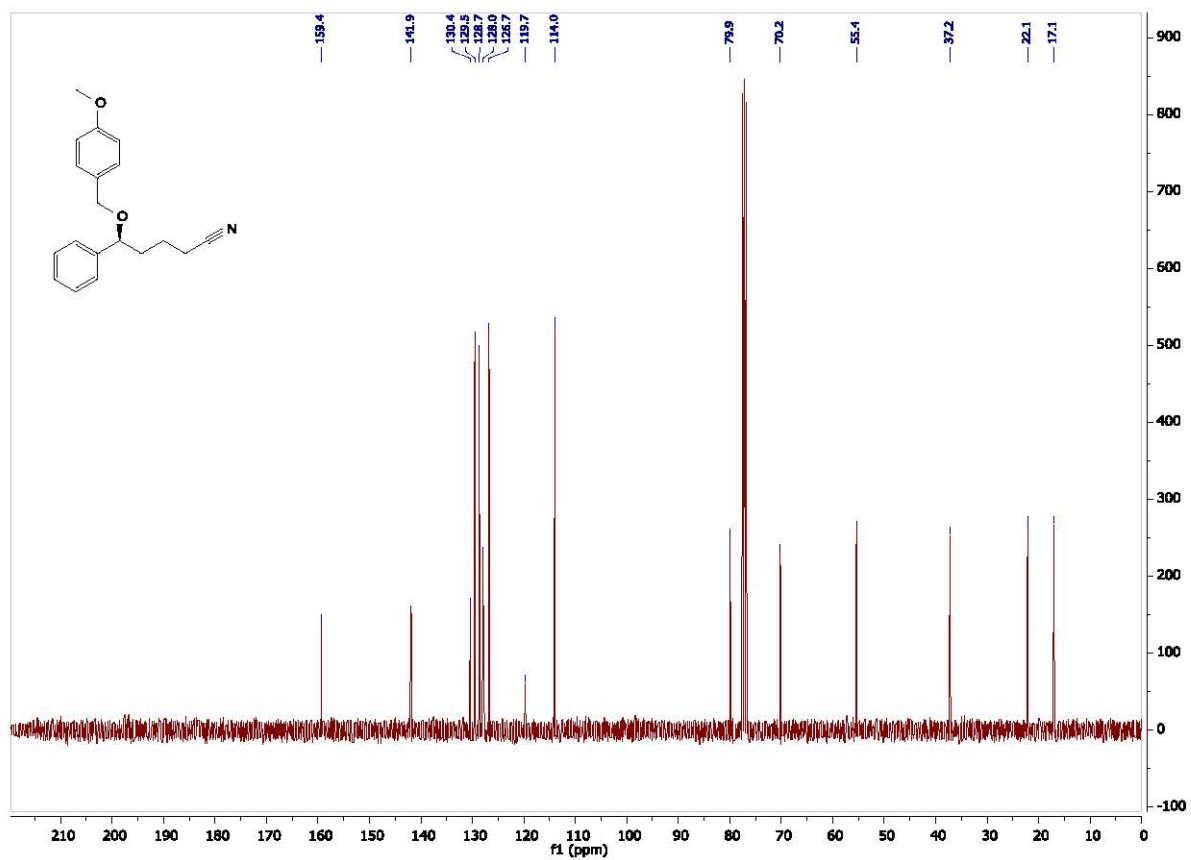
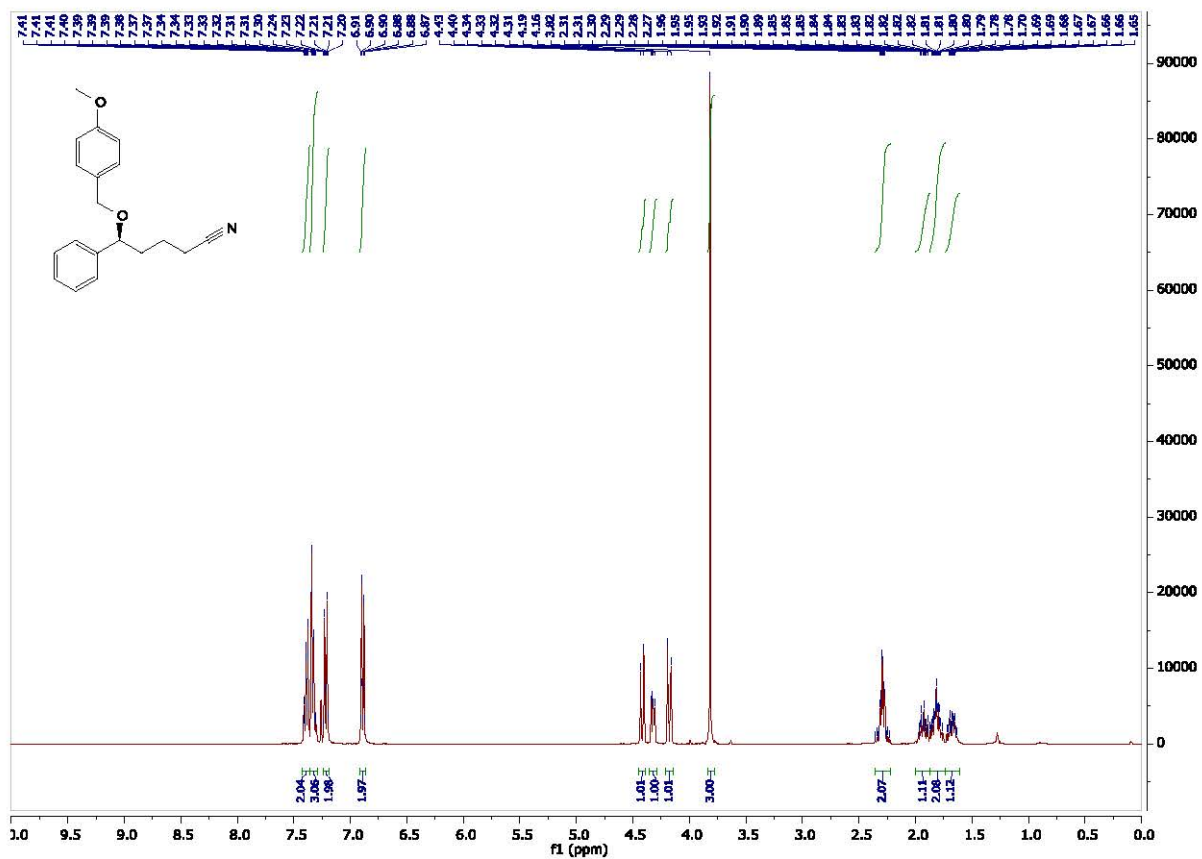


Figure S12. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound 1.



**Figure S13.** <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound **2**.

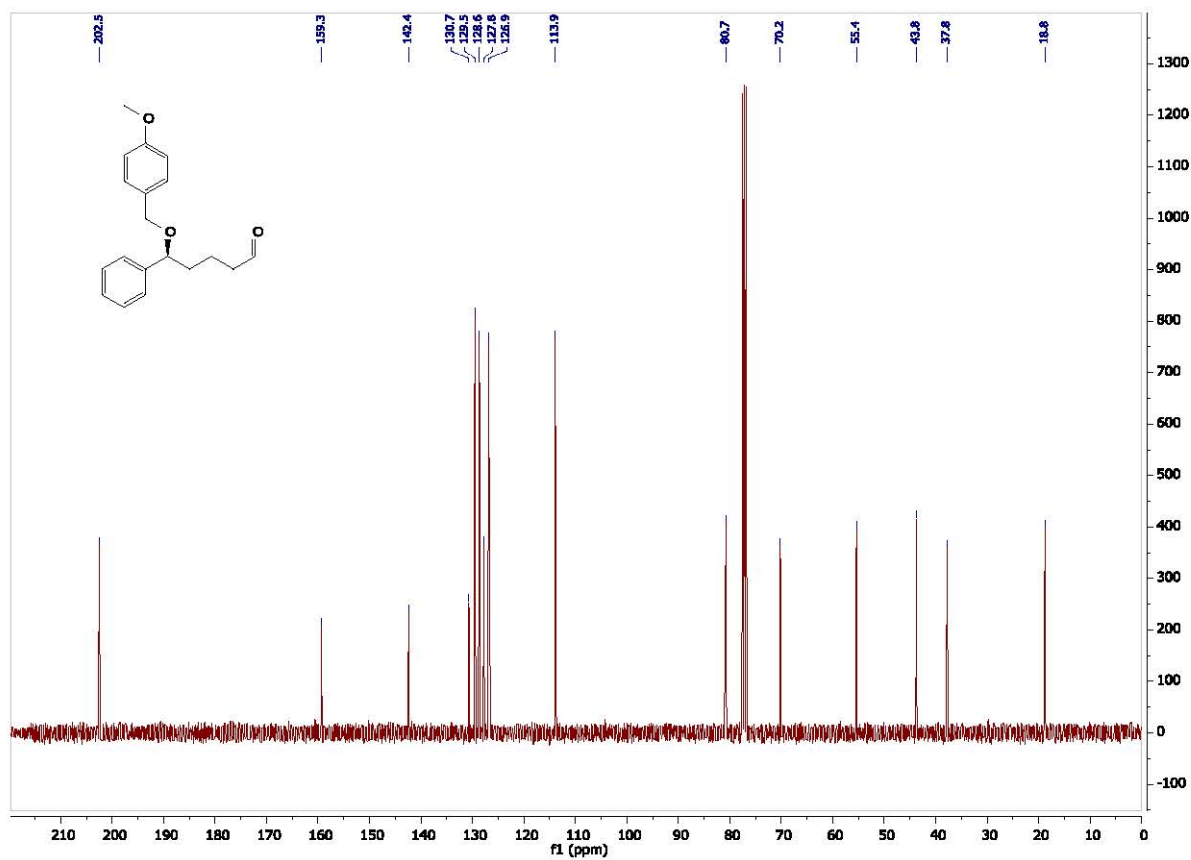
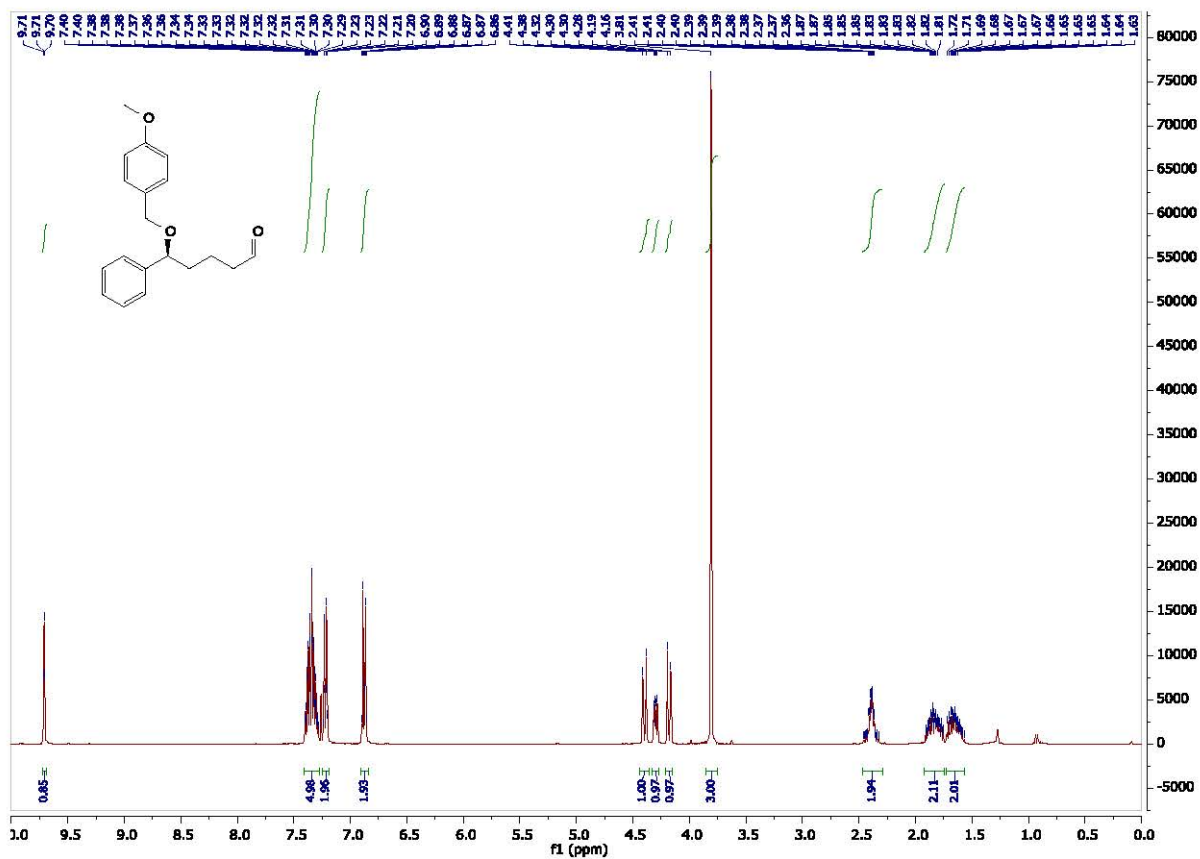


Figure S14. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound 3.

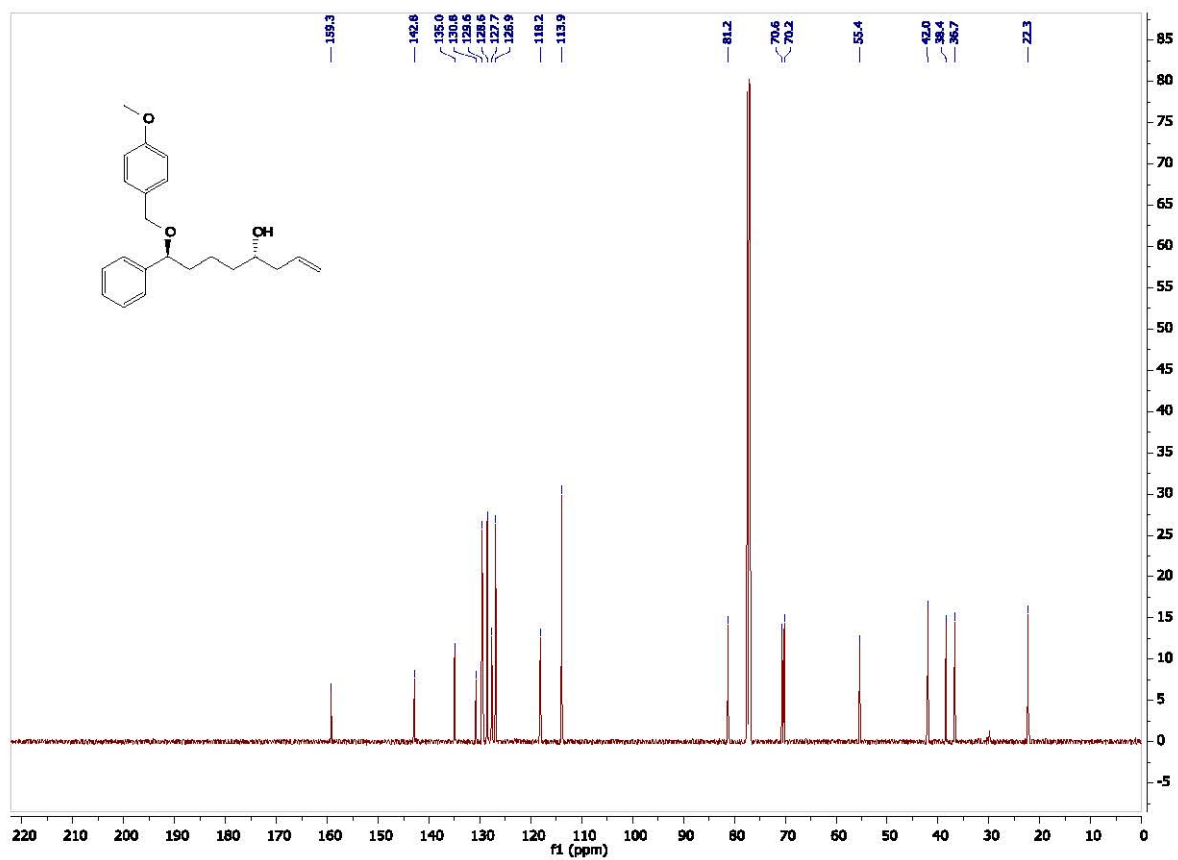
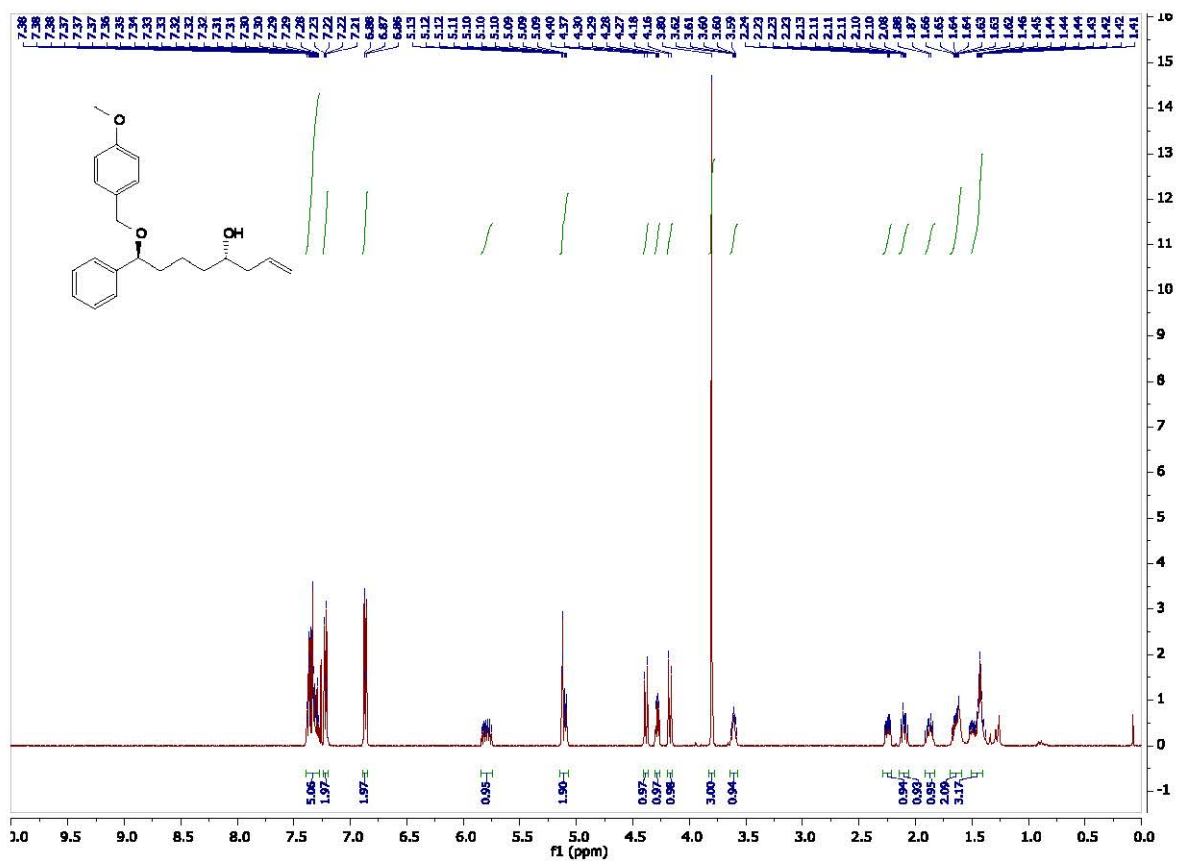
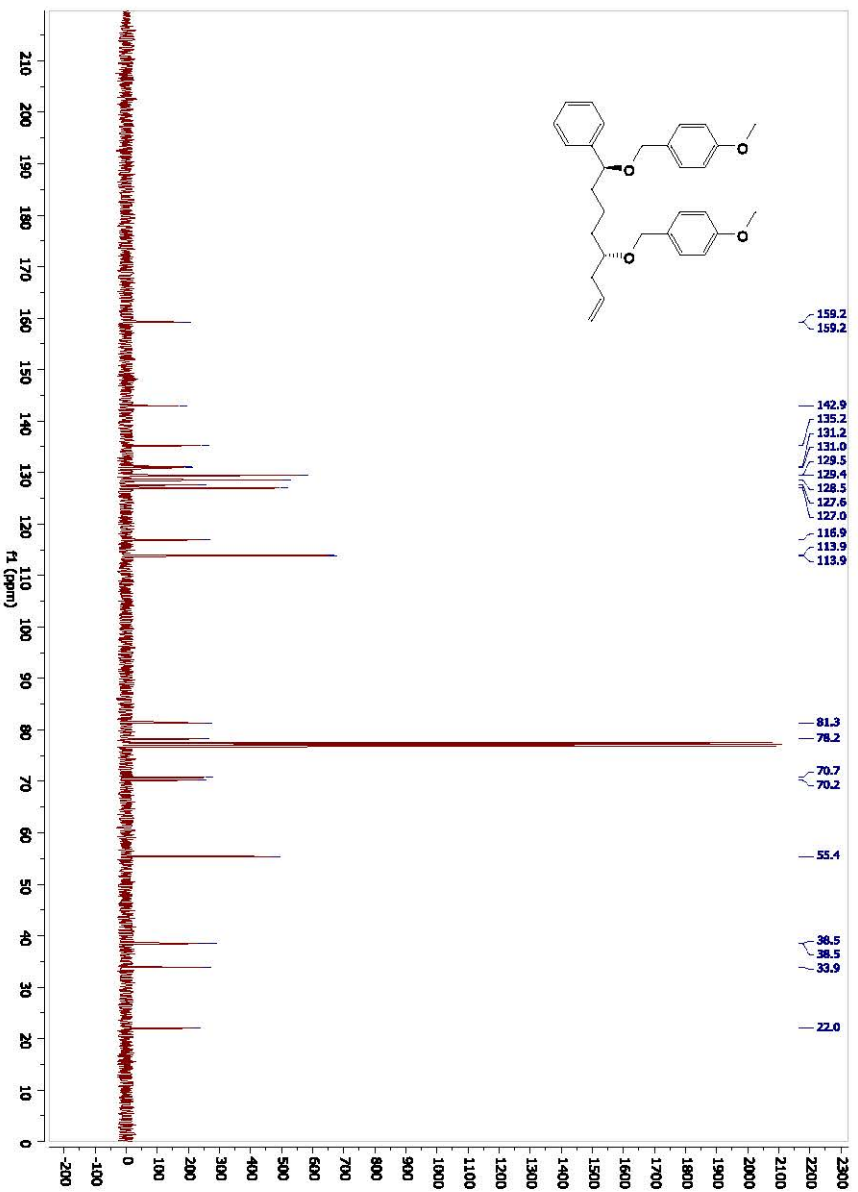
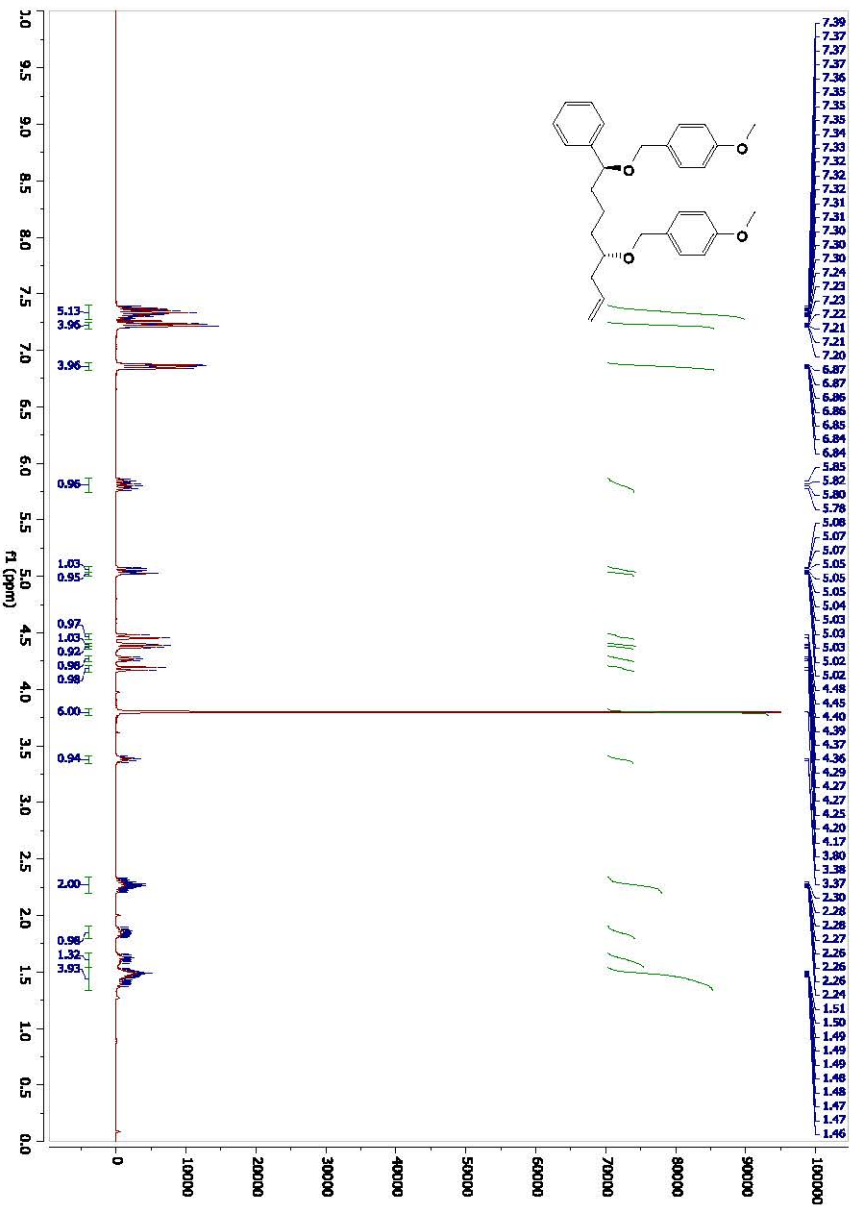


Figure S15. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound SI-2.



**Figure S16.** <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound 4.

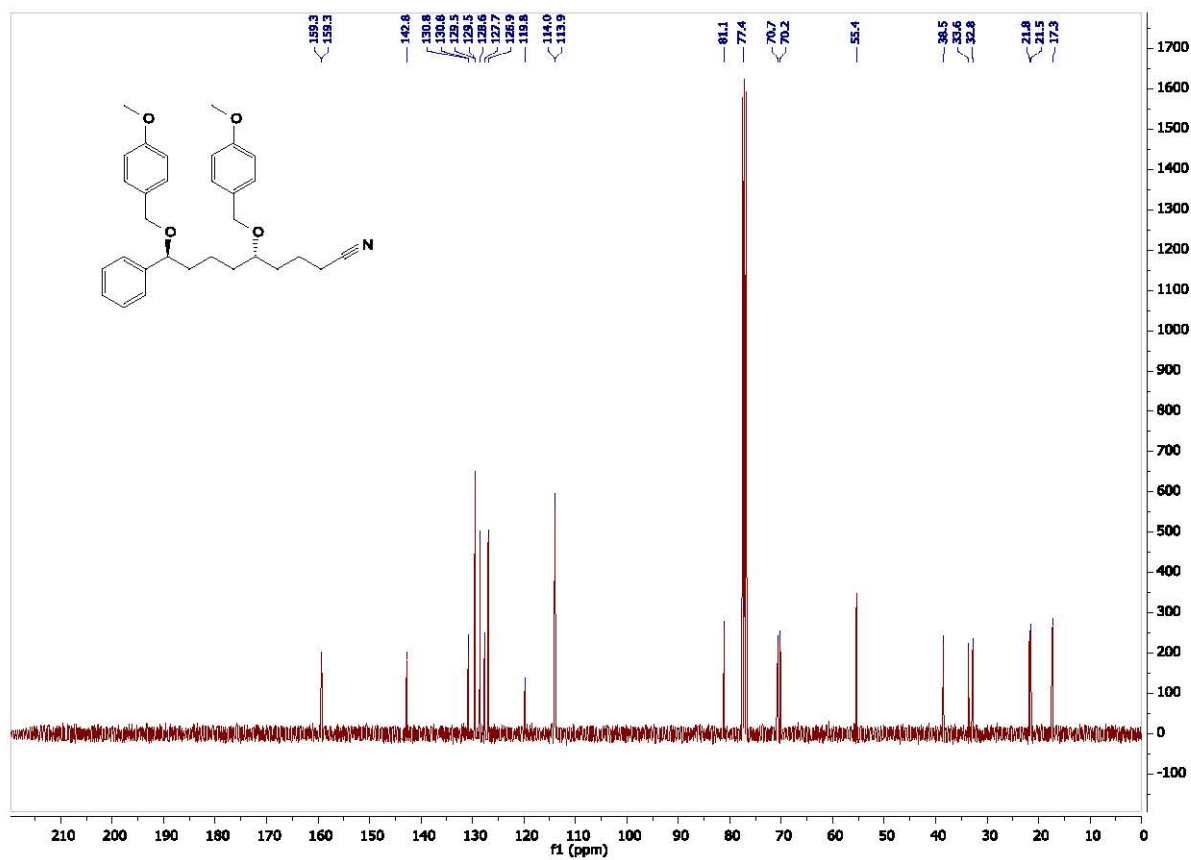
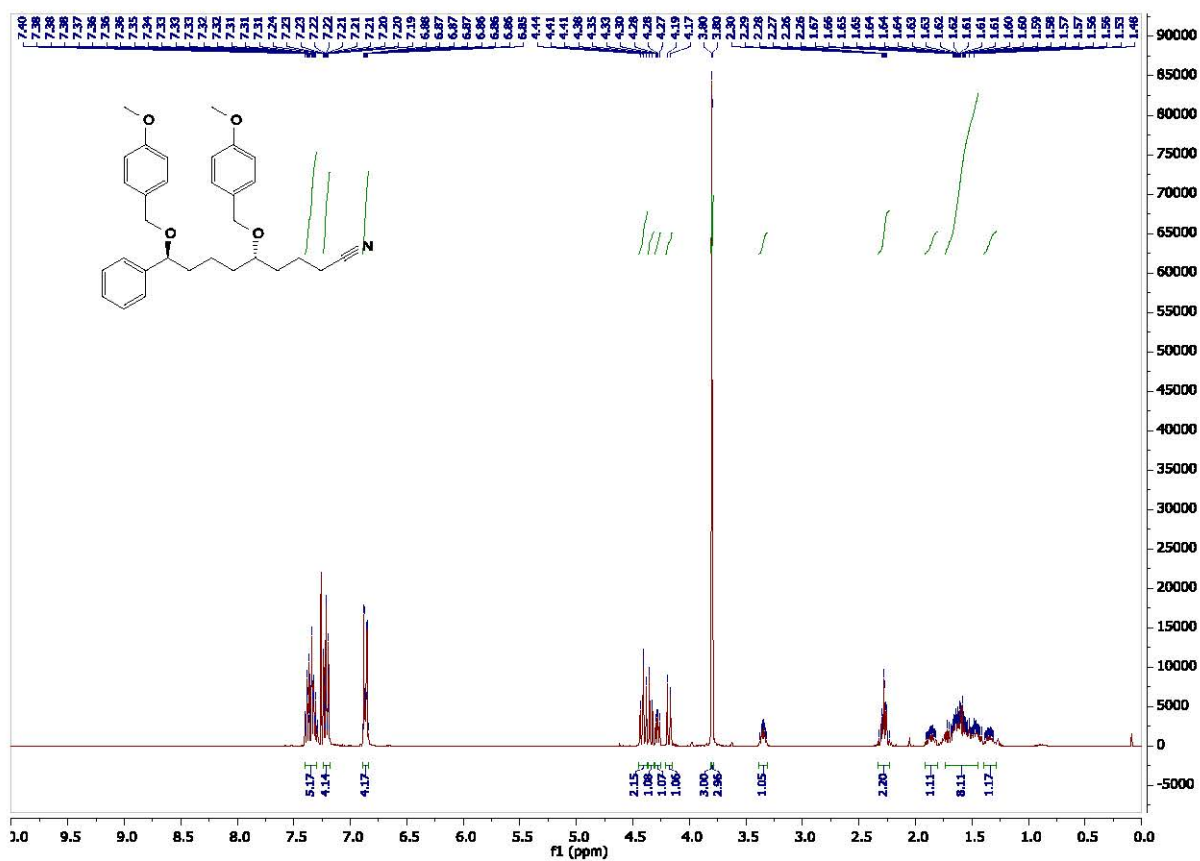


Figure S17.  $^1\text{H}$  NMR (top) and  $^{13}\text{C}$  NMR (bottom) spectra of compound 5.



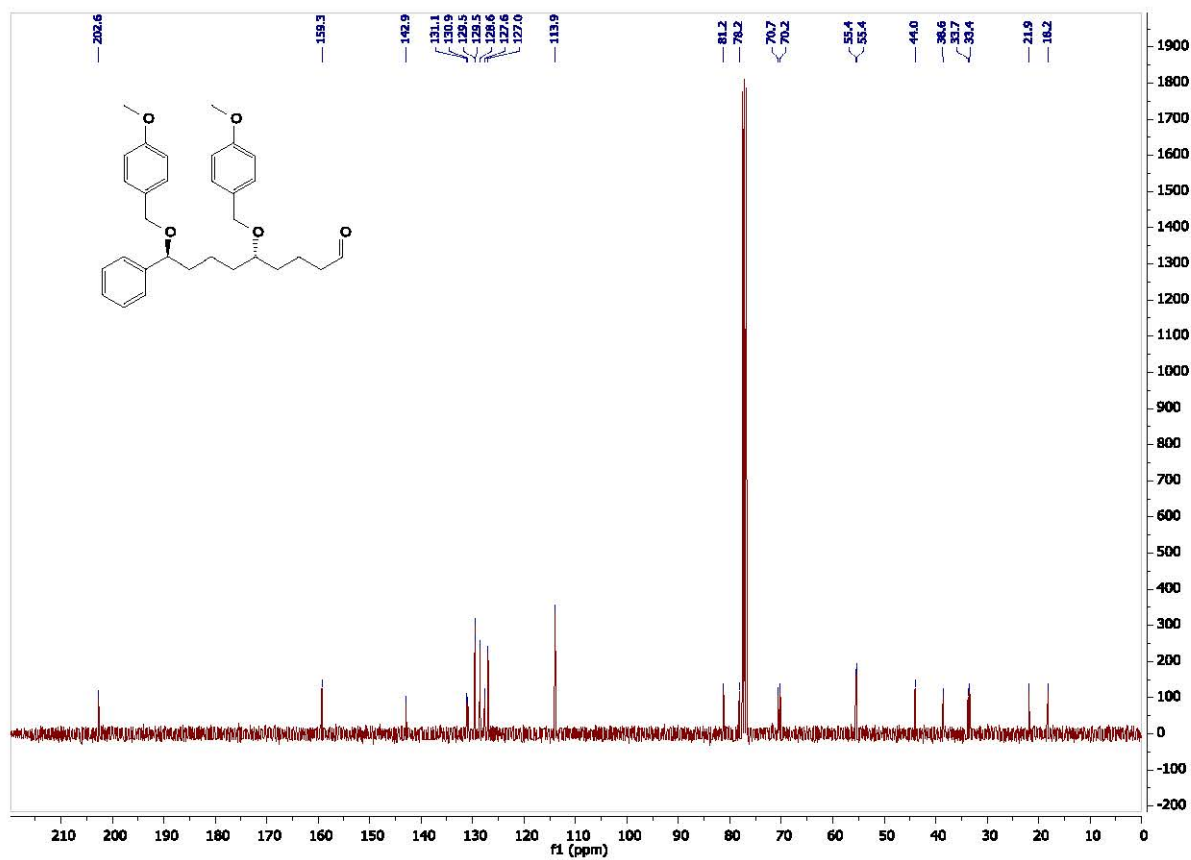
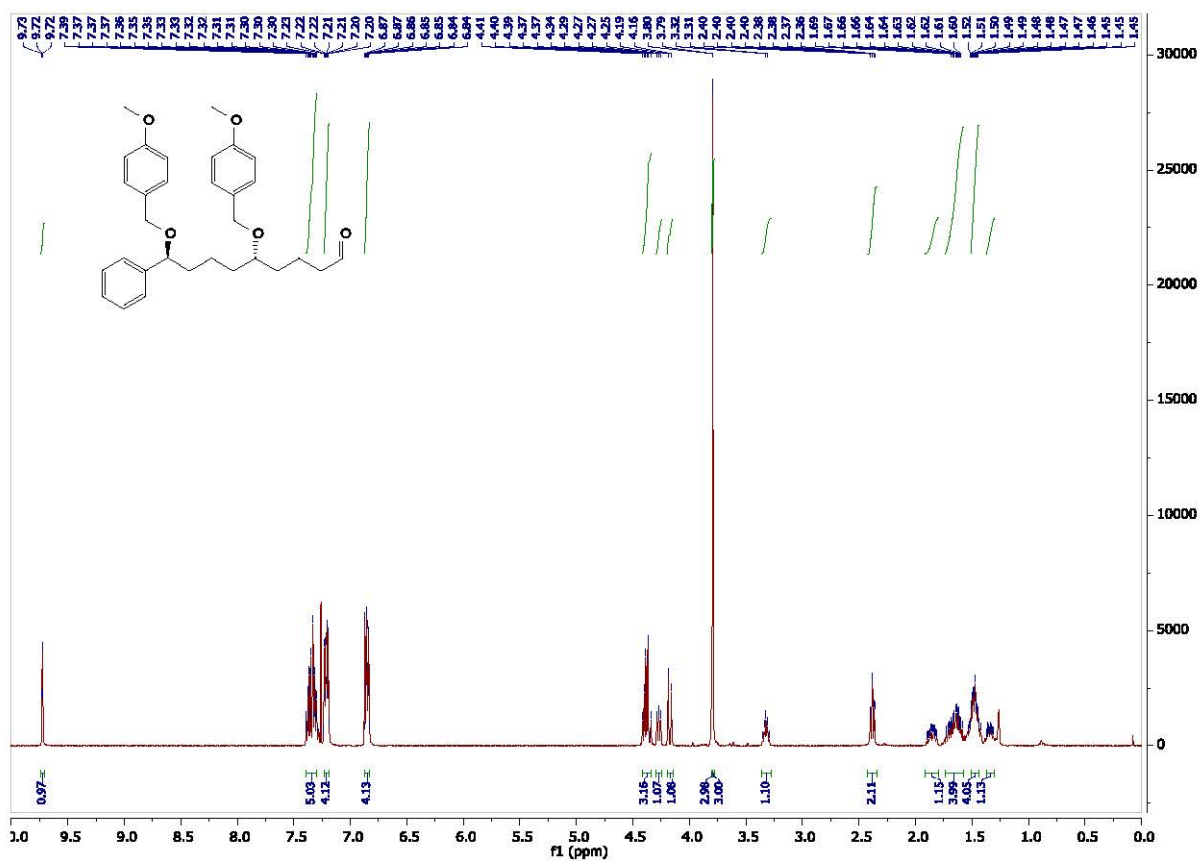


Figure S18. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound 6.

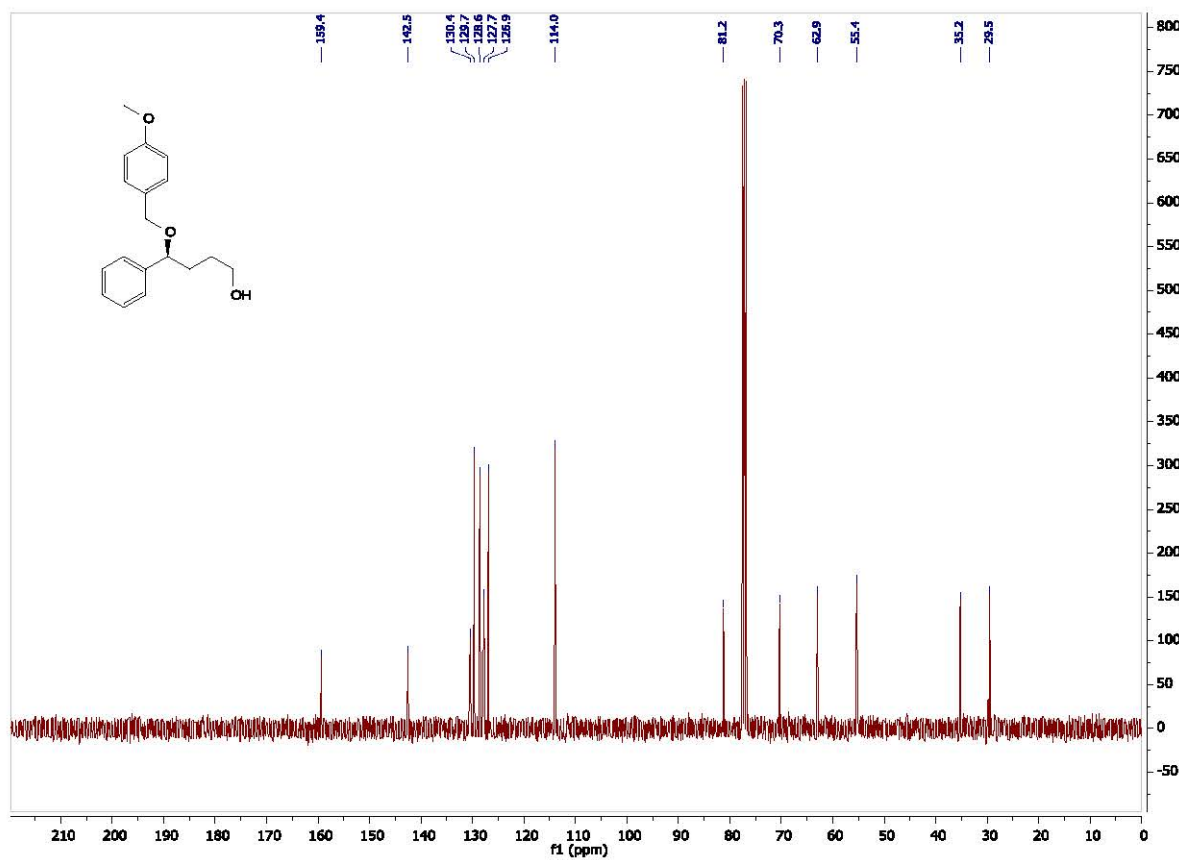
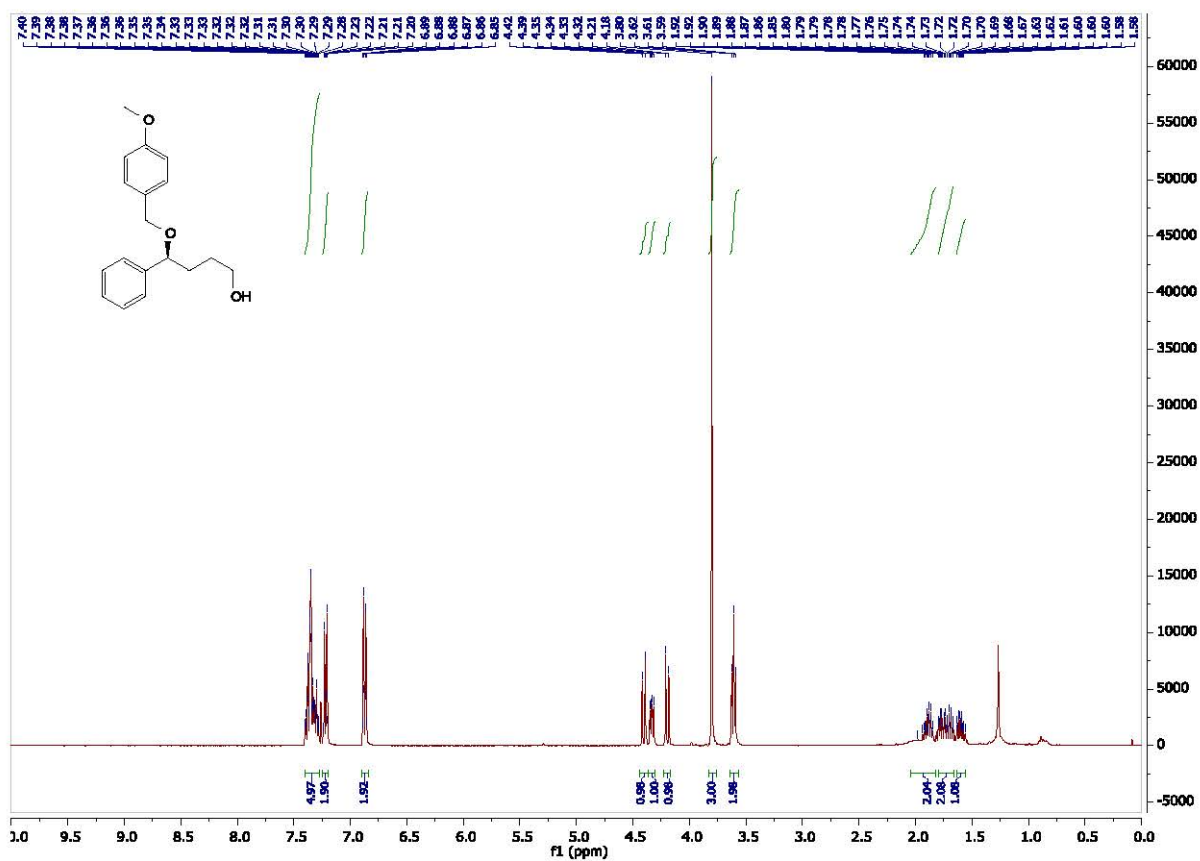


Figure S19. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound 7.

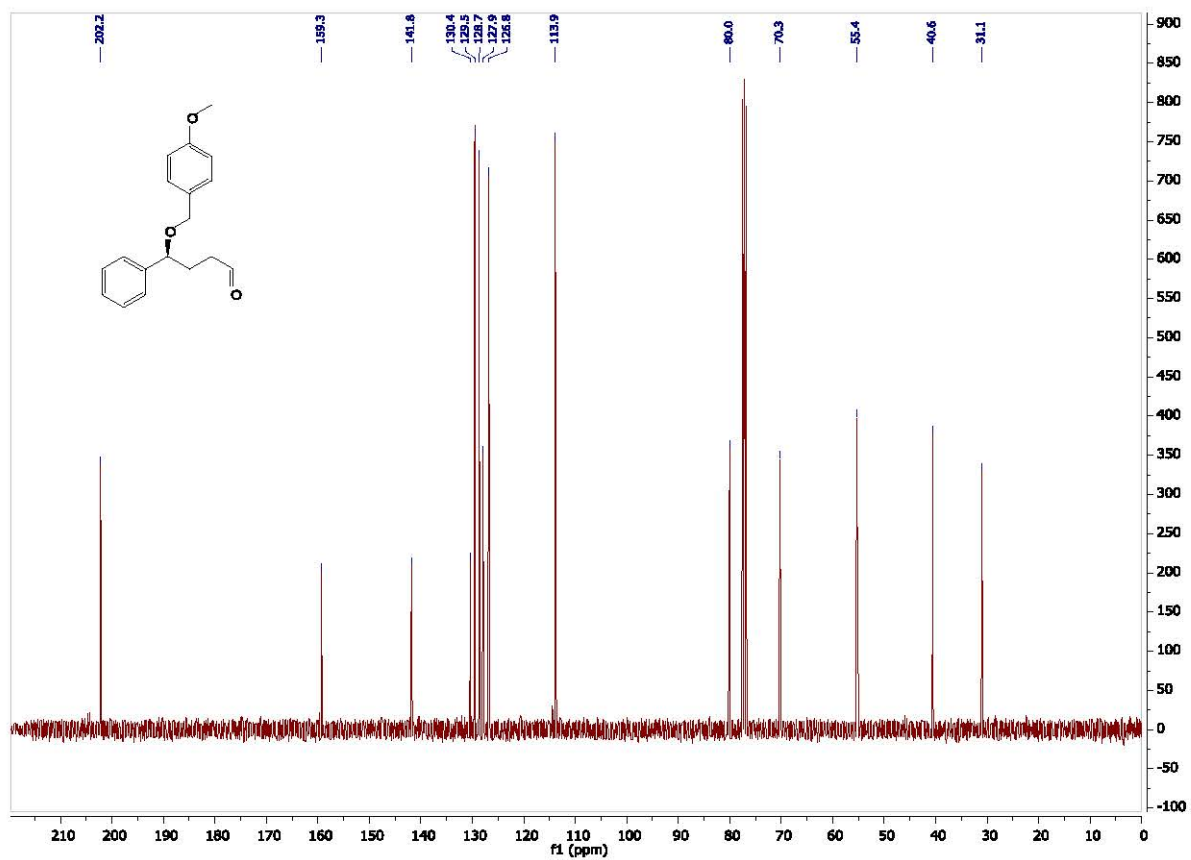
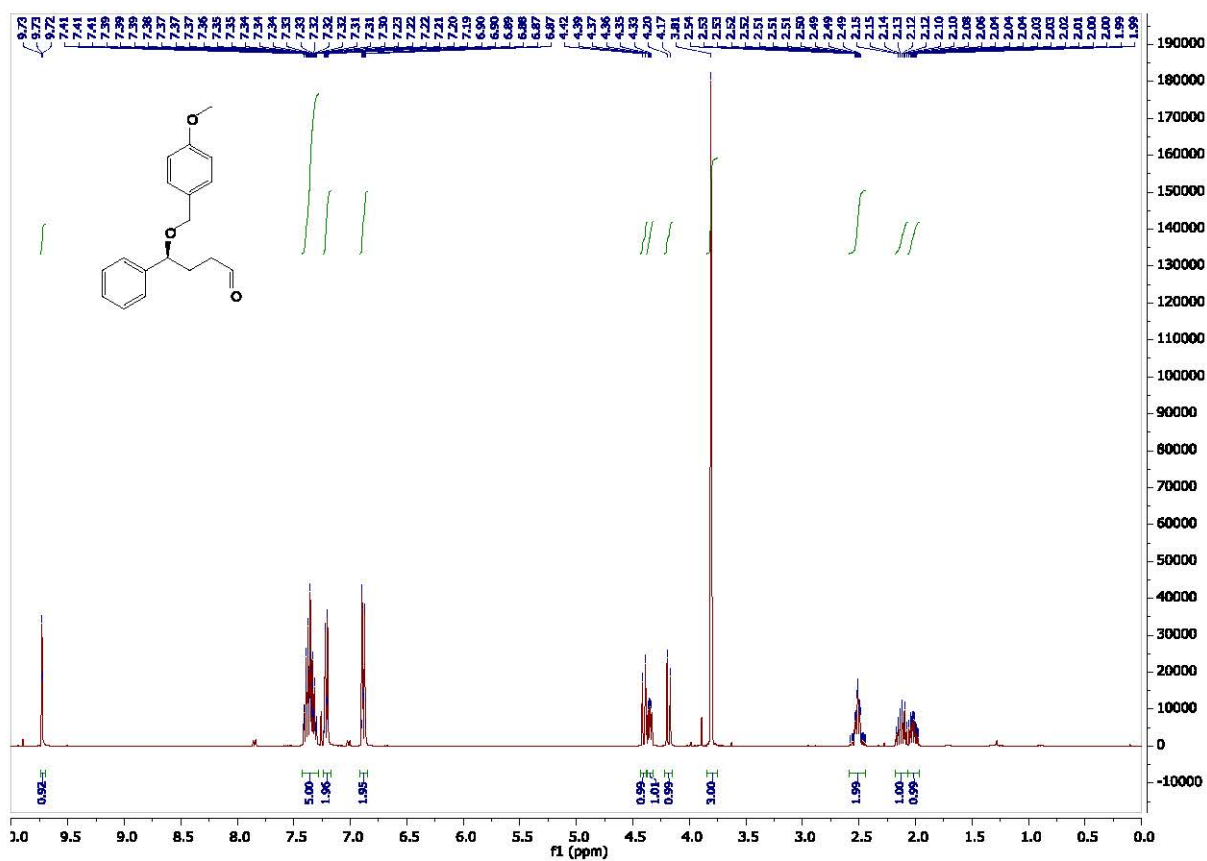


Figure S20. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound 8.

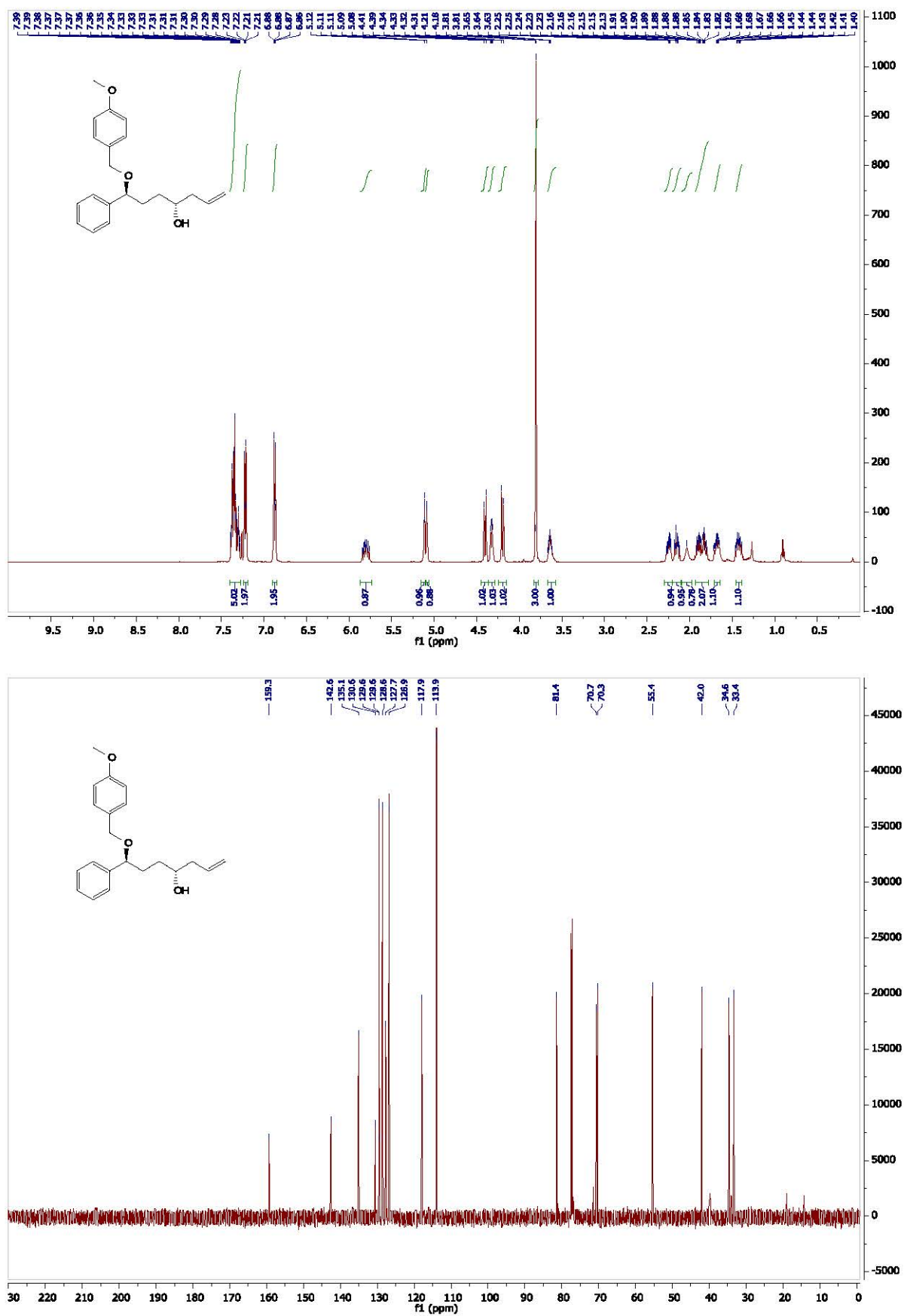
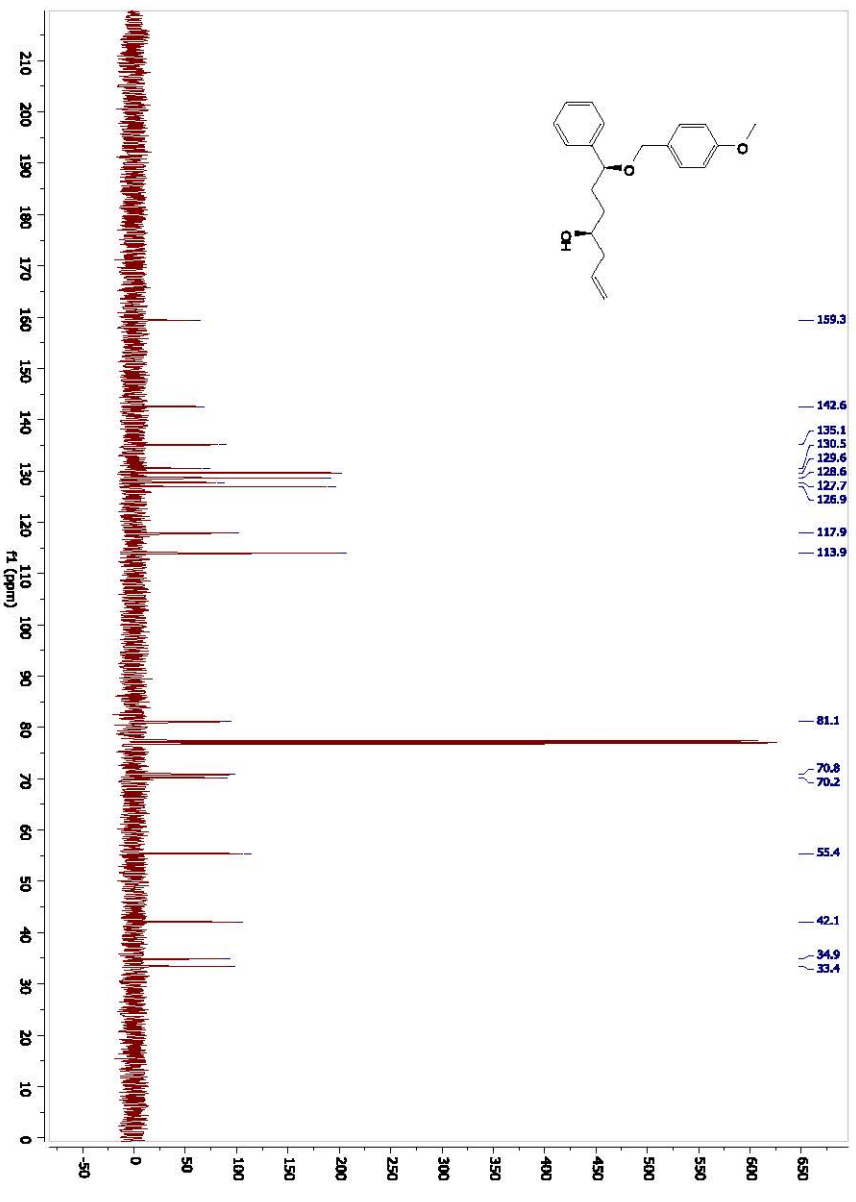
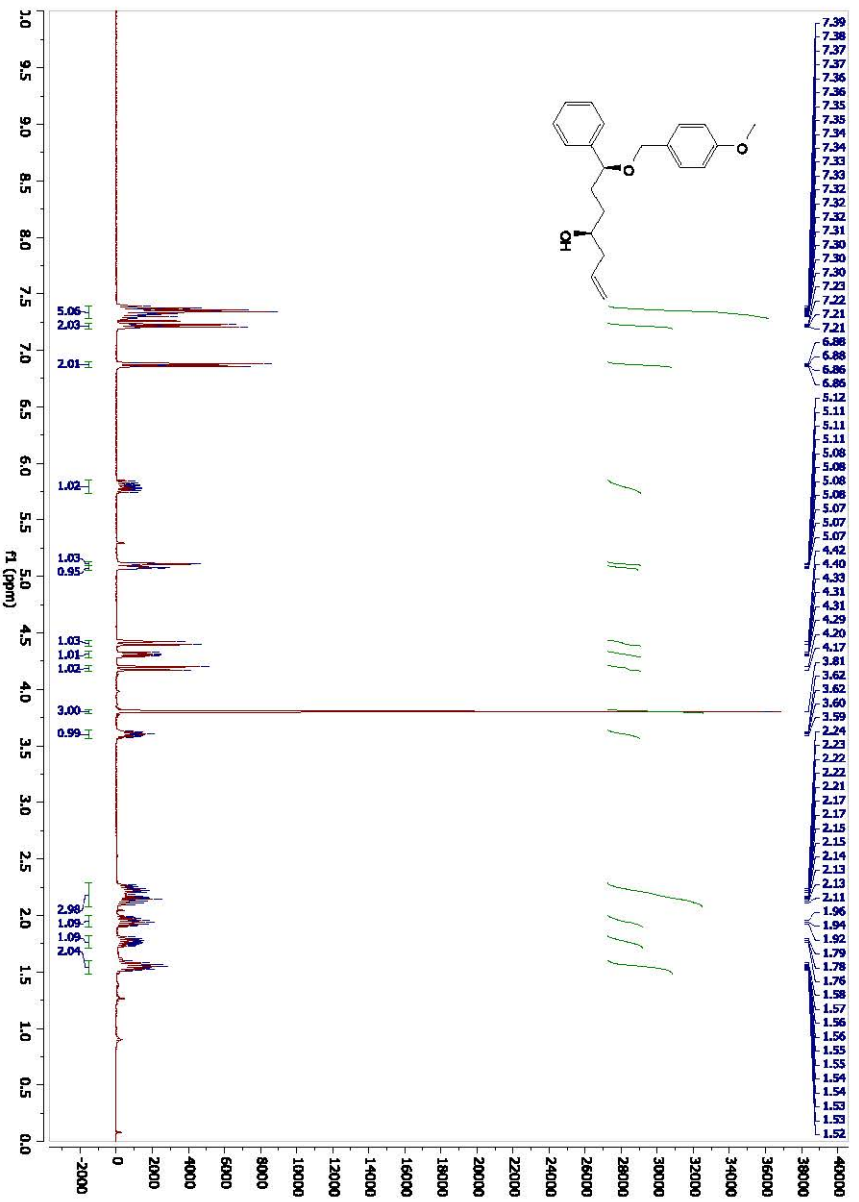


Figure S21. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound **9**.



**Figure S22.** <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound SI-3.



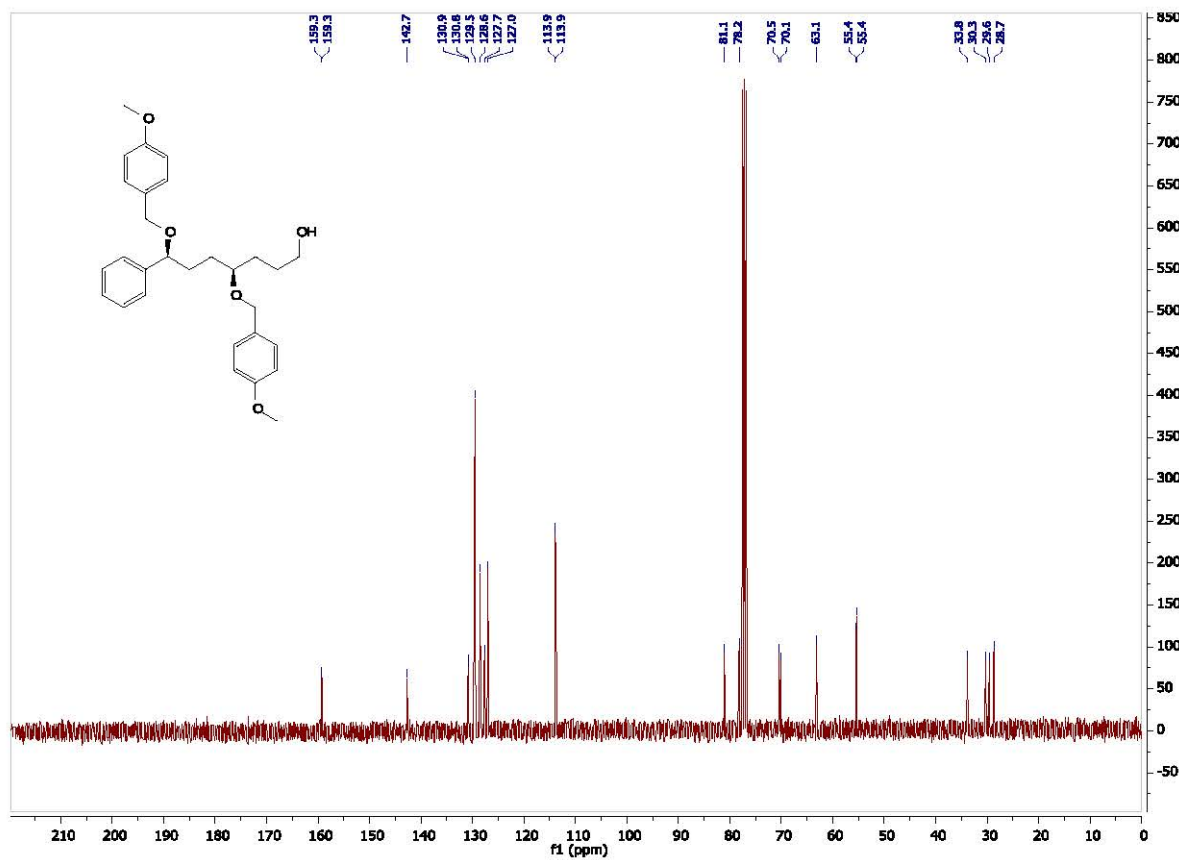
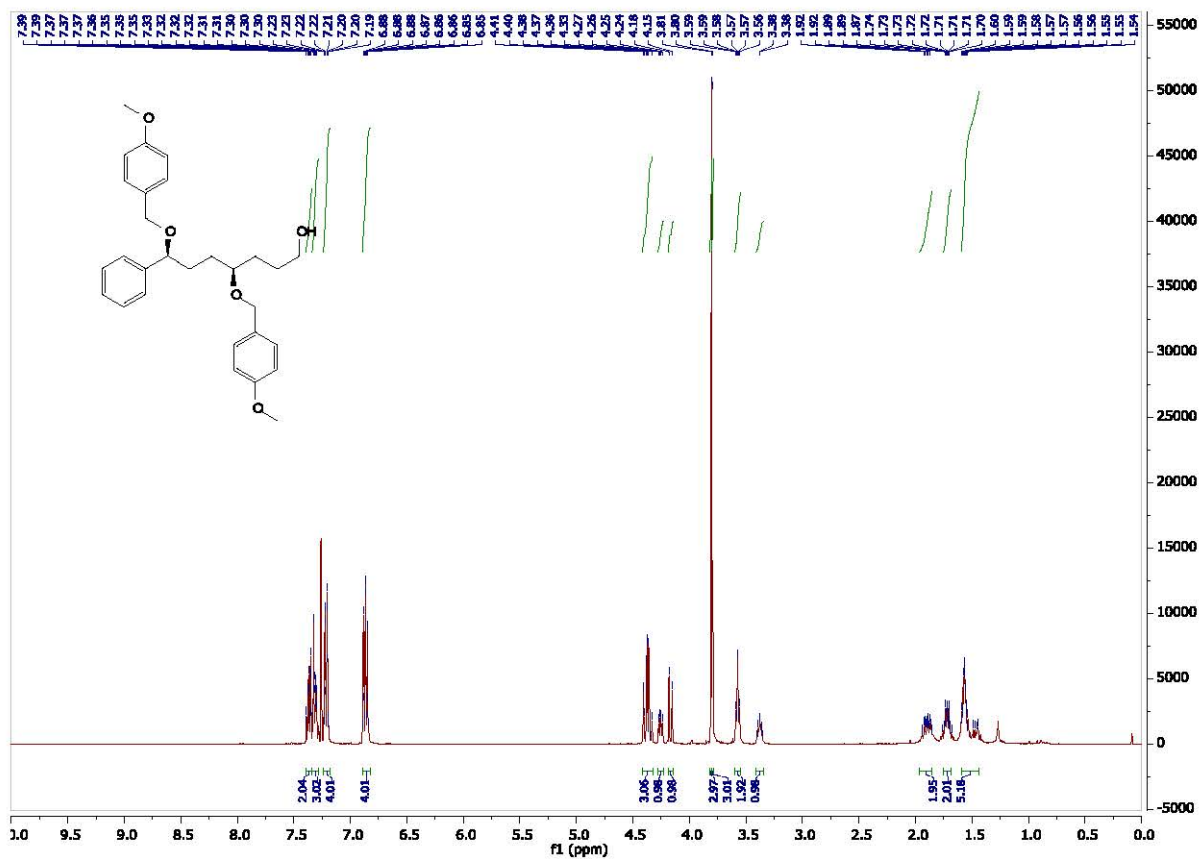


Figure S24. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound 11.

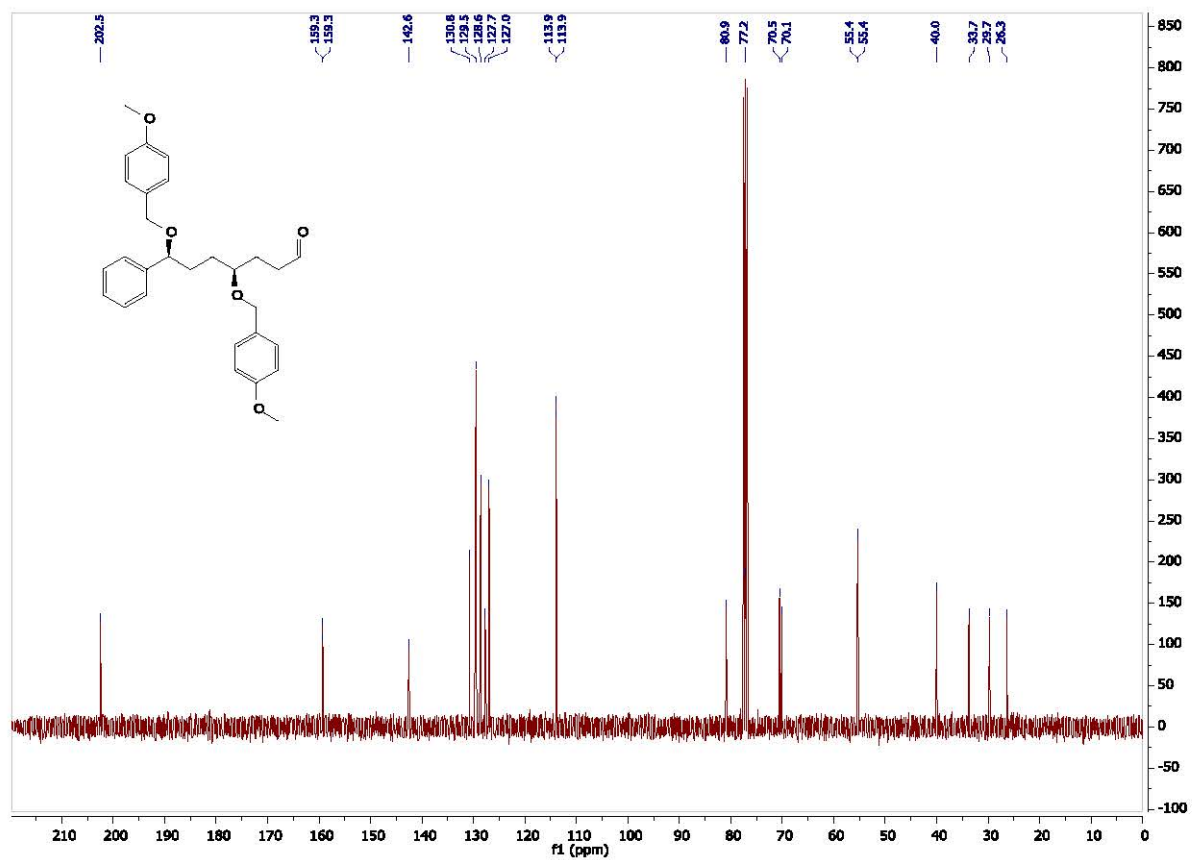
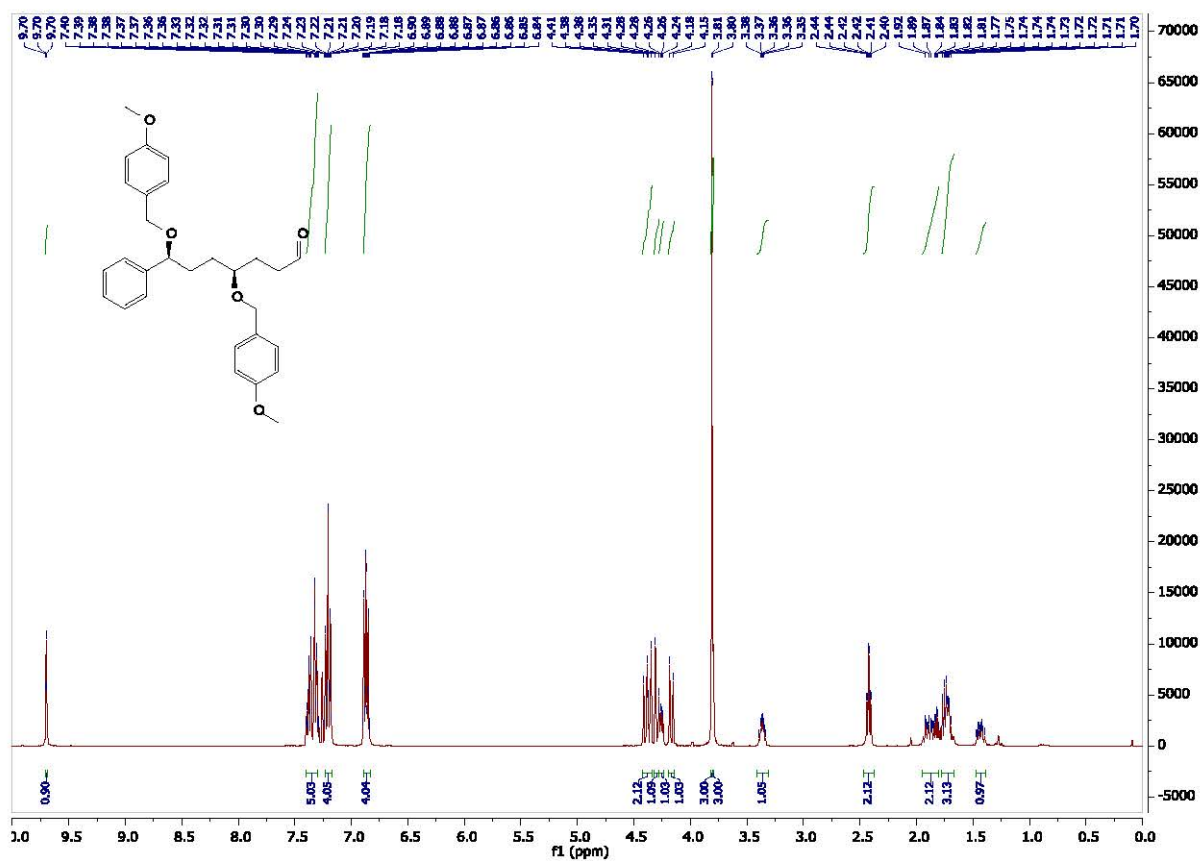


Figure S25. <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound 12.



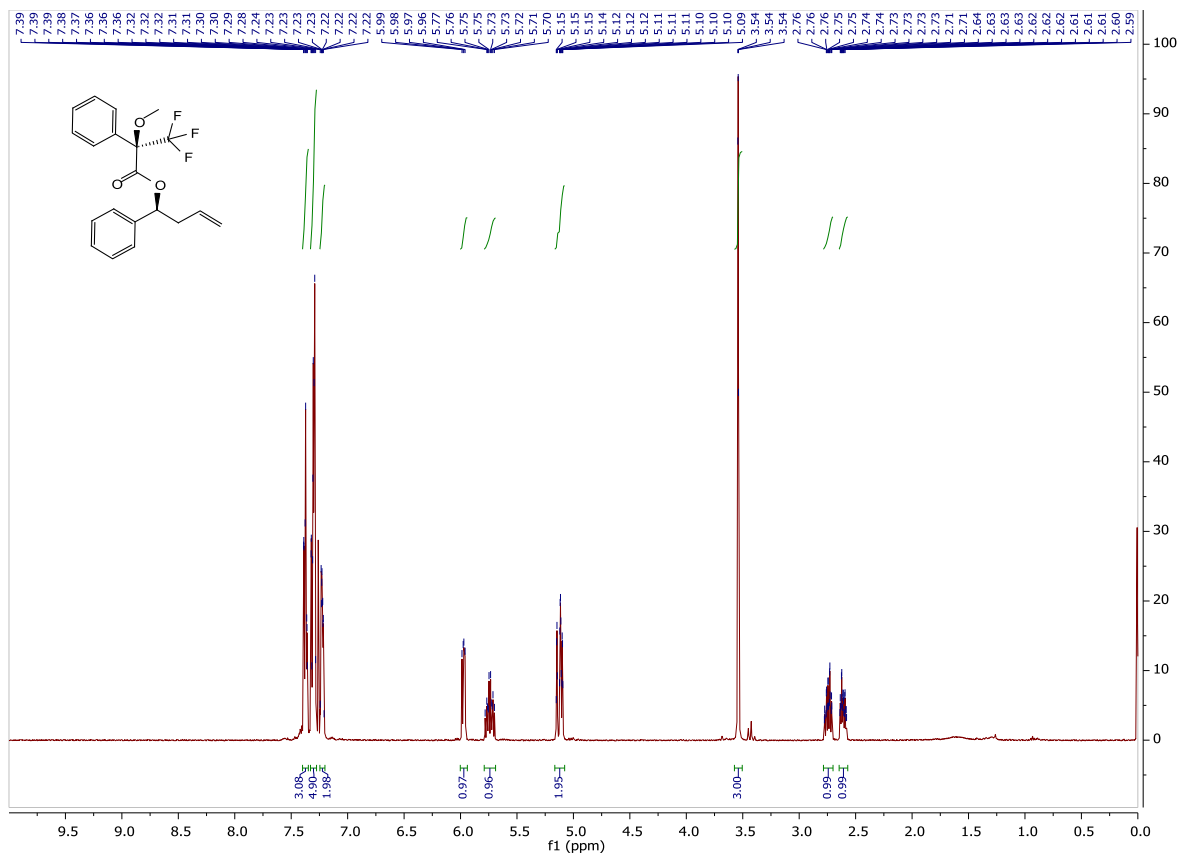


Figure S26. <sup>1</sup>H NMR spectrum of compound SI-4.

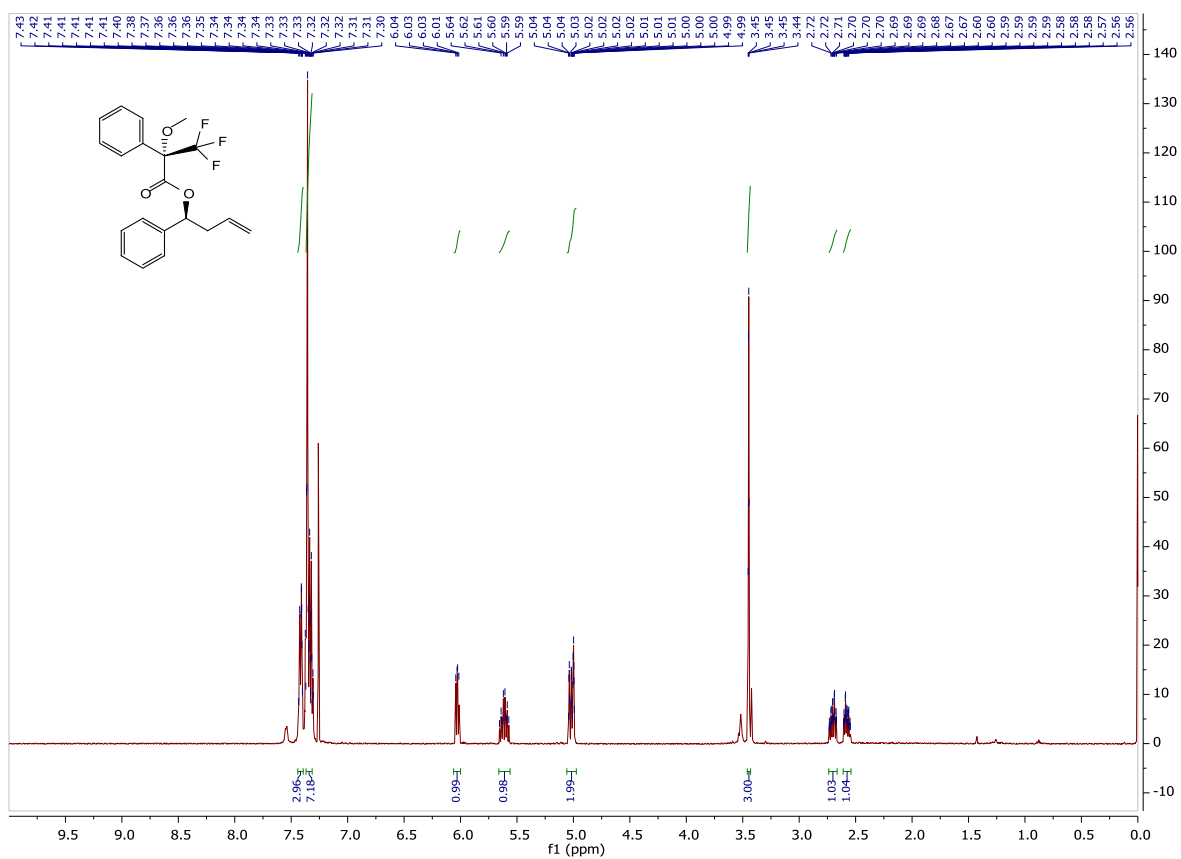


Figure S27. <sup>1</sup>H NMR spectrum of compound SI-5.

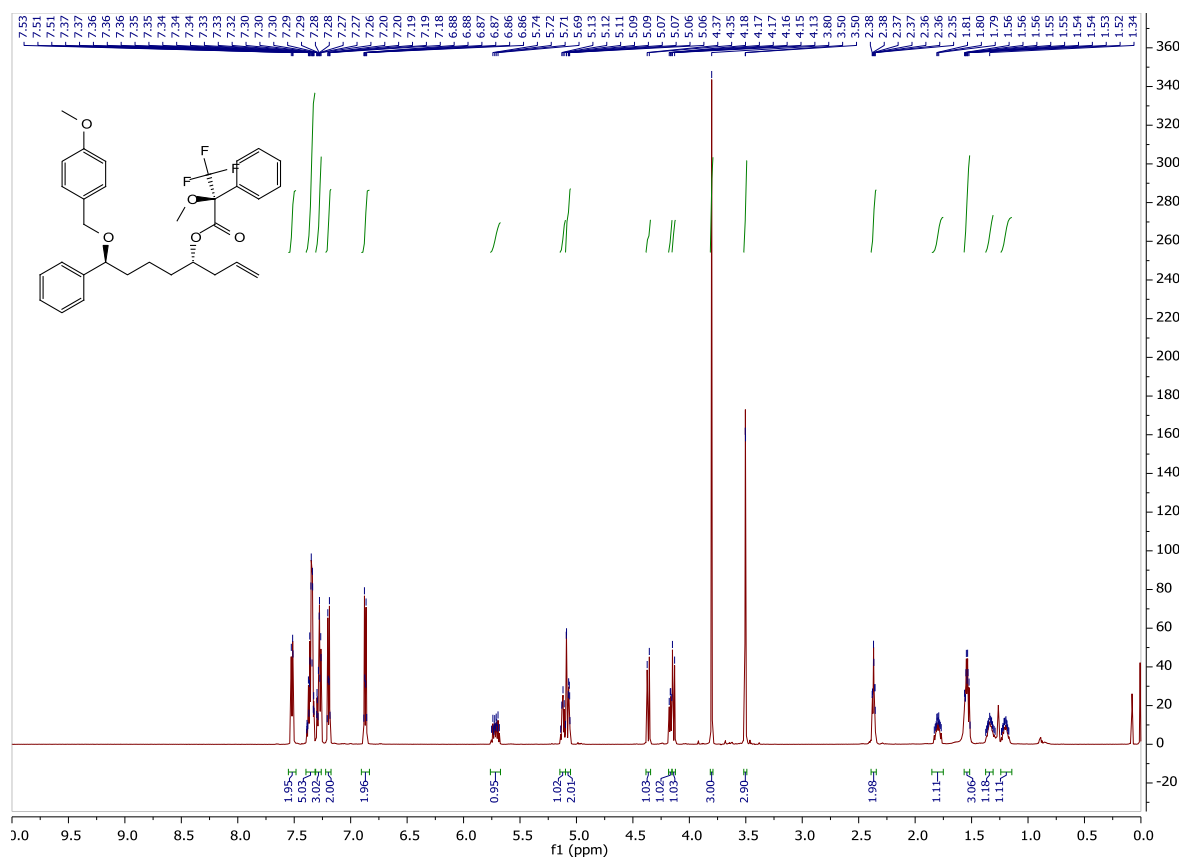


Figure S28. <sup>1</sup>H NMR spectrum of compound SI-6.

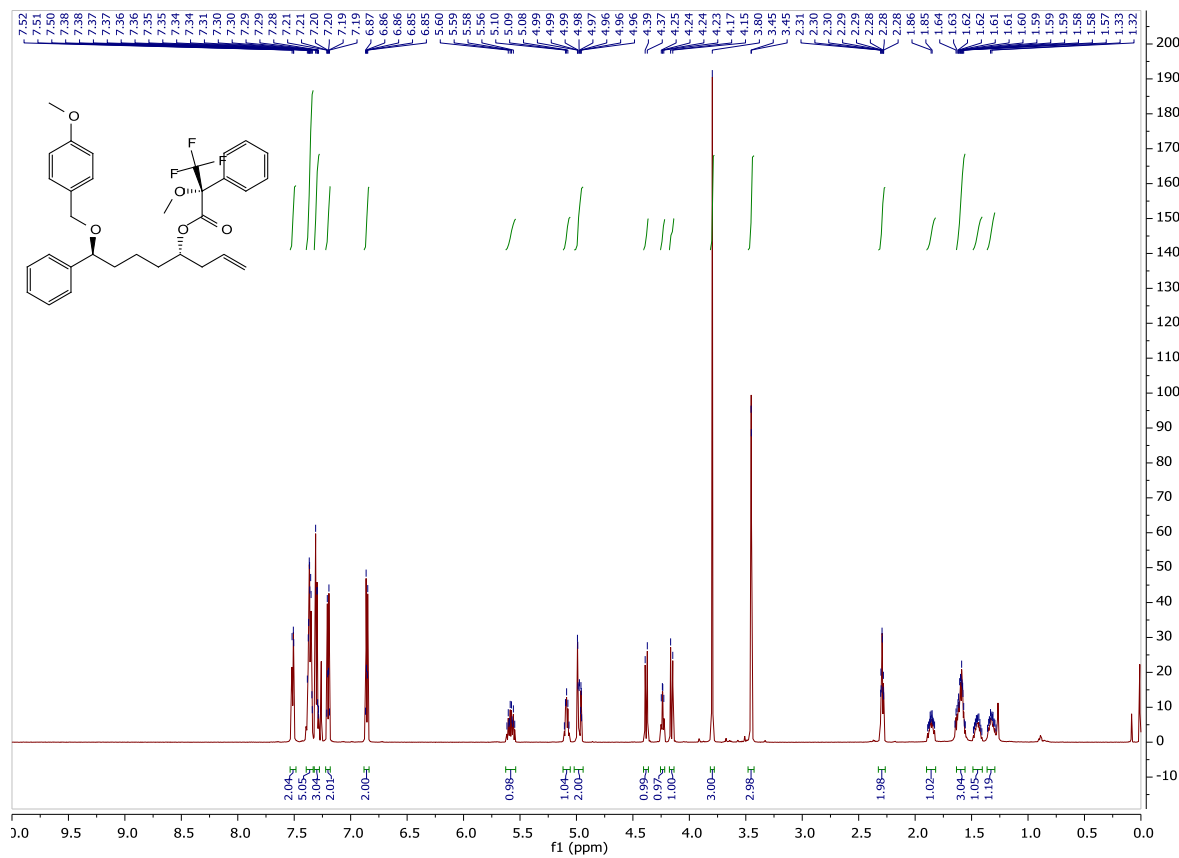


Figure S29. <sup>1</sup>H NMR spectrum of compound SI-7.

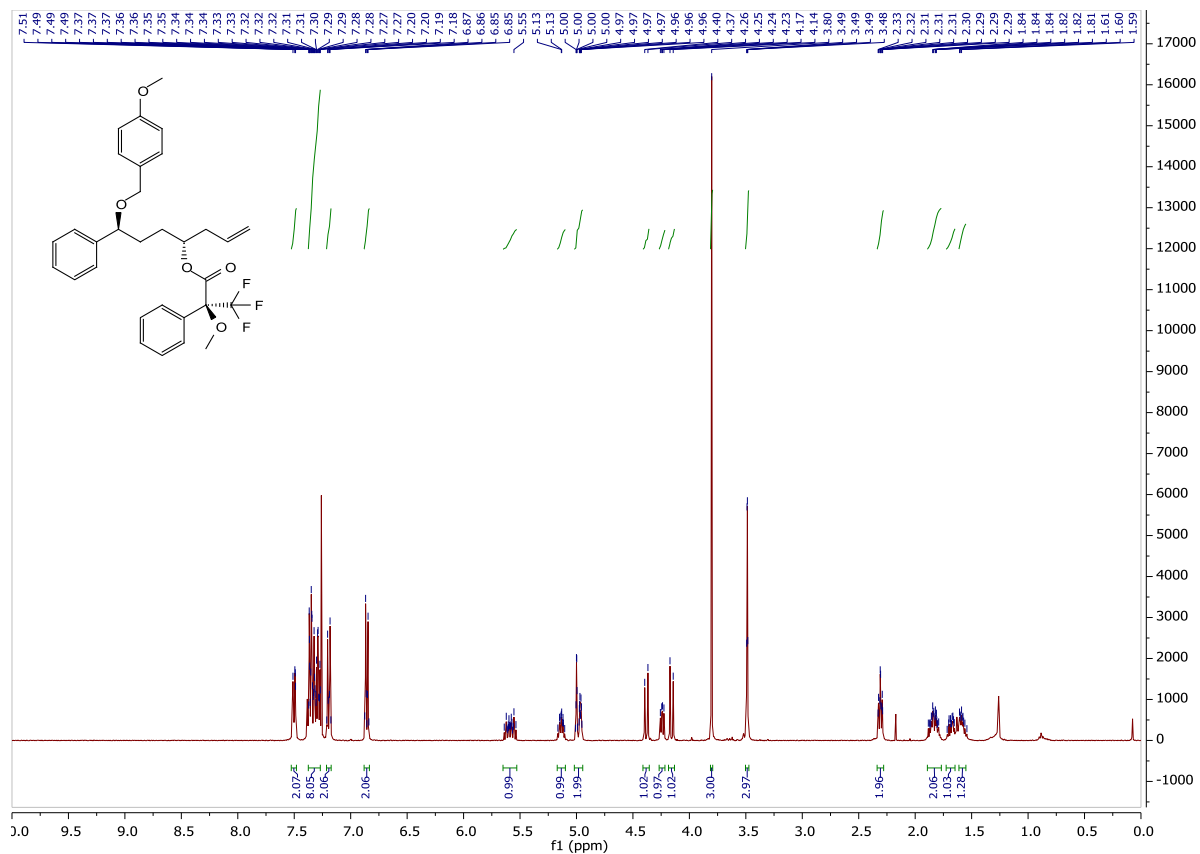


Figure S30. <sup>1</sup>H NMR spectrum of compound SI-8.

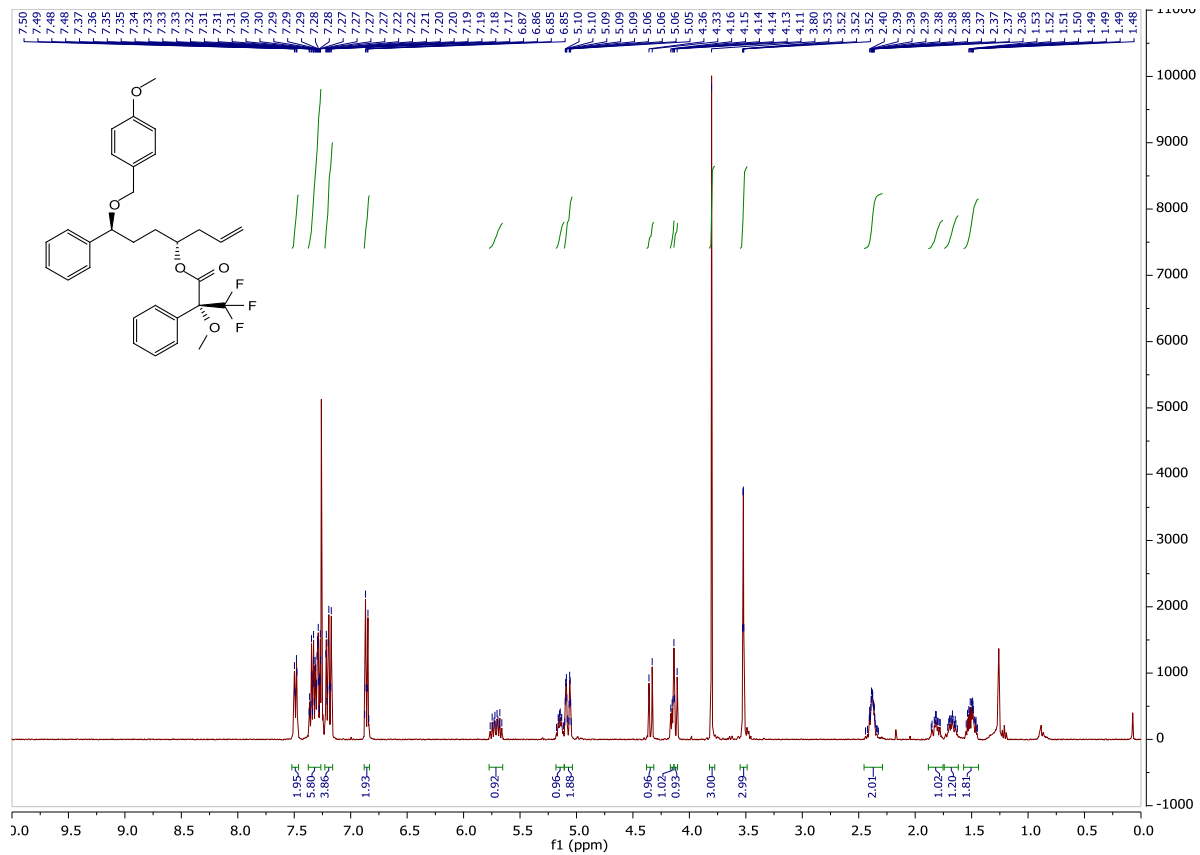


Figure S31. <sup>1</sup>H NMR spectrum of compound SI-9.

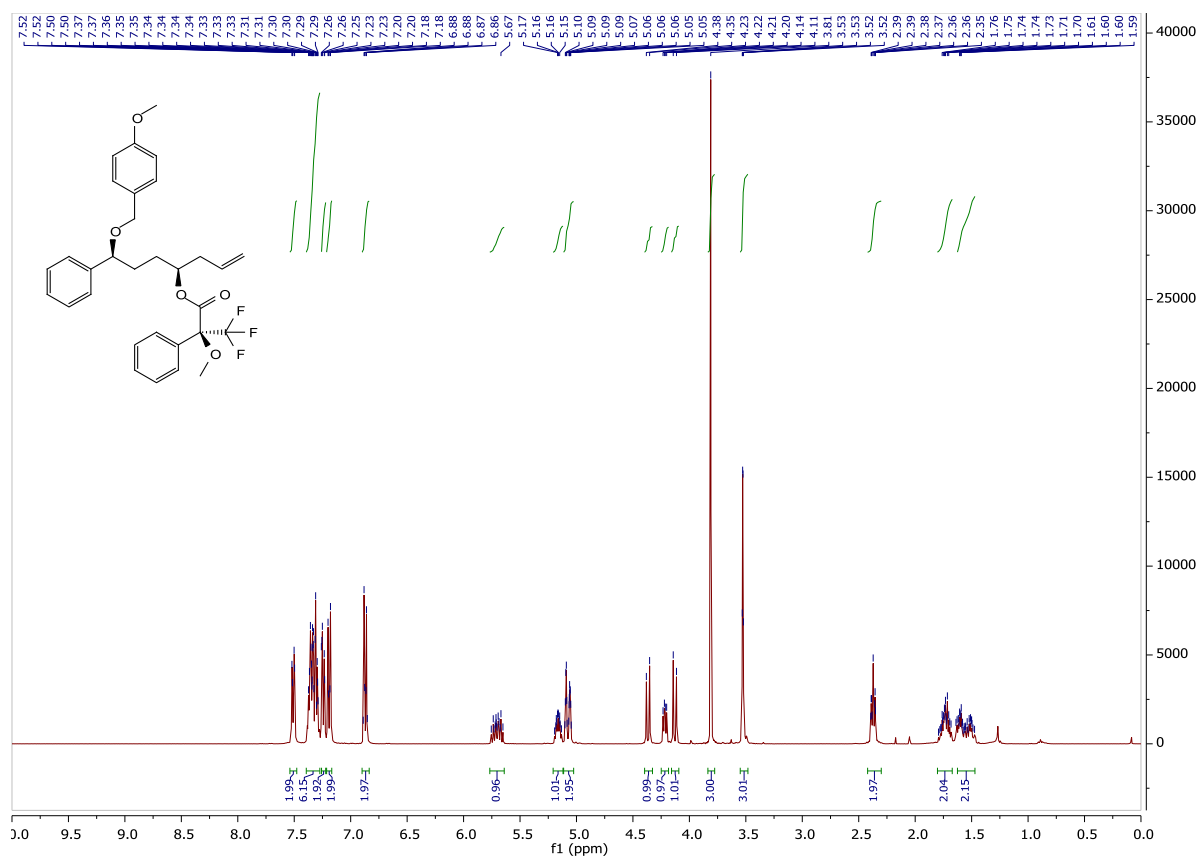


Figure S32. <sup>1</sup>H NMR spectrum of compound SI-10.

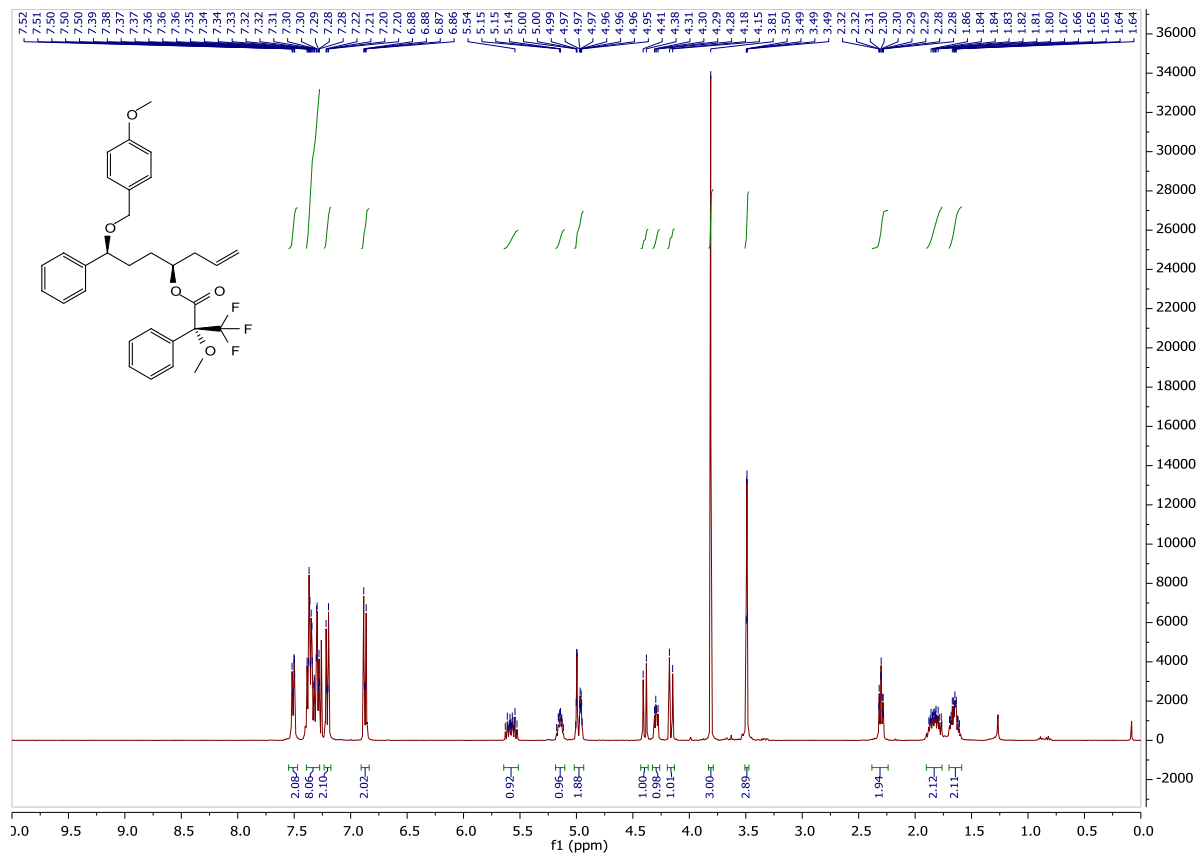
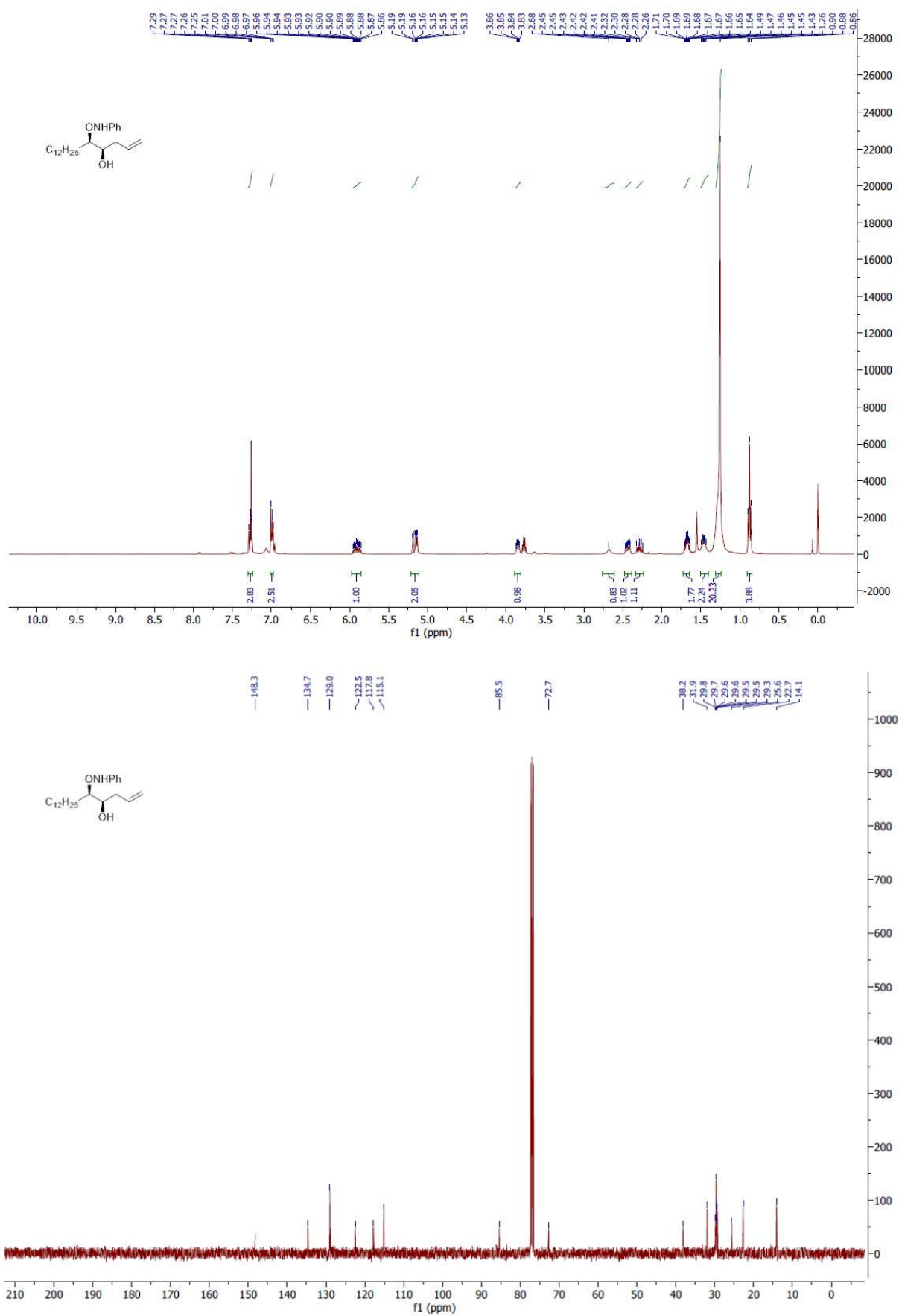
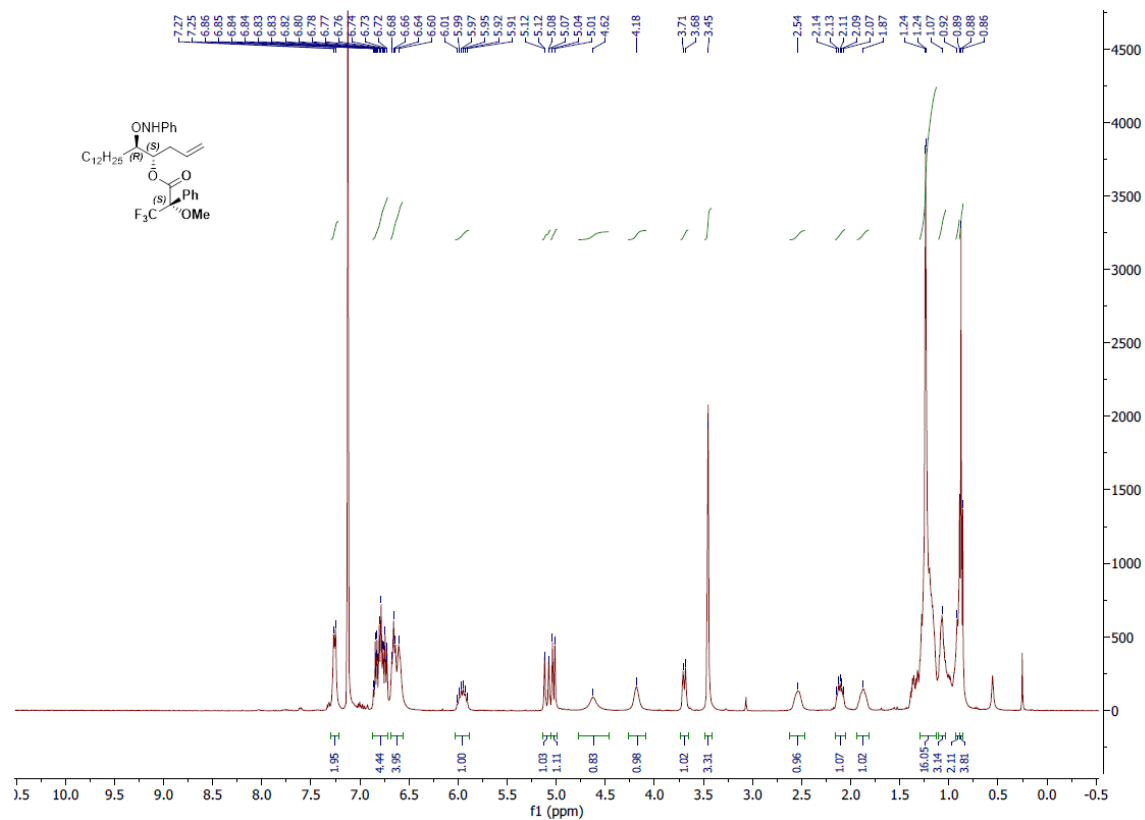


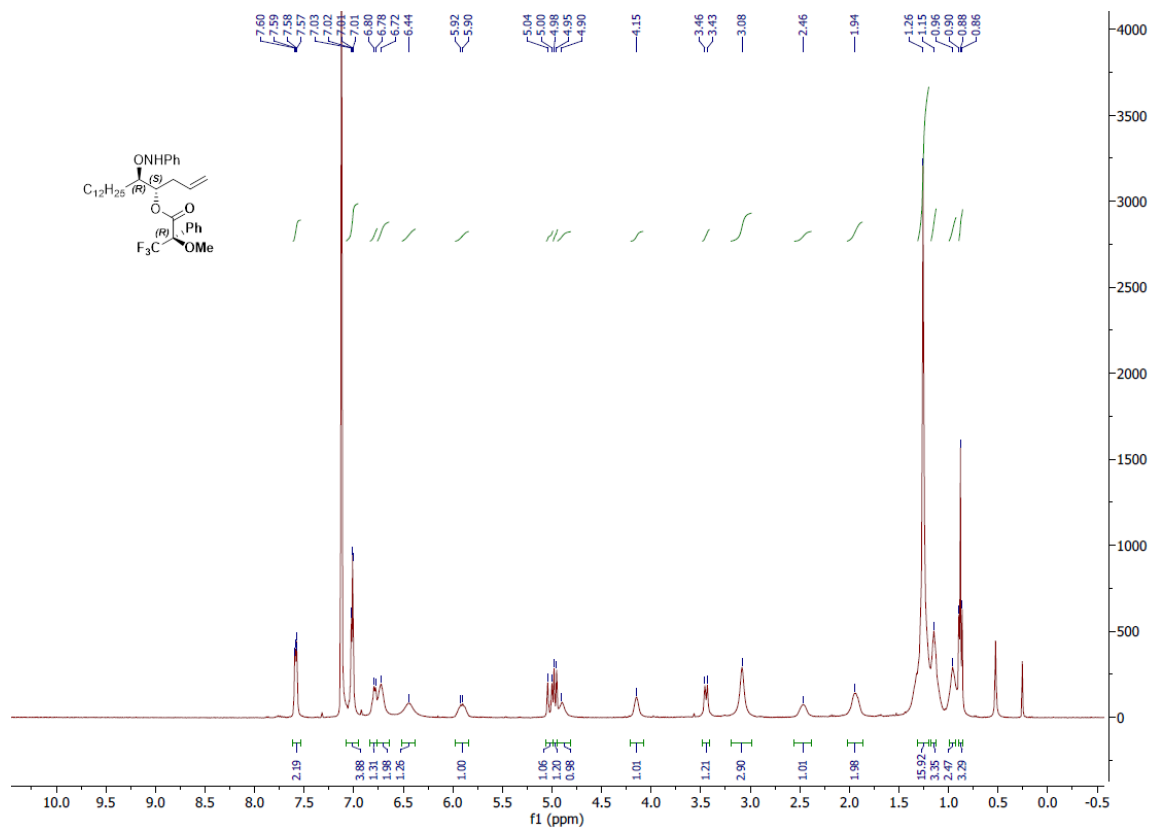
Figure S33. <sup>1</sup>H NMR spectrum of compound SI-11.



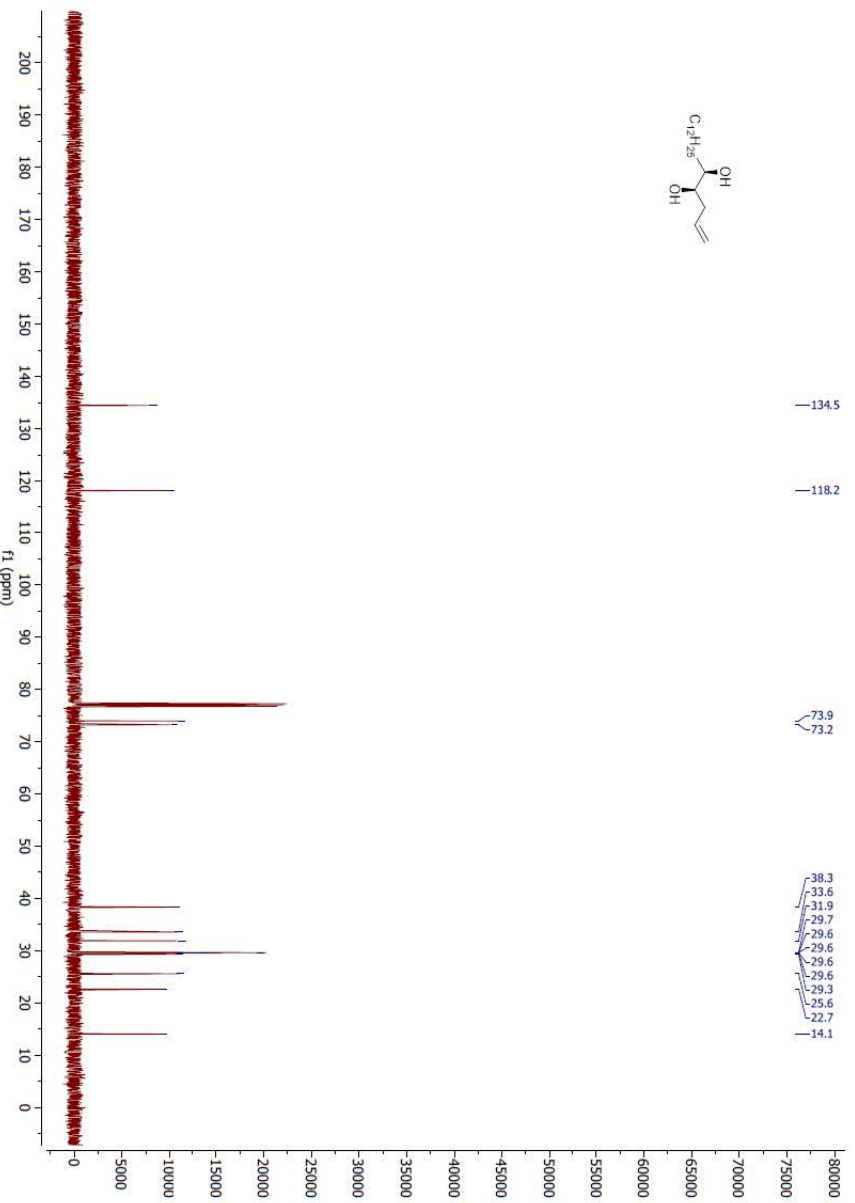
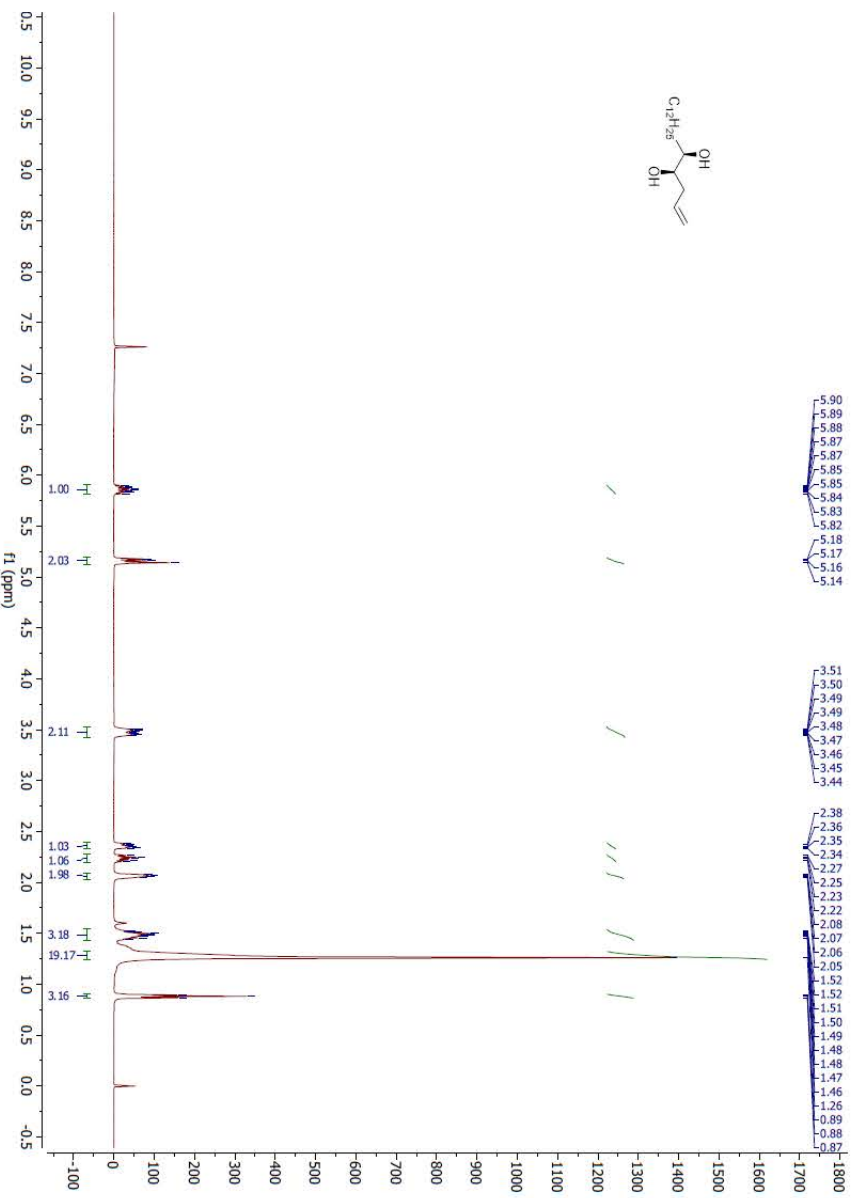
**Figure S34.** <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound **13**.



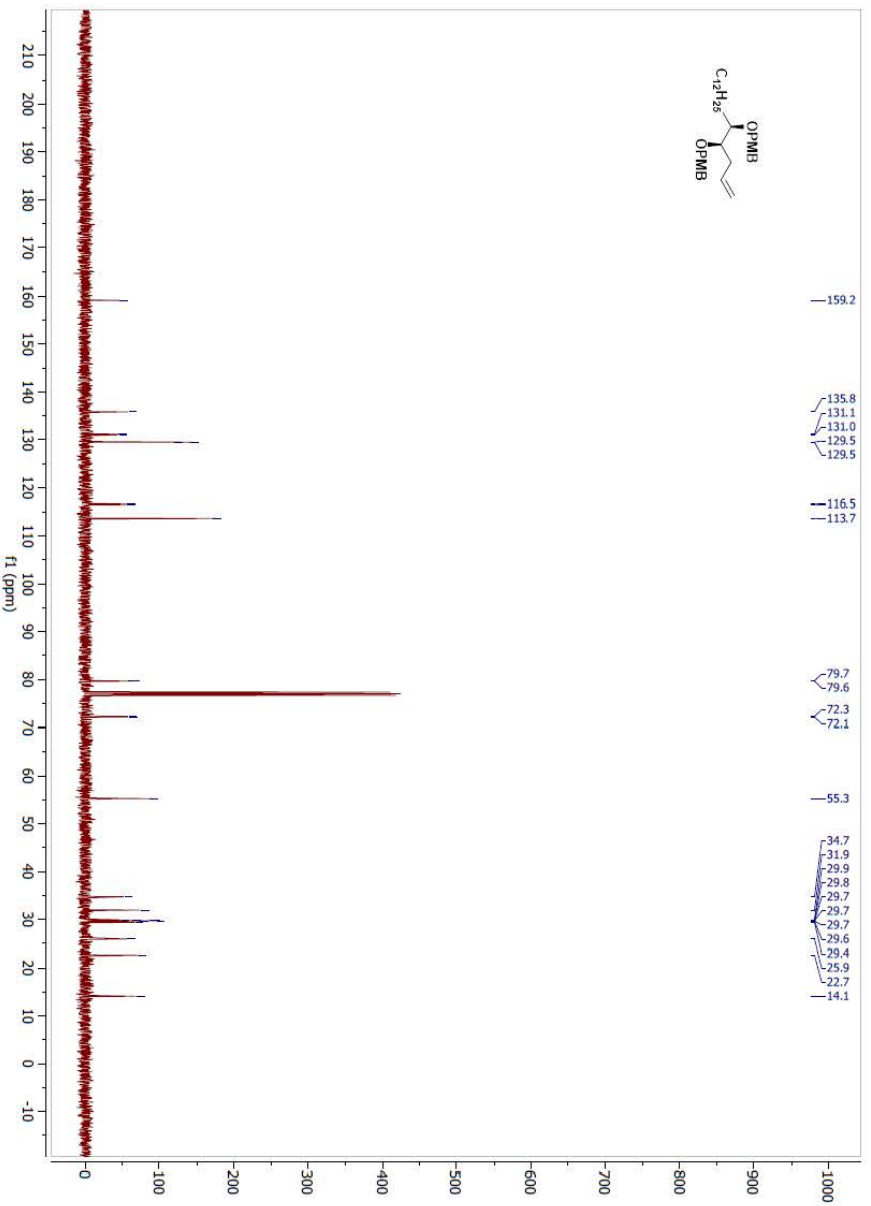
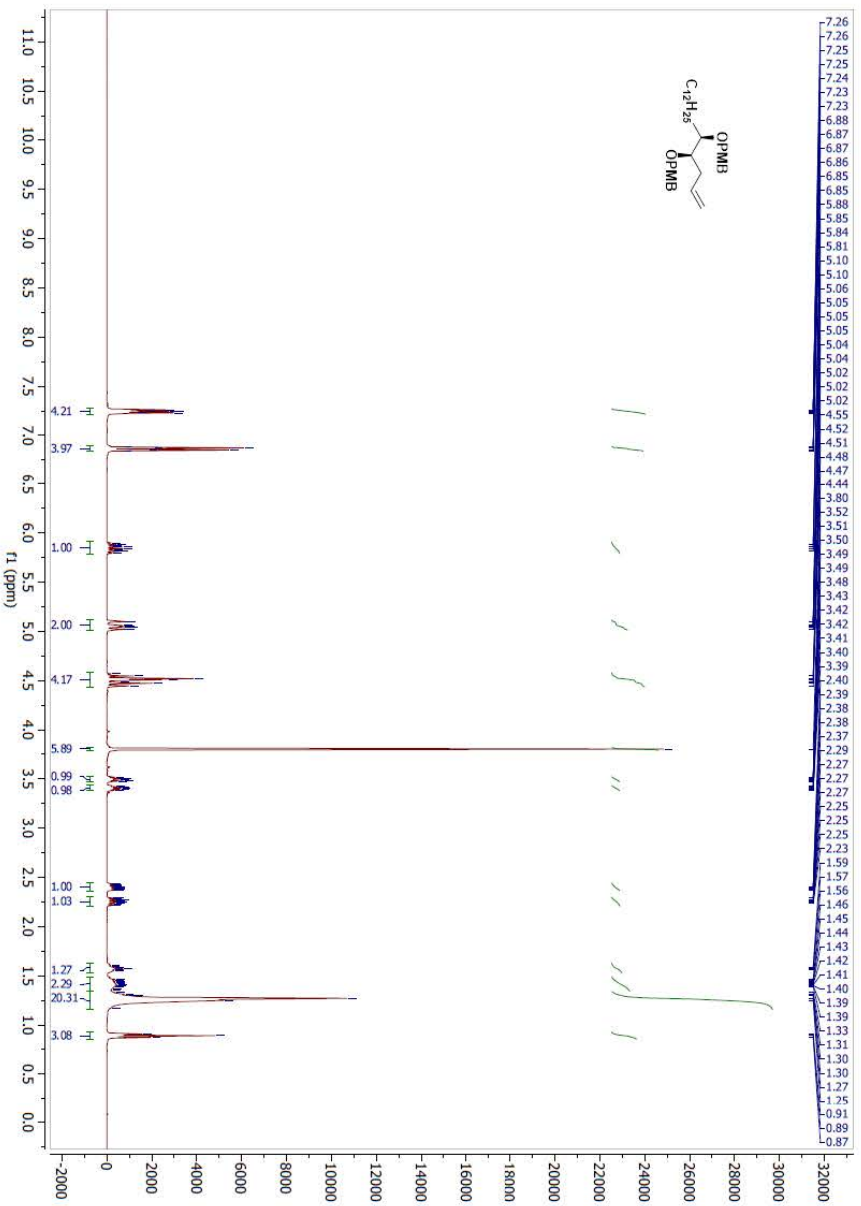
**Figure S35. <sup>1</sup>H NMR spectra of compound SI-12.**



**Figure S36. <sup>1</sup>H NMR spectra of compound SI-13.**

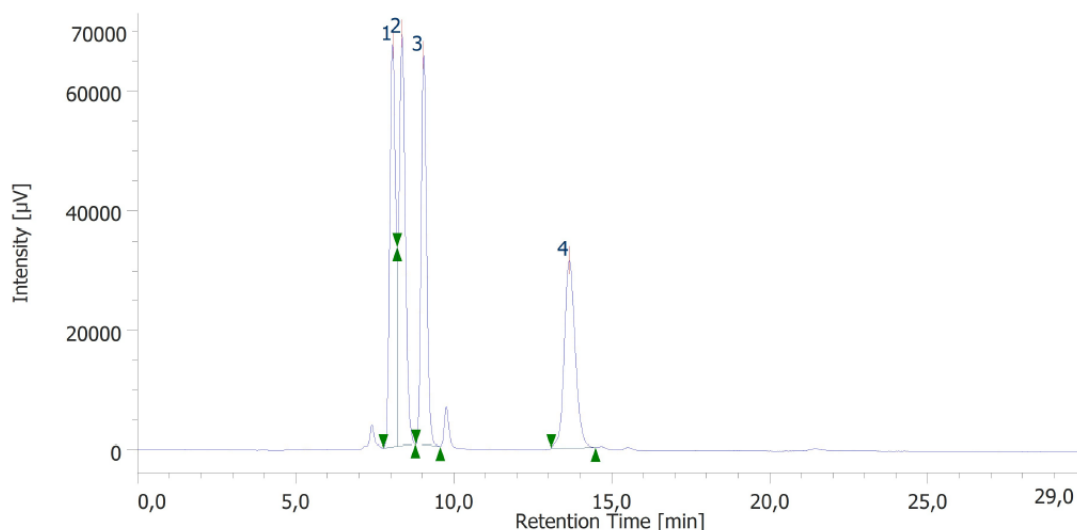


**Figure S37.**  $^1\text{H}$  NMR (top) and  $^{13}\text{C}$  NMR (bottom) spectra of compound **SI-14**.



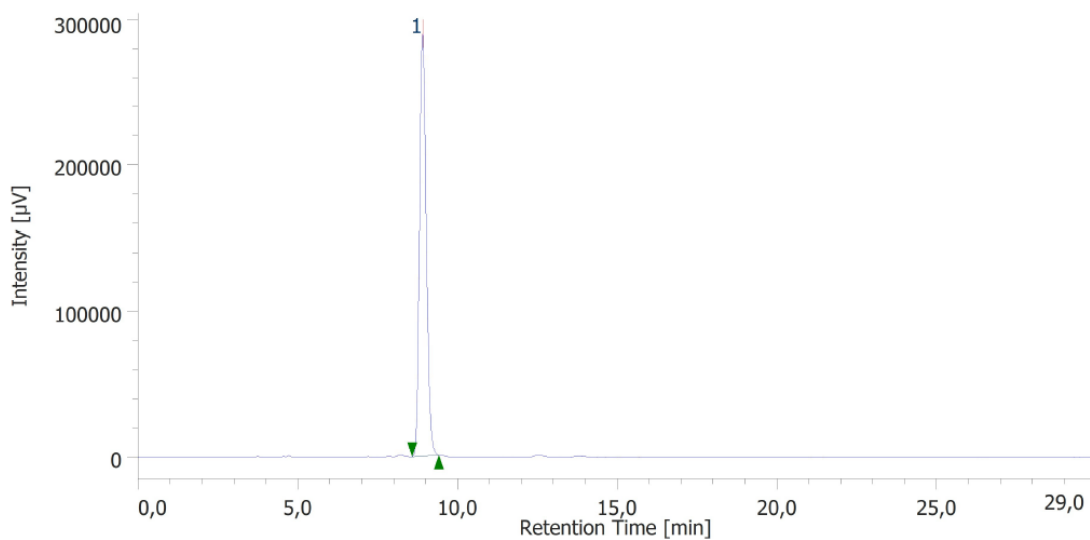
**Figure S38.**  $^1\text{H}$  NMR (top) and  $^{13}\text{C}$  NMR (bottom) spectra of compound **14**.





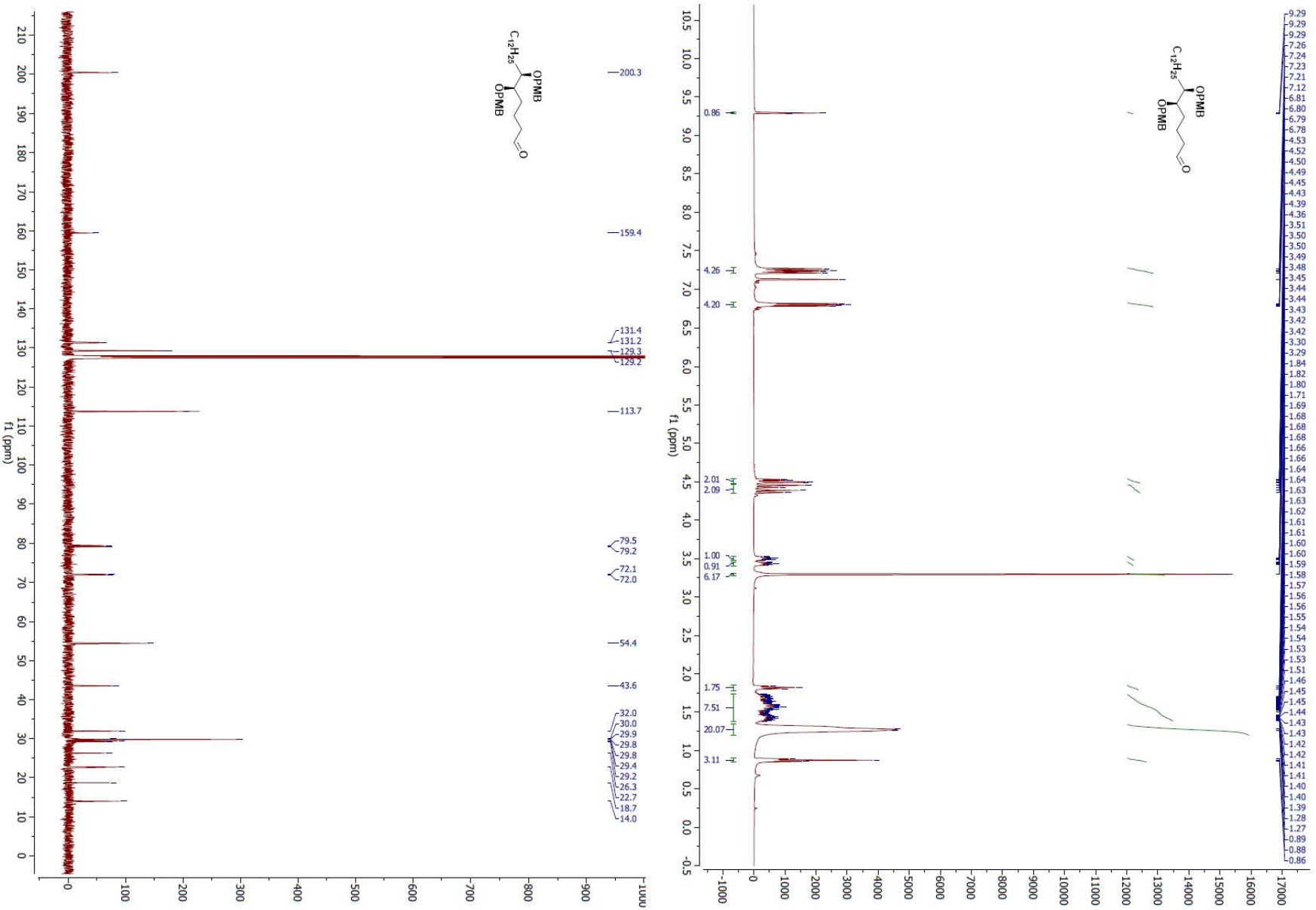
#	Peak Name	CH	tR [min]	Area [µV·sec]	Height [µV]	Area%	Height%	Quantity	NTP	Resolution	Symmetry Factor	Warning
1	Unknown	1	8,067	859766	67212	25,409	28,899	N/A	6226	0,709	N/A	
2	Unknown	1	8,358	976141	68870	28,848	29,612	N/A	6467	1,874	N/A	
3	Unknown	1	9,042	790407	65053	23,359	27,971	N/A	13152	9,907	1,202	
4	Unknown	1	13,650	757437	31438	22,385	13,517	N/A	7819	N/A	1,127	

**Figure S39.** HPLC spectra of compound **14** racemate.

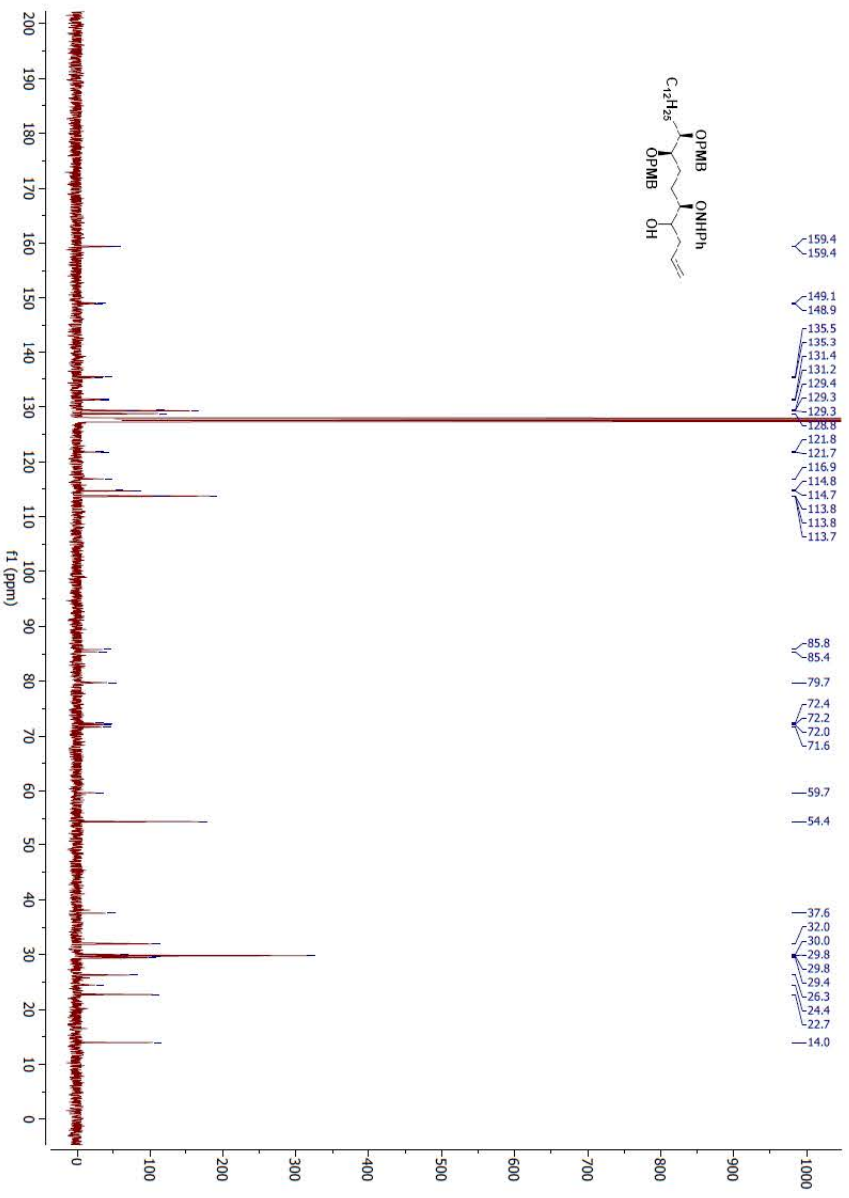
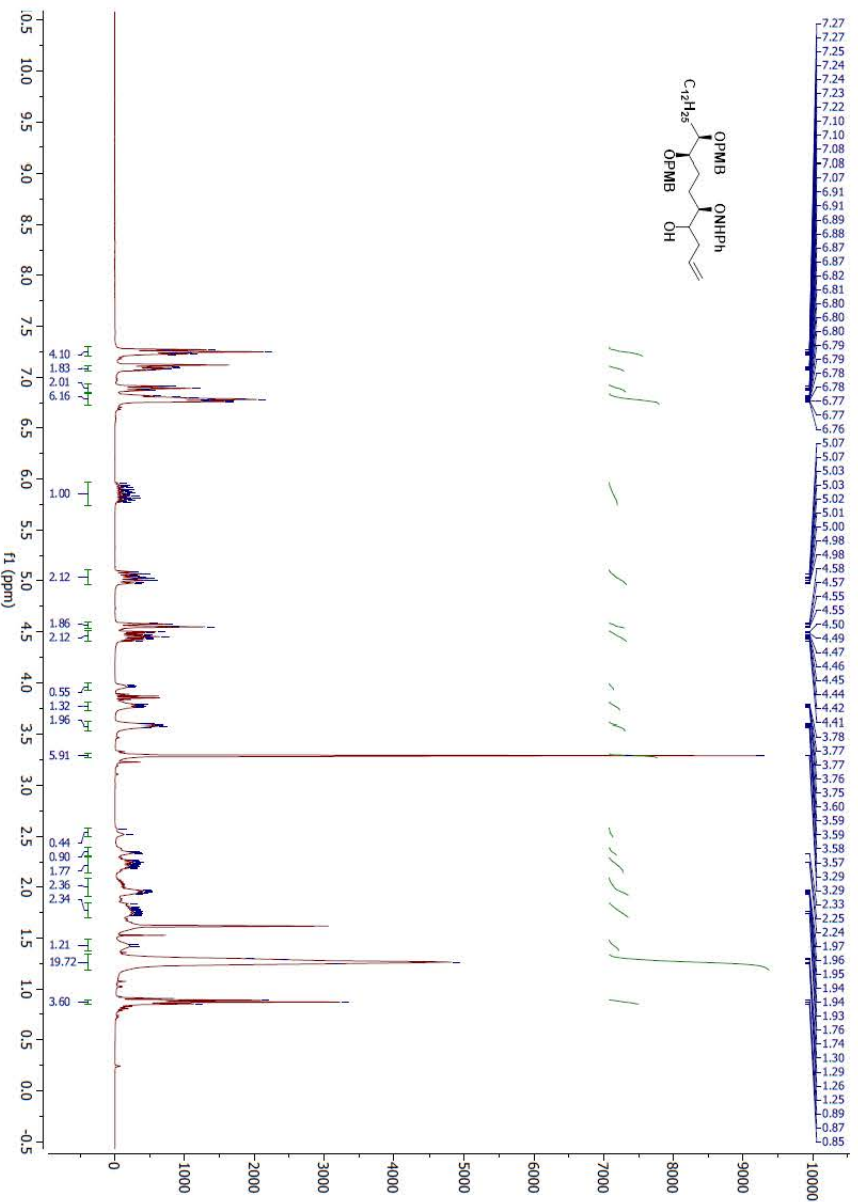


#	Peak Name	CH	tR [min]	Area [µV·sec]	Height [µV]	Area%	Height%	Quantity	NTP	Resolution	Symmetry Factor	Warning
1	Unknown	1	8,908	4255104	289141	100,000	100,000	N/A	8560	N/A	1,159	

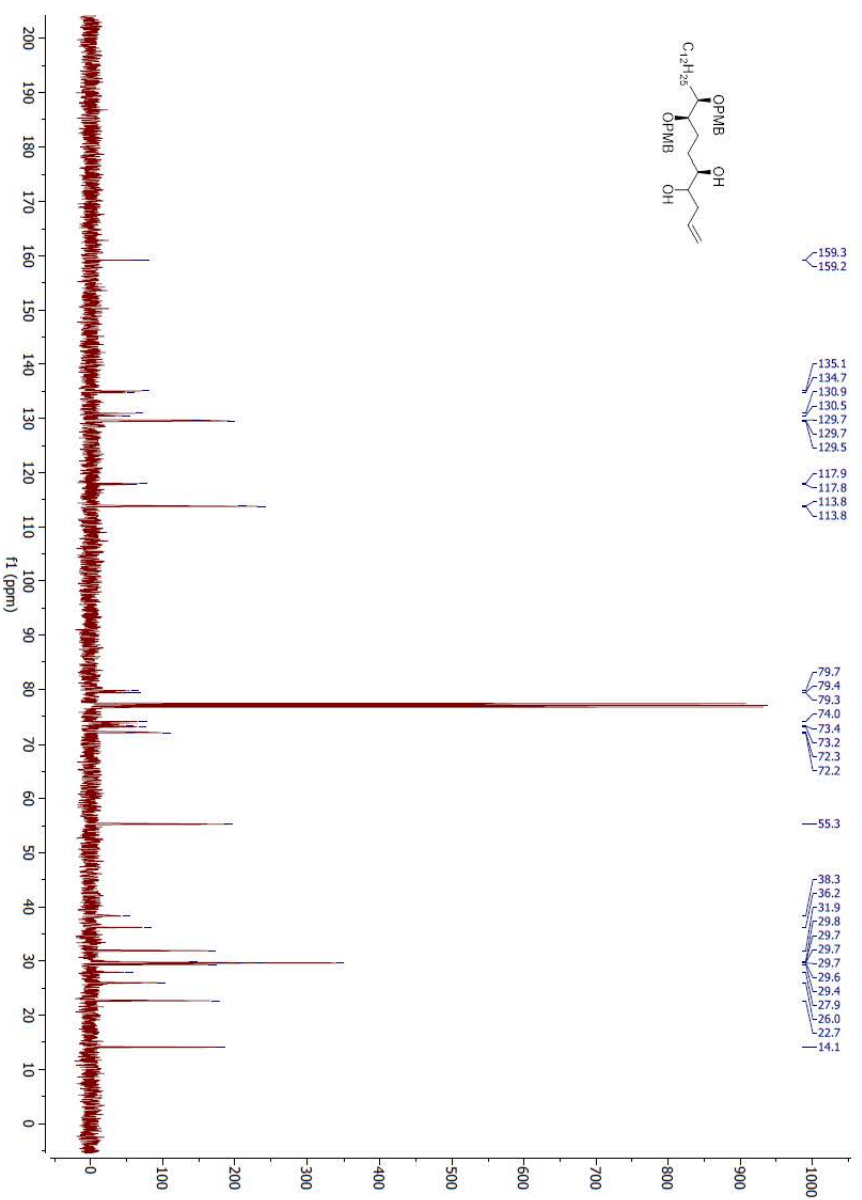
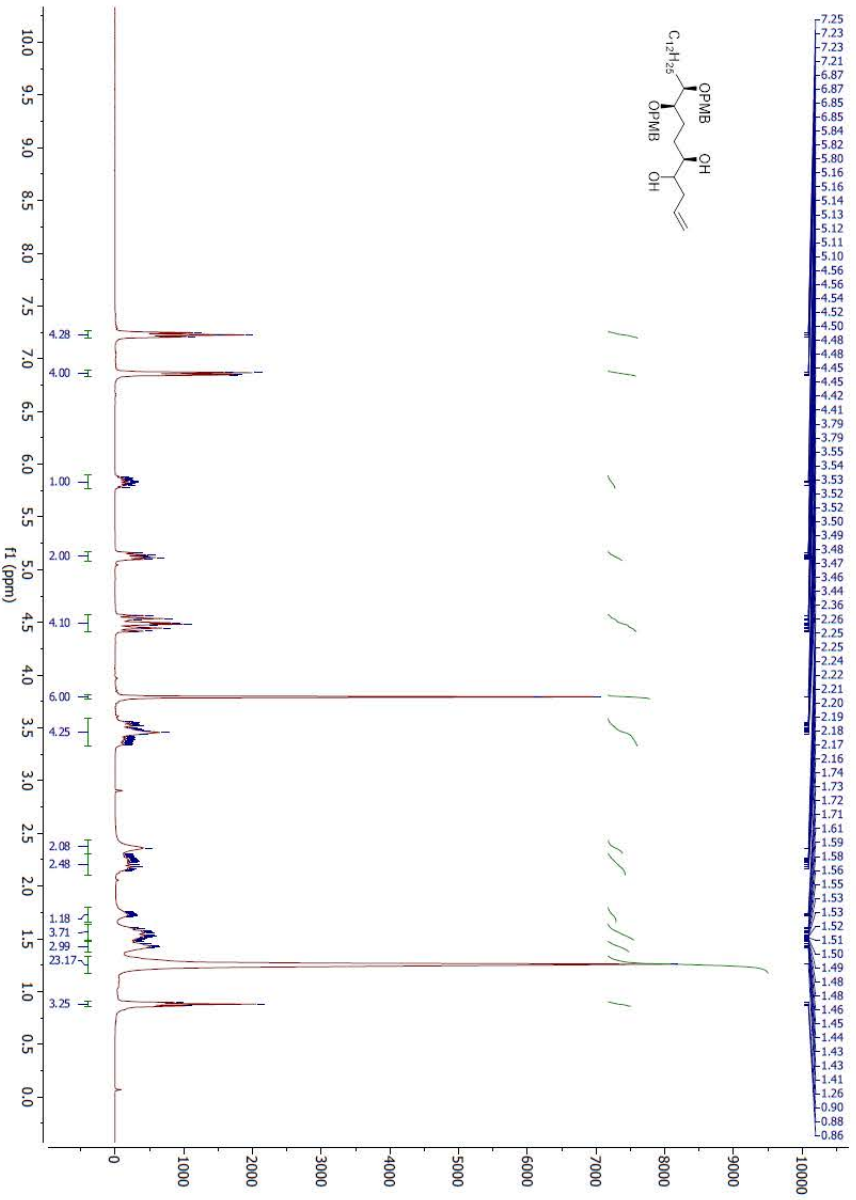
**Figure S40.** HPLC spectra of compound **14**.



**Figure S41.**  $^1\text{H}$  NMR (top) and  $^{13}\text{C}$  NMR (bottom) spectra of compound **15**.



**Figure S42.** <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound **16**.



**Figure S43.**  $^1\text{H NMR}$  (top) and  $^{13}\text{C NMR}$  (bottom) spectra of compound **SI-15**.

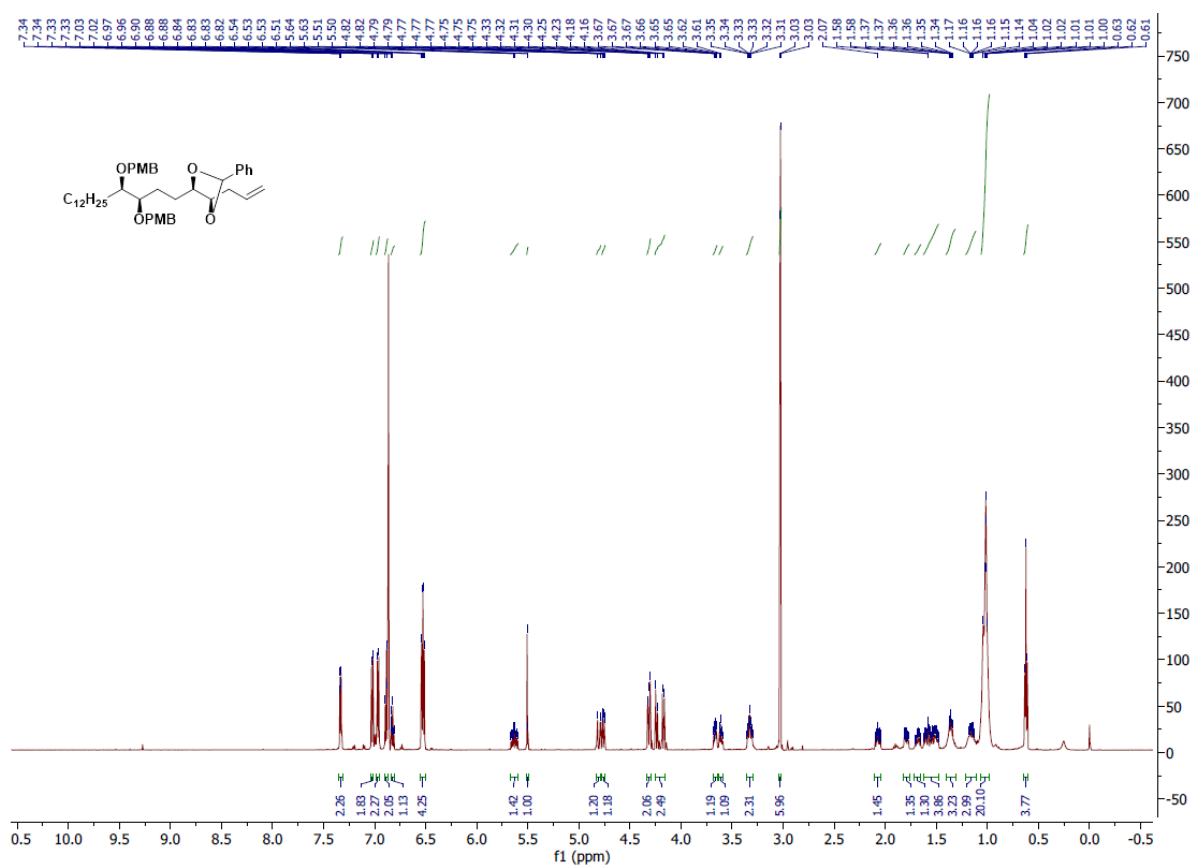


Figure S44.  $^1\text{H}$  NMR spectra of compound SI-16.

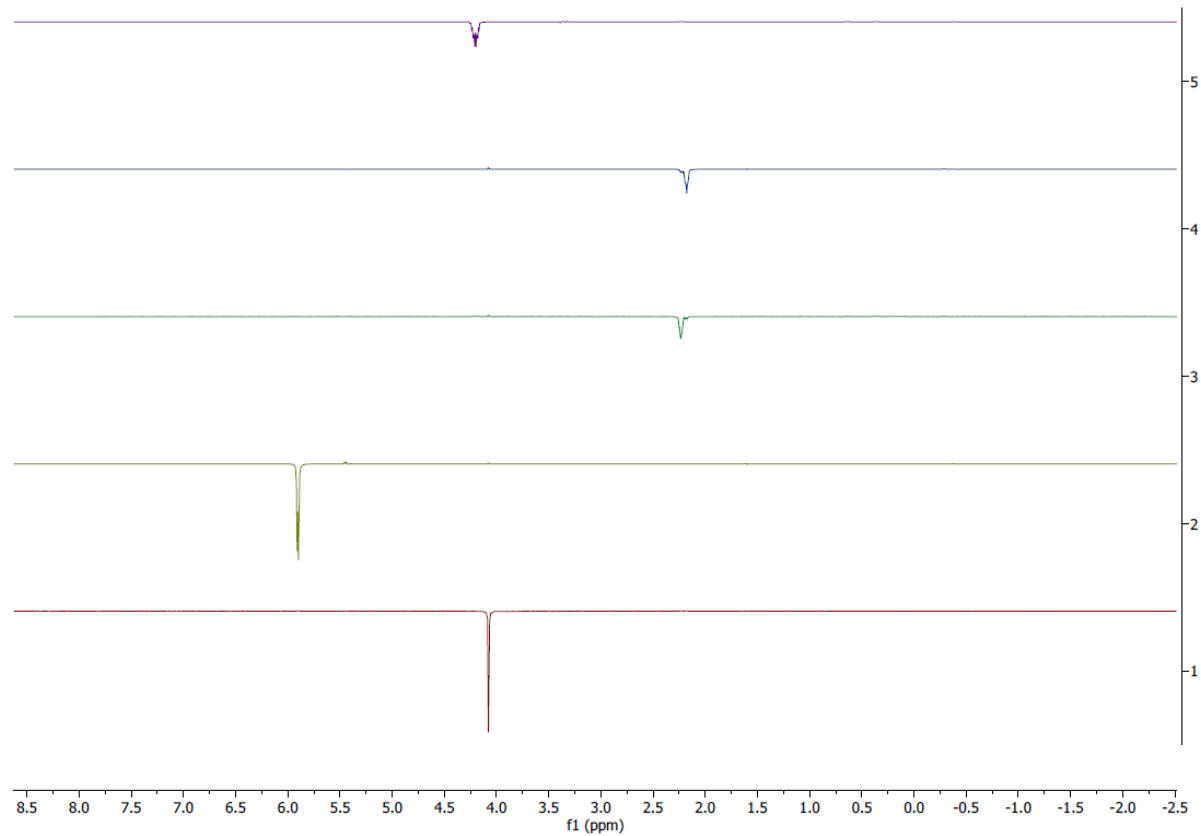


Figure S45.  $^1\text{H}$  NMR NOE spectra of compound SI-16.

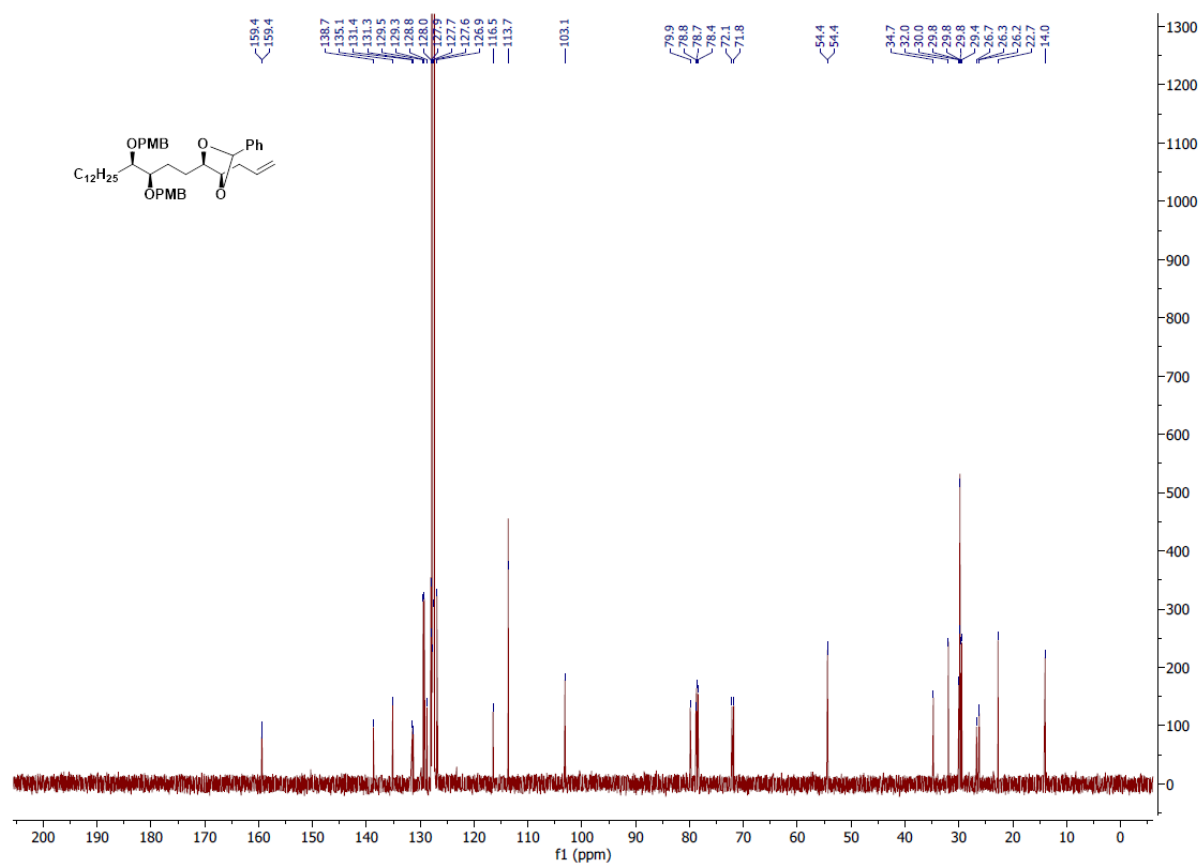
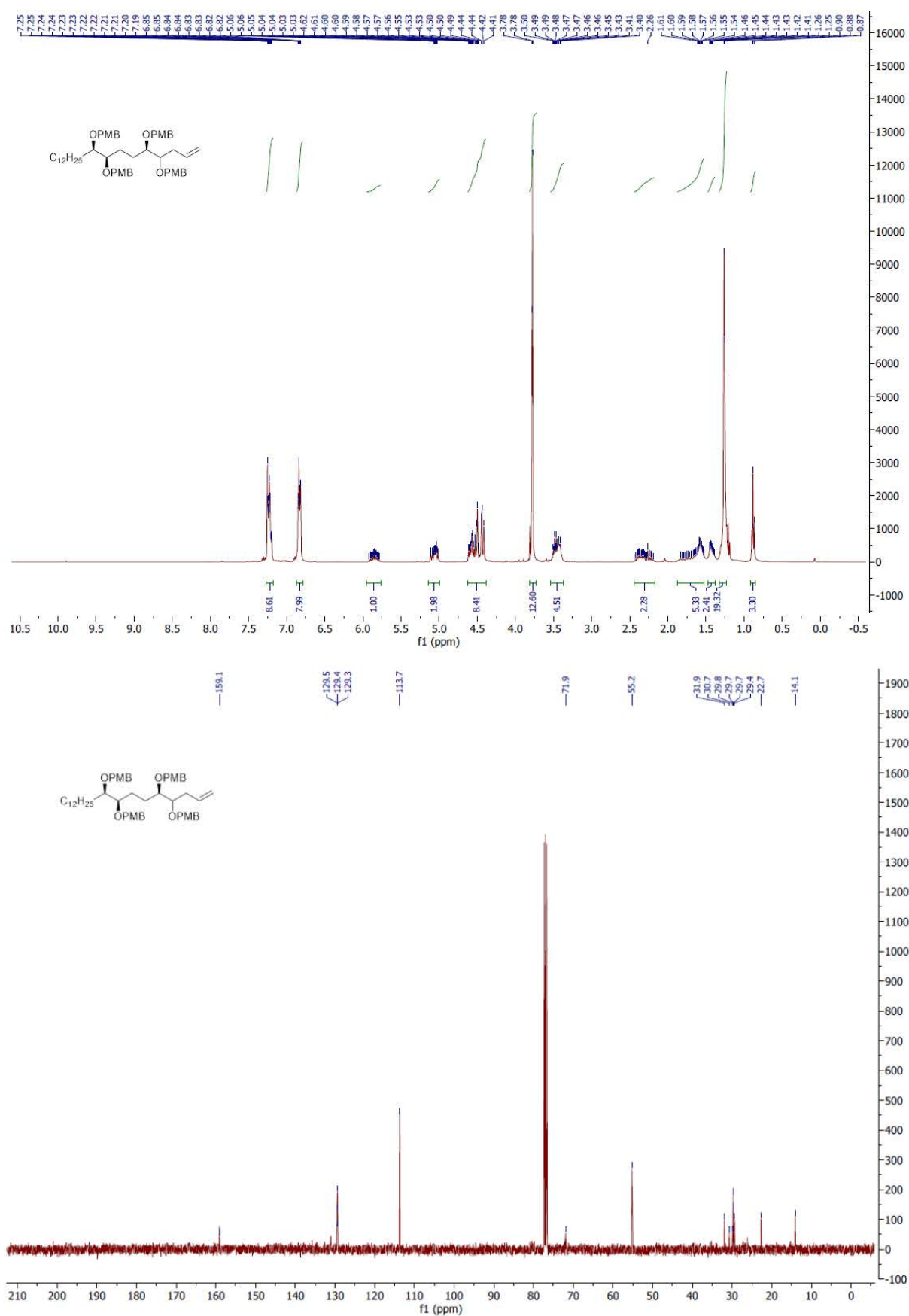


Figure S46. <sup>13</sup>C NMR spectra of compound SI-16.



**Figure S47.** <sup>1</sup>H NMR (top) and <sup>13</sup>C NMR (bottom) spectra of compound 17.

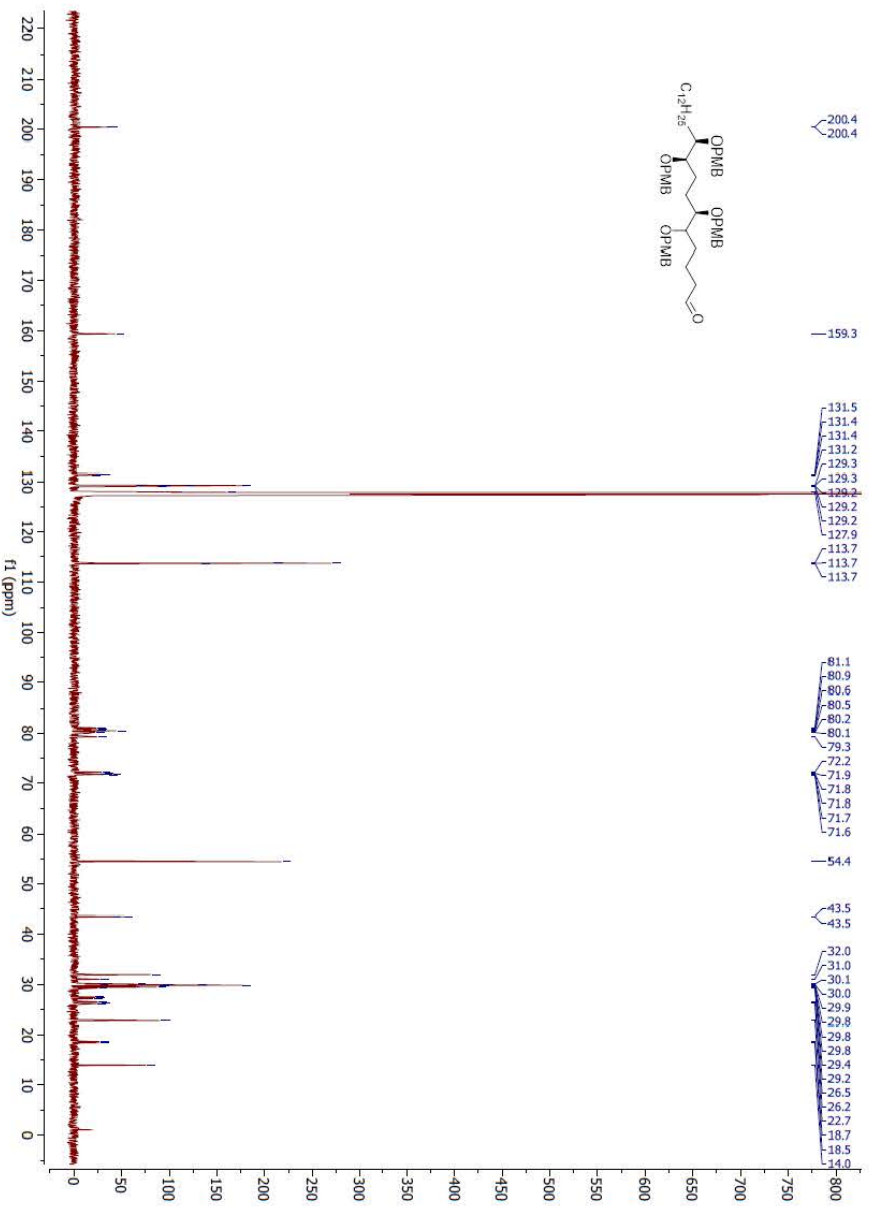
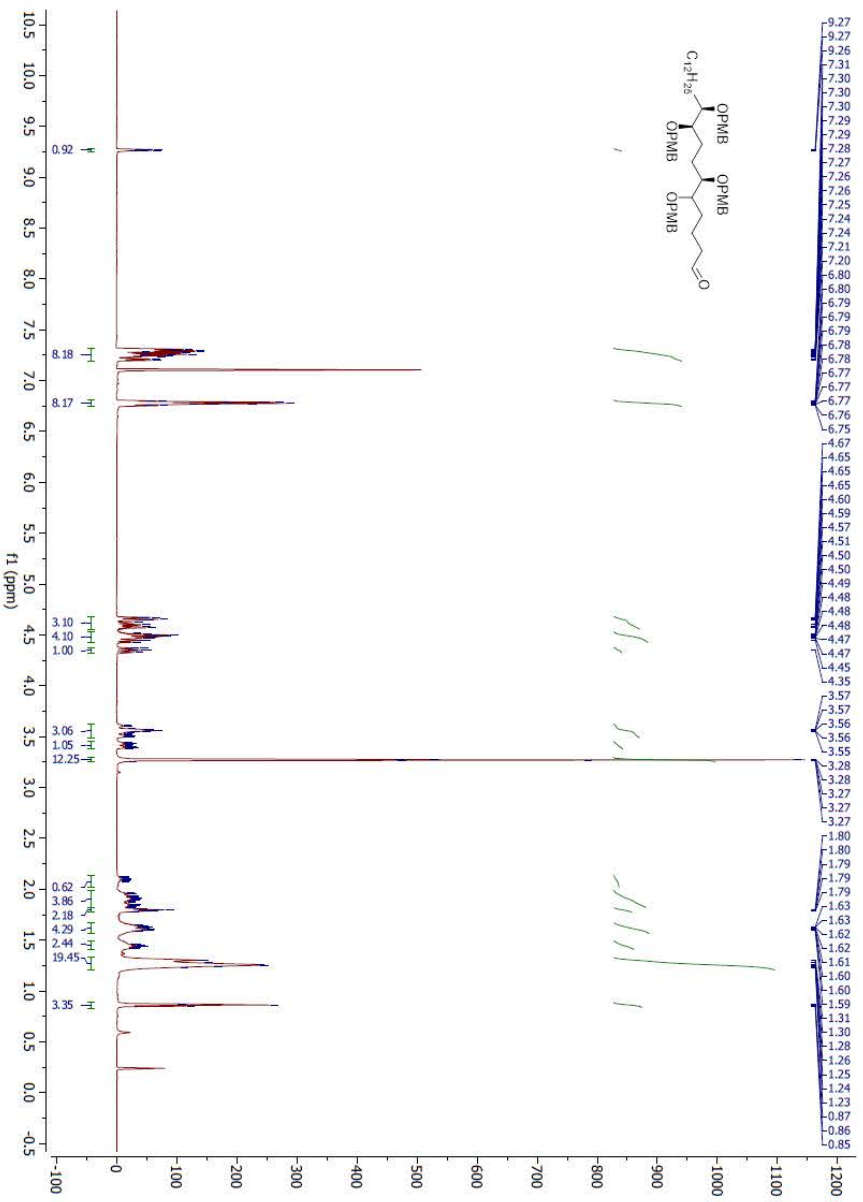


Figure S48.  $^1\text{H}$  NMR (top) and  $^{13}\text{C}$  NMR (bottom) spectra of compound 18.

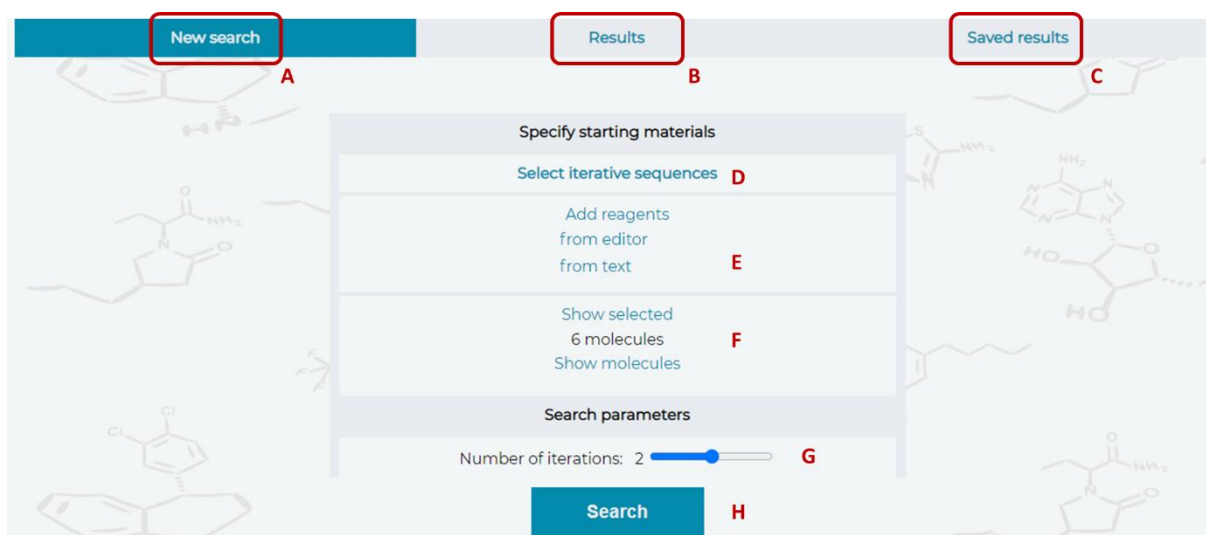


## Section S12. User Manual for Allchemy’s “Iterator” module

### S12.1. Basic Information

Allchemy’s “Iterator” module is freely available to academic users at <https://iterator.allchemy.net>. For optimal performance, we recommend using Google Chrome or other web-browsers supporting SVG2. To register a new account, send an e-mail to [admin@alchemy.net](mailto:admin@alchemy.net) from your academic address. Each user should create individual account. To start using the software, please log in using a valid username and password. After logging in you will see a window providing some technical information. Due to limited capacity of our servers and the fact that iterative searches can easily explode for large number of products (especially when larger numbers of substrates are used) the searches are limited to three iterative loops and three additional, user defined starting materials.

The main control panel visible after logging in is divided into three sections: the first one enables starting new searches (**A** in **Figure S49**; described in detail in **Section S12.2**), the second tab allows for displaying recent results and currently performed operations (**B** in **Figure S49**; described further in **Section S12.3**), while the third one provides access to the previously stored results (**C** in **Figure S49**; described further in **Section S12.4**).



**Figure S49. Main control panel of Allchemy’s “Iterator” module.** **a**, Setting up a new search is available under “New search” (**A**) tab; **b**, Preview of currently performed operations (including checking the position in server’s calculation queue and termination of searches) is available under “Results” (**B**) tab. When no calculations are currently performed, the last calculation’s results are displayed; **c**, Results of the previously performed searches can be retrieved under “Saved results” (**C**) tab; **d**, Selection of iterative sequences; **e**, Panel for adding user-specified building blocks. Adding up to three additional substrates to the calculation is allowed; **f**, Preview of currently selected substrates; **g**,) Number of iterations to be performed; **h**, Launch “Search” button

### S12.2 Starting a new search

Starting a new calculation requires:

- i) Choosing types of iterative sequences to be used during network generation (button **D** in **Figure S49**). The user may select up to three sequences from the displayed list (**Figure S50a**). Once these sequences are selected, their background will change color to light blue. After selecting desired sequences, push “Apply” button to confirm selection.
- ii) Selecting proper building blocks/reagents for selected iterative sequence(s)

**Note 1:** For user’s convenience, basic set of appropriate reagents will be proposed automatically for each iterative sequence selected (**Figure S50b**).

- The user can remove some of the simple starting materials proposed by the software (or replace them with more complicated ones) using “Keep selected” and “Remove selected” controls (**I** in **Figure S50b**)
- The user will be asked to confirm the removal operation (**Figure S50c**). The calculation will not work properly if indispensable substrates are removed (e.g., allylating reagents necessary for Krische’s/Brown allylation or carbon monoxide necessary for the hydroformylation step).

**Note 2:** Optionally, the user can add up to three of his/her own starting materials (**E** in **Figure S49**). The molecules can be added using structure-drawing editor (**recommended**, one can draw multiple molecules at once) or from SMILES string (separated by full stops). The software checks if the added molecule(s) match any of the selected iterative sequences (non-matching ones will not be added to the list). We recommend previewing the selected molecules (**F**) before launching the calculation, especially if the list of reagents was modified.

- iii) specifying the number of iterations to be performed (**G**). User can select up to three iterations to be performed during a single search.
- iv) Pushing the “Search” (**H**) button.

**a)** Select up to 3 iterative sequences from the list.

Krische allylation - Ozonolysis
Krische allylation - Hydroboration
Krische allylation - Hydroformylation
Krische crotylation - Hydroboration
Krische crotylation - Hydroformylation
Krische prenylation - Hydroboration
Krische prenylation - Hydroformylation
Brown alkoxyallylation - Hydroboration
Brown alkoxyallylation - Hydroformylation
Brown allylation - Ozonolysis
Brown allylation - Hydroboration
Brown allylation - Hydroformylation
Brown crotylation - Ozonolysis
Brown crotylation - Hydroboration
Brown crotylation - Hydroformylation
Brown prenylation - Hydroboration
Brown prenylation - Hydroformylation
Homocrotylation of aldehydes - Ozonolysis
Homocrotylation of aldehydes - Hydroboration

**b)** all Proposed building blocks Your compounds

<chem>[O-]C#C</chem>	<chem>C=CC=C</chem>	<chem>C=CC=C</chem>
<chem>CC(C)(C)SiCl</chem>	<chem>C=CC=C</chem>	<chem>C=Cc1ccccc1Cl</chem>
<chem>CCCC=O</chem>	<chem>CC(=O)OCC=C</chem>	<chem>O=Cc1ccccc1</chem>

**c)** You are about to remove at least one proposed molecule, which may cause the iterative sequence to not proceed properly. Do you want to continue?

Continue Cancel

**Figure S50. Starting a new search.** **a**, The user is allowed to select up to three types of iterative sequences from the displayed list. The selected sequences are highlighted in light-blue. **b**, After confirming the choice with “Apply” button, a basic set of matching substrates and reagents is suggested and displayed. User can remove any of the proposed molecules (e.g., to replace them

with more complicated ones) using **J** controls. **c**, Removal of a key reagent will return empty result as the sequence will not be calculated properly. To avoid such outcomes, the user will be asked to confirm the removal operation.

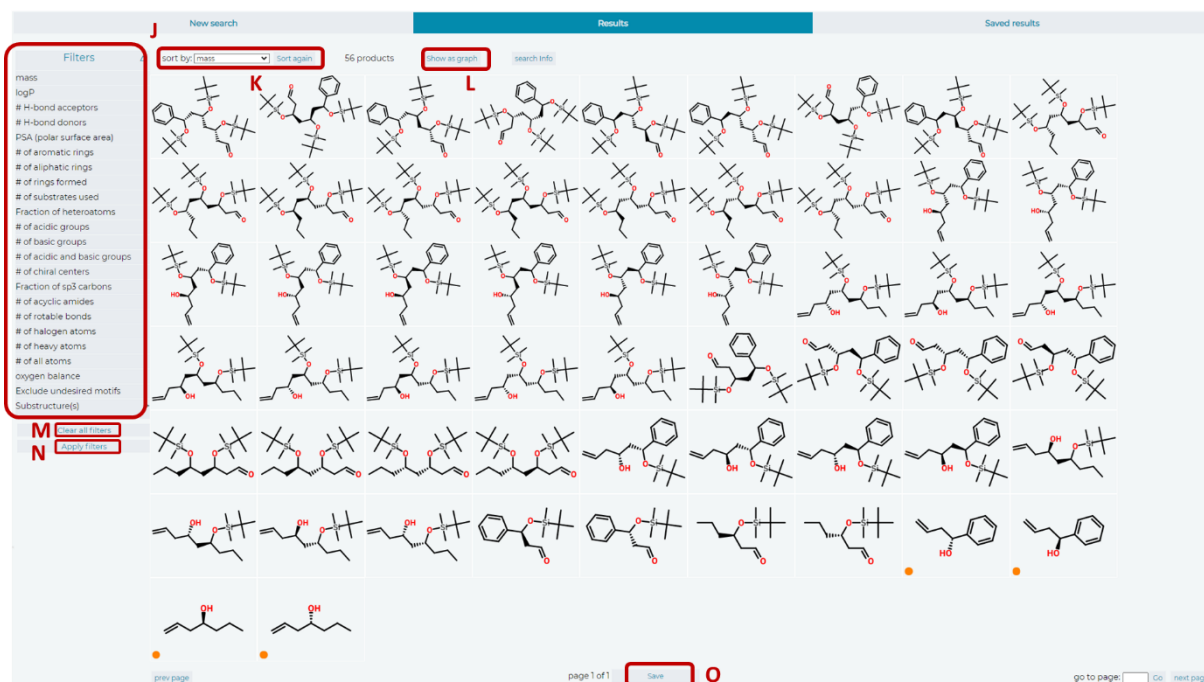
### S12.3 Analysis of results

After launching calculations, the user is transferred to the Results (**B**) tab. The calculation may take from few seconds up to several minutes and the results will be displayed automatically. The search is limited to 30 minutes due to server capacity. Longer calculations will be stopped automatically and already generated results will be returned.

In the default view, results are displayed as a panel of molecular structures (**Figure S51**). As the number of generated products may be in the thousands, we implemented filtering and sorting functionalities (**Figure S51 J** and **K**, respectively) to facilitate their analysis. In particular, the user can sort generated products according to their mass, number of rings, number of stereocenters, etc. using the **K** drop-down menu confirming the new choice with “*Sort again*” button. By default, the molecules are sorted according to their molecular mass. Panel **J** allows for filtering out products that do not meet user-specified structural criteria (molecular mass, number of rings, number of stereocenters, number of H-bond donor/acceptors, number of basic/acidic groups, number of halogens, Polar Surface Area (PSA), etc.). Additionally, the user can filter out molecules according to the number of substrates used to make them or by substructure. The latter filtering option can be used both for retaining or excluding products containing a specific structural motif. The structural motifs should be input in the SMARTS notation.

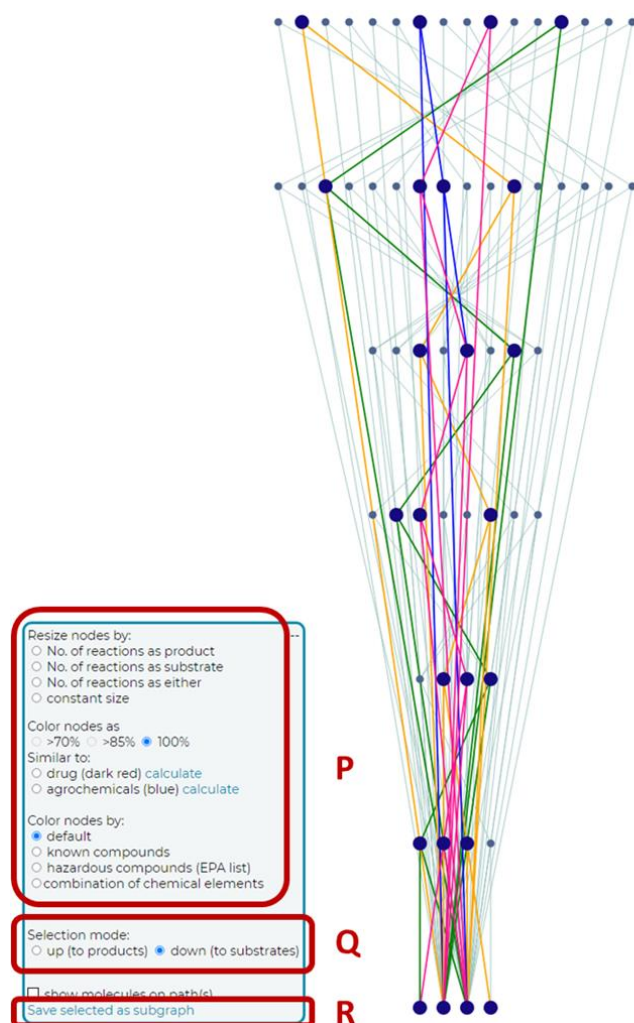
Filters are activated by clicking “*Apply filters*” button (**N** in **Figure S51**). To remove any applied filters, the user should use “*Clean all filters*” button (**M**). Left-clicking on any structure, will display details of the iterative synthetic pathway leading to this molecule. Each reaction in this plan is accompanied by reaction name, typical conditions, typical solvents and literature references (with DOIs as hyperlinks).

The generated results can be saved under user-specified name using “*Save*” button (**O** in **Figure S51**). These saved results will appear in the “*Saved results*” tab (**C**).



**Figure S51. Analysis of results.** In the default mode, the results are displayed as a panel of molecular structures sorted by molecular mass in descending order. The user may change sorting criteria using drop-down menu **K**. After changing these sorting criteria, the choice is confirmed by clicking “Sort again” button. Additionally, the user may apply filtering of products according to their masses, numbers of rings, numbers of stereocenters, etc. using list of filters in panel **J**. The filters are applied by using “Apply filters” **N** button whereas removing any applied filters is possible with “Clear all filters” **M** button. The results are saved under a user-given name by using “Save” (**O**) button. To change the view mode to network, the user should use “View as graph” (**L**) button.

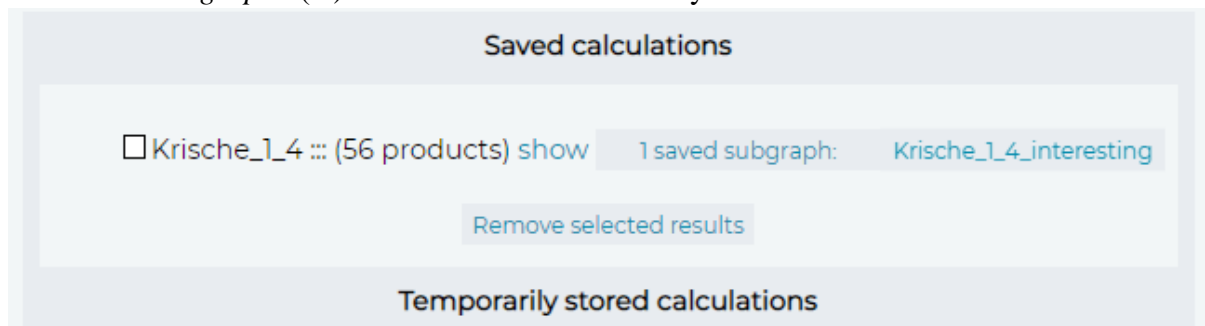
Finally, the results can be displayed in network format (using **L** button in **Figure S52**). In this mode (**Figure S52**) molecules are represented as nodes. The nodes are layered according to the synthetic generation in which they are produced, with substrates in the first row at the very bottom. Hovering over any node displays a structure of a molecule while left-clicking on any node displays the synthetic pathway leading to this molecule. Right-clicking on the node selects the pathway leading to given molecule or from a given molecule, depending on the chosen selection mode **Q**.



**Figure S52. Network view of results with molecules represented as nodes.** In this view, the substrates are located at the bottom. The user can select the synthetic pathway(s) leading to or from a given compound(s) (mode chosen with **Q** control) using right-mouse-button click on any molecule. The selected pathways can be saved as a subgraph using **R** control – the saved search appears in the “*Saved results*” tab (**C**). Left-mouse button click displays the synthetic pathway leading to a given molecule. Additionally, the **P** panel allows for coloring the nodes according to their status (known/present in EPA list/combination of chemical elements), for resizing nodes according to their popularity in the network, or for calculating their similarity to drugs or agrochemicals.

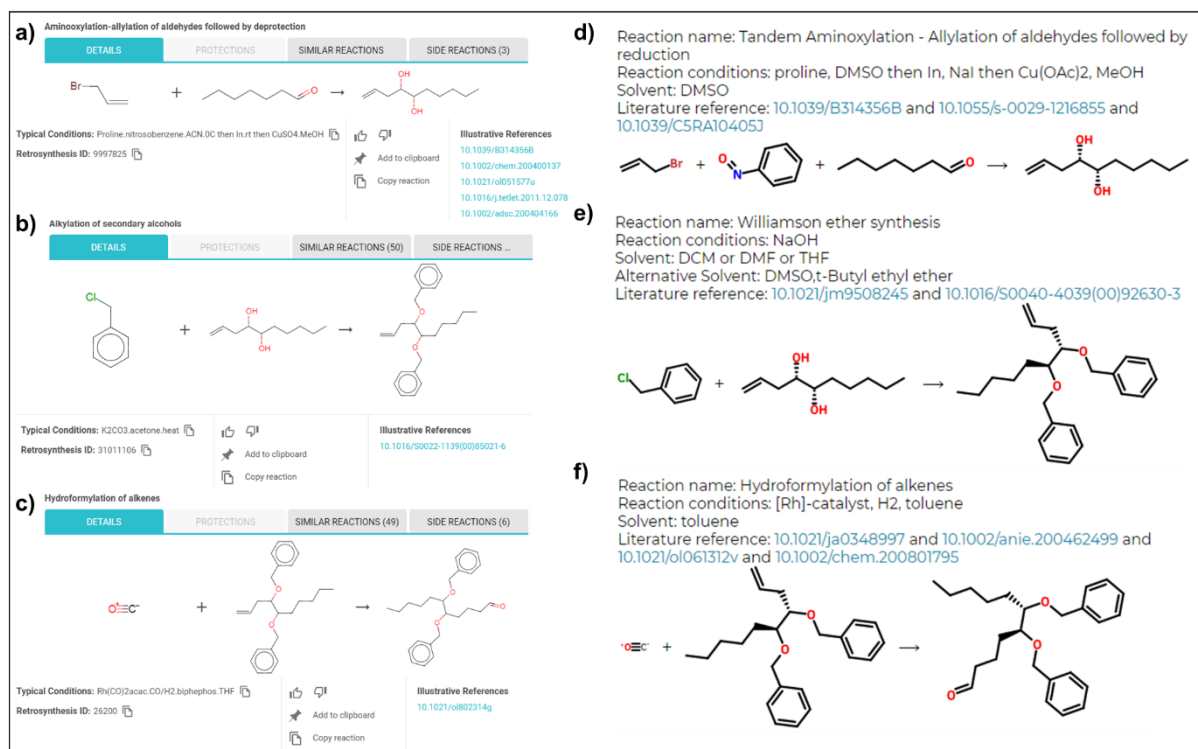
## S12.4 Managing results

The results of searches are displayed automatically after being calculated in the Results (**B**) tab. Any unsaved results calculated during current session are available under “Temporarily stored calculation” in “*Saved results*” tab (**C**). These unsaved calculations are permanently lost after logging out. Calculations saved using “*Save*” button (**O**) and sub-networks saved using “*Save selected as subgraph*” (**R**) are stored unless deleted by the user.

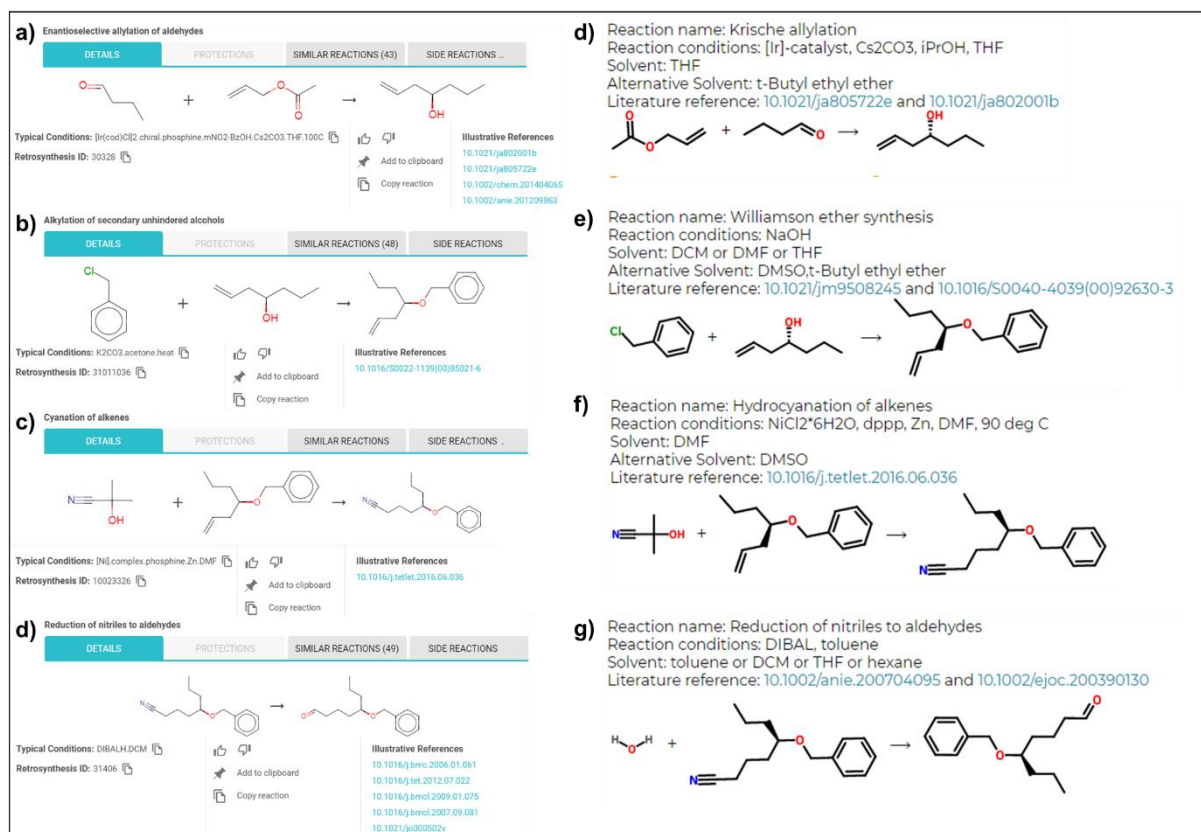


**Figure S53. Saved results tab.** Calculations are saved in “*Saved calculations*” section of “*Saved results*” (**C**) tab (saving is performed with **O** button) under user-given names and are not removed after logging out. The saved subgraphs (saved with **R** button) are available close to parent searches. The results stored in “*Temporarily stored calculations*” section are available only during current session and will be lost after logging out.

## Section S13. Selection of conditions for iterative sequences.

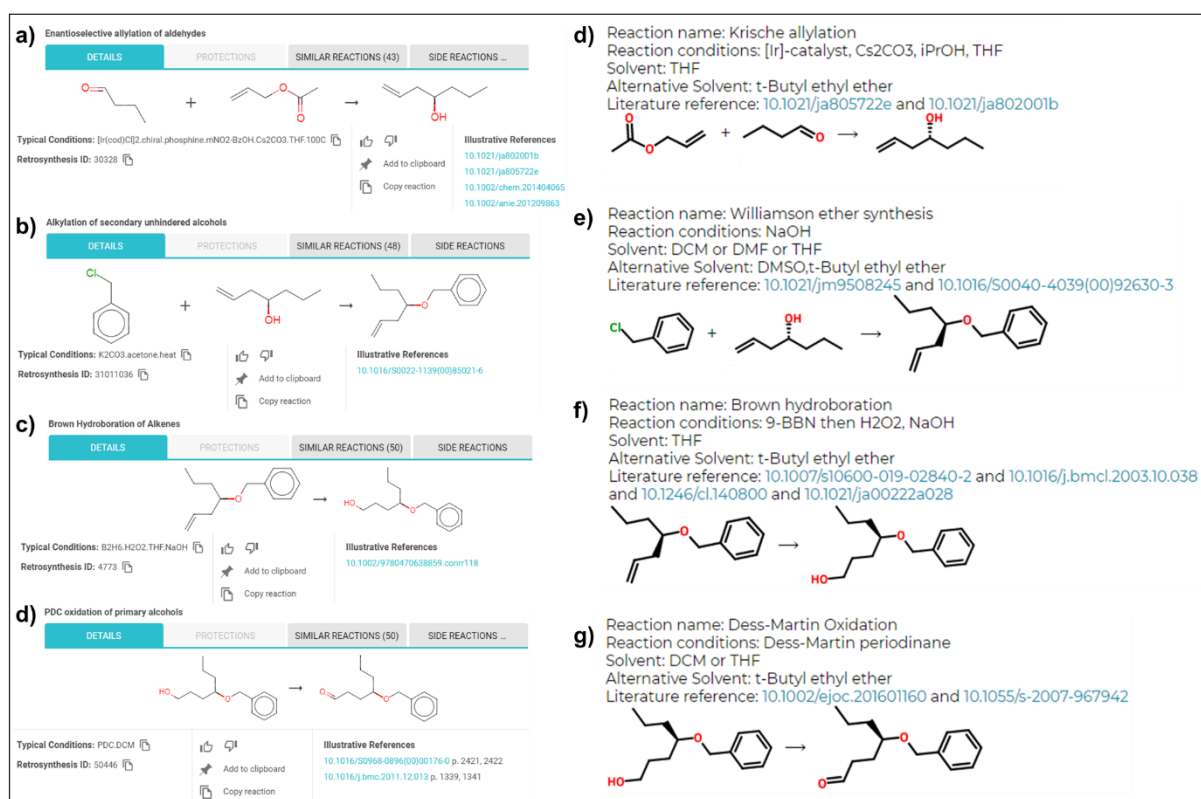


**Figure S54.** Conditions for iterative sequences are sourced from individual reaction rules taught to the computer. This example shows screenshots from **a-c**, Chemica and **d-f**, Allchemy programs for individual steps of the iterative synthesis of 1,2,5,6-polyols via aminoxylation-allylation. Each step lists suggested conditions and provides hyperlinks to illustrative references (please note that Chemica and Allchemy are coded independently of one another and so illustrative literature links are generally not identical). As to the most noticeable adjustments in the computer-proposed conditions, for the cleavage of N-O bond step corresponding to panels (**a-Chemica**, **d-Allchemy**), we used the Zn/AcOH system to achieve 91% yield while the Cu<sup>2+</sup> salts (as suggested by Chemica/Allchemy based on, e.g., <sup>R7</sup>) produced the desired product in 45% yield. Additionally, the Breit's<sup>R15</sup> catalytic system (Rh/6-DPPon) was used rather than the proposed Rh/Xantphos system, because it allowed us to replace the highly flammable and toxic H<sub>2</sub>/CO mixture with HCOOH as the formylating agent<sup>R16</sup> during hydroformylation of alkenes. Finally, in this and other examples from Figure 5, we used the NaH in DMF as a base for protection of secondary alcohols with PMBCl rather than K<sub>2</sub>CO<sub>3</sub> in acetone proposed by Chemica.



**Figure S55.** Conditions for iterative sequences are sourced from individual reaction rules taught to the computer. This example shows screenshots from **a-d**, Chematica and **e-g**, Allchemy programs for individual steps of the iterative synthesis of 1,5-polyols via Krische's allylation (cf. main-text **Figure 5a**). Each step lists suggested conditions and provides hyperlinks to illustrative references (please note that Chematica and Allchemy are coded independently of one another and so illustrative literature links are generally not identical). As to the differences between computer-suggested and experimental conditions, in allylation step corresponding to panels (**a-Chematica**, **d-Allchemy**), we used the improved Krische's catalyst that we happened to have on the shelf – with 4-CN-3-NO<sub>2</sub>-BzOH instead of 3-NO<sub>2</sub>-BzOH ligand proposed by Chematica. In the same spirit of shelf-availability, the hydrocyanation step (**c-Chematica**, **f-Allchemy**) was performed with simple Zn(CN)<sub>2</sub> salt rather than proposed acetone cyanohydrin.





**Figure S56.** Conditions for iterative sequences are sourced from individual reaction rules taught to the computer. This example shows screenshots from **a-d**, Chematica and **e-g**, Allchemy programs for individual steps of the iterative synthesis of *1,4*-polyols via Krische's allylation (**Figure 5b**). Each step lists suggested conditions and provides hyperlinks to illustrative references (please note that Chematica and Allchemy are coded independently of one another and so illustrative literature links are generally not identical). Regarding the differences between computer-suggested and experimental conditions, in the allylation step corresponding to panels (**a-Chematica**, **d-Allchemy**), we used the improved Krische's catalyst that we happened to have on the shelf – with 4-CN-3-NO<sub>2</sub>-BzOH instead of 3-NO<sub>2</sub>-BzOH ligand proposed by Chematica.

## Section S14. Pseudocode for the algorithm to identify iterative sequences.

---

```
1: function GENERALFILTERING(seq)
  ▷ seq - considered reaction sequence
  ▷ returns False if basic conditions for considered reaction sequence are not
  satisfied

2:   if seq.mt = seq.ms then return False
3:   if seq.ms.isTrivial() then return False
      ▷ removes substrates like water etc.
4:   if commonAtoms(seq.mt, seq.mi) = ∅ then return False
5:   if commonAtoms(seq.mi, seq.ms) = ∅ then return False
   return True

6: function GETFRAGMENTS(m)
  ▷ returns set of substructures (fragments) within radius R = 2 for atoms of
  molecule m

7: function FINDITERATIVES(seq)
  ▷ seq - considered reaction sequence, has the following fields:
  ▷ seq.r1 - reaction closer to target
  ▷ seq.r2 - reaction further from target
  ▷ seq.mt, seq.mi, seq.ms, - target, intermediate, considered substrate
  ▷ seq.minters - all intermediates (ie. intermediate and its 'siblings' from r1)
  ▷ seq.msubs - all substrates (ie. considered substrate and its 'siblings' from
  r2)
8:   if not generalFiltering(seq) then return False
9:   frt ← getFragments(seq.mt)
10:  fri ← getFragments(seq.mi)
11:  frsubs ← getFragments(seq.msubs)
12:  Fi ← (fri - frt) ∩ (fri - frsubs)
13:  if (Fi = ∅) then return False
14:  frs ← getFragments(seq.ms)
15:  Fts ← (frt - fri) ∩ (frs - fri)
16:  if findCoreLoop(seq, Fts) then return True
17:  if findABLoop(seq) then return True
18:  return False
```

---

**Figure S57:** A general scheme of detecting iterative sequences (function *findIteratives*; lines 7-18). As input, the algorithm takes sequence *seq* of individual reaction steps (and/or sequences of steps, like FGI, see main text). Reactions entailing any incompatibilities – as determined

based on the list of groups incompatible with a given reaction rule – are excluded. In the ‘general filtering’ phase (`generalFiltering`; lines 1-5) the algorithm removes trivially useless pairs of reactions sequences such as simple loops (*substrate* = *target*, e.g., reduction and then oxidation of the same group), or sequences in which *target* or *substrate* have no atoms common with the sequence’s *intermediates*. In the ‘structural fragment A regeneration filtering’ phase (lines 9-13) the algorithm retains sequences if there exists any fragment (functional substructure within radius R=2 of the molecule’s atom; `getFragments`; line 6) present exclusively in the *intermediate*. Furthermore, iterative sequences are to be identified by the following two functions: `findCoreLoop` (cf. also Figure 53) and `findABLoop`; cf. also **Figure S54**).

---

```

1: function ACCEPTEDFTS( $r_1, m_t, m_{inters}, F_{ts}$ )
  ▷ at least one member of  $F_{ts}$  has only one matching to target and overlaps
  with core of  $r_1$  applied on  $m_t$  producing  $m_{inters}$ 
2:   for  $f_{ts}$  in  $F_{ts}$  do
3:     if ( $m_t.countSubstructures(f_{ts}) \neq 1$ ) then continue
4:     if ( $m_t.substructure(f_{ts}) \cap core(r_1, m_t, m_{inters}) \neq \emptyset$ ) then return
      True
5:   return False

6: function GETSYNTHON( $r, m_1, m_2, m_{lst}$ )
  ▷ returns synthon of reaction  $r$  that, when applied to product  $m$  returning
  substrates  $m_{lst}$ , corresponds to substrate  $m_2$ 

7: function GETSYNTHONFORREACTIONCLUSTER( $r_1, r, m_1, m_2, m_{lst}$ )
  ▷ returns synthon of reaction  $r_1$  analogous to getSynthon( $r, m_1, m_2, m_{lst}$ )
  ▷ assumption:  $r$  and  $r_1$  are from the same reaction cluster (have the same
  name, number of products and number of synthons)

8: function CLOSELOOPCONDITION( $seq$ )
  ▷  $seq$  - considered reaction sequence
  ▷ condition checking the possibility of iterating more times with  $seq$ 
9:    $synt_2 \leftarrow getSynthon(seq.r_2, seq.m_i, seq.m_s, seq.m_{subs})$ 
10:  if not  $seq.m_t.hasSubstructure(synt_2)$  then return False
11:   $m_{fi} \leftarrow applyReaction(reverted(seq.r_2), seq.m_t)$   ▷ generating forward
  intermediate
12:  if  $m_{fi}.atomCount() \leq m_i.atomCount()$  then return False
13:  for  $r$  in  $getClass(seq.r_1)$  do
14:     $synt_1 \leftarrow getSynthonForReactionCluster(r, seq.r_1, seq.m_t, seq.m_i, seq.m_{inters})$ 
15:    if  $m_{fi}.hasSubstructure(synt_1)$  then return True
16:  return False

17: function FINDCORELOOP( $seq, F_{ts}$ )
  ▷  $seq$  - considered reaction sequence
  ▷  $F_{ts}$  - structural fragments present both in the target and the substrate,
  but is absent in the intermediate of considered sequence
18:  if not  $acceptedFts(seq.r_1, seq.m_t, seq.m_{inters}, F_{ts})$  then return False
19:  if not  $closeLoopCondition(seq)$  then return False
20:  return True

```

---

**Figure S58:** Pseudocode of CoreLoops (for example see main-text **Figure 2a**) identification (findCoreLoop; lines 17-20). The algorithm selects sequences of reactions having a substructure (structural motif) present both in the *target* and the *substrate*, but absent in the

*intermediate*. This structural motif overlaps with the “core” of reaction  $r_1$  (‘structural fragment B regeneration’). More precisely, we define collection of substructures  $F_{ts} = (fr_{target} - fr_{intermediate}) \cap (fr_{substrate} - fr_{intermediate})$  (cf. **Figure S52** lines 9-10 and 14-15), and retain sequences in which at least one member of  $F_{ts}$  has only one matching to *target* and overlaps with the  $r_1$  core (acceptedFts; lines 1-5). In addition, in order to assign a sequence of transformations to this category, the following close-the loop conditions have to be satisfied (closeLoopCondition; lines 8-16): *target* contains  $r_2$  core synthon, *forward intermediate*, i.e., product of  $r_2$  applied to the *target* in the forward direction (“reverted  $r_2$ ”) contains core of the synthon from the  $r_1$  class (i.e., expert-coded reaction rule from the same chemical category having the same name, number of products and number of synthons as transformation  $r_1$ ); *forward intermediate* must have more non-hydrogen atoms than *intermediate* to avoid, e.g., unproductive iterations (line 12).

---

```

1: function COUNTNONTRIVIAL( $m_{lst}$ )
  ▷ counts members of  $m_{lst}$  without trivial molecules (eg. water, iodine mono-
  bromide, carbon dioxide)

2: function CROSSINCOMPATIBLE( $seq$ )
  ▷  $seq$  - considered reaction sequence

3:   for  $g$  in  $seq.r_1.incompatibilities \cup seq.r_1.protections$  do
4:     if  $seq.m_s.hasSubstructure(g)$  then return True
        ▷ substrate not compatible with  $seq.r_1$ 

5:   for  $g$  in  $seq.r_2.incompatibilities$  do
6:     if  $seq.m_t.hasSubstructure(g)$  then return True
        ▷ target not compatible with  $seq.r_1$ 

7: function ADDITIONALCHEMICALFILTER( $seq$ )
  ▷  $seq$  - considered reaction sequence
  ▷ returns True for sequences with Grignard reagent as an intermediatea-
  mong intermediates, where a functional group incompatible with the syn-
  thesis of organomagnesium compounds was found among substrates (apart
  from cases, where this group overlapped with core of  $r_2$  creating considered
  Grignard reagent)

8: function FINDABLOOP( $seq$ )
  ▷  $seq$  - considered reaction sequence
9:   if  $countNonTrivial(seq.m_{inters}) < 2$  then
10:    if  $countNonTrivial(seq.m_{subs}) < 2$  then return False
11:    if  $hasProtection(seq.r_1, seq.m_t, seq.m_{inters})$  then return False
        ▷ reactions with incompatibilities already removed while generating
        candidates for reaction sequences
12:    if  $crossIncompatible(seq)$  then return False
13:    if  $additionalChemicalFilter(seq)$  then return False

```

---

**Figure S59:** Pseudocode of ABLoops (for example see main-text **Figure 2b**) identification (findABLoop; lines 8-13). The algorithm filters out the following groups of sequences: (a) one-molecule sequences, i.e., only one non-trivial (countNonTrivial; line 1) synthon resulting from  $r_1$  and  $r_2$  (lines 9-10), (b) having protection on  $r_1$  (line 11; note sequences with incompatibilities at this step are removed earlier), (c) satisfying cross-incompatibility condition, i.e., *target* with incompatibility from  $r_2$ , or *substrate* with either protection or incompatibility from  $r_1$  (crossIncompatible; lines 2-6) (d) returning *True* when applying additionalChemicalFilter (line 7) .

## Section S15. References

- R1)** I. S. Kim, M.-Y. Ngai, M. J. Krische, Enantioselective iridium-catalyzed carbonyl allylation from the alcohol or aldehyde oxidation level using allyl acetate as an allyl metal surrogate. *J. Am. Chem. Soc.* **130**, 6340–6341 (2008).
- R2)** D. Ghosh *et al.*, Synthetically amenable amide derivatives of tosylated-amino acids as organocatalysts for enantioselective allylation of aldehydes: computational rationale for enantioselectivity. *Org. Biomol. Chem.* **11**, 3451–3460 (2013).
- R3)** F. Hessler, A. Korotvička, D. Nečas, I. Valterová, M. Kotora, Syntheses of a Flobufen metabolite and Dapoxetine based on enantioselective allylation of aromatic aldehydes. *Eur. J. Org. Chem.* **2014**, 2543–2548 (2014).
- R4)** J. Cossy, C. Willis, V. Bellosta, S. BouzBouz, Enantioselective allyltitanations and metathesis reactions. Application to the synthesis of piperidine alkaloids (+)-Sedamine and (–)-Prosopphylline. *J. Org. Chem.* **67**, 1982–1992 (2002).
- R5)** G. Wang, X. Xie, W. Xu, Y. Liu, Nickel-catalyzed highly regioselective hydrocyanation of alkenes with Zn(CN)<sub>2</sub>. *Org. Chem. Front.* **6**, 2037–2042 (2019).
- R6)** T. R. Hoye, C. S. Jeffrey, F. Shao, Mosher ester analysis for the determination of absolute configuration of stereogenic (chiral) carbinol carbons. *Nat. Protoc.* **2**, 2451–2458 (2007).
- R7)** G. Zhong, Tandem aminoxylation–allylation reactions: a rapid, asymmetric conversion of aldehydes to mono-substituted 1,2-diols. *Chem. Commun.* 606–607 (2004).
- R8)** M. Simek, K. Bartova, R. Pohl, I. Cisarova, U. Jahn, Tandem anionic oxy-Cope rearrangement/oxygenation reactions as a versatile method for approaching diverse scaffolds. *Angew. Chem. Int. Ed.* **59**, 6160–6165 (2020).
- R9)** A. G. M. Barrett, W. W. Doubleday, K. Kasdorf, G. J. Tustin, Stereochemical elucidation of the pentacyclopropane antifungal agent FR-900848. *J. Org. Chem.* **61**, 3280–3288 (1996).
- R10)** Huang, H. *et al.* Design, synthesis, and biological evaluation of novel nonsteroidal farnesoid X receptor (FXR) antagonists: Molecular basis of FXR antagonism. *ChemMedChem* **10**, 1184–1199 (2015).
- R11)** Schöllkopf, U. & Schröder, R. 2-Unsubstituted oxazoles from  $\alpha$ -metalated isocyanides and acylating Agents. *Angew. Chem. Int. Ed.* **10**, 333–333 (1971).
- R12)** J.K. Augustine, V. Vairaperumal, S. Narasimhan, P. Alagarsamy & A. Radhakrishnan, Propylphosphonic anhydride (T3P®): an efficient reagent for the one-pot synthesis of 1,2,4-oxadiazoles, 1,3,4-oxadiazoles, and 1,3,4-thiadiazoles. *Tetrahedron* **65**, 9989–9996 (2009).

**R13)** Khan, A. T. *et al.* Synthesis of tetra-substituted pyrroles, a potential phosphodiesterase 4B inhibitor, through nickel(II) chloride hexahydrate catalyzed one-pot four-component reaction. *Tetrahedron Lett.* **53**, 4145–4150 (2012).

**R14)** Laha, J. K., Tummalapalli, K. S. S. & Gupta, A. Palladium-catalyzed domino double *N*-arylations (inter- and intramolecular) of 1,2-diamino(hetero)arenes with *o,o'*-dihalo(hetero)-arenes for the synthesis of phenazines and pyridoquinoxalines. *European J. Org. Chem.* **2013**, 8330–8335 (2013).

**R15)** Kemme, S. T., Šmejkal, T. & Breit, B. Combined Transition-Metal- and Organocatalysis: An Atom Economic C3 Homologation of Alkenes to Carbonyl and Carboxylic Compounds. *Chem. - A Eur. J.* **16**, 3423–3433 (2010).

**R16)** Ren, W. *et al.* An Effective Pd-Catalyzed Regioselective Hydroformylation of Olefins with Formic Acid. *J. Am. Chem. Soc.* **138**, 14864–14867 (2016).

