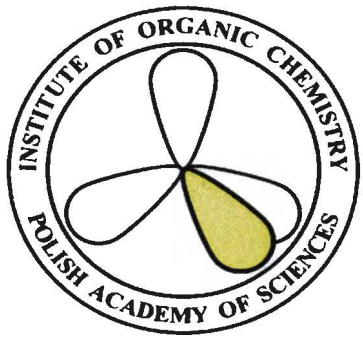
PRACA DOKTORSKA

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES



DOCTORAL THESIS

Teaching the computer reactivity rules and strategies of automated retrosynthetic planning

mgr inż. Sara Szymkuć

Ph.D. thesis in the form of a collection of research articles presented to the Scientific Council of the Institute of Organic Chemistry of the Polish Academy of Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemistry.

Supervisor: Prof. Bartosz A. Grzybowski

Warsaw 2018

I would like to thank:

Prof. Bartosz Grzybowski who taught me how to ask meaningful scientific questions.

The team XI for support and outstanding atmosphere.

My parents who always encouraged me to follow my dreams.

Streszczenie w języku polskim/ Abstract in Polish

Nauka komputera projektowania ścieżek syntezy związków organicznych stanowi jedno z najstarszych wyzwań chemii obliczeniowej. Pierwszy program podejmujący się rozwiązania tego problemu został opracowany już w latach 60. XX wieku. Wiele innych programów było rozwijanych na przełomie lat 80. i 90. lecz żaden z nich nie sprostał wymaganiom chemików organików, co przyczyniło się do utraty zainteresowania tą dziedziną nauki w latach 2000. Było to niezwykle niefortunne, gdyż w międzyczasie komputery opanowały inne umiejętności analityczne, dotychczas uważane wyłącznie za domenę ludzkiego intelektu oraz kreatywności np. symboliczne rozwiązywanie złożonych równań różniczkowych (Mathematica) czy opanowanie gier strategicznych (szachy, Go) na poziomie przewyższającym ludzkich mistrzów. W mojej pracy doktorskiej, czerpałam inspirację (oraz nadzieję) z tych dokonań, przez kilka lat rozwijając platformę obliczeniową *Chematica*, która planuje ścieżki syntezy związków organicznych.

Pierwszym krokiem w nauce komputera chemii było przyjęcie właściwego formatu danych dla reakcji chemicznych oraz cząsteczek w postaci zrozumiałej dla maszyny. Alfanumeryczna notacja SMILES/SMARTS została wybrana ze względu na szybkość wykonywania operacji pojedynczych reakcji chemicznych oraz możliwość inkorporacji szczegółowych informacji na temat stereochemii.

Kolejnym etapem była nauka komputera reakcji organicznych. Dla każdej klasy transformacji, szczegółowo badałam jej mechanizm i dokładnie wyznaczałam rdzeń na który składały się motywy strukturalne wraz z dopuszczalnymi podstawnikami. Każda reakcja zawiera także informację kontekstową opisującą niekompatybilne bądź wymagające protekcji grupy funkcyjne znajdujące się poza rdzeniem oraz typowe warunki dla danej transformacji. Sformalizowana przeze mnie struktura bazy danych oraz reguły reakcji stanowią podwalinę wiedzy chemicznej programu *Chematica* na którą obecnie składa się 60 tysięcy reakcji, z których osobiście zakodowałam około 15 tysięcy.

Nadrzędnym celem mojej pracy była nauka komputera samodzielnego projektowania ścieżek syntezy. Kluczową rolę odgrywało zdefiniowanie funkcji oceny pozwalającej algorytmowi wyszukiwania na rozpoznanie czy porusza się we właściwym kierunku oraz jak daleko znajduję się od substratów. Zaproponowałam dualną funkcję, oceniającą zarówno skomplikowanie cząsteczkowe jak i szansę realizacji poszczególnej reakcji.

Niestety, nawet prawidłowe, pojedyncze reakcje nie gwarantują jeszcze utworzenia rozsądnej ścieżki syntezy. Mając świadomość tego problemu, skoncentrowałam się na identyfikacji najbardziej obiecujących sekwencji reakcji oraz eliminacji jałowych kombinacji.

Niezależnie od nowatorskości idei kryjącej się za programem retrosyntetycznym, jego użyteczność powinna zostać zweryfikowana. Początkowo przeprowadziłam „walidację na papierze", w której program niezależnie odtworzył opublikowane ścieżki syntezy dla wybranych związków organicznych. Po tym teście przyszedł czas na weryfikację proponowanych przez program syntez w laboratorium. Osiem ścieżek syntetycznych zaprojektowanych przez Chematicę zostało wykonanych przez chemików z firmy Sigma-Aldrich, Uniwersytetu Northwestern oraz ICHO PAN. Wszystkie zakończyły się sukcesem dokumentując pierwszą znaną walidację eksperymentalną programu retrosyntetycznego.

Następnie moje zainteresowania naukowe skierowały się w stronę chemii systemów. W tym duchu ostatnia część rozprawy doktorskiej opisuje komputerowe odkrywanie cykli reakcji chemicznych, m.in. wzorujących się na cyklach biologicznych i potencjalnie użytecznych jako metoda recyklingu katalizatorów czy autoamplifikacji użytecznych związków.

Abstract

Teaching computers to design syntheses of organic molecules has been one of the oldest challenges of computational chemistry. First software packages aiming to solve this problem were developed already in the late 1960s. Many other programs were created in the 1970s and 1980s but none of them met the expectations of organic-synthetic chemists and the effort was largely abandoned by the 2000s. This is quite unfortunate given that, in the meantime, computers have mastered many other analytical skills that had been considered exclusive domains of human intellect and creativity – for example, they can solve complex differential equations in symbolic forms (Mathematica) or can play games of strategy (chess, Go) better than human champions. In my doctoral thesis, I have taken inspiration (and hope) from these advances and for several years have been developing a computational platform known as *Chematica* that could finally plan efficient chemical syntheses.

The first step in teaching computer chemistry was to employ proper machine-readable data format for reactions and molecules. SMILES/SMARTS alphanumeric notation was chosen largely because of the speed with which it can process reaction operations and also because it was possible to augment it with detailed information about stereochemistry.

Equipped with this suitable notation, I undertook the challenge to teach the machine a nearly complete selection of organic reaction types. For each reaction class, I ventured deep into the underlying mechanism and delineated carefully the reaction core encompassing structural motifs and admissible substituents as well as "contextual" information describing incompatible functional groups, need for protection outside the core, and information about typical conditions. The rules and database fields I formalized underlie *Chematica*'s knowledge base of over 60,000 reactions of which I personally coded ca. 15,000.

The ultimate goal of my work was to teach the machine how to plan complete synthetic pathways without any human intervention. The key element here was to define proper scoring function enabling the search algorithm to estimate whether it is "moving" in a promising "synthetic direction," and how far it is from starting materials. To this end, I proposed a dual scoring function that assesses "synthetic positions" based on both molecular complexity and reaction feasibility.

Unfortunately, even correct but logically isolated synthetic steps do not necessarily make up for a sensible pathway. Recognizing this problem, I focused on how to identify the most promising reaction sequences and eliminate those that are unproductive or problematic.

The ultimate value and usability of any retrosynthetic software lies in experimental validation of its predictions. Initially, I performed "paper validation" whereby the program blindly recreated some published synthetic routes. The next step was the wet-lab validation. In this ultimate test, *Chematica* designed eight syntheses that were subsequently executed by chemists at Sigma-Aldrich, Northwestern University, and in our own laboratory at ICHO PAS. All computer designs were confirmed experimentally establishing the first-ever validation of a retrosynthetic software.

In the meantime, my scientific interests have been gradually shifting to new areas, especially to systems chemistry. In this spirit, the last chapter of my thesis describes recent work on the computational discovery of chemical-reaction cycles, akin to those used by biological systems and potentially useful as a means to recycle catalysts or to autoamplify valuable chemicals.

Summary

1. List of publications

Publications included in the doctoral thesis:

[P1] Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, **Sara Szymkuć**, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532 **(cover art)**.

[P2] Michał Bajczyk, Piotr Dittwald, Agnieszka Wołos, **Sara Szymkuć** and Bartosz A. Grzybowski "Discovery and Enumeration of Organic-Chemical and Biomimetic Reaction Cycles within the Network of Chemistry" *Angew. Chem. Int. Ed.* **2018**, 57, 2367- 2371 **(VIP paper)**.

[P3] Bartosz A. Grzybowski, **Sara Szymkuć**, Karol Molga, Ewa P. Gajewska, Agnieszka Wołos "Synthetic design with the *Chematica* program – the importance of accurate rules and of higher-order logic" *CHIMIA* **2017**, 71, 512.

[P4] Grzegorz Skoraczyński, Piotr Dittwald, Błażej Miasojedow, **Sara Szymkuć**, Ewa P. Gajewska, Bartosz A. Grzybowski, Anna Gambin "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? " *Sci. Rep.* **2017**, 7, 3582.

[P5] **Sara Szymkuć**, Ewa Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk and Bartosz A. Grzybowski "Computer-assisted synthetic planning: The end of the beginning" *Angew. Chem. Int. Ed.* **2016**, 55, 5904-5937.

[P6] Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, **Sara Szymkuć**, Piotr Dittwald, Karol Molga and Prof. Bartosz A. Grzybowski "A Priori Estimation of Organic Reaction Yields", *Angew. Chem. Int. Ed.* **2015**, 54, 10797-10801.

Other publications:

1. Michał Woźniak, Agnieszka Wołos, Urszula Modrzyk, Rafał L. Górski, Jan Winkowski, Michał Bajczyk, **Sara Szymkuć**, Bartosz A. Grzybowski, Maciej Eder "Linguistic measures of chemical diversity and the "keywords" of molecular collections" *Sci. Rep.* **2018**, *8*, 7598.

2. Bartosz A. Grzybowski, **Sara Szymkuć**, Ewa P. Gajewska, Karol Molga, Piotr Dittwald, Agnieszka Wołos, Tomasz Klucznik, "Chematica: A Story of Computer Code That Started to Think like a Chemist" *Chem*, **2018**, 4, 390-398.

3. Szymon Kłossowski, Adam Redzej, **Sara Szymkuć**, Ryszard Ostaszewski "Studies towards enzymatic kinetic resolutions of 1,3-diol peptidomimetics obtained via the Ugi reaction" *Arkivoc*, **2013**, iv, 134-143.

4. **Sara Szymkuć**, Ryszard Ostaszewski , "Biocatalytic methods for preparing of nonracemic arylallylic alcohols", *Wiadomości Chemiczne*, **2012**, 66, 93.

2. Conferences and workshops

1. Poster presentation: *"Chess-like algorithms behind Chematica's retrosynthetic planning"*, 250th American Chemical Society National Meeting & Exposition 16-20.08.2015 Boston, USA. **Poster awarded with 2015 CINF Scholarship for Scientific Excellence.**

2. Workshop: *"AI for Scientific Progress: Bringing Digital Control to Physical Matter"*, 30.09-2.10.2016, Palo Alto, USA.

3. Oral presentation: *"Automatic discovery and enumeration of new tactical combinations"*, 256th American Chemical Society National Meeting & Exposition 19-23.08.2018 Boston, USA.

3. Hypothesis and purpose of the work

The main hypothesis of my doctoral work has been that by combining proper representation of organic-chemical knowledge with the power and methods of modern computing, it would finally be possible – after over five decades of effort by various groups – to design a software system capable of autonomous planning of synthetic routes leading to arbitrary target molecules. In validating this hypothesis, my work focused on teaching the computer the rules of organic-chemical reactivity in machine readable format and augmenting such rules with various heuristics fine-tuning the reactivity patterns beyond the reaction cores. In the end, my work laid ground for what is today is known as the *Chematica* platform for computer-assisted synthetic planning.

4. Introduction and historical background

Documented origins of the Computer Assisted Organic Synthesis (CAOS) can be traced as far back as 1963 when a relatively poorly known Russian émigré into the United States, Vladimir Vleduts published a paper in which he envisioned computers able to design synthetic routes [1]. Such planning would work in "backwards" direction from the target towards simpler intermediates until ultimately reaching starting materials available from *"the set of initial compounds"*. In this approach, the putative routes generated would constitute branches of a synthetic tree of all possibilities, and alternative pathways should be compared and ranked according to some user-specified criteria. Vleduts also astutely stressed that a "strategy" of sorts should guide the searches so that solutions could be found in reasonable times. Two years later, in 1967, these general principles were further codified by Prof. E.J. Corey [2], who christened this methodology as "retrosynthesis" (or "retrosynthetic analysis"). Corey's ideas forever revolutionized the way in which chemists approach synthetic planning. Shortly afterwards, in 1969, Corey and Wipke disclosed the first software for retrosynthetic analysis called OCSS (Organic Chemical Simulation of Synthesis) [3]. This program was not automated, however, in the sense that that the user had to manually choose between the options the machine generated at each step. Down to some technical detail, the chemical knowledge (i.e., database of reaction rules underlying OCSS) was separated from the program's code. This dichotomy proved to be quite flexible in terms of any changes/updates and was widely used in future retrosynthetic programs. OCSS subsequently split in two different projects: LHASA (Logic and Heuristics Applied to Synthetic Analysis) lead by Corey and SECS (Simulation and Evaluation of Chemical Synthesis) by Wipke.

LHASA remained an interactive, design tool, in which human operator navigated the trees of synthetic possibilities step-by-step. Chemical rules were written in an English-like notation called CHMTRN (CHeMistry TRaNslator) along with SMILES-like notation called PATRAN (PATtern translator) and were stored in a database separated from the source code [4]. The transforms (2271 rules as of 2004, version 20.3) [5] were divided into two subgroups: (i) the so-called goal transforms simplifying the structure (mainly carbon-carbon bond formation) and (ii) sub-goal transforms, not simplifying skeletal but allowing for the manipulation of functional groups [6]. Combinations of goal and sub-goal transformations were referred to as "tactical combinations" [7]. In addition to the knowledge base and some protection data, the program encompassed five design strategies:

(i) transform-based; (ii) structure-goal (S-goal); (iii) topological; (iv) stereochemical; and (v) functional group oriented [8] that along with tactics guided the analysis and limited the size of a retrosynthetic tree. While LHASA was an ingenious contribution well ahead of its times and has generated considerable interest in the community, its predictions – like those of so many other programs we will see in this introduction – were never validated in synthetic laboratory practice (or, at least, there are no published reports of any such validation).

The other offspring of OCSS, Wipke's SECS, was conceptually similar to LHASA and required human interaction to select synthetic pathways [9]. SECS's knowledge base was written in ALCHEM language, an English-like machine-readable format [10]. A substantial effort during this program's development was put into recognizing and analyzing stereochemistry of reactions and molecules [10,11,12]. An offshoot of SECS, called CASP (Computer Aided Synthesis Planning) was used and financed by a consortium of Swiss and German pharmaceutical companies. It had a considerable reaction knowledge base and introduced graphical representation of chemical rules. The project's funding was ultimately disconnected for reasons that were never disclosed.

The first retrosynthetic program aiming at fully automatic – that is, without step-by-step navigation under user's control – retrosynthetic design was Gelertner's SYNCHEM. It used WLN (Wiswesser Linear Notation) representation for molecular structures and its reaction knowledge base comprised some 1000 general-type reactions (called "schemas") as well as a collection of few thousand of available starting materials from Aldrich's catalogue. The algorithm recognized promising "synthemes" (functional groups or structural motifs) in the target molecule and applied appropriate transformations ("schemas") from the reactions' database corresponding to a chosen "syntheme". The library of reaction schemas was grouped into chapters corresponding to the syntheses of a given "syntheme". When the program failed to design a route to a given target, missing reaction rules were often added *a posteriori* and the program's performance was re-checked [13]. If no serious reactivity conflicts were detected, an intermediate ("sub-goal") was generated. Sub-goal molecules were then scored and ranked according to both estimated complexity of the sub-goal molecule and reaction's merit/feasibility. The most promising candidates were further expanded, and the expand-score-expand cycle was repeated until the program reached molecules from the database of starting materials. Unfortunately, the algorithm was unable to keep track of any "global history" of the putative syntheses which made its tactics "short-termist" [12,13]. Such problems were compounded by the scopes of the reaction "schemas" being too broad which often resulted in chemically naïve routes or intermediate molecules that could simply not exist (e.g., those violating Bredt's rules). Further development of SYNCHEM was ultimately abandoned, largely because the underlying WLN notation could not handle stereochemistry. The authors then focused their efforts on the development of its successor, SYNCHEM2. In one of the improvements, the SLING notation replacing WLN allowed for at least rudimentary handling of stereochemistry [4,13]. The chemical rules in the new version could be applied two-ways, in both retrosynthetic and forward directions. Inverse application of reaction schema was intended to evaluate selectivity of particular steps in a synthetic pathway predicting possible stereoisomeric byproducts and estimating yields for each reaction. The software was also capable of performing reactions at multiple reaction sites [14]. During project's

development, the authors also made first documented attempts to automatically extract reaction rules from databases of literature reactions, rather than to code these rules laboriously by hand [15]. Still, the program was not widely adopted by the community and the last publication appeared in 1998 (it described parallelization of the SYNCHEM search algorithm [16]).

Next notable development was the SYNLMA software for automated synthetic design developed by P.Y. Johnson's group. A distinctive feature of the system, as emphasized by the authors, was its division into three autonomous parts: a "reasoning" component, a knowledge base, and a user interface. In this way, different representations of chemical knowledge and different reaction databases could be readily tested without the need to redesign the entire software. SYNLMA was capable of planning non-stereoselective syntheses for low complexity drugs like Ibuprofen or Darvon and some bicyclic systems (e.g., cocaine). On the flipside of the coin, the knowledge database contained only 200 select reaction rules and 50 starting materials. The authors themselves pointed out that the software's synthetic trees were generated and navigated in a "naïve or inefficient manner" (e.g., producing structurally impossible intermediates). Attempts to plan a synthetic route for more complex molecules or connecting with large commercial databases of starting materials failed. For more complicated targets, the program generated retrosynthetic trees too large for it to navigate. Authors planned to remedy the situation by redesigning the software, introducing "planning strategies," and changing the structure of the reaction-rules database, but no such improved version was ever disclosed [17].

Another notable contribution was SYNGEN, a program developed by Hendrickson and aiming at automated design of economically optimal, convergent synthetic routes. The program identified a set of ordered bonds to be disconnected based on the target's scaffold, thus defining the general "direction" of the synthetic route. The main idea was the primacy of constructing molecular *skeleton* (σ C-C bonds) over introducing functionalities. The authors introduced a concept of an *"ideal synthesis"* that employs only skeletal reactions and does not entail any re-functionalizations of intermediates. Such skeleton-centered approach considerably pruned the size of a retrosynthetic tree but more constraints were still needed to avoid combinatorial explosion of synthetic possibilities, especially for less-trivial targets. In this spirit, the algorithm considered only *convergent* routes in which the retron was disconnected into two synthons. Maximally two bonds per iteration could be disconnected and not more than six bonds could be cut over the entire pathway. Additionally, to avoid "asymmetric" retrosynthetic trees, the smaller substrate from the first generation had to contain at least 25% of target's carbon atoms. The number of retrosynthetic steps was limited to two with additional restriction that all four substrate scaffolds had to be present in the database of 6,000 commercially available starting materials. While SYNGEN's chemical transformations could be applied both in retro- and forward directions, they did not handle stereochemistry which was regarded as secondary with respect to skeletal considerations [18]. To test if a given reaction was applicable to a particular molecule it was subject to a set of "Mechanistic Tests" inspecting potential reactivity conflicts, requirements for activating groups, etc. Unfortunately, strict ban on re-functionalizations resulted, for many molecules, in an empty set of results. To address this issue and allow for some re-functionalizations and generation of diverse derivatives of a given target, a software called FORWARD was also being developed by the same

group – apparently, it was never brought to fruition and the last paper describing it was published in 1990 [19].

In contrast to programs using hard-coded reaction rules, IGOR (Interactive Generation of Organic Reactions) [20] employed the Dugundji-Ugi (DU) model to describe chemical reactions as R matrices and molecules or ensembles of molecules (EM) as BE (bond-electron) matrices. Reactions R matrices corresponded to valence electron redistribution patterns and were obtained by subtracting BE matrix of substrates from BE matrix of products. IGOR was not restricted to retrosynthetic analysis and could also be applied in the forward direction, predicting new reactions. Unfortunately, the matrix notation turned out to be problematic when working in a retrosynthetic direction, since to generate a reaction it required the knowledge of all products and even the simplest byproducts, including water molecules, chloride ions, etc. To overcome this complication, a separate program called STOECH was developed to generate byproducts. Still, even with this improvement, IGOR required a well-trained chemist as an operator to correctly fine-tune the search parameters [21]. The software was more suitable to explore the space of possibilities and explore novel, unprecedented reactions than to design synthetic routes.

WODCA (Workbench for the Organization of Data for Chemical Application) developed by Johann Gasteiger also departed from the synthon–based approaches limited to literature-known reactions [22]. Instead, it modelled chemical reactions based on physicochemical properties of chemical bonds and atoms (polarity, inductive effects, resonance and polarizability effects). Molecules were presented as BE (bond-electron) matrices as in IGOR software. Analyses were then performed in an interactive step-by-step manner whereby each intermediate had to be accepted by the user. The software comprised four strategies for identifying strategic bonds in the target molecule. Each of those employed different general reaction types (e.g., carbon-heteroatom bond formation, synthesis of aliphatic bonds, aromatic substitution or synthesis of polycyclic compounds). In order to verify a proposed retrosynthetic disconnection, WODCA was interfaced with the database of known reactions looking for the closest literature precedent. In addition to the strategy-based search, the user could try to identify a starting material based on the similarity to the target. WODCA was able to assist both in synthesis planning and in substructure searches within combinatorial libraries [23]. On the other hand, the program was more of an idea-generator rather than a fully automated tool for planning complete synthetic routes. Active development of the software ceased in 2005.

Continuing our survey, Hanessian's CHIRON [24] was a software designed to identify accessible chiral starting materials, either commercially available or otherwise known in the literature. The approach aimed at minimal modification of the target's stereochemistry and functionalities with respect to the substrates. The program compared the target's structure with the database of starting materials looking for the maximal overlap using Morgan's algorithm. If an exact match was not found, a series of functional group interconversions was applied to achieve the best possible match between one of the substrates with the target or target's fragment. Although the software was capable of cleaving a double bond or a diol and evaluate the reshaped precursor, it was unable to combine two starting materials (e.g., to create a six-membered ring in a Diels-Alder reaction from two substrates). Parts of a precursor's molecule requiring modifications were tagged with appropriate keywords describing "chemical events" (e.g., annulation,

oxidation, reduction) that should produce the desired target's substructure. CHIRON was restricted to propose only starting materials without a detailed synthetic plan – the choice of specific reactions was left to chemist's creativity.

All of the platforms described so far are no longer under active development. Despite many ingenious ideas behind them, the effort and initial optimism put into their creation gradually dissipated – perhaps these developments came too early, at the time when computers were not up to the mark. Nevertheless, it is undeniable that these early attempts formalized the problem of computer-assisted retrosynthesis, identified its most difficult aspects, and paved the way to the revival of the field in recent years. Foreshadowing my discussion of *Chematica*, it is important to highlight some of these modern developments.

Developed since the 2000s, ChemPlanner® by Wiley (previously known as ARChem Route Designer by SymBioSys [25]) is a commercially available web application for retrosynthetic planning based predominantly [26] on the large number of chemical rules machine-extracted from databases of published reactions (in ARChem it was Reaxys database, ChemPlanner uses ChemInform and has recently merged with SciFinder [27]). This is the same conceptual approach as in SYNCHEM2 in 1970s – of course, modern reaction repositories are far more voluminous than decades ago so the knowledge base of this software is much richer, around 100,000 transforms. Another component of the knowledge base are the catalogs of the commercially available chemicals (from various suppliers) that serve as stop points of the searches. On the other hand, the machine-extracted rules are not very accurate, as they do not come with detailed protection or conflict information (other than negative information on the lack of conflicting groups in published examples). Also, since the rules are extracted as "reaction cores," they do not necessarily capture stereochemistry. ChemPlanner® returns complete synthetic pathway but their length is limited to four steps [28].

In a similar genre, the commercially available IC_{SYNTH} derives its chemical knowledge from automatically extracted chemical rules [29], with only limited prowess in handling stereochemistry [30]. According to InfoChem's tutorial video [31], when initiating a search, the user is able to choose among different libraries of chemical rules (categorized by the source of origin), define the size of retrosynthetic tree, and pick a construction strategy. The program generates a multistep (up to 250 precursors at 1st level) complete retrosynthetic tree of results up to 10 generations. Unfortunately, the program does not produce specific pathways which are left for the user to manually pick and choose from the tree. An interesting aspect of the software is that is can also be applied in the forward direction to predict reactivity patterns of a given substrate molecule.

Finally, the most recent examples of retrosynthetic design based on machine-extracted rules come from the Waller group [32]. In the recently published article, these authors described the use of deep neural networks and the so-called Monte Carlo Tree Searches to construct synthetic plans leading to some medicinally relevant targets. What is impressive in this approach is the speed with which the machine constructs the routes. On the other hand, the lack of detailed contextual chemical information (protections, conflicts, admissible placement of unsaturations in ring systems, etc.) in the reaction rules results in chemical inconsistencies in the pathways. The authors also mention their approach cannot handle stereochemistry adequately and is not expected to work with complex natural products for

which simple strategies of just "cutting into smaller fragments" are doomed to fail. Still, despite a problematic handling of the underlying chemistry, the rapid search algorithms used in this program are definitely a notable advance.

Summarizing, for over more than half of a century, various creative and, without exception, valuable approaches to teaching computers synthetic planning were proposed and tested. Although the majority of these methods were probably premature and never came to fruition, recent rapid advances in computer hardware and algorithms substantiate hope that we might be finally able to attack this challenging problem. It was this hope that motivated me to start working on *Chematica* back in 2013. As a chemist, I was most concerned with the need to teach the machine proper and general-scope rules of organic-chemical reactivity. I reasoned that only with such correct input will the machine ever – even with the most advanced algorithms – be able to produce chemically sensible pathways. In subsequent sections, I will narrate in detail of how this vision was implemented and how it culminated in complete synthetic pathways autonomously designed by the computer and then, in an unprecedented demonstration, validated in the laboratory.

5. (Re)defining chemical rules

5.1. Data format

(*for detailed description, see reference [P5], Section 3.3 and reference [P1], Supplementary Information, Section S.3.1*)

The cornerstone of any synthesis-design software is the representation of the underlying chemical knowledge in a format that is not only machine-readable, but also general in scope, flexible to account for various structural variations, reactivity conflicts and protection requirements, and rapid in the execution of reaction transforms. We decided to employ SMILES/SMARTS notation [33] which represents chemical reactions or molecules as alphanumeric strings. The decision was motivated in large part by the fact that operations on strings are much faster than on matrices (e.g., as in mol files). Additionally, the notation allows to track stereocenters by @ or @@ symbols while configurations of double bonds are marked with // /\ signs. Unfortunately, a well-known limitation of SMILES/SMARTS and libraries such as RDKit is that the @ or @@ configuration encoded in a string is not absolute but only reflects the "local chirality". The symbols indicating configuration on a double bond are also "local" and can lose their proper chemical meaning when the substituents on the double bond change across the reaction. To overcome these problems, I participated in the development of in-house written modules called STEREOFIX and REGIOFIX that can handle reactions in which stereochemistry appears or changes. These modules use rules of substituent "preference" to correctly transmit the symbolic information (@, @@, //, /\) as well as ordered (by the masses of substituents) lists of bonds neighboring each atom mapped in the reaction between the retron and the synthons. In other words, these lists keep track of the masses of substituents changing upon bond breaking or making and overall order the neighboring bonds according to these changes. Although these operations increase the time of transform execution, they ensure that stereochemistry of a reaction is determined properly by the consensus of the bond list orders and by the stereochemical symbols present in the SMARTS notation.

5.2. The database of chemical knowledge

(for detailed description, see reference [P5], Sections: 3.2.1, 3.2.2, Supplementary Information, Sections S7, S8 and reference [P1], Supplementary Information, Section S2)

As we have seen in Part 4, several approaches to collecting the reaction/transform knowledge have been attempted before: from chemically accurate but laborious curation of transform libraries by expert synthetic chemists, through general descriptions based on physicochemical properties of atom and bonds, to massive machine extractions from databases of published reactions.

The last of these options is by far the least time consuming and tens of thousands of reactions can be readily processed within an hour. In fact, this approach seemed all the more tempting based on the statistics of reaction types we had initially collected. Figure 1 below plots how many times reactions of certain types were used in literature-reported reactions (rank #1 is the most popular reaction type/class, #2 is the second best, etc.). As seen, the dependence of popularity on rank is linear on a double logarithmic scale (i.e., linear with both the horizontal and vertical axes logarithmic) indicating the presence of the so-called power law seen to describe many types of natural phenomena in which even the infrequent occurrences in the distribution's tail matter. In our case, the power law signals that even some specialized and rarely used reactions are important and cannot be neglected – as we know, this is often the case in the synthesis of complex targets, as a particular "specialized reaction" might be the only method to synthesize a given class of compounds. What this observation means in the context of teaching the computer reaction rules is that it has to be taught tens of thousands of them – if we only teach the machine a limited number of "popular" transforms, it might tackle simple targets, but will fail in the vast majority of non-trivial cases.

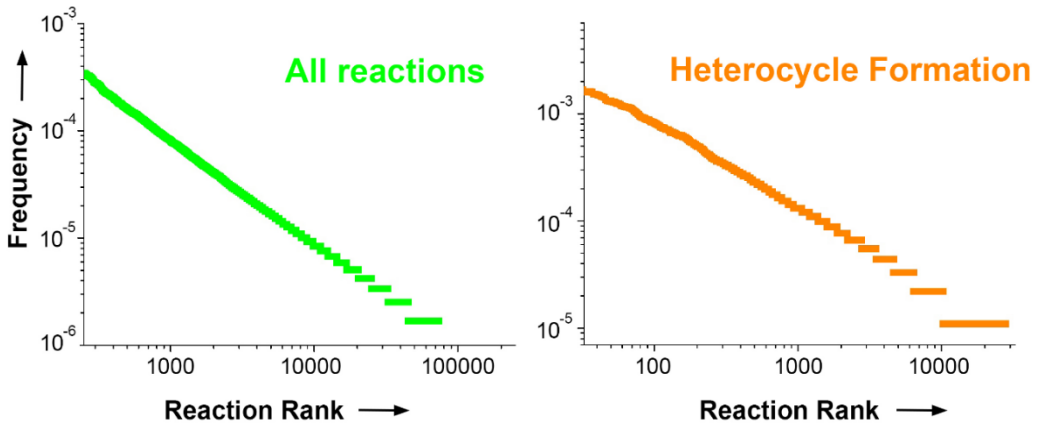


Figure 1. The frequency–rank plots of distinct reaction types. The left plot is based on the analysis of 1.2 million literature precedents. The right plot is for the formation of aromatic heterocycles. In both cases, the distributions are power laws (i.e., linear on a doubly-logarithmic scale) indicating the relative importance of reactions that are infrequent. Reaction rank = 1 indicates the most popular reaction, 2 is for the second-most popular, etc. Figure and caption reproduced from [P5].

In light of these statistics, we were initially not eager to code this myriad of rules manually and we hoped that the machine extracted rules would suffice. We extracted and categorized more than 100,000 classes of such rules but their performance in subsequent synthetic design was very poor. Having inspected a large number of results, I can attribute this failure to factors such as: (i) large number of erroneous entries in the collections of published reactions from which the rules were extracted; (ii) fundamental problems with rules that should account for distant electronic or steric effects (e.g., in Friedel-Crafts acylations, the substituent(s) dictating the reaction outcome might be far away from the reaction center); (iii) inability to properly define stereochemistry/regiochemistry within the reaction core (when to truncate the core?); (iv) lack of the information about "molecular context" in a given chemical transformation (potential conflicts or protection requirements can be deduced only indirectly by the lack of examples in literature). Although for popular reaction types with thousands of literature examples some of these (and other, see summary in Section S8 of the Supporting Information to ref. [P5]) problems might be alleviated, no statistical, machine learning approach can help with advanced and not-so-popular transforms. Given these considerations – and fully understanding the magnitude of the challenge that lied ahead – we decided to use the reaction rules coded by chemists.

5.3. Reaction decision trees – defining reaction's core and scope

(*for detailed description, see reference [P1], Supplementary Information Sections S3.2-3.4 and reference [P5], Section 3.3 and Supplementary Information Section S9*)

I began my own effort in this direction by formalizing procedures involved in reaction coding in the form that can be ultimately represented by decision trees such as the one in Figure 2. In brief, one has to first define the reaction's core as well as the scope of substituents and/or atom types. The core needs to be defined – based on extensive literature studies and considering the reaction mechanism and stereoelectronic effects, etc. – very precisely such that all relevant atoms that might influence reaction outcome are accounted for. At the same time, spurious additional atoms should not be included as they can unnecessarily limit the scope of the transform. Admissible extensions beyond published literature precedents, but conforming to mechanistic requirements for a given reaction type, should be allowed. Each core is coded in the SMARTS notation with atom numbering reflecting the mechanism and with all stereochemical and regiochemical information. Each rule also comes with the typical reaction conditions, which are crucial for defining any applicable protection chemistries. In Figure 2, this logical flow is applied to the coding of streodifferentiating condensation of aldehydes with esters and is represented by the aforementioned decision tree (for more examples, see publication [P1], Supplementary Information, Sections S3.2-S3.4). In the course of my doctoral studies, I personally coded ca. 15,000 of such rules, out of ~ 60,000 currently present in *Chematica*.

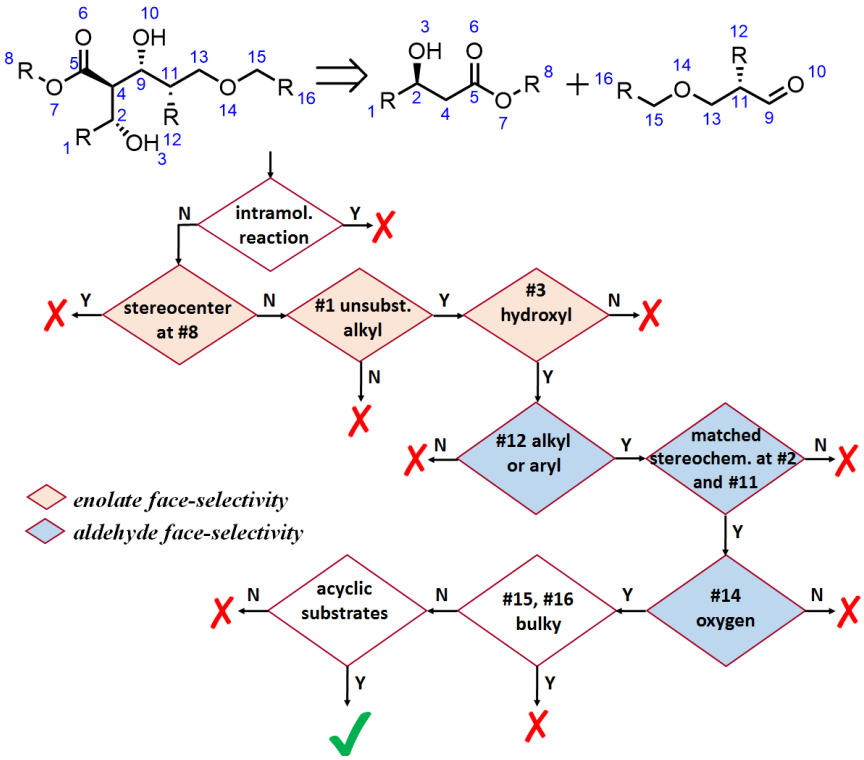


Figure 2. A decision tree capturing various conditions considered while teaching the machine streodifferentiating condensation of aldehydes with esters, one of *Chematica*'s ca. 60,000 reaction rules. The hierarchical sequence of conditions reflects various factors that need to be taken into account to produce chemically relevant outcomes when such a rule is later applied during synthetic planning. The first requirement prescribes intramolecular character of the reaction. To ensure face selectivity of the enolate, conditions for the substituents in positions #8, #1, and #3 are considered. Conditions in positions #12, #2, #11, #14 safeguard proper face selectivity of the aldehyde. The last two conditions are common for both the ester and the aldehyde. These substrates should be acyclic as the cyclic structures might distort the aldehyde-titanium chelate conformation or alter face selectivity of the ester enolate. The last requirement concerns the consonant selectivity at both substrates to yield desired stereoselecitivity. Typical conditions for this reaction class entail TMSCl and LDA for enolate formation from the ester, followed by $TiCl_4$ and aldehyde addition. Figure reproduced from [P1].

5.4. The importance of molecular context

(*for detailed description, see reference [P1], Supplementary Information Section S4 and reference [P5], Section 3.2.3)*

While the meticulously coded reaction rules capture the key effects at and near reaction center, they remain uninformed of the influence of more distant functional groups. In fact, capturing this "molecular context," as we dubbed it, is perhaps the key difficulty in synthetic planning and the reason why a "locally" defined reaction rule might work in one case but fail in another, where far-away functionalities present insurmountable reactivity conflicts or may need to be protected. One simple example is shown in Figure 3 below: The reaction in (a) will proceed without complications, but that in (b), sharing the same reaction core, is flawed because the distant aldehyde group will preferentially react with the organomagensium compound.

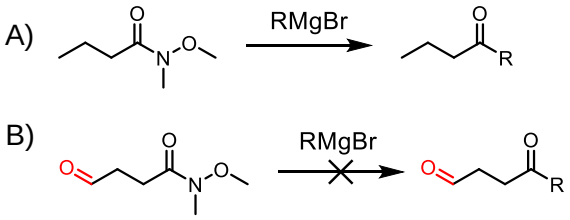


Figure 3. Importance of molecular context. In the first example, Grignard reaction will yield a ketone. However, in the second example, the "conflicting" aldehyde will preferentially undergo Grignard addition. Coding all such effects of remote substituents at the level of reaction cores is impossible. Figure reproduced from [P5].

To account for the effects of such "unwanted" functionalities outside of the reaction cores, I prepared collections – different for each reaction rule – of groups that (i) required protection under the reaction conditions specified as "typical" for this reaction class; such transforms could be executed conditionally provided that the groups in question were protected; and (ii) were always cross-reactive and could not be protected; such transformations were not applied at all during synthetic searches.

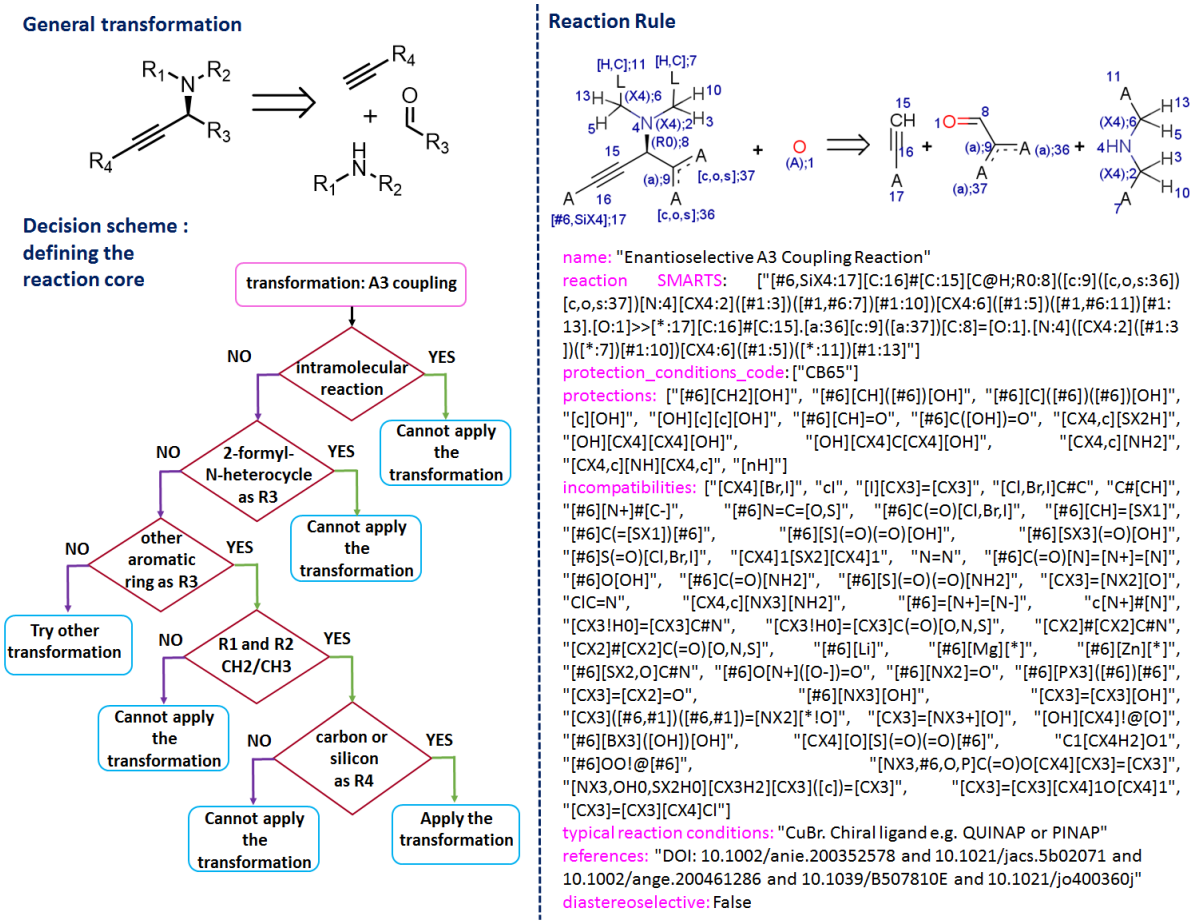


Figure 4. One of *Chematica*'s complete chemical rules. Left column has the decision tree defining the reaction core for enantioselective A3-coupling. Right column shows the complete reaction record as coded in the SMARTS notation and with all contextual information (cross-reactive groups, groups to be protected, class of typical reaction conditions) as well as some illustrative literature links. Figure reproduced from the Supplementary Information of the reference [P1].

With such additional information added to every reaction rule in *Chematica*'s knowledge base (see Figure 4), the algorithm for detecting protections and conflicts is summarized in the block diagram in Figure 5.

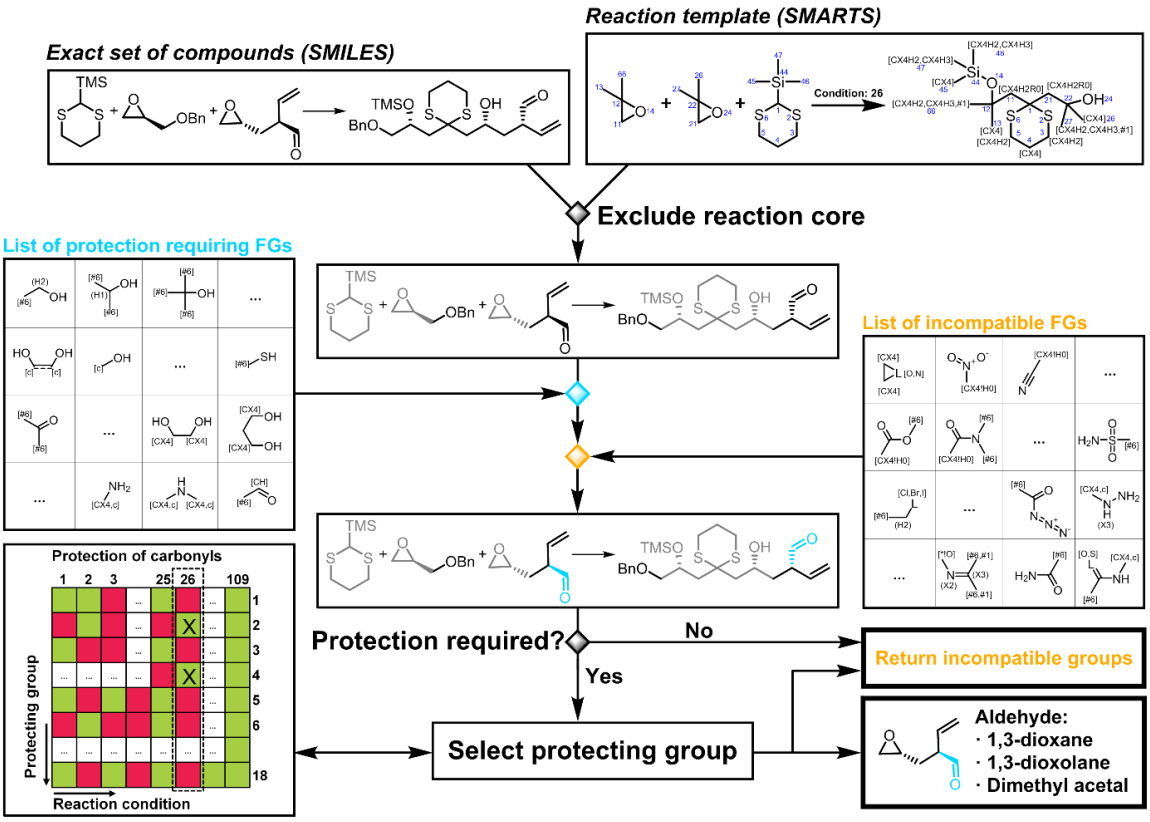


Figure 5. Block diagram illustrating detection of protections and/or conflicts. The process starts with exclusion of a reaction core from reaction's target and substrates. In this step, it is necessary to avoid false-positives whereby a group detected as interfering is itself part of the reaction template. After this step, the algorithm identifies whether the detected group presents an insurmountable incompatibility or requires protection. In case of incompatibility, the reaction is removed from the set of results. If a reaction entails group(s) requiring protection, the algorithm based on the specified conditions selects the most convenient protecting groups. Figure reproduced from the Supplementary Information of the reference [P1].

6. Navigating retrosynthetic trees manually

(*for detailed description, see reference [P5], Section 3.3.1 and Supplementary Information Section S14)*

The simplest functionality of *Chematica* we developed allowed for manual, step-by-step searches of the retrosynthetic trees – this was, in fact, similar to Corey's LHASA but served an important purpose as it allowed us not only to inspect the results but also develop metrics that would rank the proposed reactions according to various chemical criteria. To perform such ranking, we created a scripting language that evaluates synthetic steps according to a set of predefined variables defining certain features of reactions or molecules involved. My contribution was to define these variables. For example, variable MREL was defined as a mathematical function reaching maximum when the masses of synthons of a given reaction

were identical – this variable promoted "cut-in-half" disconnections and penalized peripheral ones. Variables STEREO and RINGS counted and favored reactions in which, respectively, new stereocenters or new rings were created. Variable PROTECT assigned penalties for every protection that had to be applied in a reaction, whereas variables BUY/KNOWN promoted reactions that used substrates that were, respectively, either commercially available (in *Chematica*, from Sigma-Aldrich catalogue) or previously described in literature (based on *Chematica*'s internal repository of ca. 7 million known molecules).

Importantly, *Chematica*'s user is able to combine these variables into any desired algebraic expressions and rank the reaction candidates accordingly. These expressions were the first "scoring functions" in *Chematica* and with their help I identified several interesting pathways including the synthesis of epicolactone (before it was published, since the target was given to us as a challenge from Prof. Dirk Trauner; for details see Section S14 of the Supplementary Information to ref. [P5]). Above all, the manual searches based on these scoring functions gave us a tool to query *Chematica* for the knowledge it still lacked. Personally, they also taught me some intuition of proper scoring as so they paved the road to the scoring functions used in fully automated searches (i.e., searches without any human intervention).

7. Moving towards automatic planning of synthetic routes

(*for detailed description, see reference [P5], Sections 3.4, 3.4.1 ,3.4.2 and reference [P1], Supplementary Information Section S6.4)*

Fully automated route design has been my ultimate goal and the toughest challenge. Because with a large knowledge base of reaction rules, there are also large numbers of options available at each step (in *Chematica*, currently, ca. 100 as we estimated in [P1] and [P5]), the trees of possible syntheses are extremely large – indeed, within *n* steps there are ca. 100^n options to explore. Examining all such options exhaustively is clearly beyond the power of any computer and one needs to search this space of synthetic solutions in an intelligent manner. To enable development of appropriate algorithms by our group's mathematicians, I undertook to define the chemically meaningful variables from which appropriate scoring functions could be constructed. Because the pathways are comprised of several (many) individual reaction steps, it was necessary to construct two such functions – one scoring the substrates/synthons created in each step (Chemicals Scoring Function, CSF), and one evaluating the reactions already performed to reach these substrates (Reaction Scoring Function, RSF).

It is worth observing that previous approaches focused mainly on identifying and evaluating the "key disconnections," paying relatively little attention to the overall synthetic feasibility of the substrates. I decided to define our "synthetic positions" as a sum CSF+RSF and evaluating not each substrate separately but the set of all substrates produced in a given reaction. In this way, we avoided situations in which the program would waste time on searching for further syntheses leading to the "easy" substrate while it would have no chance of synthesizing the other, "hard" substrate. Also, my metrics took into account both the structural complexity of the synthons as well as their commercial availability (if applicable) and/or popularity in previous, literature-reported syntheses (if substrates were previously made).

The specific variables I defined for CSF and RSF and summarized in Tables 1 and 2 below.

Table 1. List of variables available for the construction of the Chemicals Scoring Function, CSF.

Variable	Description
MASS	Mass of each substrate.
SMALLER$^{\gamma}$ or SMILES_LEN$^{\gamma}$	Sum of the lengths of the SMILES strings of all synthons, each raised to power γ. Accounts for the overall molecular complexity of the substrate sets as it includes in the lengths of the SMILES strings, e.g., information about stereocenters (represented by additional '@' or '@@'signs).
STEREO	Number of stereocenters in each substrate.
RINGS	Number of rings in each substrate.
KNOWN	+1 for a compound known in the literature, 0 otherwise.
BUY	+1 for a commercially available compound, 0 otherwise.

Table 2. List of variables available for the construction of the Reaction Scoring Function, RSF.

Variable	Description
PROTECT	+1 penalty for each functional group requiring protection.
CONFLICT	+1 penalty for each conflict detected.
NON_SELECTIVITY	+1 penalty for each non-selectivity found.
FILTERS	+1 penalty for each reaction sequence in which a fragile group is dragged along two or more synthetic steps.
YIELD	Estimation of reaction yield based on a thermodynamic model [P6].
HIDE_SEEK_ID	If used with a "+" sign, penalizes a given reaction or a set of reaction's id's; if used with "-", then promotes such id's.
HIDE_SEEK_NAME	If used with a "+" sign, penalizes a given keyword or a list of keywords; if used with "-", then promotes such keywords.
HIDE_SEEK_SMILES	If used with a "+" sign, penalizes a given SMILES or a list of SMILES's; if used with "-", then promotes such SMILES's.
HIDE_SEEK_SMARTS	If used with a "+" sign, penalizes a given SMARTS or a list of SMARTS's; if used with "-", then promotes such reaction templates.

With the variables thus defined, I proposed and implemented the typical forms of the scoring functions that are used in *Chematica* up to the present day:

for CSF:

CSF=SMALLER$^{\gamma}$ +α*RINGS+β*STEREO

Since CSF evaluates complexity of the substrate sets it should favor the simplest possible substrates at each step (i.e, in the "forward" direction, it should favor reactions generating the highest molecular complexity). In the above algebraic expression the SMALLER$^{\gamma}$ variable (in earlier versions of *Chematica*, named SMILES_LEN, see Table 1), is an indirect measure of molecular complexity. Parameter γ is typically between 1.5 to 2, and its value determines preference for more peripheral disconnections (γ = 1.5) or those leading to like-size fragments (γ = 2). This preference reflects the definition of the variable – that is, SMALLER$^{\gamma}$ is the sum over all substrates of the lengths of their SMILES strings, each raised to power γ. This function always has a minimum for equal-sized disconnections, but its slope increases with increasing γ. In this way, for a given non-equal disconnection, the value of SMALLER$^{\gamma}$ increases with increasing γ and such unequal disconnections are more heavily penalized for larger γ exponents (see Figure 6 for illustration). The other two parameters in this CSF, α and β (typically between 50-100) specify the "weights" with which each newly created stereogenic center or a ring are promoted.

for RSF:

RSF = C + β*PROTECT+ α*(NON_SELECTIVITY+FILTERS+CONFLICT)

RSF is intended to favor the shortest possible synthesis with the proviso that individual steps do not suffer from nonselectivities or conflicts. The value of C is specified by the user to denote a "constant" cost of each reaction performed and is usually between 20 and 120. Parameter β specifies an additional cost for each protection reaction that is required and is usually equal to twice the cost of a single step, $\beta \approx 2*C$. This choice is motivated by the fact that a protection requirement adds two extra steps to the synthesis: protection and deprotection. Parameter α denotes a high penalty (usually more than 5,000) for reactions in which nonselectivities, cross-reactivity conflicts, and other serious problems are detected. By being so high, this part of RSF effectively eliminates problematic reactions from consideration. Finally, not shown in the generic RSF above, the user can utilize additional variables such as HIDE_SEEK to eliminate specific molecules, reaction names, keywords specifying reaction conditions, etc. from the searches ("HIDE") or, conversely, channel the reaction to seek such solutions ("SEEK").

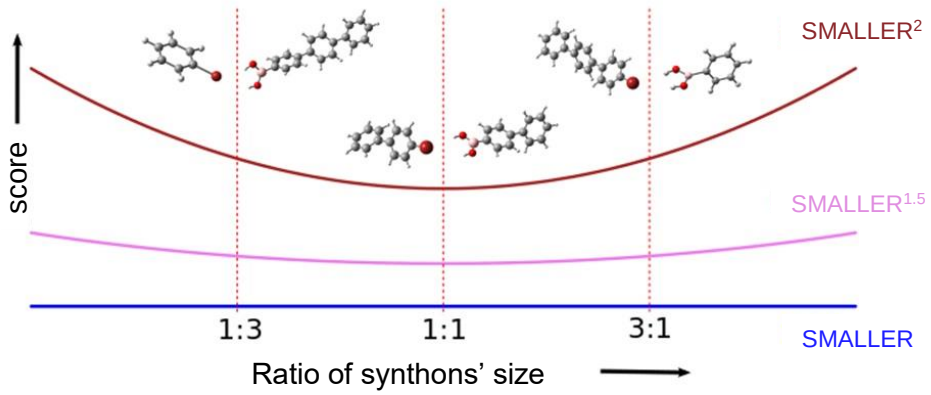


Figure 6. Graphical illustration of the SMALLER$^{\gamma}$ operation approximating the molecular complexity of the reaction's synthons and the centrality of the synthetic disconnection. The SMALLER variable itself is defined as the sum (over the substrate set) of the length of SMILES strings of substrates, each taken to some power γ which effectively specifies preference for the relative sizes of the substrates. Say γ=1 and CSF=$\Sigma_{substrates}$ SMALLER$^{\gamma}$. Disconnecting the target

(here, p-quaterphenyl, SMALLER=38) into halves (19+19) will result the same CSF value as disconnecting into unequal parts (e.g., 8+30 or 30+8). However, if $\gamma>1$, the function summing $SMALLER^{\gamma}$ over the substrates will favor equal-sized cuts – indeed, it can easily be proven that if $SMALLER_{sub1}+SMALLER_{sub2}=SMALLER_{target}$ then function $SMALLER_{sub1}{}^{\gamma}+SMALLER_{sub2}{}^{\gamma}$ is minimized if $SMALLER_{sub1}=SMALLER_{sub2}$. Because best synthetic positions minimize CSF, such equal-size disconnections will be preferred during synthetic searches. Also please note that higher values of γ yields higher CSF values and thus penalize unequal cuts more strongly than lower γ exponents (here $\gamma=2$ versus $\gamma=1.75$ curves). Figure and caption adapted from [P5].

7.1. Two approaches allowing for yield estimation

(*for detailed description, see references [P6] and [P4])*

As shown in Table 2, the Reaction Scoring Function can also use as a variable theoretical estimates of a reaction's yield. *A priori* estimation of yield of arbitrary reactions is a highly nontrivial problem. In my doctoral studies, I contributed to two efforts – one based on thermodynamic modelling and one based on Machine Learning – to attack this problem.

7.1.1. Thermodynamic model for *a priori* yield estimation

(*for detailed description, see reference [P6])*

As described in detail in the introduction to our paper [P6], the main premise of the model is the observation that most (but certainly not all!) organic reactions proceed under thermodynamic control. This assumption, supported by the statistical analysis I performed (see Section 1 of the Supporting Information to ref [P6]), relates yield to reaction free energies ΔG. In order to calculate this parameter, substrates and products molecules were divided into smaller fragments with pre-calculated Gibbs free energies of formation G_i^{form}. Summation of each group's contribution with appropriate stoichiometric coefficient ν_i gives ΔG_{calc} for the reactions. The non-idealities of the system like solvent effects or temperature were incorporated by using activity coefficients calculated at the molecular level by perturbed-chain statistical associating fluid theory (PC-SAFT). The model was iteratively optimized and trained on the total 23,000 diverse reactions with full yield and stoichiometry. The accuracy of the estimation was ±15% for reactions not included in the test set. The model proved capable of capturing yield differences related to solvent changes. While the model was certainly not ideal in term of its predictivity, it proved a valuable addition to *Chematica* as the means of rough categorization into good, average, and poor-yielding reaction. I should stress that this work was a highly collaborative effort between chemists (from our laboratory in Warsaw) and theorists (from Northwestern University in the U.S.) – while I benefited intellectually from such an interaction, I do not, by any means, take credit for the development of the theoretical model. As mentioned above, I contributed to the chemical validation of the model, and its relevance to non-trivial organic reactions.

7.1.2. Machine-Learning based approach for yield estimation

(*for detailed description, see reference [P4])*

Encouraged by the numerous examples of successful application of Machine Learning, ML, methods to various scientific problems and in collaboration with our fellow mathematicians from the University of Warsaw (group of Prof. Anna Gambin), we strived to use ML to improve the accuracy of yield prediction offered by the thermodynamic model. Unfortunately, this time, the methodology turned out to be less successful with the accuracy of binary ("high"/ "poor") yield prediction only c.a. 65±5% (i.e., error ~35%). This result did not depend on the type of ML method applied (e.g., neural networks vs. random forest classifiers), the number of molecules in the training set, or the nature and the number of descriptors used to train the model. Additionally it was proven by the so-called Bayes classifier error estimates that the obtained outcome cannot be considerably improved (max. 80% of accuracy which is less than in case of thermodynamic model) for currently available chemical descriptors. Still, this work, published in 2017 in *Scientific Reports* [P4], generated considerable interest as it has emphasized that in order for the ML methods to become chemically accurate, the underlying ways of representing molecules (i.e., descriptors) need to be dramatically improved to account, for instance, for stereoelectronic properties, three-dimensional conformations of molecules, reagents, etc. An effort to find such improved representations is still ongoing not only in our laboratory in Warsaw but also in several other laboratories worldwide. As in the case of the thermodynamic model, my role in the project was (i) to ensure that the input descriptors are chemically correct and (ii) to inspect which types of reactions offer better predictivity than others (though, ultimately, no such clear cut correlation was found). Personally, I valued the work on this project as it introduced me to ML which is one of the areas I would like to study in more depth after my graduation.

8. Overcoming local complexity maxima and the need for higher order logic

(*for detailed description, see references: [P3], [P1], Supplementary Information, Section S7 and [P5] Section 4.1)*

Although various CSF and RSF variables discussed so far help the machine to make synthetically reasonable choices at each synthetic steps, they are not guiding it to make more far reaching "strategic decisions" – that is, how to combine individual reaction steps into an "elegant" synthesis. Although "elegance" is not, by any means, a scientific criterion and its measure is highly subjective, the term is often used to describe, for instance, convergent sequences involving counter-intuitive *sequences* of reactions. To teach the machine to strategize over several steps, it has to be told which sequences of steps are not promising (e.g., as they drag along fragile functional groups), in which "local" complexification of the structure would be beneficial as it could simplify further synthesis, etc. These considerations are listed below. Unlike in the yield prediction where my role was largely auxiliary to theorists, I was playing a leading role in defining and implementing these solutions in *Chematica*.

8.1. Fragile functional groups

(for detailed description, see reference [P1], Supplementary Information, Section S7.1)

One of the fundamental premises of an "elegant" synthesis is that highly reactive groups should not be dragged along the synthetic pathway. Based on extensive analyses of classic syntheses (and other, less prominent ones), supported by the statistics of transformation combinations reported in the literature. I identified and coded over 100 classes of "fragile" structural motifs which, if they appear in an intermediate, should be used immediately, in the subsequent step. This "do-not-drag-along" heuristic has been included in *Chematica*'s RSFs under the FILTERS variable.

8.2. Strategies

(for detailed description, see references: [P1], Supplementary Information, Section S7.3 and [P5] Section 4.1)

Another situation mentioned in the introduction to this Section is when it is beneficial to "complexify" the synthons but, by doing so, open up new synthetic possibilities that result in an overall more efficient (and elegant!) synthesis. In other words, a seemingly "futile" step can sometimes set the scene for a subsequent reaction that offers a drastic structural simplification. Such sequences of chemical transformations were described by E.J. Corey as "tactical combinations" and were introduced during his development of the LHASA software (to which, unfortunately, we did not have access). In *Chematica*, I have defined over fifty combinations of general reaction classes that serve this purpose and subsequently extended them into several thousand of combinations of more specific reactions. Some examples are shown in Figure 7 which is reproduced from ref [P1].

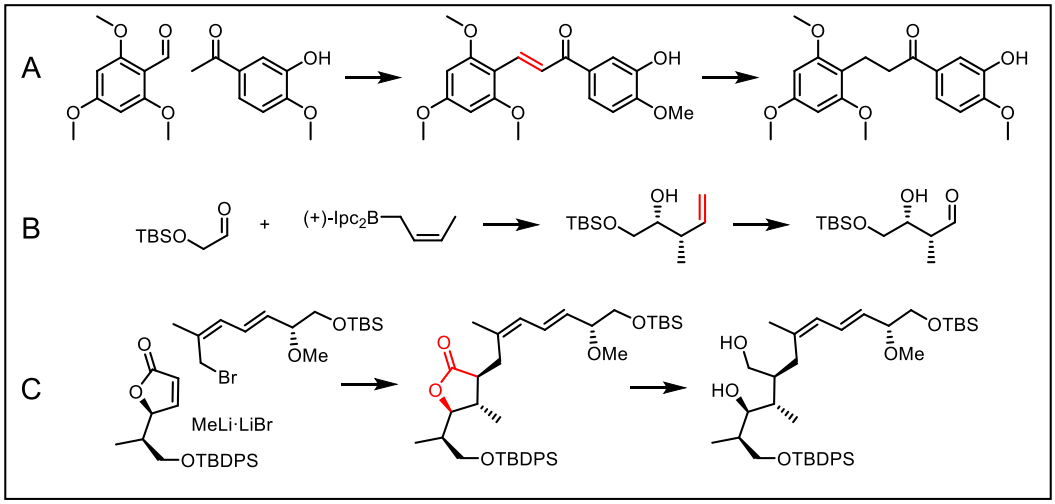


Figure 7. Examples of syntheses comprising two-step strategies. (a) Short and efficient synthesis of taccabulin AS57 relies on a condensation of benzaldehyde and acetophenone followed by hydrogenation of the double bond. In the retrosynthetic direction, introduction of the double bond might not immediately serve beneficial as it does not simplify the structure. (b) Synthesis of brevisamide: Brown crotylation is followed by oxidation of terminal alkene to aldehyde. Again, in the retrosynthetic direction, changing an aldehyde into an alkene might not be immediately seen as advantageous. (c) Halichomycin intermediate is obtained from the corresponding lactone. In the

retrosynthetic direction, formation of the ring might be counterintuitive (as it apparently complexifies the structure) – on the other hand, it introduces the electron-withdrawing group which then enables "division" of this intermediate into three synthons while installing two vicinal stereocenters. Figure and caption reproduced from [P1].

8.3. Cyclizations

(for detailed description, see reference [P1], Supplementary Information, Section S7.2)

Another multi-step consideration concerns syntheses whereby preparation of a synthon would be significantly more challenging that the retron. A case in point is a preparation of systems of smaller rings via contraction of a larger heterocycle (e.g., a 6,6 system from a 10-membered ring). As any seasoned chemist would point out, formation of medium-sized or larger rings is, in most cases, a low-yielding, slow process. Preparation of a macrocycle only to "destroy" it in subsequent moves is usually not the best synthetic approach. To avoid such syntheses, I introduced into *Chematica* a heuristics eliminating sequences in which rings larger than 8-membred are contracted in the retrosynthetic direction. This heuristics is optional, meaning that *Chematica*'s user can chose to apply it or not before starting a synthetic search.

9. Software validation

(*for detailed description of the paper validation see reference [P5], Section 3.4.4 and Supplementary Information, Section S15 and for the experimental validation, see reference [P1])*

No matter how interesting the ideas behind any synthesis-planning software, its only meaningful and ultimate test is whether it can produce synthetic plans that can be, without substantial changes, executed in the laboratory, hopefully offering some yield and/or cost improvements over previous routes (if known). In ref. [P5] I performed the first and very rudimentary validation by showing that the pathways designed autonomously by *Chematica* replicate those published in literature (of course, the machine was not "shown" these literature examples prior to this exercise). Still, real wet lab validation remained elusive, in part because our own laboratory was only being set up. It was only in 2017 that I was able to put *Chematica* validation by using it to design eight pathways – each leading to a medicinally relevant and commercially valuable target – that were subsequently executed in the laboratories of Sigma-Aldrich, our laboratory at ICHO, and at Northwestern University. Remarkably, all of these syntheses worked in practice, two enabling synthesis of targets that were not made before, and the remaining six offering – according to Sigma-Aldrich's metrics – substantial savings (in terms of synthetic cost, yield, and/or execution time) compared to previous approaches. These exciting results were described in our recent publication in *Chem* [P1]. The detailed synthetic plans with experimental yields and miniatures of the raw computer-generated synthetic plans are reproduced in Figures 8 and 9 below.

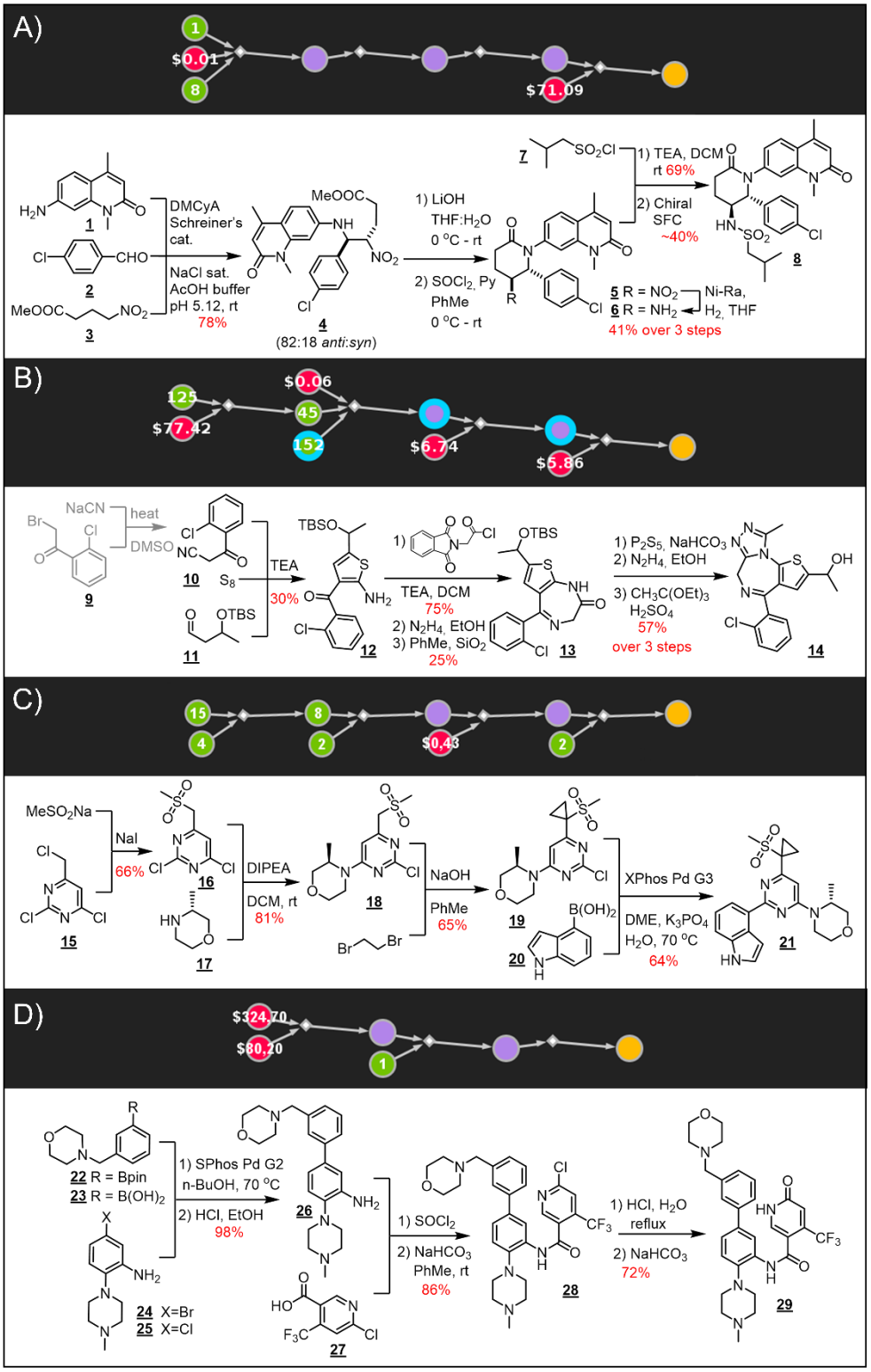


Figure 8. Syntheses planned by *Chematica* and executed in the laboratory (experimental yields are denoted by red numbers) for: A) BRD 7/9 inhibitor, B) α-hydroxyetizolam C) ATR kinase inhibitor, D) inhibitor of human acutemyeloid- leukemia cells. Synthetic graphs produced by *Chematica* are shown above each synthetic plan. Color coding of nodes: red = commercially available chemicals (prices in US$/g from Sigma-Aldrich catalog); green = molecules known in the literature; violet = unknown molecules, yellow = targets; blue halos = protection required. Figure and caption reproduced from [P1].

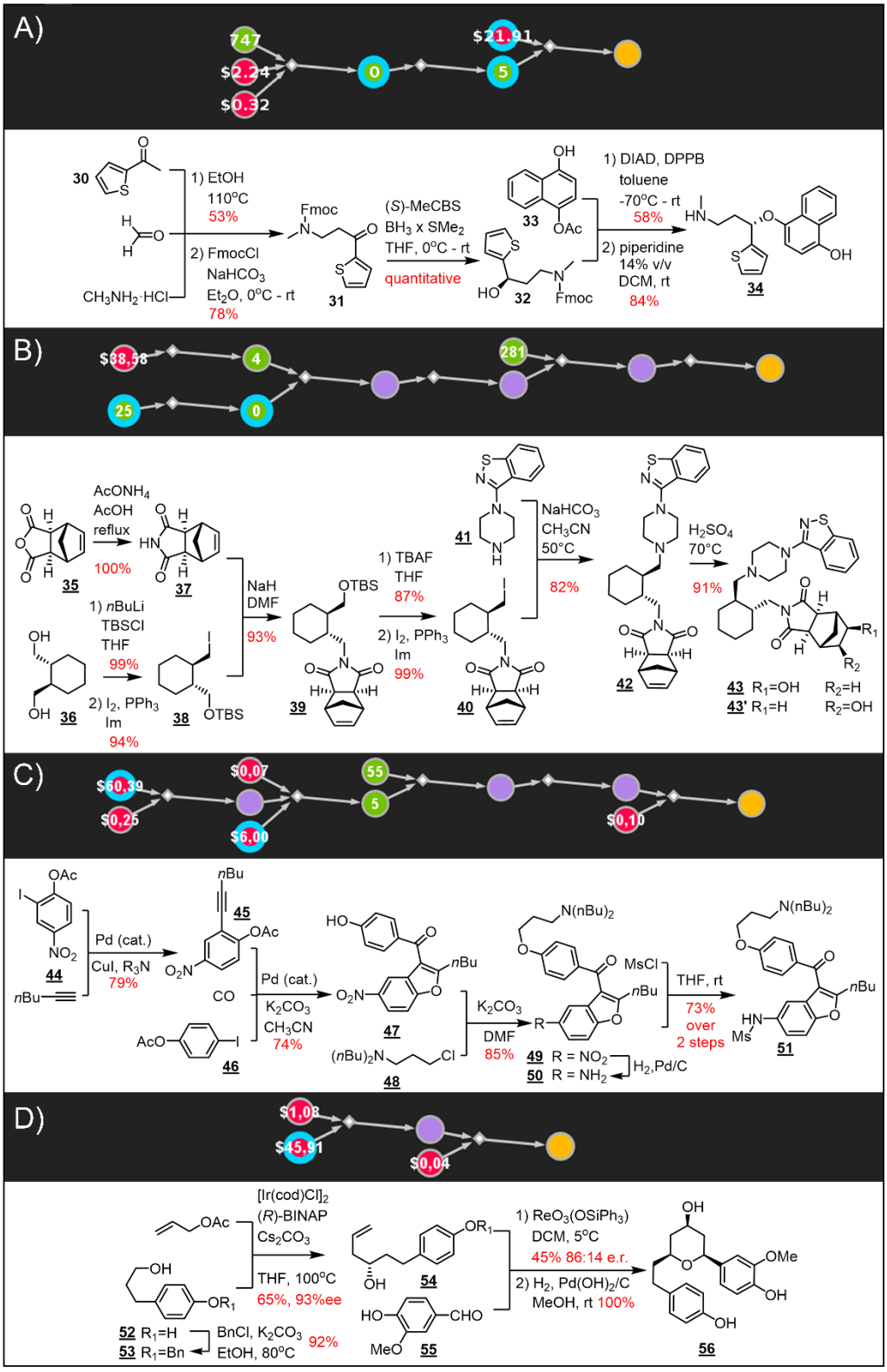


Figure 9. Syntheses planned by *Chematica* and executed in the laboratory (experimental yields are denoted by red numbers for: A) (*S*)-4-hydroxyduloxetine, B) 5β/6β-hydroxylurasidone, C) dronedarone, D) engelheptanoxide C. Synthetic graphs produced by *Chematica* are shown above each synthetic plan. Color coding of nodes is the same as in Figure 7. Figure and caption reproduced from [P1].

10. Looking forward: possibilities and challenges

(*for detailed description, see reference [P2])*

With *Chematica* now mature and under commercial development (under the auspices of the MilliporeSigma/Merck conglomerate), my interest in computer aided organic synthesis continues to broaden beyond the design of individual pathways. Indeed, one of the grand challenges I envision is the design of highly networked chemical systems that, akin to the networks of biochemical reactions in cells, could be performing different synthetic tasks depending on "inputs" (substrates) they receive. In this spirit, the last project I partook during my doctoral studies was the identification of reaction cycles – arguably, the simplest chemical systems – in the vast network of reactions published in the literature (so-called Network of Organic Chemistry constructed earlier by the Grzybowski group [34-39]). With the help of the search software written by my computer-science colleagues and using the search criteria I helped design, we were able to identify millions of cycles that chemists working at different places and at different times constructed – actually, without realizing it! The analysis I and two other fellow students performed on some of the cycle candidates revealed that among them were faithful replicas of some biochemical cycles, those that can be performed one-pot, and those that autoamplify useful chemicals. These and millions of cycles we identified are stored and can be queried in a repository called *Cyclorg* (publicly available at http://cyclorg.grzybowskigroup.pl/). Theoretical and chemical details of the work are described in detail in ref [P2].

11. Summary

My doctoral studies have been a demanding but fantastically exciting journey into the new world of computers interfacing with the needs of synthetic chemists. I came into the Ph.D. program as a beginning and scientifically naïve chemist but I leave it with a satisfaction to have made a significant contribution to an important and longstanding challenge of synthetic organic chemistry. I believe I can take credit for designing the structure of *Chematica*'s knowledge base and translating into the machine-readable format some 15,000 chemical rules. I formalized various "synthetic variables" that guide *Chematica*'s step-by-step as well as fully automated searches. I also came up with the general form of the CSF and RSF scoring system for "synthetic positions" and made contributions to various aspects of automated planning beyond individual steps (strategic sequences, filtering off sequences involving too reactive or synthetically unfeasible intermediates, etc.). I am very proud that the *Chematica* platform I helped developed performed so well in the experimental validation of its theoretical results. In fact, the results of eight complete syntheses described in the *Chem* paper [P1] are, to the best of my knowledge, the first-ever demonstration of a computer designing pathways standing the test of wet-lab validation and offering tangible improvements over previous approaches (including those of Sigma-Aldrich experts). As narrated in the last Section, I am now venturing into the realm of chemical systems and look forward to further challenges in this emerging area of synthetic research. I feel my doctoral work equipped me with the requisite knowledge, the ambition, and the intellectual curiosity to attack such challenging problems in the years to come.

12. References

[1] "Concerning one system of classification and codification of organic reactions" G.É. Vléduts, *Inf. Storage Retr.* **1963**, 1, 117.

[2] "General methods for the construction of complex molecules" E. J. Corey, *Pure Appl.Chem.* **1967**, 14, 19.

[3] "Computer-Assisted Design of Complex Organic Syntheses" E. J. Corey, W.T. Wipke, *Science.* **1969**, 166, 178.

[4] "Knowledge-based Expert Systems in Chemistry: Not Counting on Computers", P. Judson, RSC, Cambridge, **2009**

[5] http://cheminf.cmbi.ru.nl/cheminf/lhasa/ (accessed 15.03.2018)

[6] http://cheminf.cmbi.ru.nl/cheminf/lhasa/doc/lhasa191.pdf (accessed 15.03.2018)

[7] "Computer-assisted synthetic analysis. Performance of tactical combinations of transforms" A. K. Long, J. C. Kappos, *J. Chem. Inf. Comput. Sci.*, **1994**, 34, 915.

[8] "Computer-Assisted Analysis in Organic Synthesis" E. J. Corey, A. K. Long, S. D.Rubenstein, *Science*, **1985**, 228, 408.

[9] "Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques" W.T. Wipke, G. I. Ouchi, S. Krishnan *Artif. Intell.*, **1978**, 11, 173-193.

[10] "Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry" W. T. Wipke, T. M. Dyott, *J. Am. Chem. Soc.*, **1974**, 96, 4825.

[11] "Stereochemically Unique Naming Algorithm" W. T. Wipke, T. M. Dyott, *J. Am. Chem. Soc.*, **1974**, 96, 4834.

[12] "Computer-aided organic synthesis" M. H. Todd, *Chem. Soc. Rev.*, **2005**, 34, 247.

[13] "Empirical Explorations of SYNCHEM" 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.

[14] "Application of chemical transforms in synchem2, a computer program for organic synthesis route discovery" K. K. Agarwal, T. D. L. Larsen, H. L.Gelernter, *Comput. Chem.*, **1978**, 2,75-84.

[15] "Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning" H. Gelernter, J. R. Rose, C. Chen, *J. Chem. Inf. Comput. Sci.*, **1990**, 30, 492.

[16] "Distributed Heuristic Synthesis Search" D.Krebsbach, H. Gelernter, S. McN. Sieburth, *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 595.

[17] "Designing an Expert System for Organic Synthesis in Expert Systems Application in Chemistry" (Eds.: B. A. Holme, H. Pierce) P. Y. Johnson, I. Bernstein, J. Crary, M. Evans, T. Wang, ACS Symposium Series, Am. Chem. Soc. Washington, **1989**.

[18] “Approaching the logic of synthesis design” J. B. Hendrickson, *Acc. Chem. Res.*, **1986**, 19, 274.

[19] “The SYNGEN approach to synthesis design” J. B. Hendrickson, *Anal. Chim. Acta*, **1990**, 235, 103.

[20] “Computer-Assisted Solution of Chemical Problems-The Historical Development and the Present State of the Art of a New Discipline of Chemistry” I. Ugi, J. Bauer, K. Bley, A. Dengler, A. Dietz, E. Fontain, B. Gruber, R. Herges, M. Knauer, K. Reitsam, N. Stein, *Angew. Chem. Int. Ed. Engl.*, **1993**, 32, 201.

[21] “Computer-assisted bilateral solution of chemical problems and generation of reaction networks” J.Bauer, E.Fontain, I.Ugi, *Anal. Chim. Acta*, **1988**, 210, 123.

[22] “Chemoinformatics- A Textbook” J. Gasteiger, T. Engel (Eds.), **2003**, Wiley-VCH, Weinheim.

[23] “Computer-assisted synthesis and reaction planning in combinatorial chemistry” J.Gasteiger, M. Pförtner. M. Sitzmann, R. Höllering, O. Sacher, T. Kostka, N. Karg, *Persp.Drug Disov. Design*, **2000**, 20, 245.

[24] “The Pschobiological Basis of Heuristic Synthesis Planning – Man, Machine and the Chiron Approach” S. Hanessian, J. Franco, B. Larouche, *Pure Appl. Chem.*, **1990**, 62,1887.

[25] https://www.wiley.com/WileyCDA/PressRelease/pressReleaseId-110972.html (accessed 15.03.2018)

[26] “Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation” J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, H. Y. Ando, *J. Chem. Inf. Model.*, **2009**, 49, 593.

[27] https://www.cas.org/products/scifinder-n/chemplanner (accessed 28.03.2018)

[28] https://www.youtube.com/watch?v=_rdx2evzs2U (accessed 28.03.2018)

[29] http://www.infochem.de/products/software/icsynth.shtml (accessed 28.03.2018)

[30] “Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction” A. Bøgevig, H. J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Loew, C. Oppawsky, T. Rein, H. Saller, *Org. Process Res. Dev.*, **2015**, 19, 357.

[31] https://www.youtube.com/watch?v=kRR74sAkMmI (accessed 28.03.2018)

[32] “Planning chemical syntheses with deep neural networks and symbolic AI” M. H. S. Segler, M. Preuss, M.P. Waller, *Nature*, **2018**, 555, 604.

[33] http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed 28.03.2018)

[34] “Architecture and Evolution of Organic Chemistry” M. Fialkowski, K. J. M. Bishop,V.A. Chubukov, C. J. Campbell, B. A. Grzybowski, *Angew. Chem. Int. Ed.*, **2005**, 44, 7263.

[35] “The Core and Most Useful Molecules in Organic Chemistry” K. J. M. Bishop, R. Klajn, B. A. Grzybowski, *Angew. Chem. Int. Ed.*, **2006**, 45, 5348.

[36] “The 'wired' universe of organic chemistry” B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk, C. E. Wilmer, *Nat. Chem.*, **2009**, 1, 31.

[37] “Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry” M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski, K. J. M. Bishop, *Angew. Chem. Int. Ed.*, **2012**, 51, 7928.

[38] “Rewiring Chemistry: Algorithmic Discovery and Experimental Validation of One-Pot Reactions in the Network of Organic Chemistry” C. M. Gothard, S. Soh, N. A. Gothard, B. Kowalczyk, Y. H. Wei, B. Baytekin, B. A. Grzybowski, *Angew. Chem. Int. Ed.*, **2012**, 51, 7922.

[39] “Chemical Network Algorithms for the Risk Assessment and Management of Chemical Threats” P. E. Fuller, C. M. Gothard, N. A. Gothard, A. Weckiewicz, B. A. Grzybowski, *Angew. Chem. Int. Ed.*, **2012**, 51, 7933.

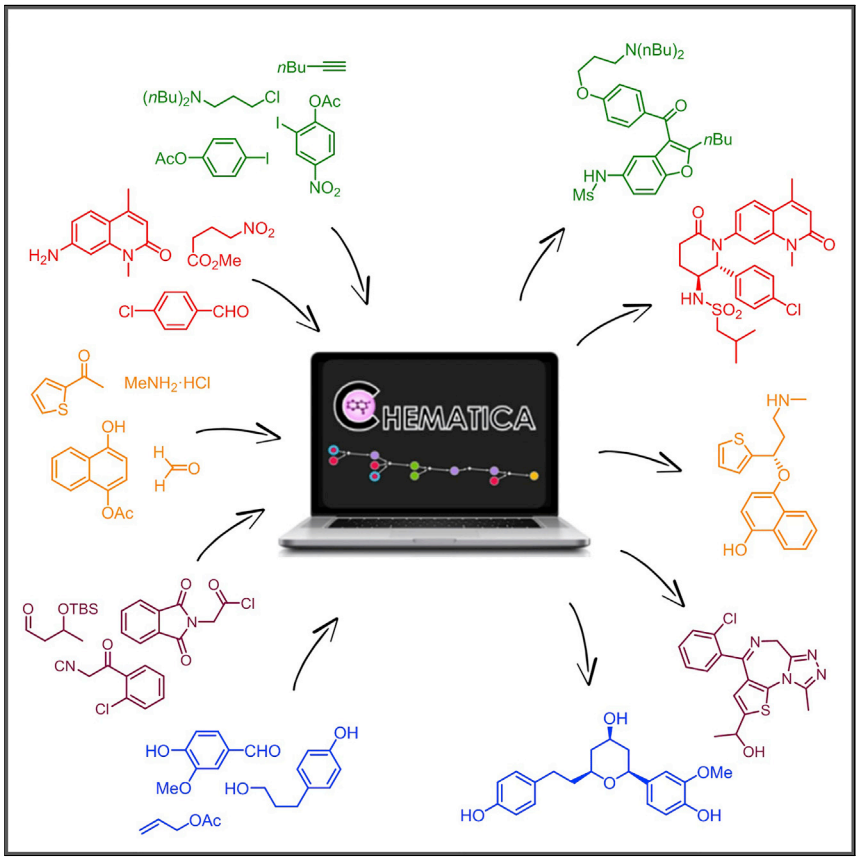
13. Reprints of publications included in the doctoral thesis

Chem

CellPress

Article

Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory



Multistep synthetic routes to eight structurally diverse and medicinally relevant targets were planned autonomously by the Chematica computer program, which combines expert chemical knowledge with network-search and artificial-intelligence algorithms. All of the proposed syntheses were successfully executed in the laboratory and offer substantial yield improvements and cost savings over previous approaches or provide the first documented route to a given target. These results provide the long-awaited validation of a computer program in practically relevant synthetic design.

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, ..., Milan Mrksich, Sarah L.J. Trice, Bartosz A. Grzybowski

milan.mrksich@northwestern.edu (M.M.)
sarah.trice@sial.com (S.L.J.T.)
nanogrzybowski@gmail.com (B.A.G.)

HIGHLIGHTS

Computer autonomously designs chemical syntheses of medicinally relevant molecules

The syntheses are successfully executed in the laboratory

The machine-designed routes improve on previous approaches

Klucznik et al., Chem 4, 522–532
March 8, 2018 © 2018 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.chempr.2018.02.002

Chem, Volume *4*

Supplemental Information

Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L.J. Trice, and Bartosz A. Grzybowski

CONTENTS:

Section S1. Overview of Chematica's key components and algorithms.

Various aspects of retrosynthetic planning in *Chematica* were described in our recent review[10] aimed at a general chemistry audience. Here, we recapitulate the main points of this discussion while placing more emphasis on some key conceptual and algorithmic issues.

We begin with the discussion of reaction rules – that is, how to teach the machine the myriad of different types of chemistries, from simple S_N2 to advanced stereoselective transformations. After some general considerations of the scope of the transforms and the crucial importance of "molecular context" extending beyond the very reaction cores (**Section S2**), we discuss (in **Section S3**) some specific examples of such rules chosen to illustrate the overall logic of translating chemical knowledge into a machine readable format. Next, in **Section S4** we discuss how some of the most important parts of the "contextual information" (notably protection chemistries and reactivity conflicts) are handled. In **Section S5**, we extend our discussion to cases where the coded transforms need to be augmented with quantum mechanical or molecular dynamics methods. Having defined the reaction rules we are then ready to discuss how these basic "synthetic moves" can be used to construct entire "games" – that is, synthetic pathways. In **Section S6**, we formalize the concepts of synthetic graphs and hypergraphs and describe the key aspects of algorithms that navigate them to find synthetically efficient – and hopefully, elegant and also diverse– synthetic routes. In **Section S7** we discuss how the one-step-at-a-time search algorithm can be further improved by introducing higher-order chemical logic defining sequences of steps that need to be promoted or avoided. These type of multistep strategies allow the searches to, for example, overcome local "hurdles" (akin to Monte Carlo algorithms overcoming local minima) and then venture into elegant branches of the space of synthetic solutions. Finally, in **Section S8**, we illustrate how the results of all of the above operations/calculations as well as some additional information are presented to *Chematica*'s user.

Section S2. Reaction rules: General considerations.

S2.1. Importance of "non-local effects". As in chess, the computer must first be taught the basic reaction rules ("moves") from which it can then construct complete synthetic pathways ("games"). To the first approximation, reaction rules specify which bonds change in a given reaction (e.g., in S_N2, Wittig, etc.). In principle, such "reaction cores" (potentially extended to include "flanking" atoms or groups, see **Figure S1**) could be extracted automatically from the millions of already published reaction precedents– and this is how we (and many others) have initially approached the problem many years ago, only to discover this approach is conceptually flawed (note: some 120,000 such machine extracted rules are still available in *Chematica* as one of the options – however, they are only of "historical" interest as their use in synthesis planning leads to utterly unreliable results). This is so because reaction cores themselves do not take into account the *context* of the entire molecule. By the context, we mean here effects influencing reactivity that originate from parts of the molecule far away from the reaction core. One example of such distant influence is illustrated in the already mentioned **Figure S1**. In another related example, the same reaction rule/core for SAMP-directed stereoselective alkylation can be applied to the substrates in **Figure S2a** (this reaction is confirmed to work in the laboratory) but will fail for substrates in **Figure S2b** (this reaction cannot be executed experimentally). The culprit for this failure is a distant nitroalkyl group (colored red) on the alkyl iodide which is incompatible with the lithiated azaenolate– these two groups will react suppressing the desired alkylation. Another untoward situation is illustrated in **Figure S2c** where the reaction would actually give a racemic mixture instead of a predicted stereopure product because the stereodirecting CH_2OMe group (not accounted for in the reaction core) is missing on the hydrazone.

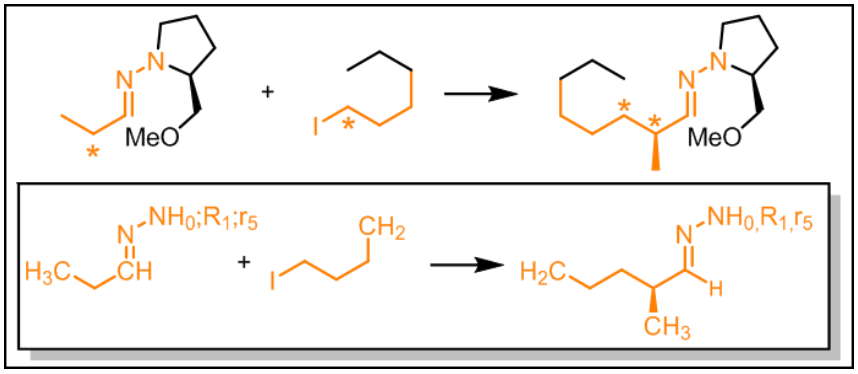


Figure S1. One of many problems with automatic extraction of reaction rules. An example of a literature-reported transformation[S1] from which the "reaction core" is extracted. The core is colored in orange, specifies some key atom types in SMARTS notation (e.g., $NH_0;R_1;r_5$ = no hydrogens on N belonging to one five membered ring), covers atoms changing their local environments (denoted with stars), and also includes flanking atoms up to three bonds away. Even with this extended neighborhood, the transform does not capture the influence of a distant (in terms of bonds but not in terms of 3D structure) stereodirecting group, CH_2OMe. Of course, one could extend the neighborhood for this particular example to 5-6 bonds, but this would make all simple transforms (in which extended neighborhoods are not needed) over-specialized and applicable only to precisely defined skeletons. Without inspection by a human expert, making *a priori* choices where the "core" should end is problematic if not outright impossible. For more examples, see the Supporting Information in ref [10].

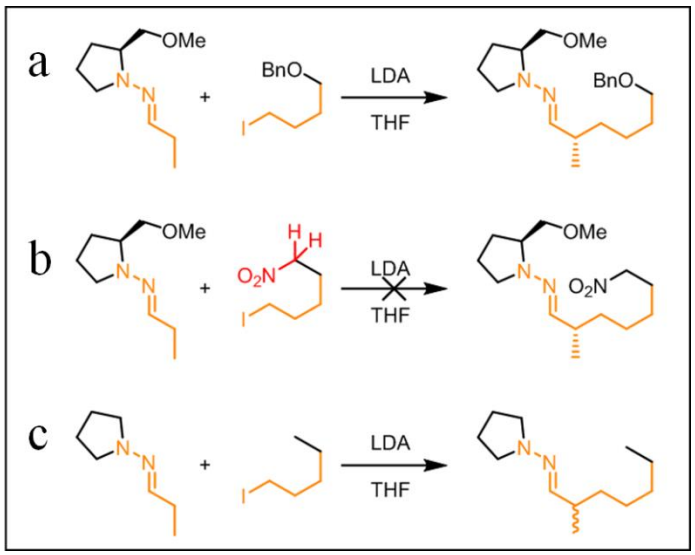


Figure S2. Application of an automatically-extracted reaction core to various synthetic targets. Reaction rule extracted from a literature precedent (see Figure S1) applied to **(a)** Epothilone A intermediate[S2] and **(b)** substrate with a nitropentane side chain. In the latter case, the reaction is not feasible since the pendant nitroalkyl group is incompatible (due to the presence of acidic H's) with lithiated azaenolate formed from the hydrazone upon the initial treatment with LDA. **(c)** In the absence of the distant stereodirecting group – not included in the reaction rule/core – the transform will still predict stereoselective outcome whereas in experiment, a racemic mixture will be obtained.

In some cases, incompatibility can be made to "disappear" by temporarily putting a "molecular invisibility hat" on one of the conflicting, "distant" groups. This is illustrated in **Figure S3** in which a primary alcohol and an organomagnesium compound are incompatible *unless* the alcohol is first

reacted with tert-butyldimethylsilyl chloride. This bulky silane serves as the so-called protecting group and makes the conflict “disappear”. With the protection in place, the desired reaction can now be easily carried out and, when done, the protecting group is removed (“deprotected”) to unmask the original hydroxyl functionality.

The above situations – and many others, involving steric and/or electronic effects stemming from substituents distant from the reaction core – are very common in organic chemistry. What they imply is that in addition to matching the reaction core, **any putative reaction rule should also take into account a range of conditionals** (e.g., “the reaction can be applied if no incompatibilities are detected” or “reaction rule can be applied only if group X is protected”).

We note that such conditional relations can sometimes reflect very subtle effects – as illustrated in **Figure S4**, two molecules might differ in only few (and sometimes just one) atoms somewhere far away from the reaction core, yet their overall reactivities might be drastically different. Teaching the machine to recognize such subtle effects requires detailed knowledge of reaction scope and mechanism. At the same time, such examples put into question the applicability of machine learning approaches which would necessarily categorize molecules having almost identical features/descriptors as having similar (reactivity) properties.

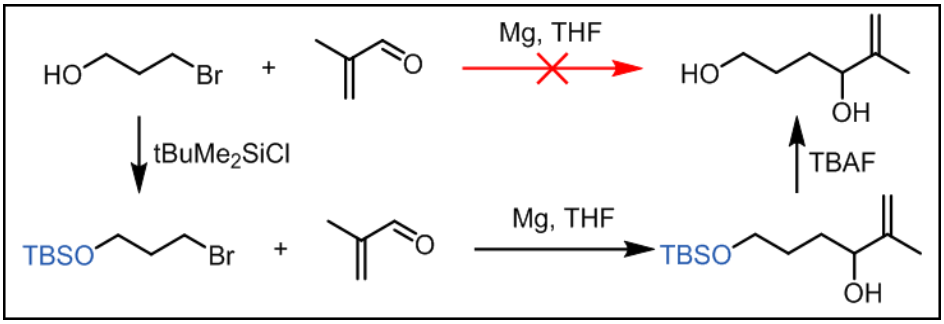


Figure S3. Non-local conflicts that can be avoided using protecting groups. During addition of an organomagnesium compound derived from 3-bromopropanol to an aldehyde, presence of a protic group (primary alcohol) causes destruction of the Grignard reagent formed – consequently, no desired product is obtained. This problem is avoided by converting the hydroxyl group to a silyl ether (*left*) which is ultimately cleaved (*right*) after performing the planned synthetic step.

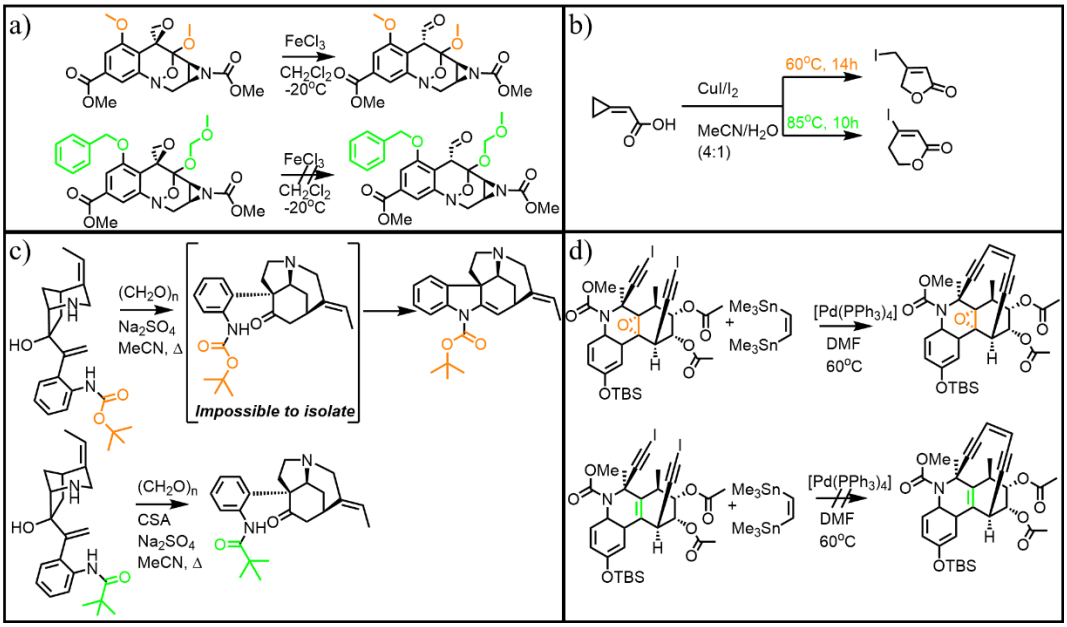


Figure S4. **Minor structural changes in starting materials can dramatically influence the reaction outcomes. (a)** Replacement of two O-protecting groups (orange OMe to green OBn and OMOM) in the intermediate in Danishefsky's synthesis of (+/-)-FR-900482 changes the lability of ether groups and prohibits rearrangement of an epoxide to an aldehyde[S3]. (**b)** Minute changes in temperature alter reaction mechanism and result in different products[S4]. (**c)** Small changes in electron density modify reactivity of N-pivaloyl and N-Boc protected anilines. The upper substrate reacts into an intermediate that is impossible to isolate and thus leads to a product that is markedly different than the one obtained from the lower substrate differing in only one atom (oxygen) [S5]. (**d)** Presence of the epoxide ring in the tricyclic moiety allows for close proximity of the terminal iodides enabling double Pd-mediated coupling. In contrast, when the epoxide is replaced by a double bond, the iodides are further apart and no cyclization is observed[S6]. Figure and caption reproduced from[S7].

S2.2. Statistics of reaction types and the importance of "black swan" chemistries. For any but trivial types of chemistries, the conditional relations discussed above can become quite involved (cf. **Section S3**). While it would be tempting to somehow machine-learn, ML, these conditionals from the examples of syntheses reported in the literature, one must consider the following facts:

(i) There are on the order of 10 million of reliable literature examples of chemical reactions; at the same time, the number of distinct reaction types and their variants is somewhere between 10,000 and 100,000. This means that, on average, there are few hundred to few thousand literature examples per reaction type on which one could attempt any machine learning. This is quite little given that the number of combinations of possible functional group/substituent "descriptors" that would have to be taken into account to train the models is easily into hundreds.

(ii) The situation is actually even worse given the distribution of reaction types plotted in **Figure S5**. In this plot, the x-axis ranks the reaction types according to the popularity (rank = 1 means the most popular reaction type, rank =2 denotes the second most popular reaction type, etc.). The y-axis gives the normalized number of times a particular reaction type was used in literature-reported reactions (reaction "popularity"). Importantly, the dependence of popularity vs rank is linear on the doubly-logarithmic scale indicating a power law. The presence of a heavy-tailed power law indicates, in turn, the importance of the low-occurrence, "black swan" events in the process underlying the distribution. In our plot, these are reaction types that are typically advanced chemistries used infrequently but still very important – as any chemist knows, such specialized transformations might be indispensable for making certain targets. For example, Meyers' synthesis of doxycycline required an unprecedented LiOTf catalyzed tandem SN' opening of the epoxide followed by ylide formation and [2,3]-rearrangement for the construction of A ring and highly diastereoselective tandem Michael–Dieckmann condensation setting the C ring[S8]. In another example, Baran's synthesis of (+)-

hapalindole Q relied on a previously undescribed oxidative coupling of indole with carvone-derived enolate[S9].

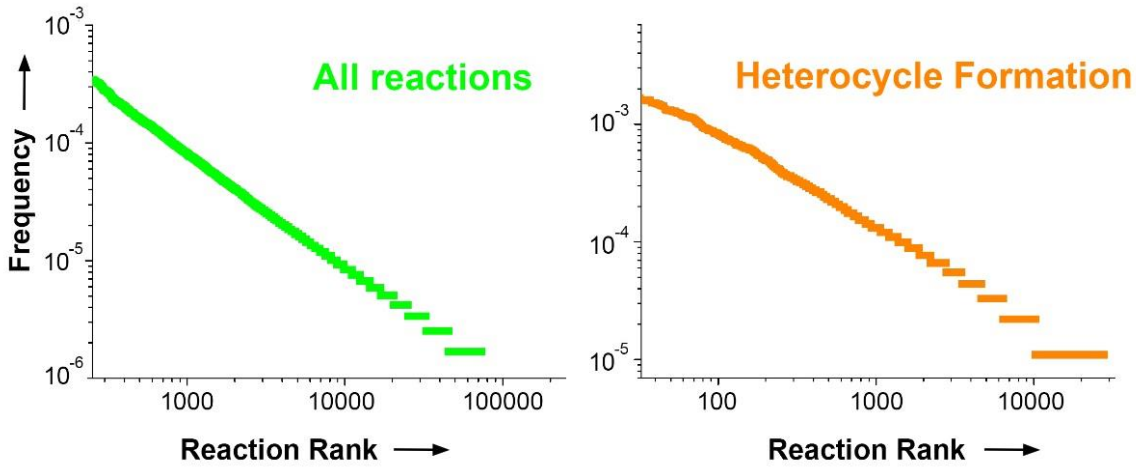


Figure S5. The frequency-rank plots of distinct reaction types. The plot on the left is based on the analysis of 1.2 million randomly chosen literature-reported reactions. The plot on the right is for the reactions forming aromatic heterocycles. In both cases, the distributions are power laws (i.e., linear on a doubly-logarithmic scale) indicating the relative importance – in entire chemistry (*left*) and in its specialized subfields (*right*) – of reactions that occur infrequently. Reaction rank = 1 indicates the most popular reaction, 2 is for the second-most popular, etc. Figure and caption reproduced from [10].

S2.3. The "philosophy" of rule coding. The above considerations (and also some other ones, like the extreme scarcity of available "negative" examples of reactions that did not work) point to our general conclusion for this section – namely, that ML approaches could possibly be used to learn simple chemistries for which the sets of literature examples are abundant (see ref [S10] where such learning was attempted) but are unlikely to capture the nuances of more advanced and less popular chemistries. In fact, recent work by Segler and Waller confirms that a neural network using automatically generated rules was much less efficient than the one using trivial set of 103 hand-coded rules unless more than 5,000 examples were available for each automatically extracted rule (see SI, Table 4 in ref [S11]). Accordingly, in our development of *Chematica* – which we aimed to become an expert system applicable not only very popular/simple chemistries – we had made an early strategic decision (i.e., after some early toying with machine extracted rules, see **Section 2.1**) that the chemical rules are to be codified by human experts and take into account nuances of admissible substituents, correct stereo- and regiochemistry, as well as reactivity conflicts, protection requirements, and selectivity issues. Of course, such an approach is very laborious and has been one of the main reasons development of *Chematica* took so many years – especially that in order to cover not only trivial chemistries, we had to code tens of thousands of reaction rules including the abovementioned "black swans" (currently, there are **50,000+ rules in *Chematica***). The specific examples illustrating transform coding are described in the next section.

Section S3. The logic and examples of reaction coding.

S3.1. Accounting for stereo- and regioselectivity. All reactions are coded in the well-known SMILES/SMARTS[S12,S13] notation which represents the molecules and reactions as alphanumeric strings (on which the operations are much faster than on matrices, e.g., in MOL files). However, before the SMILES/SMARTS notation can be employed, one must adapt it to deal with the all-important issues of stereochemistry and regiochemistry.

For the stereochemistry, the SMILES/SMARTS notation uses the @ and @@ symbols which, unfortunately, do not correspond to the absolute R and S configurations. As discussed in ref [10], in simple reactions and using software like RDKit[S14], the stereochemistry of reactions coded with all

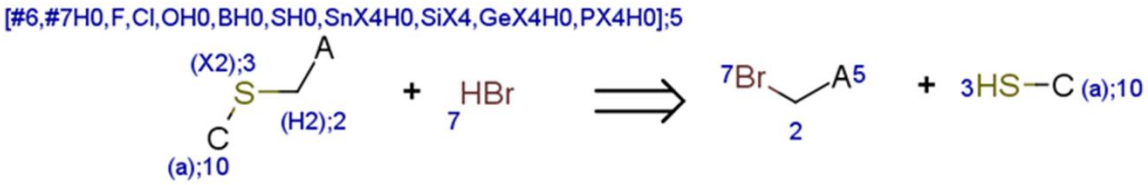
atom mappings can usually be assigned properly (e.g., @ changing into @@ or @@ changing into @ denote configuration inversion). However, for more complex reactions involving multiple stereocenters (especially, proximal ones), the existing algorithms perform poorly. The same is true for regiochemistry with symbols // and /\ specifying regiochemistry of double bonds in individual molecules – unfortunately, there have been no algorithms to keep track of these symbols over reactions' SMARTS and to ascribe proper reaction regiochemistry.

To overcome these problems, we developed two software modules (called STEREOFIX and REGIOFIX) which pass between the retrons (reaction products) and the synthons (reaction substrates) not only the (@, @@) and/or (//,/\) information, but also appropriately ordered (by the masses of substituents) lists of bonds neighboring each atom mapped in the transform. These lists keep track of the masses of substituents changing upon bond breaking or making and overall order the neighboring bonds according to these changes. While constructing the lists, it is essential to add any missing hydrogens, which aids proper ordering by avoiding ambiguity (in corner cases, next-nearest bonds might need to be taken into account). Ultimately, upon the execution of a transform, its stereo/regiochemistry is determined by the consensus of the bond list order and the stereo/regiochemistry symbols present in the SMARTS notation.

S3.2. Coding a simple transform: alkylation of aromatic thiols. Let us first consider a relatively simple reaction of alkylation of aromatic thiols (**Figure S6**). This reaction has a broad scope of admissible substituents and is a good example of how to make plausible generalizations beyond the already reported literature precedents. For instance, there are many literature examples of aromatic thiols serving as substrates in this reaction – although not all aromatic moieties have been tried experimentally, it is chemically reasonable to assume that carbon "10" belongs to any aromatic or heteroaromatic ring – hence, in the SMARTS notation it is denoted by a general-scope lower-case "c" denoting any aromatic carbon. In a similar spirit, for the primary bromide moiety, position "5" can have different types of atoms (carbon, nitrogen, boron, fluoride, chloride, oxygen, sulfur, silicon, germanium, phosphorus, etc.) – most of these substituents were reported in the literature, but there have been no examples of boron in the "5" position. Still, boron is included in the list of admissible "5" substituents since it was present in this positions in analogous reactions of alkyl thiols, phenols, and alcohols.

Other fields in the record, specify the groups present elsewhere in the molecule that need to be protected (here, thiols) or pose a serious cross-reactivity conflict (e.g., alkyl iodides, acyl chlorides, organomagnesium compounds, etc.). The record also gives typical/commonly used reaction conditions (here, DIPEA or other base, DCM), and also categorizes these conditions (here, denoted by internal abbreviation WL62). This categorization is important for selecting protecting groups most proper for this reaction (see **Section S4**). The user of *Chematica* is also provided with DOI's of relevant/illustrative references for this type of chemistry. Other fields (18 in total; not all shown here), focus on some additional nuances (e.g., whether reaction is diastereoselective) and are important for the search algorithms (**Section S6**) to ensure that the transform is properly executed for desired retrons.

We note that a similar reaction for aliphatic thiols is coded as a separate record. This is so because if carbon "10" were specified as aromatic ("c") or any aliphatic ("CX4"), then such a line would admit, for instance, trifluoromethylthiol which needs to be transformed into an active species, for which different sets of protections/incompatibilities are required and the scope of bromides is limited and depends on the specific trifluoromethyl reagent used[S15,S16]. In the record for the aliphatic thiols, carbon "10" has precisely delineated scope of substituents, [SX2:3][CX4:10]([#6,#1:1])([#6,#1:8])[#6,#1:6], meaning that it can have in its immediate environment (positions "1", "8", and "6") only carbon or hydrogen atoms (#6- carbon, #1-hydrogen). Although there are literature precedents of a heteroatom in the abovementioned positions, such cases need a more detailed reaction core (e.g., specifying if it can be linear or should be a part of a ring) which, again, needs to be coded as a separate record.



name: "Alkylation of Thiols with Primary Bromides"
reaction SMARTS: "[#6,#5&H0,F,Cl,OH0,BH0,SH0,SnX4H0,SiX4,GeX4H0,PX4H0:5][CH2:2][SX2:3][c:10].[Br:7]>>[C:2]([Br:7])[*:5].[S:3][c:10] "
protection_conditions_code : ["WL62", "BW11"]
protections: ["[CX4,c][SX2H]"]
incompatibilities: ["[CX4][Br,I]", "[#6][N+]#[C-]", "[#6]N=C=[O,S]", "[CX4!H0][N+]([O-])=O", "[#6]C(=O)[Cl,Br,I]", "[#6][CH]=[SX1]", "[#6]C(=[SX1])[#6]", "[#6][S](=O)(=O)[OH]", "[#6][SX3](=O)[OH]", "[#6]S(=O)[Cl,Br,I]", "[CX4]1[SX2][CX4]1", "[#6][SX2][SX2][#6]", "N=N", "[#6]C(=[O])OC(=[O])[#6]", "[#6]C(=O)[N]=[N+]=[N-]", "[#6]O[OH]", "[#6]OO[#6]", "ClC=N", "[CX4,c][NX3][NH2]", "[#6]=[N+]=[N-]", "c[N+]#[N]", "[#6][Li]", "[#6][Mg][*]", "[#6][Zn][*]", "[#6][SX2,O]C#N", "[#6][NX2]=O", "[CX3]=[CX2]=O", "[CX3]=[CX3][OH]", "[CX4][O][S](=O)(=O)[#6]", "[CX3]([#6,#1])([#6,#1])=[NX2H]", "[OH][CX4]!@[O]", "[CX3]=[NX3+][O-] "]
typical reaction conditions: "DIPEA or other base.DCM"
references: "DOI: 10.1021/jm051010j and 10.1039/C4CC08829H (SI, page S3) and 10.1002/cmdc.201100144 and 10.1016/j.bmc.2014.11.002 and 10.1021/jm500827t"
diastereoselective: False

Figure S6. General scheme (*top*) and part of *Chematica*'s reaction record (*bottom*) for the alkylation of aromatic thiols with primary bromides.

S3.3. Moving to more advanced chemistries: An example of an A3 reaction. The example of thiol alkylation hinted at the importance of very logical coding of the reaction rules such as to permit plausible extensions while avoiding notation that would encompass structural motifs not admissible in a given reaction. In more complex chemistries, the process of coding is greatly facilitated by constructing block diagrams summarizing all logical "yes/no" conditions. This kind of logical dissection was used, for example, by Baldwin while formulating his celebrated "Baldwin rules" when preparing a review and categorizing the available knowledge of relative feasibility of ring closure[S17].

Let us take as an example the enantioselective A3 coupling – between an aldehyde, an alkyne, and a secondary amine – whose chemical scheme (in retrosynthetic direction) is shown in the top left portion of **Figure S7**. In the accompanying block diagram, the first condition to be met is that the reaction has to be intermolecular. The chemical rationale here is that (i) proper binding of a chiral catalyst might be hampered with all components belonging to the same molecule and leading to a highly strained ("cyclic alkyne") transition state; and (ii) there are no literature examples of intramolecular A3 reactions. In the SMARTS notation (reaction record in the right portion of **Figure S7**), this requirement is encoded by indicating atom "8" as **"**R0**"** which means that it cannot be a part of a ring. The next condition serves to eliminate 2-formyl-N-heterocyclic aldehydes such as 2-pyridinecarboxy-aldehyde scaffolds, for which this reaction is unprecedented in the asymmetric variant. We note that the 2-aminomethylazole/azine that forming in such a reaction would be able to bind copper and act competitively as a ligand. Also, examination of precedent attempts of addition of organometallics to relevant aldehydes[S18] or imines[S19] clearly evidences deterioration of enantioselectivity compared to phenyl, 3- or- 4-pirydylaldehydes. To encode these conditions in SMARTS, we specified two additional atoms ("36" and "37") next to the aromatic carbon "9". Atom "36" is limited to an aromatic carbon (lower case "c") and atom "37" can be of types "c,o,s" (meaning aromatic carbon, oxygen, or sulphur), which eliminates 2-formyl-N-heterocyclic scaffolds from potential results. With the condition fulfilled, we need to decide whether the aldehyde is aromatic or not. If it is not aromatic, the transform might match other variants of A3 transform (with their own "decision trees"). If the aldehyde is aromatic, we inspect the bulkiness of substituents at nitrogen-bound carbons "2" and "6" – only non-bulky groups with two or three hydrogens are allowed [N:4]([CX4:2]([#1:3])([#1,#6:7])[#1:10])[CX4:6]([#1:5])([#1,#6:11])[#1:13]; this notation also eliminates amines with stereocenters at this position (a stereocenter could create an "opposite" stereochemical induction to that of the chiral catalyst and also act as a steric hindrance). Such "mismatch" could affect

the overall stereochemical outcome. The last condition that has to be met is the presence of carbon or silicon atom at position “17”. This condition is based on the available literature whereby only these two elements are admissible in the position of interest.

Naturally, the record for the transform contains not only the substituent scope from the decision tree but also categorization of reaction conditions, typical suggested conditions, list of incompatible groups, list of groups to protect, and other fields not shown in **Figure S7**.

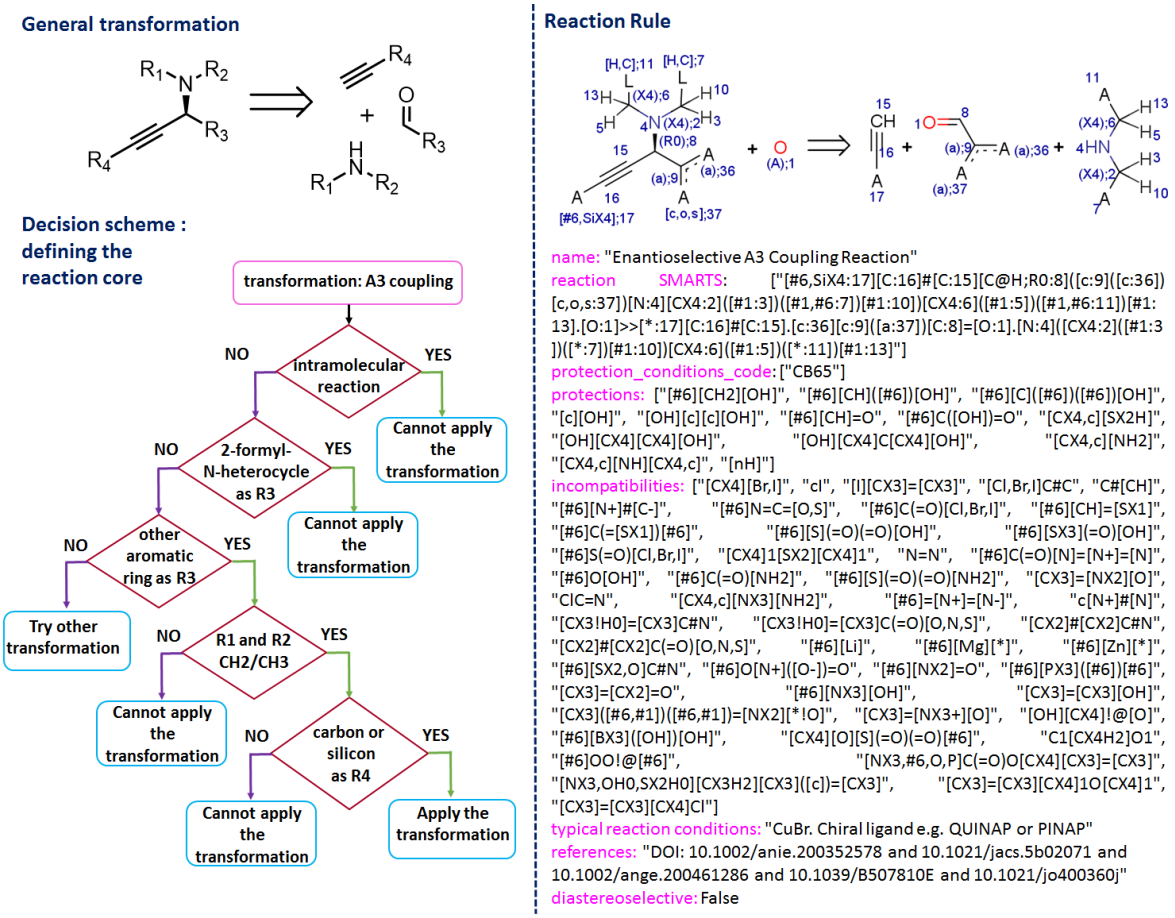
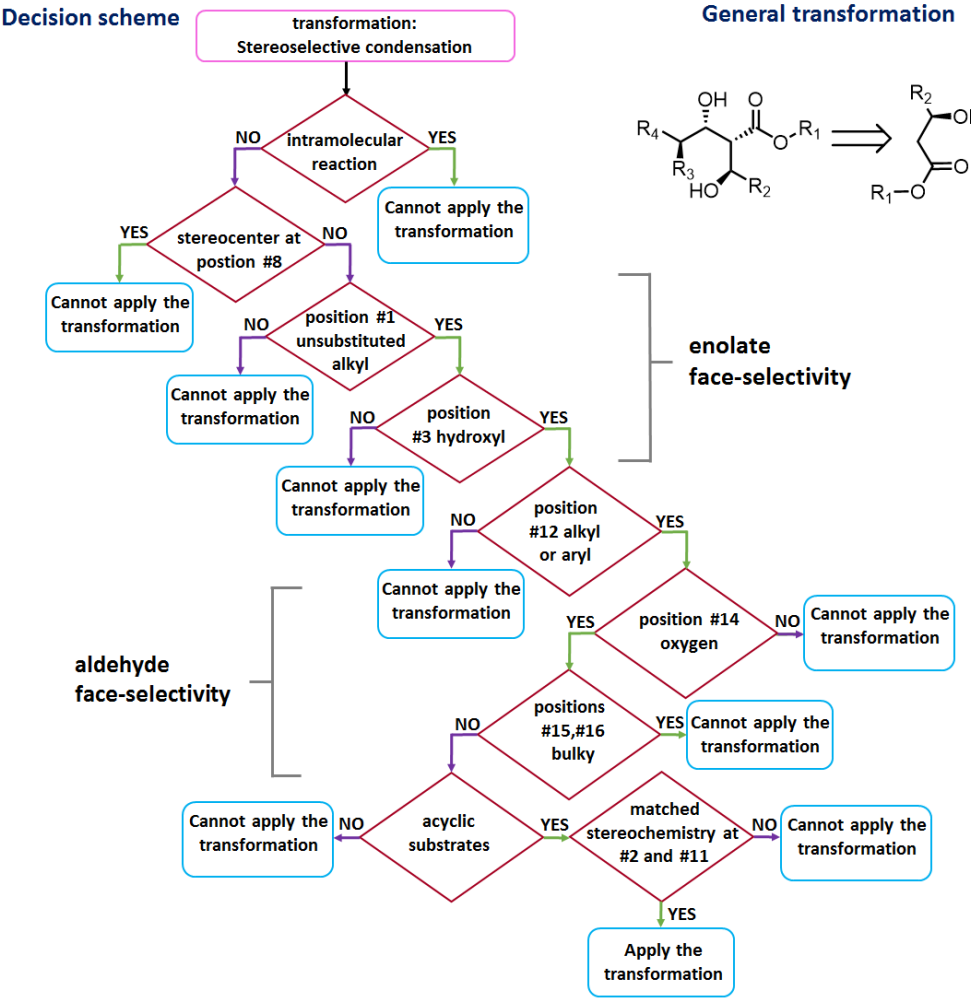


Figure S7. The left portion shown the general scheme and a “decision tree” guiding the coding of a reaction rule for the A3 coupling reaction. The key parts of the coded reaction record are shown in the right panel.

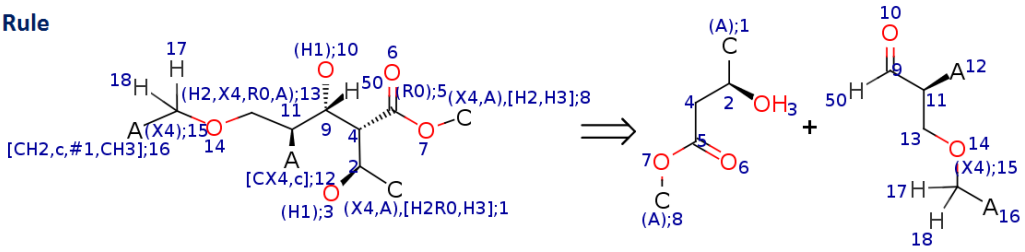
8.2.3. Coding complex rules: An example of a double stereodifferentiating condensation of esters with aldehydes. Our last example deals with a very advanced – and hard to code – reaction in which the stereoselective outcome is dictated by the arrangement of substituents on the substrate's scaffold (**Figure S8**). In this reaction, two stereocenters are created via double stereodifferentiating condensation reaction. First, we check if the reaction involves an aldehyde and an ester within one molecule (i.e., is it intramolecular). Since the first mechanistic step requires deprotonation of the alcohol and enolisation of the ester with LDA (strong base), the tethered enolizable aldehyde will cause a cross-reactivity problem (the enol ether will also form). Therefore, the reaction must be intermolecular. Next, we move to the requirements that need to be met to provide proper ester enolate face-selectivity. To avoid an additional chiral center that might have a mismatched stereodirecting induction,[S20] admissible atoms at position "8" are limited to alkyl carbons with two or three hydrogens. The face selectivity of the enolate is controlled by the conformation of the stereocenter on the β-carbon[S21,S22] (carbon "2" in the reaction scheme) which may be influenced both by dipole-dipole interaction and 1,3-allylic strain. Due to this fact and to ensure proper orientation of the ester enolate, we limit the bulkiness at position "1" to unsubstituted alkyls; we also restrict position "3" to a hydroxyl group. Next, we inspect factors that need to be considered for proper face-selectivity of the aldehyde. Addition of a nucleophile to the aldehyde occurs in a chelation-controlled manner suggesting that bulkier substituents on atom "12" would not adversely affect the stereospecificity[S22]. Accordingly – and even though a only methyl at position "12" was reported – we allow larger substituents at this position. On the other hand, we exclude from this position any EWG groups to avoid a competitive chelation site and epimerisation prone chiral 1,3-dicarbonyl motif – this leaves alkyl or aryl carbons as admissible substituents. Furthermore, we observe that presence of the oxygen atom at position "14" is crucial because it chelates to titanium providing face selectivity of the aldehyde. Also size and type of substituents at oxygen "14" have to be carefully inspected because bulky groups would prevent Ti binding. Based on these considerations, atom "15" is limited to a carbon with two hydrogens and atom "16" is restricted to an aromatic carbon or an unsubstituted alkyl. The last two conditions (at the bottom of the decision tree in Figure 90) are common for both substrates. The first one limits the scope of potential substrates only to acyclic compounds as cyclic structures might distort the aldehyde-titanium chelate conformation or face selectivity of the ester enolate. This condition is coded by denoting positions "5" and "13" as "R0". The last requirement concerns the consonant selectivity at both substrates that ensures the desired diastereoselectivity. In case of a "mismatched" pair of substrates (i.e., for an aldehyde with stereochemistry at carbon "2" as drawn in the scheme but for alcohol with "opposite" stereochemistry at carbon "11"), experiments evidence formation of a racemate at C4 – in other words, with such a stereochemical mismatch, the transformation is no longer fully stereoselective and cannot be applied.

Decision scheme



General transformation

Reaction Rule



name: "Stereoselective condensation of esters with aldehydes"
reaction SMARTS: ["[CX4H2R0,CX4H3:1][C@@H:2]([OH:3])[C@H:4]([C@@:9]([#1:50])([OH:10])[C@@H:11]([CX4,c:12])[CX4H2R0:13][O:14][CX4:15]([#1:17])([#1:18])[CH2,c,#1,CH3:16])[CR0:5](=[O:6])[O:7][CX4H2,CX4H3:8]>>[C:1][C@@H:2]([O:3])[C:4][C:5](=[O:6])[O:7][C:8].[#1:50][C:9](=[O:10])[C@@H:11]([*:12])[C:13][O:14][CX4:15]([#1:17])([#1:18])[*:16]"]
protection_conditions_code: ["SB16","SC88"]
protections: ["[#6][CH2][OH]", "[#6][CH]([#6])[OH]", "[#6][C]([#6])([#6])[OH]", "[c][OH]", "[OH][c][c][OH]", "[#6][CH]=O", "[#6]C([OH])=O", "[CX4,c][SX2H]", "[OH][CX4][CX4][OH]", "[OH][CX4]C[CX4][OH]", "[CX4,c][NH2]", "[CX4,c][NH][CX4,c]", "[nH] "]
incompatibilities: ["[#6][CH]=[SX1]", "[CX3]=[NX3+][O-]", "[CX4][O][S](=O)(=O)[#6]", "[#6]S(=O)[Cl,Br,I]", "[#6]C(=[SX1])[#6]", "[#6][SX3](=O)[OH]", "[CX4]1[SX2][CX4]1", "[#6][S](=O)(=O)[OH]", "[#6][N+]#[C-]", "[#6]N=C=[O,S]", "[#6][SX2,O]C#N", "[#6]C(=O)[Cl,Br,I]", "[#6]C(=[O])OC(=[O])[#6]", "ClC=N", "[#6]O[N+](=[O-])=O" , "[#6]O[OH]", "[#6]OO[#6]", "[#6][NX2]=O", "[CX3]=[CX2]=O", "[#6]=[N+]=[N-]", "c[N+]#[N]", "[CX3]=[NX2H]", "[CX3]=[NX2][O]", "[#6][NX3][OH]", "[CX3]=[CX3][OH]", "[OH][CX4][O]", "[#6][Li]", "[#6][BX3]([O,#6])[O,#6]", "[#6][Mg][*]", "[#6][B-](F)(F)F", "[#6][Zn][*]", "[#6][PX3]([#6])[#6]", "N=N", "[#6][SX2][SX2][#6]", "[#6][SX3](=O)[#6]", "[CX4][Cl,Br,I]", "[Cl,Br,I]C#C", "C#[CH]", "[#6][S](=O)(=O)[#6]", "[CX4]1[O,N][CX4]1", "[#6]C(=O)[N]=[N+]=[N-]", "[CX4!H0][N+]([O-])=O", "[CX4!H0]C#N", "[#6]C(=O)[NH2]", "[#6]C([NH][CX4,c])=[O,S]", "[CX4!H0][C](=[O])[OH0]", "[CX4!H0]C(=O)N([#6])[#6]", "[#6][S](=O)(=O)[NH2]", "[CX4,c][NX3][NH2]", "[CX3]([#6,#1])([#6,#1])=[NX2][*!O]", "[CX3!H0]=[CX3]C#N", "[CX3!H0]=[CX3]C(=O)[O,N,S]", "[CX2]#[CX2]C#N", "[CX2]#[CX2]C(=O)[O,N,S]", "[CX3]=[CX2]=[CX3,CX2]", "[n][c;r6]([Cl,F])[n,c]"]
typical reaction conditions: "1.LDA.THF then TMSCl 2.TiCl4.DCM"
references: " 10.1016/S0040-4039(00)82373-4 "
diastereoselective: False

Figure S8. The upper part has the general scheme (in retrosynthetic direction) of the reaction and the "decision tree" guiding the coding of a reaction rule for stereoselective condensation of esters with aldehydes. The coded record – also containing information about protections, incompatibilities, etc. – is shown at the bottom.

Section S4. Additional comments on the evaluation of protection requirements and incompatible groups.

As narrated in previous sections, one of the key features of our approach is to code the transforms along with the information about molecular "context" – in particular, about groups that are incompatible with/cross-reactive within the reaction and about those that require protection. In this section, we discuss how this contextual information is managed when the transforms are applied to specific molecules. As illustrated in **Figure S9**, the first step is to "remove" the motifs present in the reaction rule from the structures of particular retron/synthon molecules to which the rule is applied. Without this preliminary step, the reacting groups specified within the transform would themselves – quite nonsensically – be detected as incompatible ones. Next, the algorithm checks whether the remaining parts of the molecules contain motifs specified in the list of incompatible groups or groups to protect (see reaction records in **Figures S6-S8**, fields "incompatibilities" and "protections").

Detection of an incompatibility is reported to the user and marks the reaction as highly unlikely. In automated searches for complete reaction pathways, such reactions receive highly unfavorable scores or are altogether avoided (see **Section S6** for details).

Management of groups that require protection is more involved. Recall that all reaction transforms include a field that categorizes the typical reaction conditions for this reaction into one of over 100 different classes (similar but not identical to Green's tables described in [S23]). For example, condition "WL62" in **Figure S6** signifies "thiol nucleophile" whereas condition marked "CB65" in **Figure S7** stands for soft Lewis acids such as silver or copper. Depending on a specific reaction condition, the groups to be protected might require the use of different protecting groups. This information is stored in tables for each group that might require protection – in the bottom left part of **Figure S9**, the table is for carbonyls and for each possible reaction condition suggests appropriate protecting groups (entries colored in green, ranked in terms of synthetic utility; entries colored in red are the groups that would not survive the reaction). For the specific reaction shown in **Figure S9**, the algorithm detects the need to protect a carbonyl group. With the reactions conditions corresponding to the table's column L26 ("alkyllithium"), the most suitable protecting groups identified are those in rows #2, #4, #6 – they correspond to 1,3-dioxane, 1,3-dioxolane, and dimethyl acetal. This list of most suitable protecting groups is then returned to the user.

We note that for the same reaction, the algorithm might detect both a conflict and a need for protection – in such cases, both pieces of information are returned although, as noted earlier, the conflict is a much more serious/unsurmountable problem and during automatic searches for full synthetic paths, reactions involving conflicts are penalized much more heavily than those requiring protection.

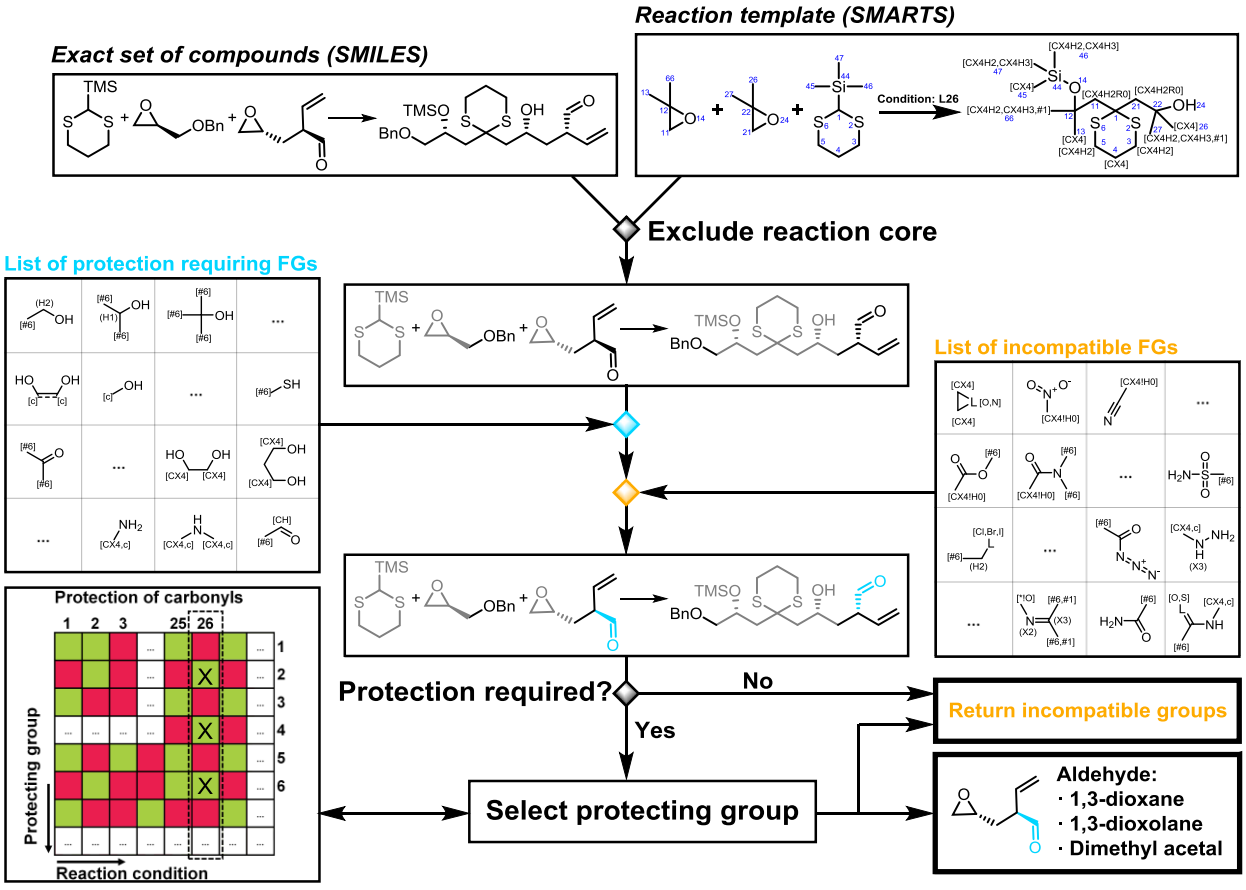


Figure S9. Block diagram illustrating how the information about protecting and/or conflicting groups is applied to specific molecules and subsequently processed.

Section S5. Evaluating applicability of transformations beyond reaction records.

S5.1. Example of electrophilic aromatic substitutions. In some cases, even the most meticulous specification of the molecular "context" is not sufficient for determining the applicability of a reaction rule. One illustrative example we focus on in this section deals with the implementation of electrophilic aromatic substitutions, EAS, which are amongst the most powerful tools in organic chemist's arsenal. The basic reaction core for EAS transformation is very simple and comprises an aromatic carbon and an attacking electrophile. Naturally, any chemist knows that such a core will not capture the all-important effects of flanking substituents. However, even for the simplest case of benzene, the number of ortho-, meta-, para- combinations one would have to consider for different electron-withdrawing and/or electron-donating groups would rapidly grow into thousands – coding each of these combinations via separate SMARTS (or even grouping them using atom lists) would be extremely tedious. Worse still, there are many other types of aromatic rings and ring systems and for such general cases, accounting for all combinations is prohibitively complicated. The way around this problem is to determine the admissible locus/loci of EAS based on physico-chemical measures such as electron densities. In thinking about this problem, the best solution would be to use high-end quantum mechanical calculations – however, one must remember that during retrosynthetic searches, *Chematica* evaluates literally millions of possibilities with a sizeable fraction of these possibilities being aromatic substitutions. Even if each QM calculation took only 1-10 sec, evaluation of, say, 10,000 EAS reactions would translate into times of several to tens of hours that are not compatible with expectations of practicing chemists using *Chematica*. Accordingly, we have developed a module that is trained/ "pre-parametrized" to estimate the loci of substitutions within milliseconds and with unrivalled accuracy (cf. below) even for complex aromatic systems (see example in **Figure S10**). In

the following we briefly narrate the existing methods and then provide an overview of our approach (which will be described in detail in a separate, upcoming publication).

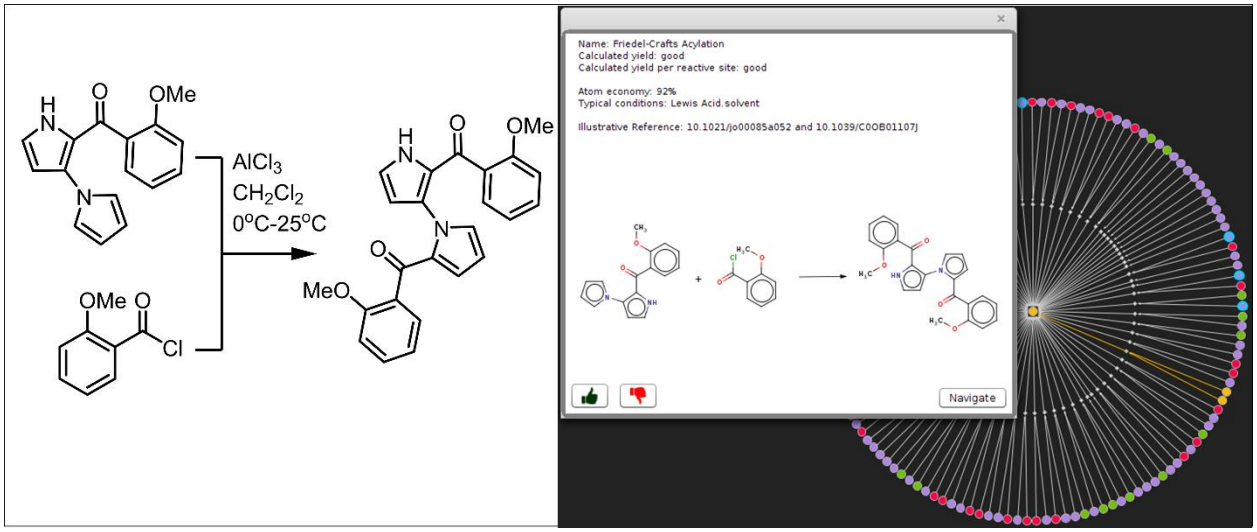


Figure S10: In electrophilic aromatic substitution reactions, EAS (e.g., Friedel-Crafts acylations, electrophilic aromatic brominations or nitrations), substitution can be performed regioselectively even in complex ring systems. In the example shown here, Friedel-Crafts acylation is performed in Nicolaou's synthesis of marinopyrrole A[S24]. Our algorithm – described in the text with outcome illustrated in the Figure by a screenshot from *Chematica*– correctly predicts the position of this substitution by comparing the activities of four aromatic rings and by considering even gentle effects of substituents present in the molecule's structure that "globally" determine the reaction's outcome. Here, the algorithm identifies atoms located in the anisole-derivative part as deactivated by the ketone substituent. For positions available in the pyrrole rings, the algorithm considers the combined influence of all substituents present, and identifies the C2 carbon in the N-substituted pyrrole as the most suitable locus for electrophilic attack.

(i) Existing methods. The numerous approaches to predicting regioselectivity of EAS reactions can be subdivided into three categories reflecting the computational cost. Two of them are based on quantum-mechanical (QM) calculations: the "density-based" QM methods and the "atom-based" ones. Methods from the former group operate directly on electron density (or electronic wave function) of the molecule. Examples of such approaches include calculations of electronic populations in the highest-occupied molecular orbital (HOMO)[S25] or of the average local ionization potential[S26]. Their drawback, aside from computational times being too long for applications such as *Chematica*, is that these molecule-wide results are generally not correlating well (as we verified and will describe separately) with specific atoms being prone to EAS.

The first "atom-based" QM methods were based on atomic charge density though the correlations between atom's partial charge and its reactivity have been only moderate[S27]. Accordingly, more modern methods have been based on the so-called electrostatic potential at nuclei (EPN) or sigma-complex approximation[S28]. The EPN was proposed by Politzer *et al.* [S29] and describes electron density around specific atoms. Sigma-complex approximation is based on an assumption that reactivity is related to the stability of the so-called sigma (Wheland) complex which is an intermediate in most EAS reactions. The simplest method developed following this idea is proton affinity (PA), the most complex one is the so-called electrophile affinity, where instead of a proton, a cationic electrophile (e.g., Cl^+ or NO_2^+) is used[S30]. While literature reports suggest that these methods are most promising to determine the loci of EAS, they are unsuitable for massive retrosynthetic analyses due to the already mentioned high computational cost.

The third group of approaches are (semi)empirical methods, which do not require ab-initio QM calculations and hence can be more suitable for automatic retrosynthetic analysis. The methods that have been developed include those based on Hückel theory[S31,S32] (which was used in the initial versions of *Chematica*[10]), ^{1}H and ^{13}C NMR shifts (which was successfully applied to predicting

regioselectivity of selected types of heterocyclic compounds[S33]), and Hammett substituent constants[S34,S35].

In assessing the accuracy of the above methods, we compared their predictions against experimentally reported data for substituted benzenes (**Figure S11**; more comparisons of this type, for diverse classes of aromatics, will be published in a separate paper). All approaches correctly predict regioselectivity in simple cases like monosubstituted benzenes (very often used as a benchmark when a new method is introduced) or disubstituted benzenes in para positions. For other cases, however, NMR-based methods have low predicting power and the Hückel method offers only a slight improvement. Reasonable precision is achieved with EPN or Hammett-constant methods. The best results are obtained using PA and electrophile affinity (we performed calculations using Cl^+ as an electrophile, hence we call it chlorine affinity, denoted as ClA). Both of these highly accurate methods (PA and ClA), however, require a series of QM calculations (one for each possible reaction site) limiting their usefulness in our retrosynthetic endeavours.

Substituents on benzene ring	Hammett	Hückel	^{1}H NMR	^{13}C NMR	EPN	PA	ClA
-Me	✓	✓	✓	✓	✓	✓	✓
-Cl	✓	✓	✓	✓	✓	✓	✓
-NO_2	✓	✓	✓	✘	✓	✓	✓
-OMe	✓	✓	✓	✓	✓	✓	✓
-Ac	✓	✓	✓	✓	✓	✘	✘
1,2-diMe	✓	✘	✘	✓	✘	✓	✓
1,2-diCl	✘	✘	✓	✓	✓	✓	✓
1,2-diOMe	✓	✘	✘	✘	✓	✓	✓
1-Me, 2-Cl	✓	≈	✓	✓	✓	✓	✓
1-Me, 2-NO_2	✓	≈	≈	✘	✘	✓	✓
1-Cl, 2-NO_2	≈	≈	✓	✘	✓	✓	✓
1-NO_2, 2-OMe	✓	≈	≈	≈	✓	✓	✓
1,3-diMe	✓	≈	≈	✓	✘	✓	✓
1,3-diCl	≈	≈	≈	✓	✓	✓	✓
1-Me, 3-OMe	✓	✘	✘	✘	✓	✓	✓
1-Cl, 3-Ac	≈	≈	✘	✓	✓	✓	✓
1-Cl, 3-OMe	✓	≈	✘	✘	✓	✓	✓
1-Ac, 3-OMe	✓	≈	✘	✘	✓	✓	✓
1-Me, 4-NO_2	✓	✓	✓	✘	✓	✓	✓
1-Me, 4-Ac	✓	✓	✓	✘	✓	✓	✓
1-Me, 4-OMe	✓	✓	✓	✓	✓	✓	✓
1-Cl, 4-NO_2	✓	✓	✓	✘	✓	✓	✓
1-Cl, 4-Ac	✓	✓	✓	✓	✓	✓	✓
1-Cl, 4-OMe	✓	✘	✓	✓	✓	✓	✓
1NO_2 4-OMe	✓	✓	✓	✓	✓	✓	✓
1-Ac, 4-OMe	✓	✓	✓	✓	✓	✓	✓
1,2-diOMe, 4-COOH	✓	✘	✘	✘	✘	✓	✓
1,2-diOMe, 4-Me	✓	≈	✘	✘	✘	✓	✓
1,3-Me, 2-OMe	✓	✓	✓	✓	✘	✓	✓
1,2-OMe, 4-CN	✓	✘	✘	✘	✘	✓	✓
1-COOH, 2-Cl, 4-F	≈	✓	✘	✓	✓	✓	✓

Figure S11. Comparison of the ability of various methods to predict regioselectivity of EAS reactions on substituted benzenes. Correct results (i.e., agreeing with literature-reported, experimental outcomes) are denoted with ✔ symbols. Incorrect predictions are marked with ✘. Symbol ≈ is used when reaction is allowed due to the so-called ortho-para rule (i.e., when two most active positions are mutually in 1,3 arrangement [ortho-para], the second most-active position is also allowed).

(ii) *Chematica*'s approach. With the aim to develop a rapid yet accurate method, we constructed a model that combines Hammett substituent constants, ring average proton affinities, the Hückel method, and various additional empirical rules. As we will describe separately, this model offers high accuracy of prediction.

In brief, for benzene rings, regioselectivity is determined by Hammett substituent constants. To overcome some known limitations of this method (e.g., underestimation of the effects of strongly donating groups), we added additional empirical/literature-result-based rules that "overrule" the raw Hammett predictions for such substituents. In case of benzene, regioselectivity is governed exclusively by substituents. In heterocyclic compounds, however, regioselectivity is dictated predominantly by heteroatoms (ring type) – consequently, we implemented zero-order heuristics to denote the most active position in every heterocyclic ring type. This leaves the cases of heterocycles in which regioselectivity is changed by the additional substituents present. 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 second "meta" ("5") position but at the "ortho" ("6") position relative to nitrogen. In another example, substituted pyrroles with an electron-donating group in the "2" position undergo EAS reactions at the "5" position. However, when an electron-withdrawing group is present in the "2" position, the most reactive position is "4". In order to include such dependencies, we supplemented these empirical rules with Hammett constants to quantitatively measure the effect of the substituents. For polycyclic aromatic hydrocarbons (PAH), detection of the most active ring and position cannot be achieved based on Hammett constants. Instead of creating rules for all possible PAH structures, we decided to use the rapid and reasonably accurate Hückel model which for this class of compounds offers good accuracy against experimental results (unlike in heterocycles, for which Hückel fails dismally).

The algorithm incorporating these rules is illustrated in **Figures S12** and **S13** and consists of two parts. The first part divides molecules into a set of single rings and the most active position within each ring is determined (based on procedures outlined above). The second part of the algorithm sequentially removes less active ring(s) and returns only those at which EAS reaction might occur. Removal of less active rings (note: different numbers might be removed depending on ring types and substituents) is itself divided into several steps. First ("Step 1" in the right portion of **Figure S12**), 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 "Step 2," 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 (note: if the heuristics judge the rings' activities being similar, no removal is done). In "Step 3", remaining rings are examined based on our own protocol we called ring average proton affinity (RAPA). As we mentioned earlier, PA is an accurate but highly time consuming method preventing its use in *Chematica*. Accordingly, we use precalculated values of PA for each position in every ring type (e.g., pyrrole, tiophene, etc.) and popular ring systems. The RAPA value is calculated as an average PA for a given ring with correction for the substituents (*RAPA_real* = *RAPA_unsubstituted* – *const*sum*(*Hammett*) + *EDG_correction*, where *RAPA_real* is RAPA for a given ring with all substituents, *RAPA_unsubstituted* is RAPA for an unsubstituted ring, sum of Hammett parameters is scaled by a constant parametrized against literature data, and *EDG_correction* corrects for underestimation of EDG by Hammett approach and is also parametrized against literature examples). Ring activities are thus quantified and the less reactive rings (below a certain preset threshold) are again removed.

After this "filtering," we change the strategy and instead of removing rings, we now try to select the most active among the remaining ones ("Step 4"). This is done by heuristics similar (but not identical) to those in "Step 2". Typically, after this stage only one ring is left and selected, along with its most active position. However, if the input molecule has two or more rings of the same type, which survived steps 1-4 and were both/all marked as "the most active ones" (based on substitution patterns, chemical environments), the final decision is taken based on Hammett substituent constants in "Step 5".

Naturally, one could argue that this multi-step protocol could be simplified and less active rings could be removed in just one step based on, for example, a combined measure of PA values and sum of

Hammett constants. As we will describe in a separate paper, however, when such simplified schemes (many) were tested, none reached the accuracy of the model described above.

Foreshadowing this upcoming publication, the accuracy of our approach has been at least 90% (and likely higher) when validated on a dataset comprising over 18,000 published reactions collected from Reaxys database and (i) matching a reaction motif for EAS but (ii) excluding entries where transition metals (Pd, Pt, Ni, Ir, Rh, Ru, Co) or strong bases (LDA, BuLi etc.) were used as reagents (indicating different reaction mechanism). Many of the incorrect predictions could be divided into four categories:

(1) Unexpected selectivity – that is, when the product reported in the Reaxys entry was in sharp contradistinction to common chemical knowledge. Such discrepancies were either due to incorrect product structure reported in a publication or due to the reaction mechanism being different from EAS. One example is the nitration of toluene in micellar media that was reported to supposedly undergo in the *meta* position – likely, dues to an error in a naming the reaction product. (example "1" in **Figure S14**; see source paper [S36]). Another example is the nitration of Loratadine, a second-generation antihistamine drug. This reaction is not proceeding through the classical electrophilic aromatic substitution mechanism but involves free radicals (example "2" in **Figure S14**; ref [S37]).

(2) Multistep mechanism – that is, reaction proceeds through a more complex mechanism involving intermediate(s) altering regioselectivity of EAS. An example of this class of errors is the nitration of *para*-substituted aniline with guanidinium nitrate. The reaction is reported to proceed in the *meta* position of aniline, although -NH_2 group activates strongly the *ortho* position. This outcome is probably a consequence of protonation of the amine group first (due to highly acidic environment), so the actual reacting partner is *meta*-directing anilinium ion (example "3" in **Figure S14**; ref [S38]). Another good example is Friedel-Crafts reaction of *N*-acetylindole which led to an unexpected substitution at position "6". It was proposed that the complex of the substrate with $AlCl_3$ (presented in the inset to example 4 in **Figure S14**) acts as the directing group [S39].

(3) Reactions controlled by conditions – that is, cases where, depending on the electrophilic reagent used, more than one regioselectively correct product of the substitution is possible. For example, bromination of 4-aminophenol takes place at positions "2" or "3" position, depending on the reaction conditions applied (examples 5a and 5b in **Figure S14**; refs [S40,S41]). In another example, 3-5-dihydroxytoluene undergoes selective bromination at "2" or "4" positions depending on the reagent used (examples 6a and 6b in **Figure S14**; refs [S42,S43]).

(4) Bibliographical errors in the testing set – that is, cases where the reaction reported in Reaxys was actually not found in the original publication/patent. An example of this class of "bibliographical entry" errors is the reaction of methyl anthranilate which, according to the original publication, takes place at position "5". In the Reaxys database, the reported product is substituted at position "6" (example 7 in **Figure S14**; correct record from the source publication is shown in the inset frame; see ref [S44]). Another example is nitration of chlorobenzene reported in Reaxys. According to the source publication, the actual starting material was 4-chlorotoluene. The correct entry reported in the publication is presented in the frame (example 8 in **Figure S14**; ref [S45]).

In summary, the 90% correctness we report now is artificially low given that the test set from Reaxys contained erroneous and/or non-EAS entries as detailed above. The true – and certainly higher – value of correctness of our method evaluated on a set of reactions that are known to proceed via EAS and are correctly input in Reaxys will be given in our upcoming paper on the topic, pending the analysis of "false negatives" such as those we described above.

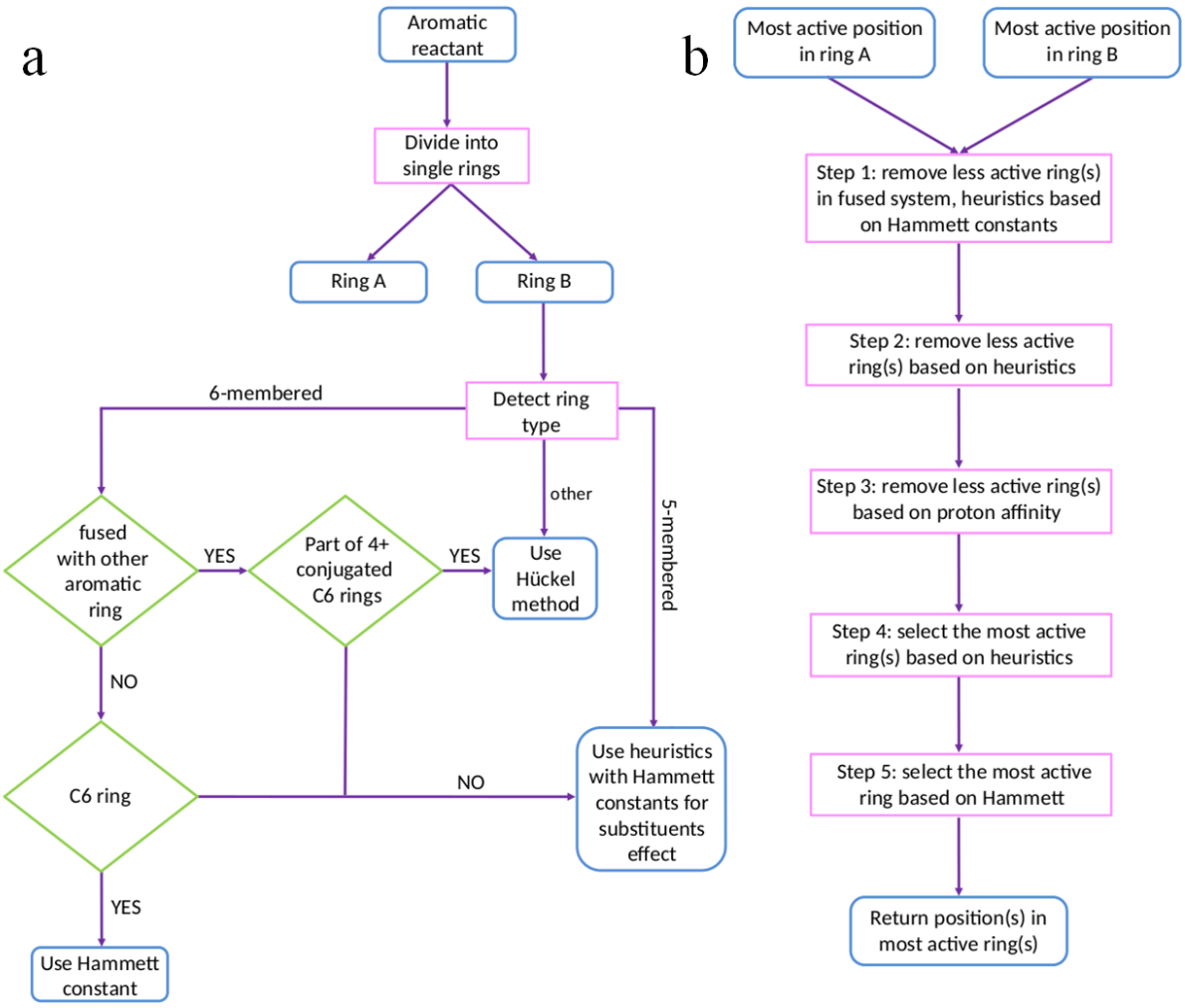


Figure S12. Simplified version of the algorithm detecting the most active position in EAS. The algorithm consists of two parts: **(a)** detection of the most active atom within a given ring. Molecule is divided into single rings. Depending on ring type, the most active position is determined by Hammett substituent constants (isolated benzene ring), Hückel model (polycyclic aromatic hydrocarbons) or heuristic approaches (all other situations). **(b)** Detection of the most active ring in the molecule. This procedure consists of five steps. Three of them are aimed to remove less active ring(s). Initial two steps are based on heuristic rules; the third step removes less active ring(s) based on proton affinity. The fourth step extracts the most active ring using heuristic rules. In rare cases, if more than one ring is retained after Step 4, final decision is based on Hammett constants.

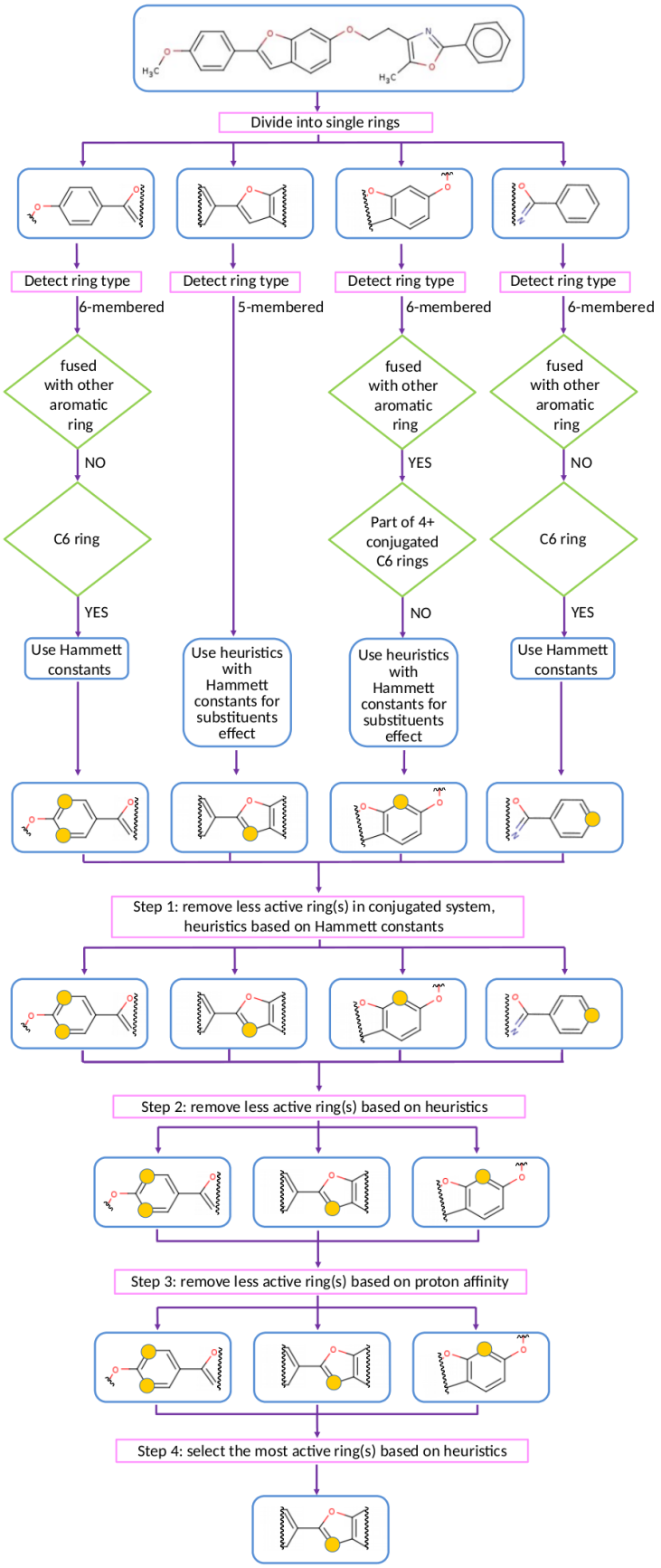


Figure S13. An example of predicting the most active atom for electrophilic aromatic substitution (EAS). Input molecule (taken from[S46]) is divided into separate rings (only rings with at least one

available position are taken into consideration). Depending on the ring type, the most active position in each ring is determined using Hammett-based model (for isolated benzenes) or more elaborate heuristics (here, for both rings of benzofuran). Subsequent removal of less active rings begins with the analysis of the fused system (here, only two rings of benzofuran are considered). In our example, this step does not remove any rings indicating that there is no major difference in rings' reactivities. In the next stage of heuristics-based analysis, monosubstituted benzene linked to oxazole ring is removed. This ring was chosen as it was marked as a significantly less reactive than remaining benzenes. Next, differences in RAPA of the remaining rings are considered (Step 3). In the example shown, these differences are not significant enough to mandate removal of any rings. Finally, the most reactive ring is selected based on heuristic rules (Step 4).

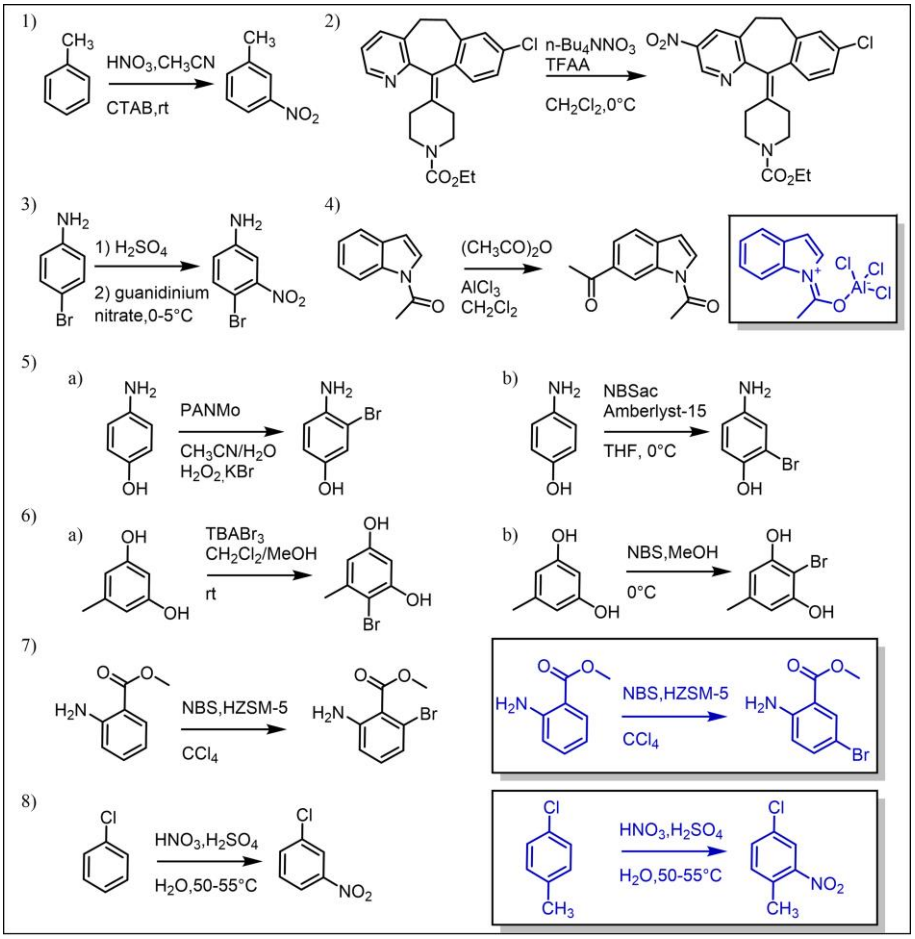


Figure S14. Examples of incorrect predictions of our aromatic filter module. These prediction errors may be due to, for example, reactions proceeding by mechanisms different than EAS, other factors altering regioselectivity (e.g. protonation of a substrate prior to the reaction), simple manual entry errors in Reaxys, etc.

S5.2. Other QM or "conformational" heuristics. The example of EAS in the previous section is just one class of reactions where heuristics can significantly fine-tune the predictions of "raw" reaction transforms. Some other important heuristics *Chematica* uses to deal with, for example, cycloadditions (Cope-type rearrangements, Diels-Alder reactions) are based on a mixed approach combining conditions derived from experimental observations with QM mechanical calculations from which parameters for specific substituents are then obtained. With the details of such calculations left for separate publications, we highlight here one other class of heuristics that deals with the conformations of the molecules *Chematica* constructs during synthetic design. The particular problem is that the reaction records alone cannot ensure that the molecules produced are not chemically unstable or conformationally too strained – for instance, the reaction record for converting a single into a double bond does not automatically "know" that such a bond should not be placed at the bridgehead of a bridged ring system, save in rings that are large enough (the so-called Bredt's rule). To eliminate synthons containing such "nonsensical" motifs from synthetic planning, we curated a library of ~600 such generalized (i.e., annotating many types of atoms with "A" or "a" meaning any atom in a particular position) motifs containing, among others, small-ring allenes, certain cyclopropyne derivatives, compounds breaking Bredt's rule, *trans*-epoxides fused to small rings, geminal triols or α-haloalcohols, and many more (for examples, see **Figure S15**). Some of these motifs are obvious to a trained chemist, for some we performed molecular-mechanics calculations to verify they are indeed strained much more than molecules that might "look" strained but actually can exist under synthetically reasonable conditions (e.g., a 10-membered cyclic alkyne). The list of nonsensical motifs is applied to every synthon *Chematica* creates and if this synthon contains at least one of these motifs, it is eliminated from further consideration. We also note that our list of forbidden intermediates together with the rule-coding philosophy discussed in earlier Sections do not exclude strained motifs participating (but immediately trapped) in some useful transformations. For example, a Diels-Alder reaction of furan with benzyne as a dienophile is coded in Chematica from a stable benzyne precursor (e.g., 1,2-dibromobenzene) and proceeds directly to the cycloadduct. Finally, in **Section S6**, we describe a functionality of the program whereby the user can calculate the strain for each molecule within the pathways Chematica produces. If the user judges the strain to be excessive, he/she may choose to mark this molecule as undesirable in future searches.

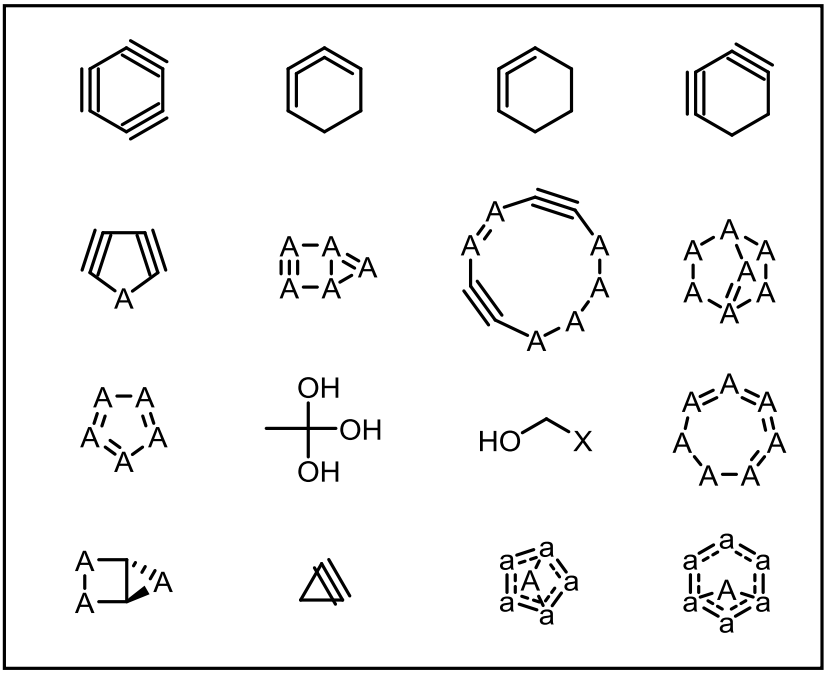


Figure S15. A small subset of some of *Chematica*'s ~600 "nonsensical" motifs. "X" = halogens, "A" = any aliphatic atom, "a" = any aromatic atom.

S5.3. Non-selective reactions. Examples in this section concern situations in which the same reaction rule can be applied at several places of the molecule leading to undesired mixture of products (as opposed to a "clean," single-product outcome). Obviously, such nonselective transformations do not depend on the reaction rule alone but also on the structure of the molecule to which the rule is applied. To detect nonselective reactions during our retrosynthetic searches, we reverse the transformations/rules (which is algorithmically non-trivial if one needs to preserve the exact scope of the reaction rule), apply them to the putative synthons, and inspect how many products are formed. If the number of products is greater than one, the tranformation is marked as non-selective and assigned a penalty durign synthetic planning. Some examples of how *Chematica* deals with non-selectivities are provided in **Figure S16**.

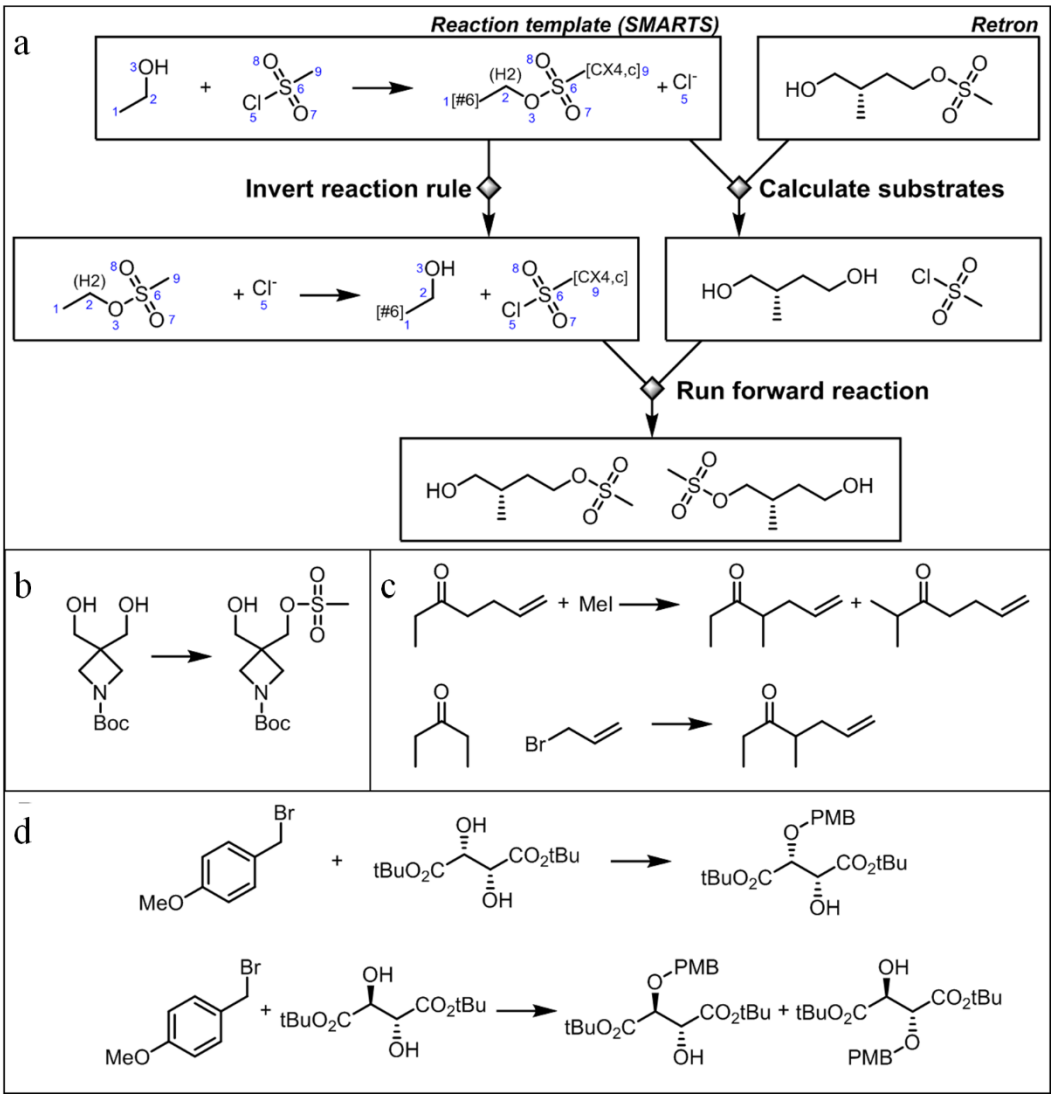


Figure S16. Evaluation of reaction selectivity in *Chematica*. **(a)** After applying the transform rule to a given retron the set of necessary synthons/substrates is generated (top-right). To evaluate selectivity, an "inverted SMARTS" rule is created and applied to the collection of substrates to generate possible products. In the example shown, two different products are obtained and the reaction is marked as non-selective. (**b**) If there are multiple but equivalent reaction sites, the algorithms does not mark them as non-selective since in most cases such transformations can be performed cleanly by adjusting molar ratios. **(C)** Two different strategies leading to an intermediate in the synthesis of sordidin, a pheromone of main banana plant pest[S47]. Methylation of unsymmetrical ketone suffers from the formation of mixture of products while allylation of 3-pentanone gives the desired compound in nearly quantitative yield. Using non-selectivity algorithm, *Chematica* can penalize the first of these reactions. **(d)** Since *Chematica*'s deals with stereochemistry of reactions (cf. **Section S3.1**), the

approached based on inverting the transforms can also detect non-selectivities originating from stereochemical effects. In the example shown, benzylation of *D*-tartrate derivative affords a unique product while attempting the same reaction for *meso*-tartrate will yield a mixture of products.

S5.4. User voting. Our final example in this Section deals with a somewhat "exploratory" modality implemented in *Chematica* but still awaiting validation of its usefulness. In brief, every synthetic option proposed by *Chematica* allows the user to "vote" on each of the individual reactions – the user can either "like" it (by clicking on the "thumb up" icon in **Figure S17**) or "dislike" it ("thumb down"). Clicking on these icons opens sub-panels in which the user can specify the reasons for his/her vote (for "likes" – "elegance" or robustness; for "dislikes" – a possible steric or strain problem, non-selectivity or reactivity conflict). This information is then sent to *Chematica*'s main server with a confirmation email also forwarded to the user. The "liked" reactions can then receive more favorable scores during searches, while the "disliked" ones are penalized. On one hand, this "chemical Facebook" can be very useful in harnessing the collective chemical knowledge of *Chematica*'s users, especially (i) to obtain information about reactions that did not work and were never published (but could be very helpful in training of our statistical models), and (ii) to fine-tune *Chematica* to the needs/practices of specific organizations using the program (e.g., in some companies, certain reactions are not possible due to the lack of infrastructure, while certain others are preferred). On the other hand, we are aware that this method can be unreliable if the information is provided by non-experts or with ill intent (*vide* the failure of Microsoft's "intelligent" bot trained by a group of users to praise one infamous dictator http://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/). As mentioned, we are in the process of evaluating this scheme and will report on the outcome when the statistically significant amount of user feedback is collected.

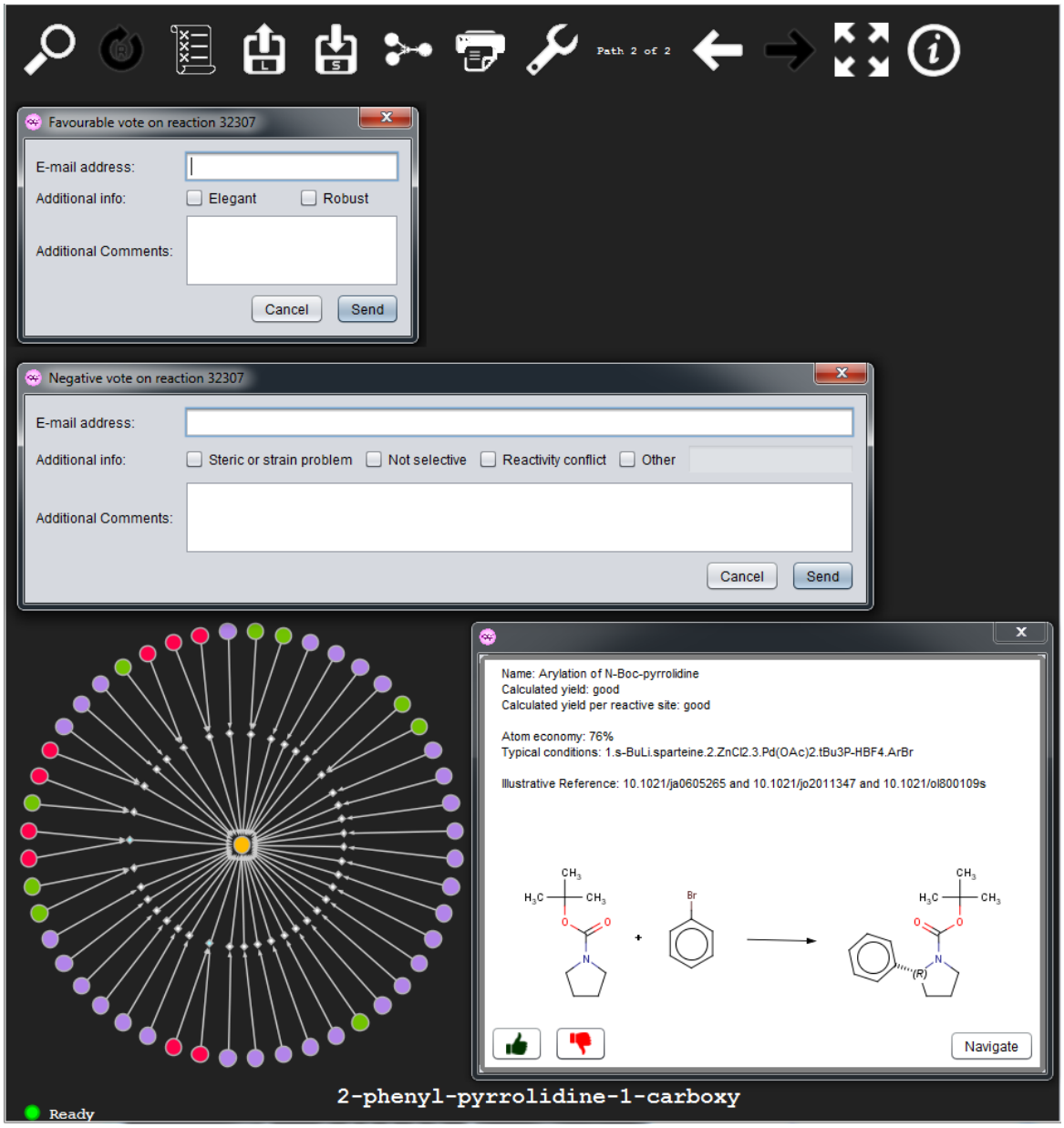


Figure S17. User voting in *Chematica*. For any transformation, the user can either "like" (green thumb-up icon in the lower-right reaction subwindow) or "dislike" (red thumb-down icon) this reaction and provide his/her specific comments (two subwindows shown at the top) which are then sent to *Chematica*'s main server.

Section S6. Searching for complete pathways.

Having the rules for individual synthetic moves is a crucial but only a preliminary step in teaching the machine the design of complete synthetic pathways. We have previously estimated[10] that with *Chematica*'s tens of thousands of rules, there are on the order of 100 possible synthetic "moves" the machine needs to consider at each synthetic step. In other words, there are, on average, ~100 reactions producing the target of interest from its immediate synthons, then there are ~100 reactions producing each of these synthons, and so on. Within *n* synthetic steps, there are ~100^n possible routes leading to the desired target – even for relatively short syntheses, such numbers of possibilities are way too large to explore in an exhaustive fashion. This problem was recognized already several decades ago and called by E.J. Corey "a combinatorial explosion of synthetic choices". The only way to avoid this complication is to teach the machine to search the space of synthetic possibilities in an intelligent manner, not venturing into or reverting from unpromising branches of synthetic options, and channeling the searches towards the most efficient, elegant sequences of steps.

Attacking this problem requires some in-depth algorithmic considerations we highlight in this section. First, we need to represent the "tree" of possible syntheses in the most appropriate graph representation. As we will see, *Chematica* uses two such representations (graphs and "hypergraphs"). The former are the so-called bipartite graphs (or Petri nets) in which the substances and reactions are represented by different types of nodes, preserving all casual relationships between the retrons and synthons (cf. **Figures S18** and **S19**). In the hypergraph representation (**Figure S20**), all synthons of a given reaction are grouped into one "supernode" – this is quite essential since when the machine evaluates a particular reaction, it has to evaluate all synthons and not only the heaviest one or the most "complex" one (e.g., in cases in which the retron is disconnected into two synthons which are neither commercially available nor known in literature, the complexity of *both* of these synthons must be taken into account and the syntheses producing *both* of them must be further planned).

Evaluation of synthon sets brings us to the all-important question of scoring the "synthetic positions" encountered during planning. In the game of chess, at a particular position (i.e., arrangement of pieces on the board), the computer considers only the "future" moves and does not have to keep track of how this position had been reached. In synthesis, we defined[10] "synthetic positions" as comprising both the current set of synthons as well as the set of reactions via which these synthons were obtained. Obviously, if we reach the same synthons by two reactions from the target vs ten reactions, the shorter, two-step solution should usually (though not always) have a better/more favorable score. Hence, we perform the scoring by two types of functions – the so-called Chemical Scoring Function, (CSF) to evaluate the synthons, and the Reaction Scoring Function (RSF) evaluating the "history" of reactions by which these synthons were reached. We will discuss here how these evaluations are performed – based on CSFs and RSFs defined by *Chematica*'s chemically-meaningful variables and augmented by various heuristics spanning **sequences** of steps – to enable "intelligent" walks over the enormous synthetic graphs, and ultimately identifying and ranking the best-scoring solutions.

S6.1. Synthesis graphs.

Let us define the *synthesis graph* G = (V, E) that will serve as a mathematical model for the search of synthetic pathways. The set of vertices will be the set of all possible chemical substances (identified with their canonical SMILES formula), along with the set of all possible reactions, that is:

V = { SMILES(x) for all chemical substances x } ⊔ { reaction nodes}

where ⊔ denotes the disjoint union of the two sets. The sets naturally divide the vertices into two classes, which will be referred to as *chemical nodes* and *reaction nodes*. The set of edges, E, consists of connections between reactions, their substrates, and products. For example, let us consider a part of the graph corresponding to the following reaction (in SMILES notation):

CCO + CC(=O)O → CC(=O)OCC + O

that is, synthesis of ethyl acetate from ethanol and acetic acid via simple esterification reaction. As illustrated in **Figure S18a**, the graph describing this reaction (and omitting the side-product water molecule, denoted O in the SMILES notation) can be represented in the so-called bipartite form comprising two types of nodes, one type for substances and one type for reactions. For our

esterification reaction, the graph comprises three *chemical nodes* (CCO, CC(=O)O, CC(=O)OCC), one *reaction node* (denoted as r_1), and three directed edges:

- CCO → r_1
- CC(=O)O → r_1
- r_1 → CC(=O)OCC

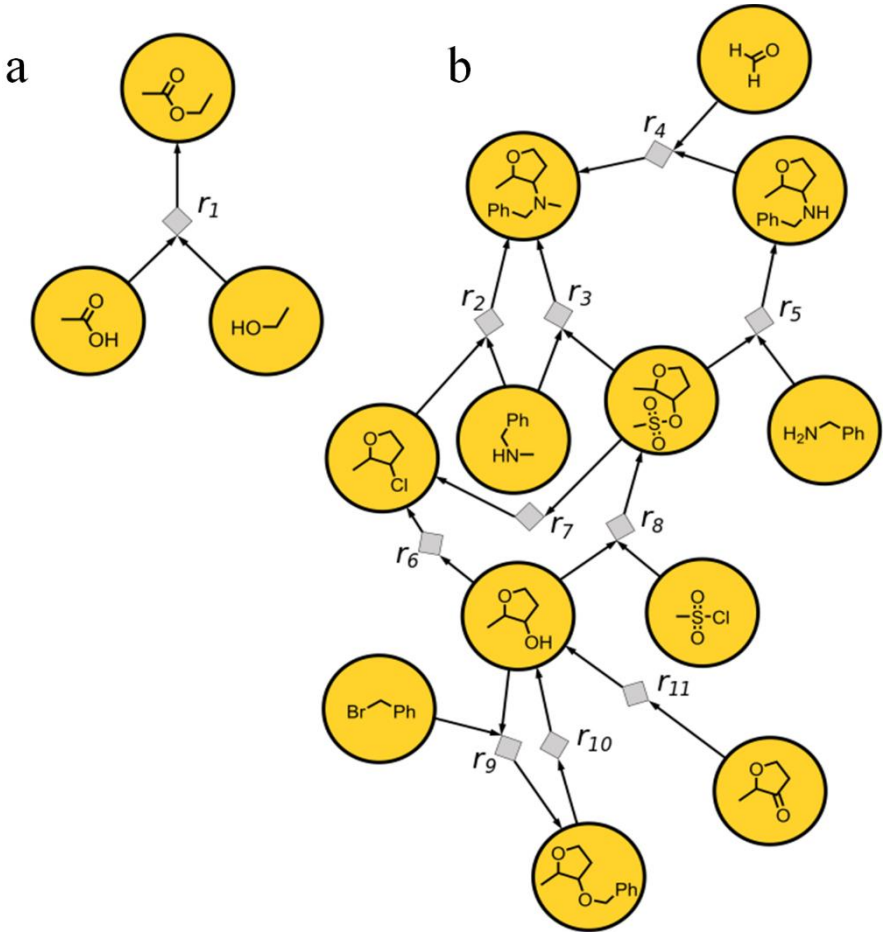


Figure S18. Examples of synthesis graphs. **(a)** In the so-called bipartite representation with two types of nodes (larger circular nodes for substances; smaller diamond nodes for reaction operations), acetic acid and ethanol substrates enter into an esterification reaction (corresponding to node denoted r_1) to produce ethyl acetate. If the bipartite representation were not used, one would have to connect each of the substrates directly to the product, which would make no chemical sense (i.e., ethyl acetate cannot be made from only ethanol or only acetic acid). (**b**) A larger bipartite graph describing several methods of making a tetrahydrofuran derivative (substance node below letter “B”).

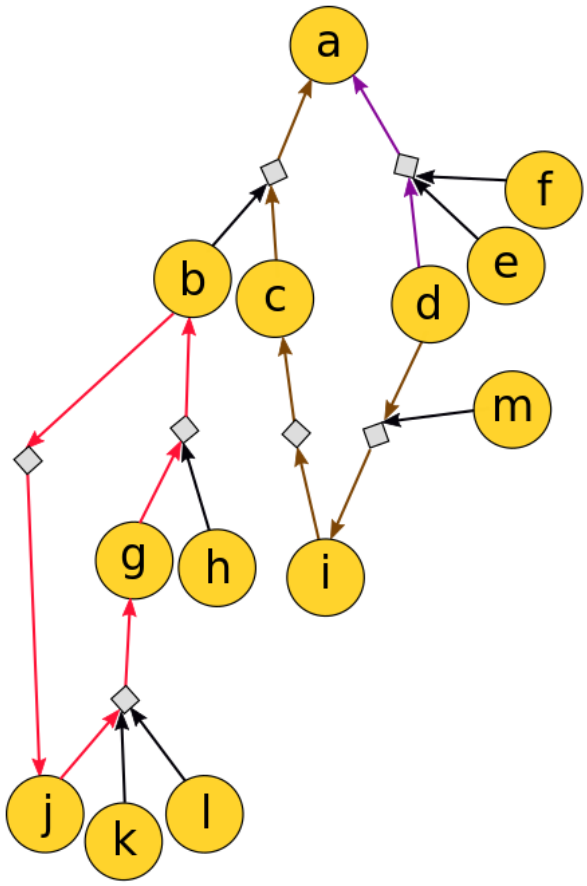


Figure S19. Topology of synthesis graphs can be complex. In this still unrealistically simple example, molecule “d,” is part of two synthetic pathways leading to target “a” – it can be directly transformed into the target “a” (violet arrows), or via sequences of reactions leading to molecules “i”, “c”, and “a” respectively (brown arrows). This example is provided to illustrate that a concept of synthetic distance (in terms of the number of reactions) from the target might not be uniquely defined as the same molecule might be at different distances depending on which synthetic plan it belongs to. Furthermore, note that the network shown is not a DAG (“direct acyclic graph”) as it contains a directed cycle involving molecules “b”, “g”, and “j” (red arrows). Such cycles are synthetically spurious and need to be avoided during searches for synthetic pathways.

As additional options for synthesis are explored, the synthesis graphs grow in size. In principle, these graphs are infinite though in practice only their finite portions are explored and kept in computer memory (in *Chematica*, up to several millions of nodes). We shall refer to the part of the graph that is explored as the *uncovered* portion of the chemical synthesis graph.

Next, we note that the graphs may contain cycles (see **Figure S19**) which are unproductive synthetic solutions, create infinite loops in the searching procedure, and need to be eliminated by any search algorithm.

Some chemical nodes in the graph are known as *terminal nodes*. In *Chematica*, these can be either commercially available chemicals or molecules whose syntheses have been already described in literature (and can be found in *Chematica*'s Network of Organic Chemistry module[13-16,S48] by traditional network-search algorithms we described in refs [10,16]). The user of *Chematica* can specify either the "buyable" or "known" substances as the terminal nodes and can further specify their attributes such as molecular weight ("continue synthetic searches until terminal nodes with MW below certain threshold are found") or price ("stop only if the prices per gram of buyable substrates are below a specified threshold"). Substances which are neither commercially available nor literature reported, and have no feasible incoming reaction pathways are called *impossible*.

With the enormous complexity of realistic graphs constructed during synthetic planning (tens of thousands to millions of vertices) and with the need to score synthetic positions not as individual substances but sets of substances at each step of synthesis (cf. the introduction to **Section S6**), we have used the synthesis graphs only for the chemically intuitive display of results but based the synthetic searches on the "hypergraphs" we describe in the next section.

S6.2. Synthesis hypergraphs.

A concept related to that of the synthesis graph is that of a synthesis hypergraph of a single substance T. It is induced by the synthesis graph, is (usually) also infinite, and is defined as the smallest graph such that:

- The singleton set {T} is a node in this graph (known as the *root node*)
- For any node N = {T_1, T_2, …, T_n} if there exists a reaction:

$S_1 + S_2 + \ldots + S_k \rightarrow T_1 + P_1 + P_2 + \ldots + P_m$

(where S_i are substrates, T_1 is the main product, and P_i are side products, side products are not considered during search) then there exists a node M = {S_1, S_2, …, S_k, T_2, …, T_n} and there is an edge between node N and M (**Figure S19**).

Please note that the **nodes in the hypergraph are sets of substances**. They are, however, not multisets – only the occurrence of a particular substance is recorded, and not its stoichiometric abundance in the reaction used.

The semantics of the graph are as follows: nodes correspond to stages of synthesis, and edges correspond to reactions: they transform one set of substances into a set that may be obtained from them using a one-step synthesis.

A hypernode is called "terminal" if all substances occurring in it are commercially available or have been synthesized before (requirements for these "stop points" are defined by the user as in the simple synthesis graphs). A hypernode is called "impossible" if at least one of the substances is impossible to synthesize.

The problem of chemical retrosynthesis therefore reduces to the *optimal path problem* widely studied in computational graph theory and here aimed at finding an "optimal" (least costly, least risky, etc.) synthetic route leading from the singleton set {T} to any terminal node.

Since the hypergraph is theoretically infinite and impossible to explore fully, only its part is ever evaluated. As such, another state of nodes in the graph is needed: a node is "known" if all substances occurring in it are either terminal (according to user-specified stop criteria) or have a computed synthetic pathway, and is "unknown" otherwise. The search has found a viable synthesis when the

root node changes status to “known” (though it may be continued in order to find additional, better syntheses).

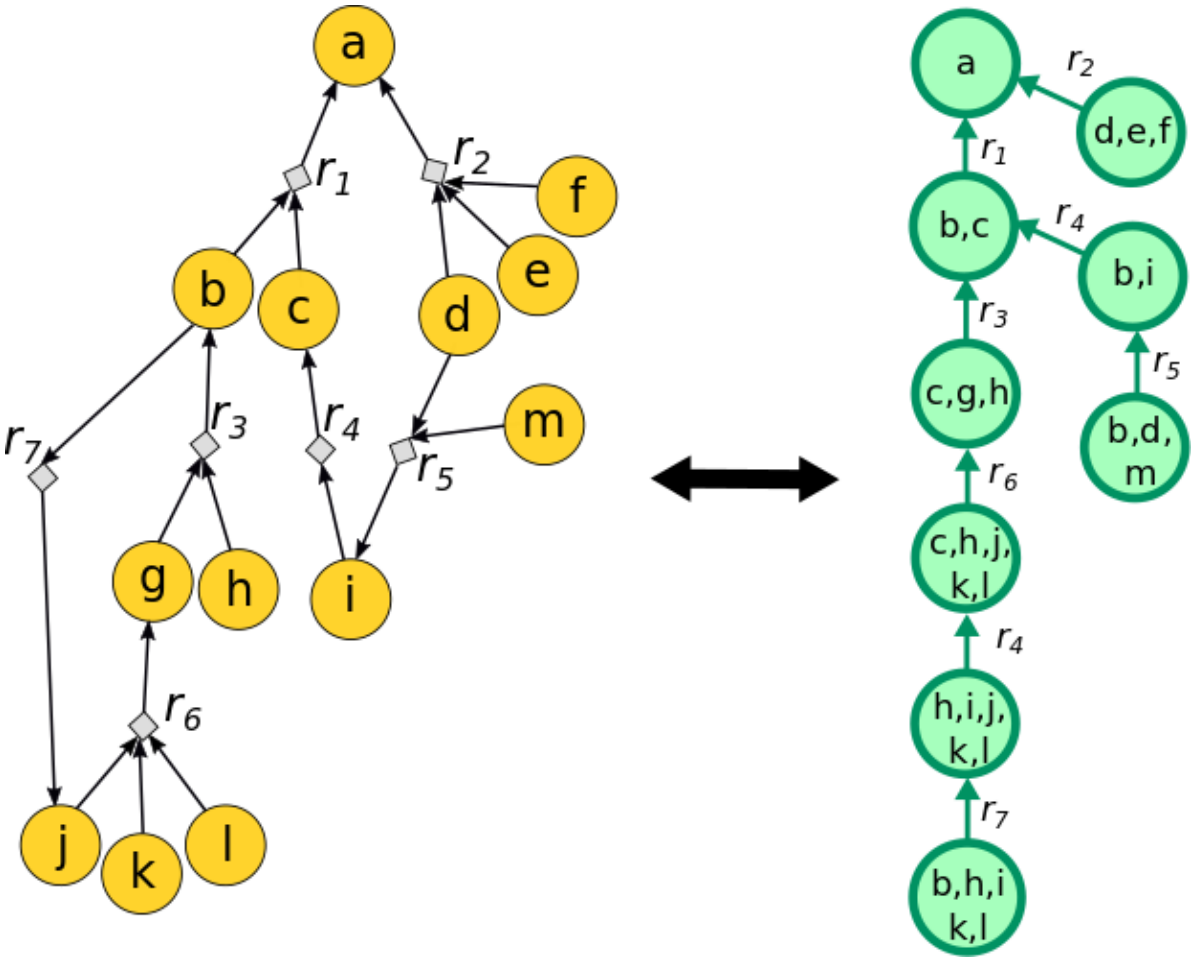


Figure S20. A schematic synthesis graph (*left*) and (a part of) the corresponding synthesis hypergraph (*right*). The labels near hypergraph edges indicate the corresponding reaction nodes in the synthesis graph. Note that more than one hypergraph edge might be related to a single reaction from the synthesis graph (as is the case here for r_4).

S6.3. Search algorithm.

The size and the overall structure of the synthesis hypergraph preclude the use of direct search for the shortest path in a BFS-like fashion for all but the very simplest of molecules. In addition, the shortest (in terms of the number of traversed edges/synthetic steps) path will not always be the most cost-effective as it may use expensive substrates, difficult reaction steps, large number of protections, etc. To quantify the preference of certain synthesis paths versus others, and to guide the algorithm toward the optimal solutions, it is therefore necessary to introduce two scoring functions.

(i) The Reaction Scoring Function, RSF, is calculated for each reaction node (that is, for every instance of reaction in the graph, taking into account a particular realization of the retrosynthetic reaction rule for a given product and substrates) and quantifies the “cost” or difficulty of performing a reaction. The function is defined based on the “chemical variables” implemented in *Chematica*. For instance, variable PROTECT assigns certain penalty (additional cost) for every reaction that requires protection (cf. **Section S4**), variable NON_SELECTIVITY penalizes reactions that can be performed non-selectively at various places of the same molecule (reducing the yield of reaction performed at a desired locus), variable CONFLICT assigns penalty (typically very large) for every reactivity conflict detected, variable FILTERS penalizes, for instance, unlikely successions of steps (see discussion

later, in **Section S7**). There are also additional variables that can promote or penalize the usage of specific types of reactions or specific molecules (e.g., using HIDE_SEEK_NAME variable with argument “aldol” would prevent the usage of any aldol-type reactions). The functions are generally linear combinations of these variables and the user can use either predefined functions or can define (**Figure S21**) his/her own ones to reflect a specific “synthetic style” – for instance, if the user wishes to construct pathways completely free of protections, he/she would assign the variable PROTECT a prohibitively high cost such that these routes will never be chosen during the search. Algorithmically, the RSF mirrors the edge cost function of “classical” graph search algorithms such as A* [S49] or Dijkstra algorithm[S50].

(i) The Chemical Scoring Function, CSF, is assigned to chemical substances and does not depend on the path from the target to this chemical. It mirrors the heuristic function of the A*-type algorithms and serves to guide the search toward less complex molecules, thus avoiding exponential branching and searching parts of the solution space that are unlikely to yield sensible synthesis pathways. Ideally, CSF should express the precise “cost” of the substance as synthesized by the cheapest possible pathway. However, such cost is not known a priori and instead CSF estimates the complexity of the synthons. In this spirit, the user can use various chemical variables promoting cuts into smaller synthons (variable SMALLER with an argument defining the relative sizes of the desired synthons; note: in ref [10], this variable was called by a less intuitive name SMILES_LEN), lowering the numbers of rings (RINGS) or stereocenters (STEREO), and few more.

Using these scoring functions, the search algorithm automatically finds a set of viable pathways while exploring the synthesis hypergraph (see above). More precisely, the algorithm is divided into two parallelized subtasks illustrated in **Figure S22**:

a) graph explorer that is “intelligently” exploring the synthesis graph and its induced synthesis hypergraph;
b) path retriever that selects diverse set of viable pathways based on the current ‘snapshot’ of the synthesis graph (expanded by explorer)

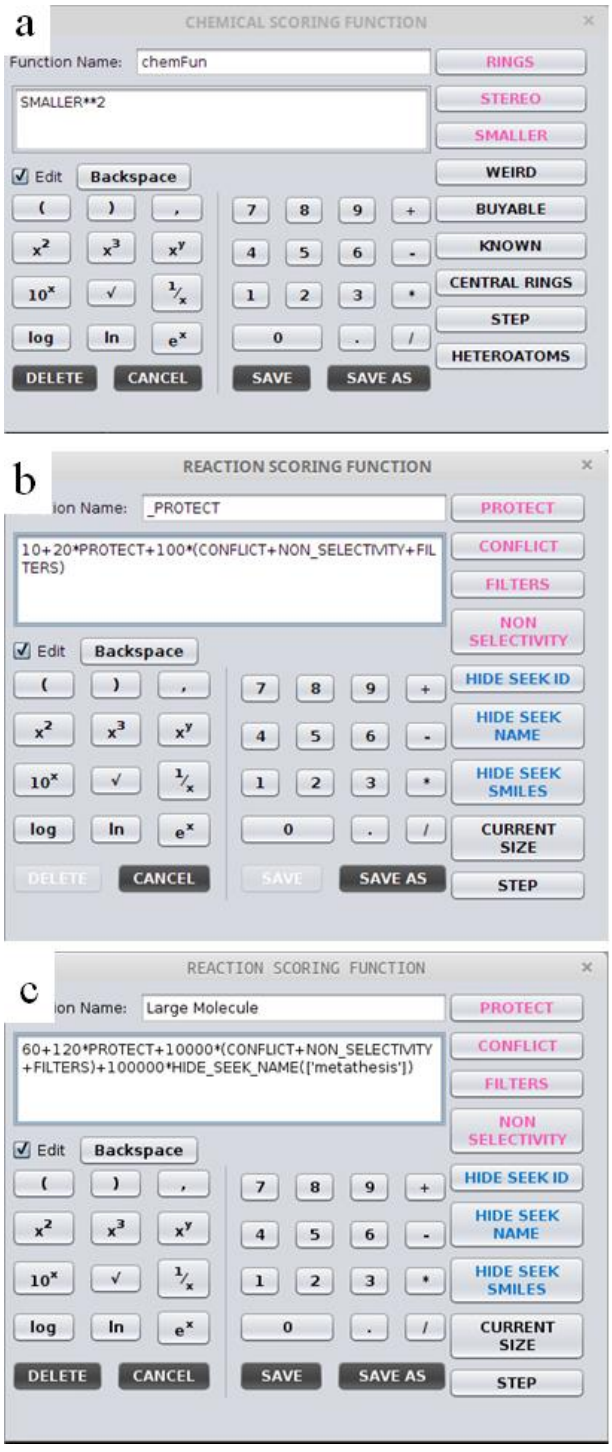


Figure S21. Screenshots of Chematica's "calculators" which allow the user to build (**a**) the CSF and (**b**) the RSF from the chemical variables in the right column. The most often used variables are in pink font; the sometimes-used in blue, and the rarely used/specialized variables in black. The rarely-used does not mean useless – for instance, variable CENTRAL_RINGS in the CSF uses the so-called largest bi-connected component[S51] to promote cuts in the central vs. peripheral rings, which is useful in the synthesis of complex polycyclic targets. (**c**) The image illustrates the syntax of the HIDE_SEEK variable – here, to penalize the use of metathesis reaction.

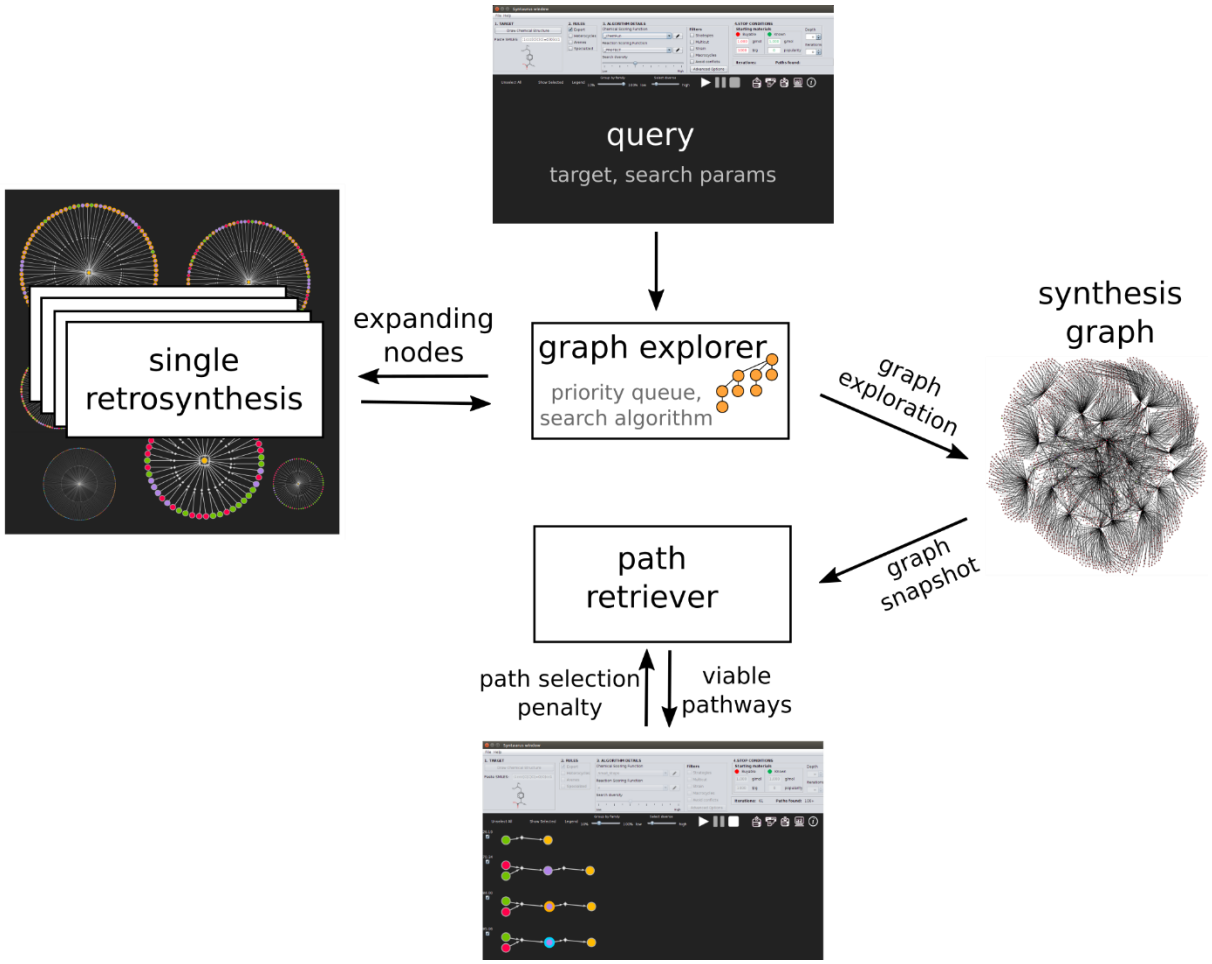


Figure S22. Overview of *Chematica*'s automatic retrosynthesis module. After the user specifies the target molecule and sets other search parameters (stop points, scoring functions), the **query** is sent to the **graph explorer**, which, based on binomial priority queue is iteratively exploring the **synthesis graph**. To do so, graph explorer asynchronously queries multi-processing service for **single retrosynthesis** analyses. Simultaneously, **path retriever** is responsible for extracting a diverse set of viable synthetic pathways based on the current state ("snapshot") of the synthesis graph. The user can specify on the flight (while search is performed) parameters that adjust the desired diversity of the pathways selected (i.e., it is generally desired to obtain many different synthetic routes rather than multiple variations of the same pathway – at the same time, requiring increased diversity can lower the "quality" of pathways as suboptimal solutions are chosen with higher likelihood).

The **exploration part** of the algorithm is similar to other shortest path search algorithms, in particular the A*. The algorithm simultaneously operates on two levels, on the synthesis graph, and the related synthesis hypergraph. The algorithm generates a sequence of nodes to be “expanded” (i.e., chemical substances for which the set of all possible one-step syntheses is to be computed). Such expansions are iteratively added to the synthesis graph, and new choices are made based on the newly-revealed graph. By keeping the priority queue, the algorithm is able to rapidly revert from unpromising synthetic “branches” into better alternatives. In addition, the algorithm assigns penalties to the regions of the hypergraph that were already explored – this allows it not to be “trapped” into one region of the solution space. These operations are illustrated in **Figure S23**.

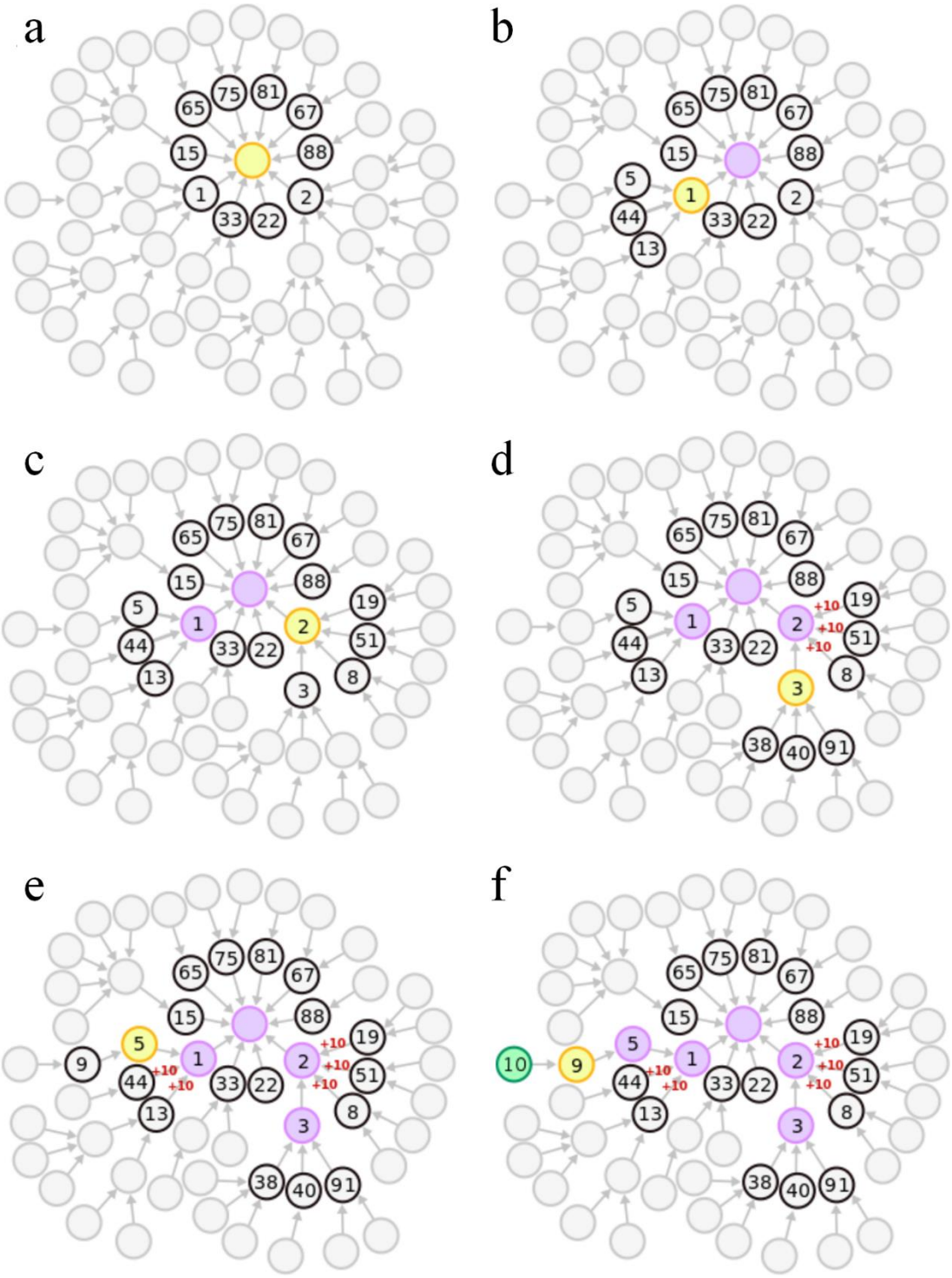


Figure S23. Scheme illustrating exploration of the synthesis hypergraph. (**a**) At the very beginning, the "currently explored" hypergraph consists of only the target (node colored in yellow) and is expanded into the first-generation nodes. All incoming nodes (unless they are *terminal* or *impossible*), are inserted into the priority queue, PQ (collection of nodes with black borders) with priority equal to the sum of RSF and CSF functions. Here, these values are 1, 2, 15, 22, 33, 65, 67, 75, 81, 88. (**b**) The previously analysed/expanded nodes are marked violet, and the currently lowest-scoring node (1) is expanded and removed from the PQ. The PQ now contains nodes with scores 2, 5, 13, 15, 22, 33, 44, 65, 67, 75, 81, 88. **(c)** Node 1 is marked as expanded. The currently most promising node in the PQ has score 2 – this node is analyzed adding nodes with scores 3, 8, 19, 51 to the PQ. Node 2 is removed from the PQ. (**d**) The best available option now has score 3. Its neighbors (nodes 38, 40, 91) are added to the PQ; 3 itself is removed from the PQ which now contains nodes 5, 8, 13, 15, 22, 33, 38, 40, 44, 65, 67, 75, 81, 88, 91). At this stage, we assign penalties (here, +10, colored red) to nodes 8, 19, and 51 which were already placed in the PQ while visiting node 3. Namely, we penalize alternative pathways having the same "exploration history" as the currently analyzed node (in our case the common "exploration history" of nodes 3, 8, 19, and 51 is related to the content of node 2 which, as we remember, is a hypernode). This penalization is done to limit the number of explorations in the same regions of the hypergraph. (**e**) We set node 5 (minimal priority from PQ) as currently analyzed, remove it from PQ, add to the PQ node 9, and penalize nodes 13 and 44 (having the same 'exploration history' as node 5) (**f**) We analyze node 9 (remember, node 8 was already penalized by +10, resulting in its increased total score of 18), and remove it from the PQ. The only not expanded neighbour of 9 turns out to be a terminal node (say, a set of commercially available substrates) – this stop point completes the first synthetic pathway traversing nodes 1,5,9,10. This pathway is kept for further consideration by path retriever algorithm and the analysis continues further, to find additional and possibly better solutions.

At the beginning of the search, a single-node synthesis graph G = ({t}, {}) – comprising only target t – is initialized. A priority queue PQ is created and initialized with the single set {t}, with priority 0.

Then, the main loop of the algorithm proceeds as follows:

- First, extract a set S of chemicals with lowest priority from the PQ, store the priority in variable q.
- If the set does not contain a node that has not yet been queued for expansion (or contains an impossible node) drop it and proceed to next iteration.
- Drop all nodes which have already been queued for expansion from S.
- Choose one as yet unexpanded node s from the set S, and expand it.
- For every reaction r: $s_1 + s_2 + \ldots + s_n \rightarrow s$ create a set $S' = S \setminus \{s\} \cup \{s_1, s_2, \ldots, s_n\}$ and insert it into Q with priority $q - CSF(s) + RSF(r) + CSF(s_1) + CSF(s_2) + \ldots + CSF(s_n)$.

Each node-expansion step ("**single retrosynthesis**") is computationally intensive, and the algorithm is parallelized and selects several substances to be expanded simultaneously. **Single retrosynthesis** steps are provided by a multi-processing service. There are several main processes, running independently, responsible for receiving queries through an asynchronous communication channel, and returning the results when ready. Processing of the queries is performed with user specified search parameters (scoring functions, stop points, etc,) and involves the following subtasks:

a) Queried SMILES is matched against SMARTS of all available reaction rules (this check is relatively fast and allows for early filtering of the majority of non-matching reactions);
b) A proprietary library designed to "perform" *in silico* reactions accurately is applied to deal with the issues of stereo- and regioselectivity (see **Section S3.1**). This step is computationally significantly more demanding than step (a);
c) Various heuristics/filters (**Sections S5**) are applied to remove "false-positive" reactions;
d) Additional information about protections and incompatibilities (**Sections S4**) is added to the output.

We note that while the implemented search algorithm is similar to A*, there is no need to explicitly store the hypergraph, only the (much smaller) graph. A synthesis pathway (i.e. a hyperpath through the hypergraph) is still implicitly retrievable from the synthesis graph.

As the exploration of the graph G (and the induced hypergraph) is guided by the constantly updated sums of CSFs and RSFs, the priority queue, PQ, stores the hypergraph's vertices (i.e., sets of currently available substances that further need to be synthesized; for optimality reasons, we skip here the terminal nodes to reduce the size of Q). We note that PQ is implemented as a binomial heap[S52] which is a data structure offering high efficiency of insert operations (amortized time of $O(1)$). The cycles are avoided without the need to store a set of visited nodes (unlike in similar algorithms) through the concept of expanded vs. unexpanded nodes.

While performing the search, we maintain additional information referring to the state of the nodes within the graph. In particular, for chemical nodes we define the following statuses:

1. exploration status (as the exploration algorithm communicates asynchronously with single retrosynthetic service [i.e., expansion of individual nodes], this status allows us not to query the same molecule multiple times):
    a. EXPLORED: node already successfully queried for retrosynthetic steps;
    b. EXPLORATION-IN-PROGRESS: node queried for retrosynthetic steps but results not yet recovered;
    c. UNEXPLORED: node not yet queried for retrosynthetic step

2. synthesizability status (can be propagated "upwards" while status of any node has changed, based on its definition)
    a. COMPUTED: this node already has (at least one) computed synthesis, i.e., it is a terminal node or there exists a reaction r from the set of substrates S and producing it, such that all s from S are COMPUTED;
    b. NON-SYNTHESIZABLE: this node has no incoming syntheses or for any reaction r from the set of substrates S and producing it, all s from S are NON-SYNTHESIZABLE;
    c. NOT-COMPUTED: otherwise.

    In particular, the upkeeping of these statuses for the target enables us to detect when the first synthesis is reported (this is achieved when target becomes COMPUTED).

3. actual synthesis cost, defined as the best yet-explored cost of synthesis for a given chemical

When a new computed chemical is found, the costs of all derivatives of this chemical (that is, the substances synthesizable from it) are updated, propagating the information upward through the graph as necessary.

For terminal nodes, the CSF of the chemical is replaced with its actual cost. As actual synthetic costs of nodes become known as computation progresses, these are propagated "upwards" through the graph, gradually replacing the CSFs (which, during searches, serve as estimates of the real cost).

As the algorithm identifies new pathways, it stores them in a network format as illustrated in the main-text **Figure 1b**. When large numbers of pathways are viable (and sometimes these numbers are in millions), the network storing them becomes quite large and one faces an additional challenge of which of these viable pathways are to be shown to the user. For instance, there might be many nearly-top-scoring pathways which, however, differ only in individual steps (typically, the trivial steps near the stop points). Obviously, not all such variations on the same theme should be shown to the user who would likely prefer to see as many as possible chemically diverse routes. To deal with this issue, *Chematica* uses the **path retrieving algorithm** which entails several iterations of (i) cost propagation from terminal nodes to the target; and (ii) generation of the next-best-scoring path solutions (starting from the optimal one). The key element of the algorithm are the penalties (of specific values determined by the user) for reactions that are present in the already-retrieved routes – these penalties help eliminate similar pathways in which the same reactions are being reused (**Figure S24**).

More precisely, the path-retrieving algorithm is operating on G' being a subgraph of G induced by COMPUTED chemical nodes (NON-COMPUTED and IMPOSSIBLE nodes cannot, by definition, become members of viable pathways) or related reaction nodes. The algorithm considers graph G' as fully expanded with each node having a well-defined cost:

a) For terminal nodes, the cost of commercially available substances corresponds to their catalog (MilliporeSigma) prices in dollars per gram. For other stop points (substances with known syntheses), the cost is proportional to the "synthetic popularity" of such a known substance (i.e., in how many ways this substance has been made before[13,14,S48]). Interestingly, as we showed before [10,13], the monetary cost and synthetic popularity are correlated, which allows us to convert the latter into "real" dollars.

b) The cost of reaction node r with substrates s_1 , …, s_n is calculated as RSF(r) + cost(s_1) + … + cost(s_n)

c) For chemical nodes that are not terminal, their cost is calculated as minimal cost for all incoming reactions $r_{1, , ..., } r_n$.

Therefore, the costs of syntheses leading to the target can be calculated by bottom-up cost propagation – the optimal pathway will correspond to the lowest cost. Ideally, we would like to do this calculation in an inverse topological order, starting from terminal nodes. However, recall, that G (and also G') in general is not a DAG (as can have directed cycles), and therefore cannot be topologically ordered. Fortunately, this complication can be remedied by considering a graph of strongly connected components, GSCC (a strongly connected component is a subgraph in which there exists a directed path between any of its two vertices). In the GSCC, the nodes themselves are SCC graphs, and edges are said to link two SCCs (say, A and B) if there exist an edge from any node in A to any node in B. In particular, terminal nodes in our solution graph G' are one-member SCCs, as they are not reachable from any other node (**Figure S25**). On such a graph, we can readily propagate costs starting from SCCs corresponding to terminal nodes. When a given node of GSCC is composed of a single chemical/reaction node from G', the cost propagation is straightforward (directly from definition). For larger nodes (SCC composed of more than one node from G'), we are still able to compute cost in finite time (since SCCs considered are finite and limited by [and typically much smaller than] the size of G', and costs are guaranteed to be positive, as necessary for the shortest-path algorithm).

After successful update of the cost for the target node, we are able to retrieve the first (and best) path that is a realization of this cost, by applying A*-like search (as a heuristic function we use actual synthesis costs of chemicals as calculated while updating costs, which guarantees that the algorithm will immediately be steered to the optimal solution). After finding the best solution, penalties are applied to the already-used reaction-substrate pairs, and the costs of nodes in the graph are recalculated. Such a recalculation is achieved by marking modified nodes as "dirty" and performing the so-called relaxations of the costs of dirty nodes (as necessary), and possibly marking further nodes as "dirty" in the process. The relaxations are performed in a manner similar to the relaxations employed by the Bellman-Ford algorithm[S53] or its well-known variant called Shortest Path Faster Algorithm (SPFA). Subsequent pathways are produced through the successive use of the A*-like algorithm on the recomputed graph.

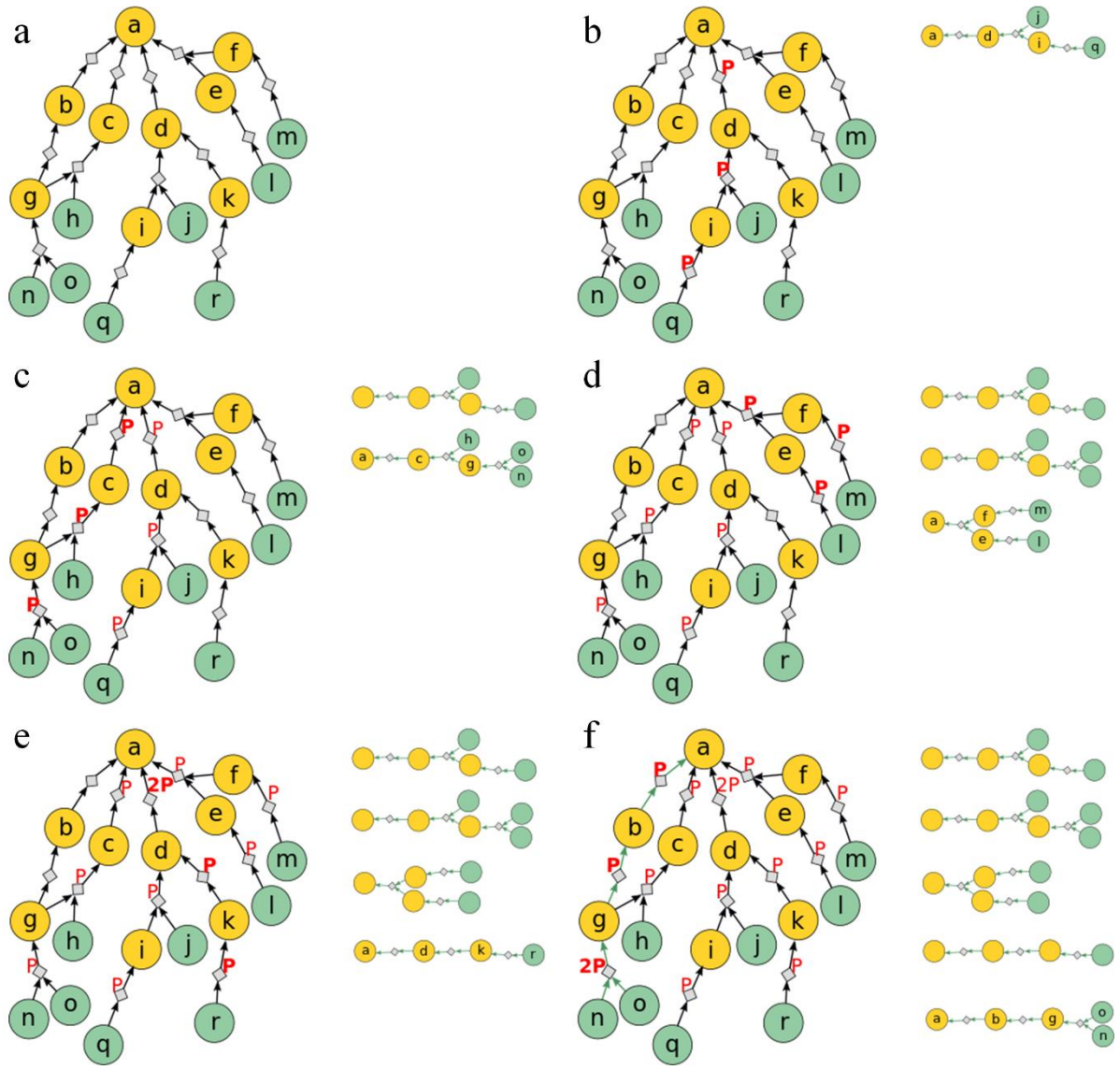


Figure S24. The figure illustrates general idea of the path-retrieving algorithm. (**a**) The algorithm operates on the subgraph of the already-expanded synthesis graph induced by COMPUTED chemical nodes (which belong to at least one viable pathway). Terminal nodes are colored green. (**b**) The algorithm finds the best-scoring path (nodes a, d, i, j, q; shown as a miniature on the right), and assigns penalties P (red) to the reactions used within this pathway. **(c,d)** In this way, the already-used, penalized reactions are less likely to be used in other paths retrieved and the algorithm is more likely to produce chemically diverse routes. (**e**) Penalties over the edges of the graph retraced several times (here, reaction between nodes a and d) are cumulated (now, 2P). (**f**) Another case of cumulated penalties (reaction between nodes g, o, and n).

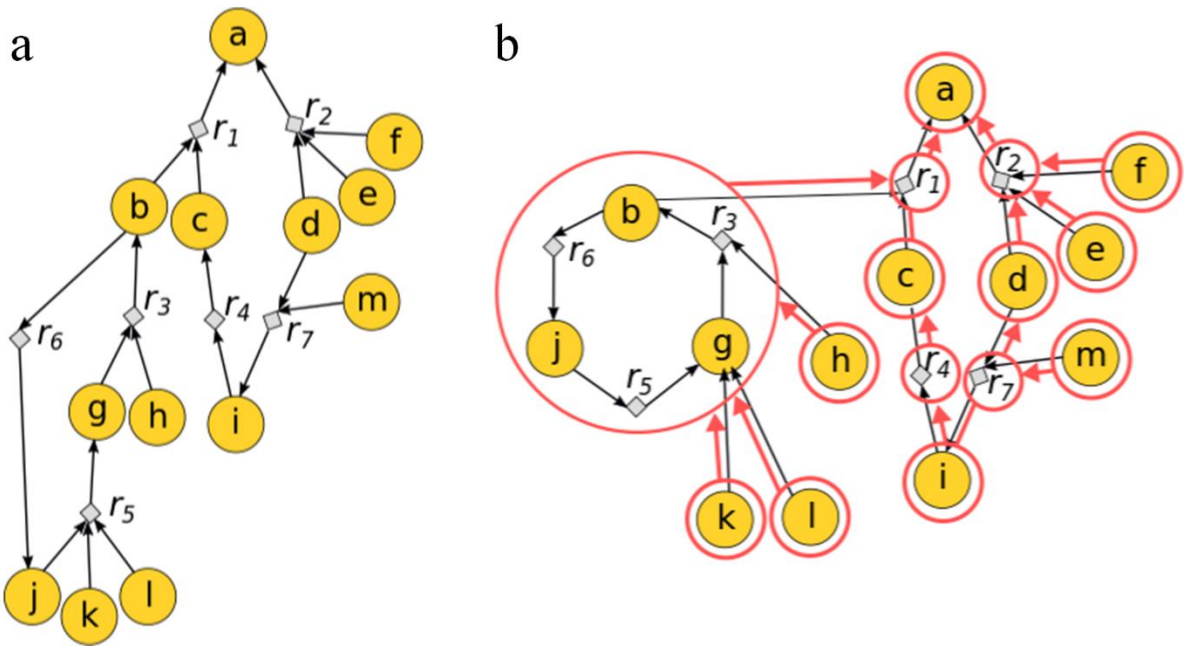


Figure S25. (a) To illustrate the basic scheme of updating costs consider the already discussed example (c.f. **Figure S20**) of solution graph G' containing directed cycles. This graph cannot be traversed in topological order. (**b**) However, the related graph composed of strongly connected components (SCC) of G' can be traversed in such order. SCC are enclosed by red circles. Note that these SCCs are mostly composed of one-element sets with original nodes from G'. The only exception is the SCC being a set of nodes of G' creating a cycle (i.e. chemical nodes b, g, j , plus reaction nodes r_3, r_5, r_6).

S6.4. Searches with constraints.

The searches described above make use of Chematica's entire knowledge base. Sometimes, however, it is desirable to restrict the searches to avoid (or promote) certain chemicals, reagents, or specific reaction types. This is done by using the family of HIDE_SEEK variables that can be incorporated into the scoring functions. Entering these variables with a positive sign assigns a penalty to the specific argument (structure or a keyword) whereas a negative sign promotes the use of the argument. For example, in the function in **Figure S21c**, the HIDE_SEEK_NAME assigns a large ("10,000") penalty ("+" sign) for every use of the metathesis reaction – in effect, the search algorithm will try to find pathways that do not use metathesis. Naturally, a much more meaningful use of this functionality would be to avoid heavy metals, or certain toxic substances, or reagents or solvents from the list of suggested reaction conditions we provide for every reaction rule.

As some examples of the HIDE_SEEK syntax consider:

- HIDE_SEEK_NAME(['Cl2']) will penalize steps requiring usage of gaseous chlorine
- HIDE_SEEK_NAME(['Synthesis of trifluoromethyl arenes from aryl boronic acids']) will avoid the specified reaction type during synthetic planning
- HIDE_SEEK_SMILES(['Nc1ccc(cc1)-c1ccc(N)cc1']) will exclude all pathways requiring the usage of benzidine
- HIDE_SEEK_SMARTS(['[#6][I]']) will prohibit the use of any iodides
- HIDE_SEEK_SMARTS(['[#6][N]=[N+]=[N-]']) will prohibit the use of any azides

All in all, the HIDE_SEEK variables might be useful to process chemists, although they are certainly insufficient to deal with all intricacies of process planning (waste disposal, ability to crystallize intermediates/products, etc.).

Section S7. Higher-order "chemical logic" and multi-step strategies.

The reaction rules and the search algorithms described in previous sections rely on the evaluation of individual synthetic steps. On the other hand, every seasoned synthetic chemist knows that planning one-step-at-a-time might be shortsighted as truly creative synthetic approaches benefit from the ability to "see several steps ahead". For the algorithmic point of view, one-step planning does not preclude the algorithm from ultimately finding the more inspired routes but the times to do so will be longer as the search will not be channeled into the desired, elegant sequence of steps and might spend some (unproductive) time examining other options (this is especially the case when the first step is the sequence does not look very promising but leads to subsequent, very elegant steps, see **Section S7.3** below).

To take multistep planning into account, we have implemented in *Chematica* several modules evaluating sequences of steps – below, we provide three most illustrative example.

S7.1. Labile, highly reactive groups. One of the cornerstones of synthetic planning is that highly reactive groups should not be dragged along multiple steps – one option is to protect them, the other is to introduce them only to immediately transform into other, more stable functionalities. For instance, once prepared, an organomagnesium species should be immediately added to an aldehyde, ketone, etc.; it is generally a bad idea to make it but then try to perform reactions on other parts of the molecule while this reacting species is "hanging around". Such considerations underlie *Chematica*'s module eliminating sequences of steps in which the labile, highly reactive groups are present for more than one step. Also, it is typically undesirable to drag along groups dramatically increasing the polarity and thus rendering subsequent work with such compounds problematic (e.g., boronic acids that are typically made to be immediately used in various metal-catalyzed couplings). Both types of groups can be identified based on expert chemical knowledge which in this particular case is well supported by the statistics of reaction sequences described in the literature (see **Figure S26**). Currently, *Chematica* uses 106 types of groups not to be "dragged along" (organomagnesiums, isocyanates, acyl halides, acyclic anhydrides, ketenes, boronic acids, etc.).

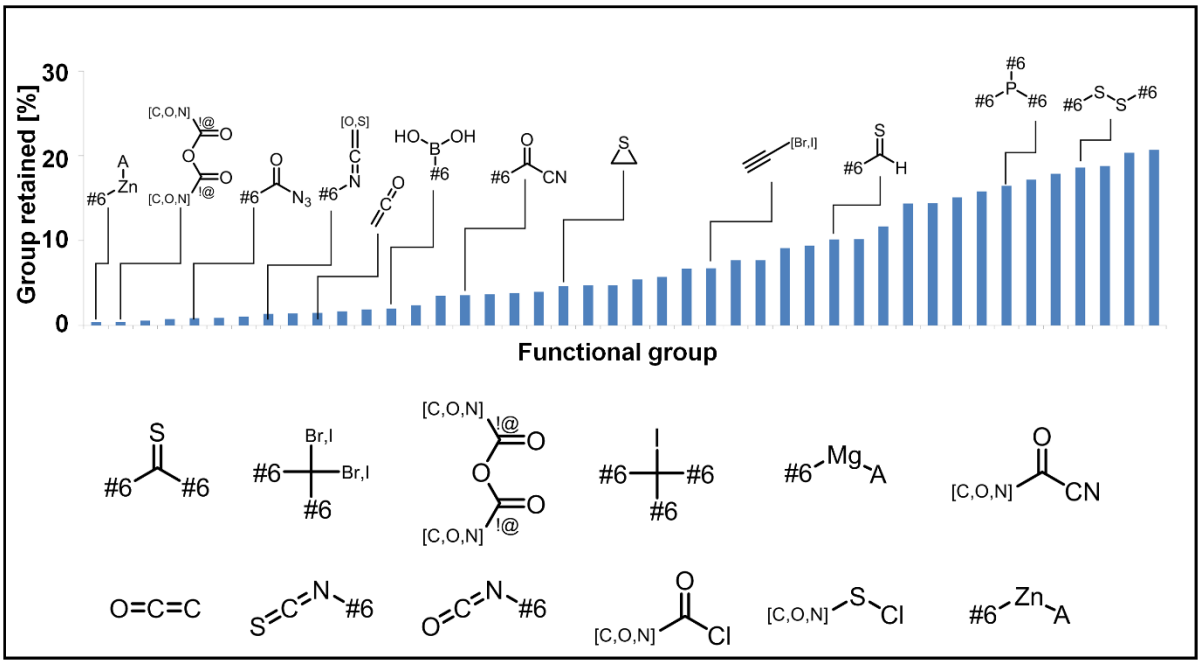


Figure S26. Histogram quantifying the fraction of two-step synthetic sequences in which specific functional groups are made in the first step and retained in the second step. The statistics is based on the two-step sequences from the Network of Organic Chemistry[10,13-16] comprising ca. 7 million published reactions. The labile groups are located at the left part of the plot. Some of the labile motifs are shown below the plot (!@ denotes non-cyclic bond, #6 stands for any carbon, and A for any aliphatic atom).

S7.2. Cyclizations. Another case where we need to look beyond individual steps is that of sequences in which a larger ring is first made and then contracted into one or more smaller rings. Such an approach is usually – but not always, see elegant synthesis of loline by Trauner's group[S54] – unjustified due to the effort to make the larger macrocycle acting only as an intermediate to a smaller-ring system. In fact, in the vast majority of cases, much simpler intermediates can be used. One example is illustrated in **Figure S27** which has three potential ways of making monomorine I. The approach involving creation of a nine-membered macrocycle that is then contracted to a 5-6 ring system is much more laborious than the other two approaches shown (one of which was actually demonstrated by the Higashiyama's group[S55]). In *Chematica*, the choice whether to prevent such macrocycle contractions is left to the user (i.e., it is optional, especially for those who would like to follow in the masterful footsteps of Professor Trauner).

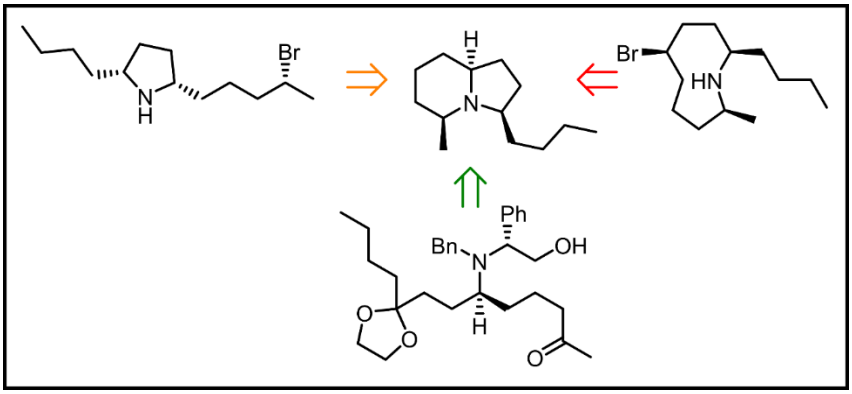


Figure S27. Different possible approaches to the synthesis of a bicyclic target, monomorine I (a pheromone of pharaoh ants[S56]). The red arrow marks the most laborious approach involving making a nine-membered precursor. This approach is also somewhat risky since making 9-membered systems is often synthetically challenging and the ensuing intramolecular cyclization would require the ring to assume proper conformation. The other two approaches shown appear much more plausible – in fact, the one marked by the green arrow was demonstrated experimentally[S55].

S7.3 Strategies. Finally, we deal with situations whereby the first step (in retrosynthetic direction) does not appear promising but, if taken, might enable subsequent elegant/effective transformations. A classic example here is the introduction of a double bond into a cyclohexane ring – by itself, this transformation does not simplify the structure and the algorithm evaluating such a step by a CSF function would not score it as promising. However, every chemist knows that the point of this preliminary "move" is to set the scene for a Diels-Alder reaction disconnecting the ring into a diene and a dienophile. An analogy that can be made here is to the Monte-Carlo methods of statistical physics whereby it is often useful to allow some uphill moves to overcome local "hurdles"/maxima, "escape" from local minima, and ultimately find the global minimum – colloquially put, it is sometimes good to walk uphill to then discover a valley of new opportunities.

In *Chematica*, sequences of steps in which the first one is a preliminary/ "sacrificial" move setting the scene for a subsequent key transformation are called "strategies" – their role is very important as they allow the search algorithm to explore syntheses involving "counterintuitive" sequences of steps that would not be easily found with one-step-at-a-time planning. There are currently **several hundred thousand strategies** implemented in *Chematica* – some examples, mirroring strategies used in the literature, are shown in **Figure S28**.

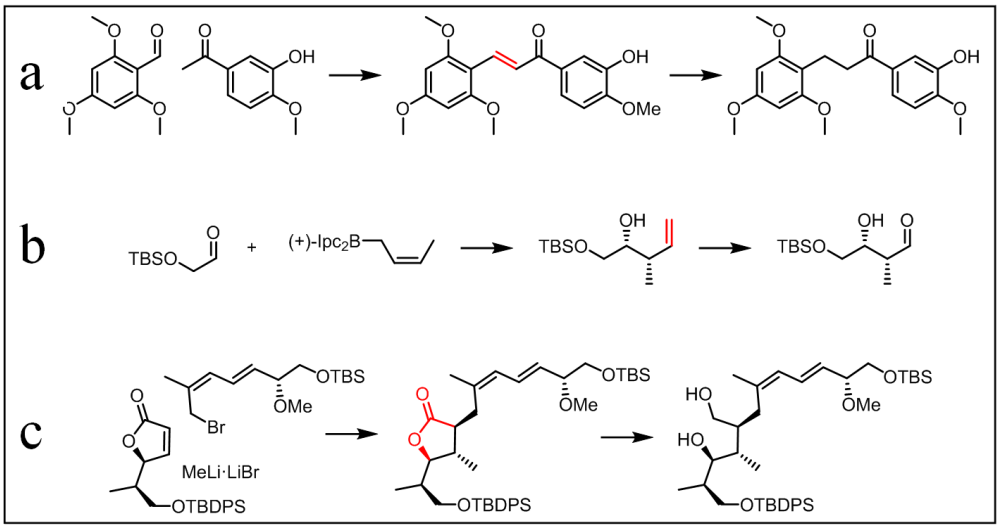


Figure S28. Examples of syntheses comprising two-step strategies. **(a)** Short and efficient synthesis of taccabulin A[S57] relies on a condensation of benzaldehyde and acetophenone followed by hydrogenation of the double bond. When planning this synthesis (i.e., thinking in the retrosynthetic direction), introduction of the double bond does not offer any immediate gains but is necessary for the condensation step. **(b)** When making an intermediate in the synthesis of brevisamide[S58], the so-called Brown crotylation is followed by oxidation of terminal alkene to aldehyde. In the retrosynthetic direction, changing an aldehyde into an alkyne might not be immediately seen as advantageous. We note that compared to a direct cross-aldol coupling between two aldehydes (allowing only for the *anti* product), the strategy shown here is highly reliable and allows for making both *anti* as *syn* products – in fact, allylation/crotylation strategy is nowadays the method of choice for making chiral β-hydroxyaldehydes. **(c)** Halichomycin intermediate[S59] is obtained from the corresponding lactone. In the retrosynthetic direction, formation of the ring might be counterintuitive (as it apparently complexifies the structure) – on the other hand, it introduces the electron-withdrawing group which then enables "division" of this intermediate into three synthons while installing two vicinal stereocenters.

We note that incorporating strategies into the search scenario does not require any modifications to the parameters of the CSF/RSF scoring functions when setting up the search. Instead, strategies are activated with single checkbox (red circle) in the control panel shown in **Figure S29** below.

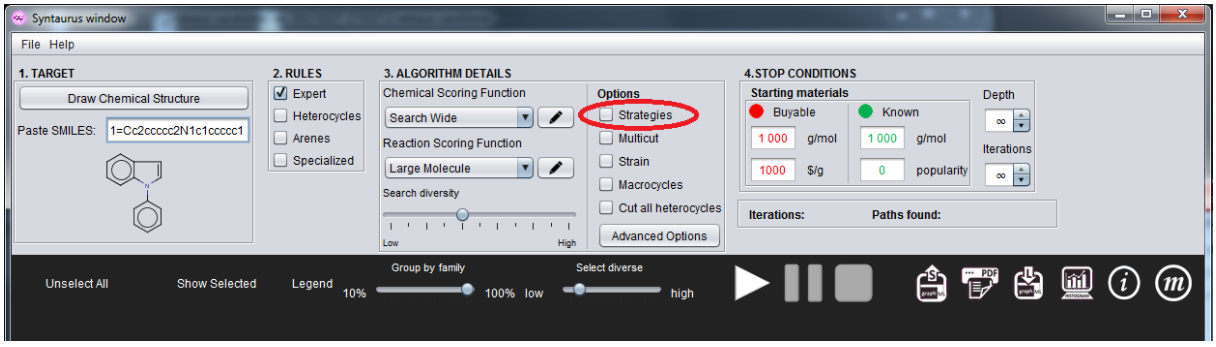


Figure S29. Turning on the "Strategies" window in *Chematica*'s main control panel.

Section S8. Typical raw output from Chematica.

Chematica's key output is a set of complete synthetic pathways ranked by the score combining the cost of executing reactions involved (Reaction Scoring Function) and the actual prices of starting materials. Each substance involved can be inspected in more detail via *Chematica*'s Molecular Mechanics module (visualizing 3D conformers, calculating bond, angle, and dihedral angle energies, etc.; **Figure S30**). More importantly, each individual reaction step in a given route is also accompanied by suggestions of the typical reaction conditions (solvent, catalyst type, illustrative literature reference describing details of a particular class of reactions), as well as information about which groups need to be protected under reaction's conditions and with what protecting groups. This information was used as provided when executing the syntheses described in the main text. Additionally, Chematica provides a list of other, similar reaction precedents described in the literature (note: these precedents are not used in *Chematica*'s reaction rules and/or synthetic planning which and coded as described in **Sections S2** and **S3**; this modality was not used – in fact, not yet available – when the syntheses described in the main text were planned; still, many current users find this extra option to "consult the literature" useful).

The screenshots in **Figures S30-S32** below illustrate output of *Chematica* for pathways leading to Engelheptanoxide C.

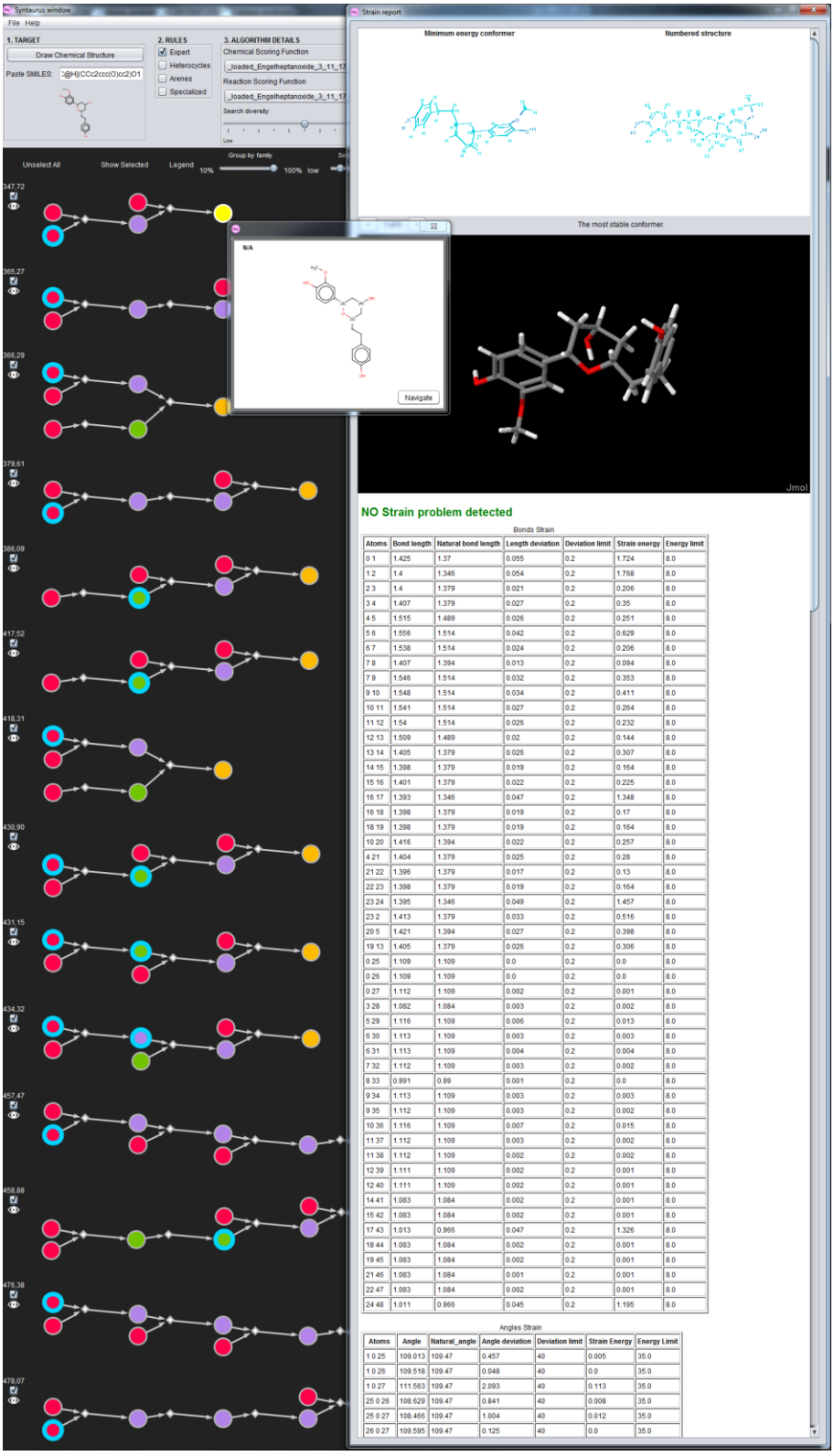


Figure S30. For each molecule in the pathways found (e.g., pathways leading to Engelheptanoxide C shown in the *left* portion of the figure), Chematica generates a compendium of structural information. Shown here is the "Strain Report" functionality coloring the bonds in the molecule according to local strain (*top-right* image; blue color signifies no strain), displaying lowest energy conformers (*middle-right*; up to five lowest-energy conformers can be displayed), as well as a list of all bond lengths, angles, and dihedral angles along with threshold parameters above which excessive strain is reported (*lower-right*; no strain problems are detected for the molecule inspected). Creation of such a "report" for each molecule (upon right clicking the desired molecule node) takes on the order of 1-2 sec.

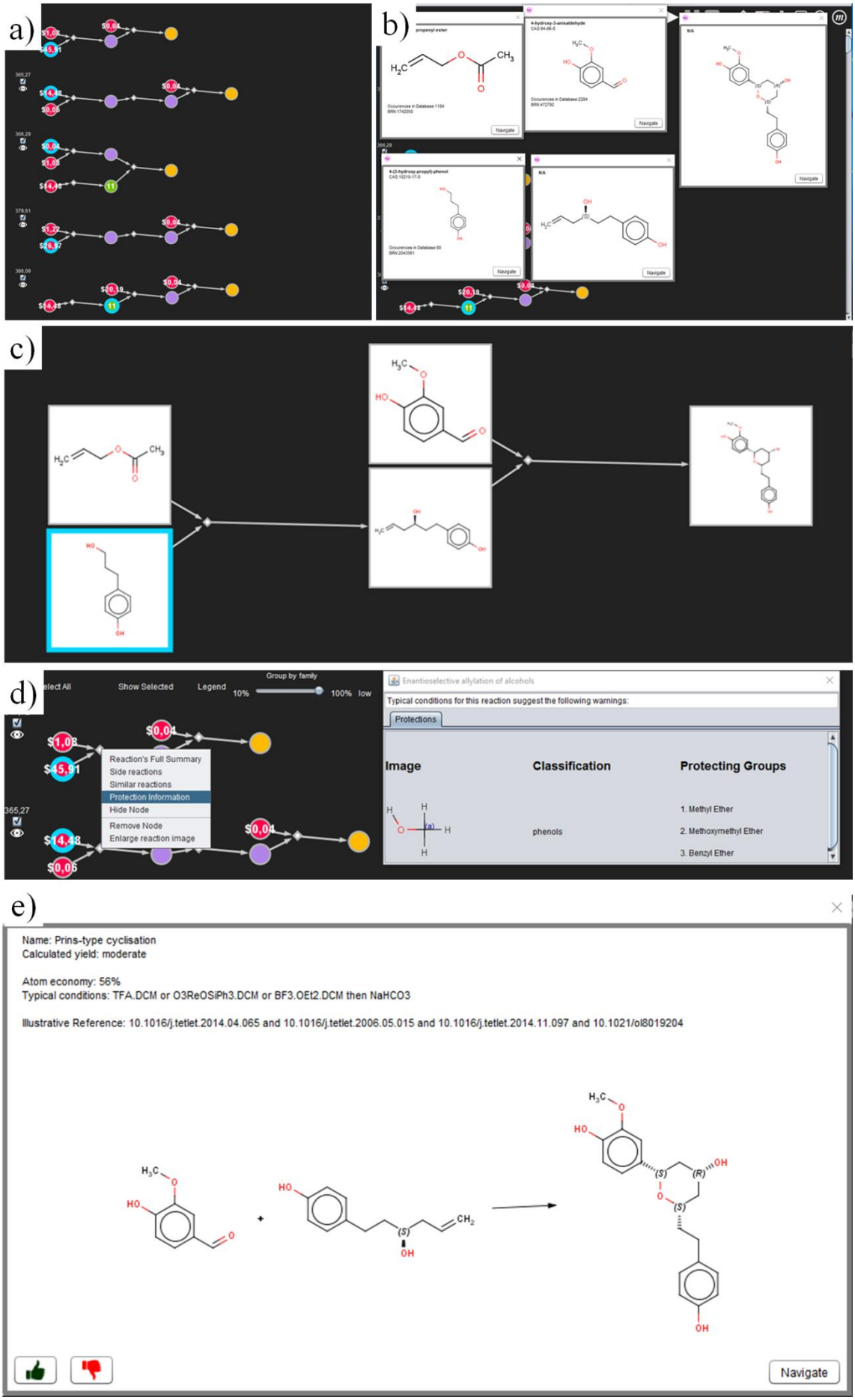


Figure S31. (a) A screenshot from *Chematica* showing multiple pathways generated for Engelheptanoxide C. Synthetic routes are represented as a bi-partite graphs with circular nodes corresponding to substances (yellow = target; violet = unknown in literature; green = known in literature; red = commercially available) and smaller, diamond-shaped nodes corresponding to reactions. The best-scoring pathway at the top of the list was carried out experimentally as described in the main text. All molecules can be displayed as structures upon **(b)** clicking on and "opening" individual nodes or **(c)** clicking an "eye" icon that displays all molecules in a selected pathway. **(d)** A blue halo surrounding a molecule node signifies need for protection. Details about which functional group should be protected as well as the list of protecting groups compatible with conditions of a specific reaction can be accessed by a right click on the reaction node and selecting the „Protection „Information" option. In the first step of Engelheptanoxide C synthesis, need to protect phenol is detected and the program suggests methyl ether, methoxymethyl ether, and benzyl ether as plausible protecting groups. The last one was actually used in the synthesis. **(e)** Window displaying the details of the second reaction step (Prins-type cyclization). From the suggested choices, $ReO3(OSiPh_3)$ catalyst was used although, for completeness, we also showed that the TFA conditions worked (tried once, yield ~20%; we did not try $BF3{\bullet}OEt_2$ as can improve the enantiomeric purity – which we found already sufficient with Re catalyst – but is known to do so at the expense of yield).

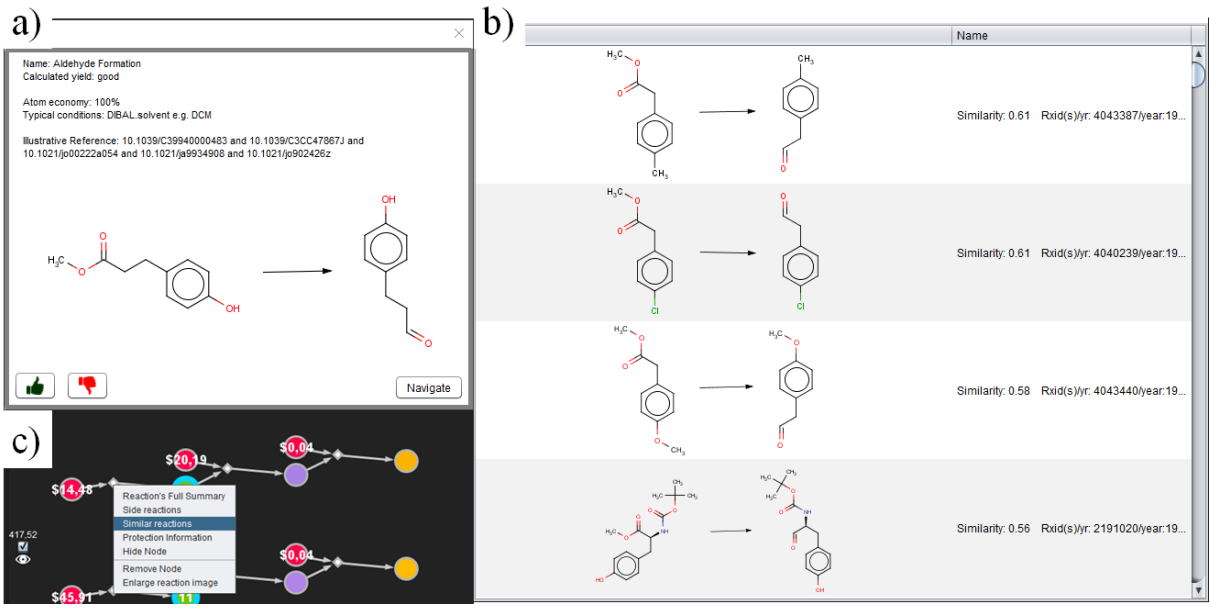


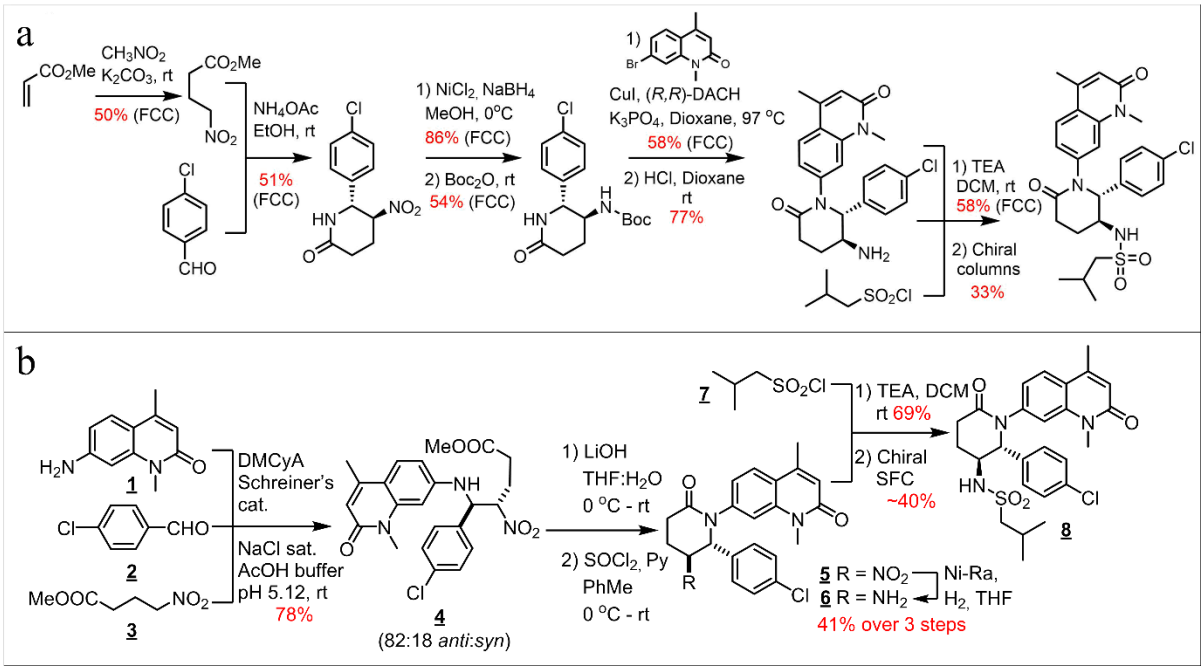
Figure S32. The "Similar reactions" functionality. The example here is for an alternative, lower-scoring pathway leading to Engelheptanoxide C (i.e., not the pathway executed) and involving reduction of methyl 3-(4-hydroxyphenyl)propanoate to a 3-(4-hydroxyphenyl)propanal (reaction shown in **(a)**). Window showing similar reactions (i.e., closest literature precedents) is shown in **(b)** and is opened by right-clicking on the pertinent reaction node in the synthetic plan **(c)**.

S9. Summary.

All of *Chematica*'s modules reflect various aspects of synthetic design employed by humans, and are synergistically important for the program's overall success – the large number of reaction rules is necessary to endow the system with a requisite base of "chemical knowledge," search algorithms are crucial for identifying full pathways tracing all the way to available starting materials, and various heuristics and multistep strategies are indispensable for ensuring these pathways are logically coherent and, hopefully, also "elegant". Whereas in our narrative we focused on the conceptual basis for these methods, there are also numerous interesting problems we have had to address at the level of algorithm optimization to (i) process very large numbers of data efficiently and (ii) deliver results within times acceptable to the user chemists. In the latter context, we have made a "sociologically" interesting observation that while in their everyday practice chemists can spend long times tinkering with pathway design and can tolerate large proportion of experimental failures, they tend to expect the results from *Chematica* to be delivered almost instantaneously and without any room for error. One might object to such harsh criteria (and we did, especially in *Chematica*'s toddler years), but ultimately this is the correct attitude – after all, the synthetic community does not need any "toy programs" dealing with only simple chemistries but a system that can of real help in attacking synthetically non-trivial targets. The examples of syntheses we described in the main section give us hope that *Chematica* is reaching this level of maturity and upon its wide dissemination, will soon become an indispensable *in silico* companion of synthetic chemists.

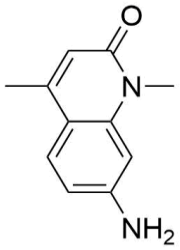
Section S10. Synthesis of the inhibitor of BRD proteins 7 and 9, 8.

S10.1 Previous vs. current synthetic routes.



Scheme S1. (a) The original preparation of **8** from the main-text reference [20]. For comparison, **(b)** shows the Chematica route (same as in the main-text **Figure 2a**).

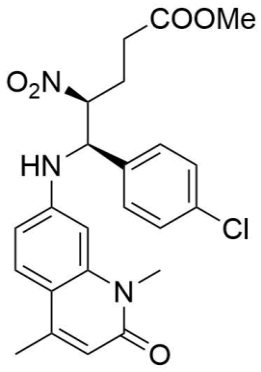
S10.2. Synthetic details.



***7-amino-1,4-dimethylquinolin-2(1H)-one* 1**

In a 250 mL round bottom flask fitted with a stir bar and nitrogen inlet, sodium hydride (60% NaH in mineral oil; 1.09 g, 27.2 mmol) was added to a solution of 7-amino-4-methyl-1H-quinolin-2-one (4.3 g, 24.7 mmol) in *N,N*-dimethylformamide (32 mL) and tetrahydrofuran (112 mL) at room temperature. After 1h, methyl iodide (1.8 mL, 30 mmol) was added to the reaction mixture and the reaction was stirred at room temperature while monitored by TLC. After 45 min, the resulting white solid was filtered off and filtrate was evaporated to dryness. The crude material was suspended in diethyl ether (200 mL), and the solid mass was broken into fine particles by sonication followed by stirring for 1h. The resulting off-white solid was filtered and treated with 100 mL of water, then dried under high vacuum to yield compound **1** (3.1g, 66% yields) as off white solid.

^{1}H NMR: (400 MHz, DMSO-d_6) δ 7.43 (d, *J* = 8.9 Hz, 1H), 6.53 (d, *J* = 7.7 Hz, 2H), 6.09 (s, 1H), 5.88 (s, 2H), 3.46 (s, 3H), 2.30 (s, 3H).
^{13}C NMR: (101 MHz, DMSO-d_6) δ 161.87, 152.04, 146.96, 142.03, 126.90, 114.44, 111.43, 110.20, 97.28, 28.93, 18.84.
LC-MS: m/z = 189.2 (M+1)

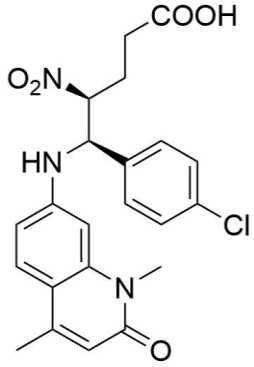


***N-[{rac-4-nitro-5-(4-chlorophenyl)-methylpentanoate-5-yl}]-7-amino-1,4-dimethylquinolin-2(1H)-one* 4**

A 250 mL round bottom flask fitted with a stir bar and the nitrogen inlet and containing 25 mL saturated sodium chloride buffer solution (NaAcO/AcOH; pH 5.12; 20 mM) was placed into a 0°C water-ice bath stirred at 100-200 r.p.m. Sequentially added were 4-chlorobenzaldehyde (**2**; 0.75 g, 5.3 mmol), methyl 4-nitrobutanoate (**3**; 13.6 mL, 106.2 mmol), *N,N′*−bis[3,5−bis(trifluoromethyl)phenyl]thiourea (532 mg, 1.1 mmol), *N,N*-dimethylcyclohexanamine (159 uL, 1.1 mmol), and 7-amino-1,4-dimethylquinolin-2(1*H*)-one (**1**;1.0g, 5.3 mmol) to form a biphasic solution. After 15 min, the stirring rate was increased to 1000 r.p.m. to break the biphasic system into small droplets. The reaction was allowed to warm to the room temperature and allowed to stir under these conditions for 72 h. The stirring was stopped and the reaction mixture was extracted with dichloromethane (3 × 100 mL). The combined organic extracts were washed with water (100 mL), brine (75 mL), dried over $MgSO_4$, filtered, and evaporated to give crude product. The crude product was purified by FCC (3:1 EtOAC:EtOH/ hexanes) to yield compound **4** (1.91g, 78% yields) as light yellow solid.

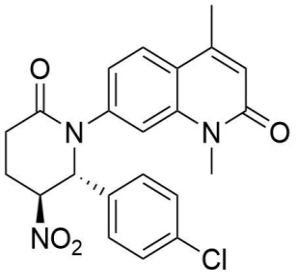
^{1}H NMR: (400 MHz, $CDCl_3$, mixture) δ 7.44 (dd, *J* = 8.7, 3.0 Hz, 1H), 7.40 – 7.33 (m, 2H), 7.33 – 7.28 (m, 2H), 6.50 (dd, *J* = 8.7, 2.2 Hz, 1H), 6.39 – 6.29 (m, 2H), 5.11 (s, 1H), 5.07 – 4.89 (m, 2H), 3.71 (s, 2.5H), 3.69 (s, 0.5H), 3.53 (s, 0.5H), 3.50 (s, 2.5H), 2.59 – 2.46 (m, 1H), 2.46 – 2.27 (m, 6H).

^{13}C NMR: (101 MHz, $CDCl_3$, major *trans isomer* from mixture) δ 172.55, 162.68, 147.76, 146.61, 141.44, 135.37, 134.84, 129.46 (2C), 128.39 (2C), 126.59, 116.79, 113.96, 109.25, 97.66, 90.70, 59.70, 52.12, 29.92, 29.05, 24.96, 18.83.
^{1}H NMR: (400 MHz, $CDCl_3$, *trans* isomer) δ 7.44 (d, *J* = 8.7 Hz, 1H), 7.40 – 7.28 (m, 4H), 6.51 (dd, *J* = 8.7, 2.2 Hz, 1H), 6.33 (s, 2H), 5.21 (d, *J* = 5.3 Hz, 1H), 4.99 (d, *J* = 7.6 Hz, 2H), 3.71 (s, 3H), 3.50 (s, 3H), 2.60 – 2.45 (m, 1H), 2.35 (s, 3H), 2.45 – 2.29 (m, 3H).
^{1}H NMR: (400 MHz, $CDCl_3$, *cis* isomer) δ 7.44 (d, *J* = 8.7 Hz, 1H), 7.40 – 7.27 (m, 4H), 6.50 (dd, *J* = 8.7, 2.2 Hz, 1H), 6.40 – 6.29 (m, 2H), 5.21 (d, *J* = 8.2 Hz, 1H), 5.09 – 4.89 (m, 2H), 3.69 (s, 3H), 3.53 (s, 3H), 2.34 (s, 3H), 2.59 – 2.27 (m, 3H), 2.15 – 2.05 (m, 1H).
^{13}C NMR: for minor *cis* isomer was not recorded.
LC-MS: m/z = 458.2 (M+1).



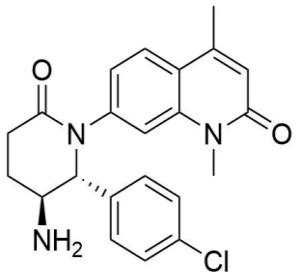
***N-[{rac-4-nitro-5-(4-chlorophenyl)-pentanoic acid-5-yl}]-7-amino-1,4-dimethylquinolin-(1H)-one* SI-1**
In a 100 mL round bottom flask fitted with a stir bar, lithium hydroxide, monohydrate (69 mg, 1.6 mmol) was added portion-wise to a solution of *N*-[{*rac*-4-nitro-5-(4-chlorophenyl)methylpentanoate-5-yl}]-7-amino-1,4-dimethylquinolin-2(1*H*)-one (**4**; 0.50 g, 1.09 mmol) in a mixture of tetrahydrofuran (12 mL) and water (4 mL) solvents at 0 °C (ice-water bath). The reaction mixture was allowed to warm to room temperature and was stirred overnight. After 16 h, the organic solvent was removed under the reduced pressure and water (15 mL) was added to the residue. The aqueous solution was slowly acidified by adding 1N aq. HCl solution (to pH 4.5-5.5) and the solvent was evaporated. The solid residue was suspended in water (15 mL) and stirred for 1 h followed by sonication. The remaining solid was filtered was dried over P_2O_5 in vacuum oven to yield 0.431g (89%) of the crude acid **SI-1** as off white solid. The crude acid was used in the next reaction without any further purification.

LC-MS: m/z = 444.3 (M+1).



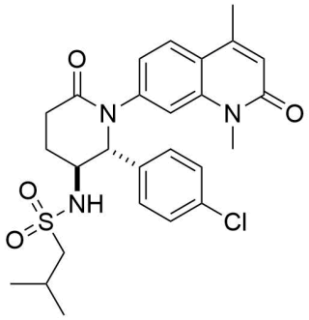
***7-[rac-3-nitro-2-(4-chlorophenyl)-6-oxopiperidin-1-yl]-1,4-dimethylquinolin-2(1H)-one* 5**
In a 100 mL round bottom flask fitted with a stir bar and nitrogen inlet, thionyl chloride (0.2 mL, 2.8 mmol) was added drop-wise to a solution of *N*-[{*rac*-4-nitro-5-(4-chlorophenyl)pentanoic acid-5-yl}]-7-amino-1,4-dimethylquinolin-2(1*H*)-one (**SI-1**, 0.431 g, 0.97 mmol) and pyridine (0.78 mL, 9.7 mmol) in toluene (12 mL) at 0 °C. The reaction was continued for 4 h at the same temperature and then quenched by the addition of water (5 mL). The solvent was removed under reduced pressure. Toluene (15 mL) was added to the solid residue and evaporated to dryness. The solid material was stirred in water (15 mL) for 1h followed by sonication (10 min). The remaining solid was filtered and dried over P_2O_5 in vacuum oven to yield 400 mg (97%) of the crude oxopiperidine **5** as light yellow solid. The crude compound was used in the next step without any further purification.

^{1}H NMR: (400 MHz, $CDCl_3$, *trans* isomer) δ 7.63 (d, *J* = 8.5 Hz, 1H), 7.43 (d, *J* = 8.5 Hz, 2H), 7.33 (d, *J* = 8.5 Hz, 2H), 7.17 (d, *J* = 1.8 Hz, 1H), 7.04 (dd, *J* = 8.5, 1.9 Hz, 1H), 6.54 (s, 1H), 5.82 (s, 1H), 4.82 (dd, *J* = 6.8, 3.8 Hz, 1H), 3.56 (s, 3H), 2.90 – 2.79 (m, 2H), 2.79 – 2.66 (m, 1H), 2.39 (s, 3H), 2.44 – 2.30 (m, 1H).
^{1}H NMR: (400 MHz, $CDCl_3$, *cis* isomer) δ 7.63 (d, *J* = 8.5 Hz, 1H), 7.50 – 7.38 (m, 2H), 7.38 – 7.28 (m, 2H), 7.17 (d, *J* = 1.9 Hz, 1H), 7.04 (dd, *J* = 8.5, 2.0 Hz, 1H), 6.56 (d, *J* = 1.0 Hz, 1H), 5.87 – 5.77 (m, 1H), 4.81 (dd, *J* = 6.8, 3.9 Hz, 1H), 3.57 (s, 3H), 2.91 – 2.79 (m, 2H), 2.79-2.65 (m, 1H), 2.39 (s, 3H), 2.47 – 2.30 (m, 1H).
^{13}C NMR: (101 MHz, $CDCl_3$, *trans* isomer) δ 168.20, 161.93, 145.81, 143.04, 140.52, 135.36, 135.21, 129.82 (2C), 128.12 (2C), 126.25, 121.55, 120.72, 120.63, 113.15, 83.95, 65.73, 29.20, 27.64, 20.41, 18.90.
^{13}C NMR: for *cis* isomer was not recorded.
LC-MS: m/z = 426.3 (M+1).



***7-[rac-3-amino-2-(4-chlorophenyl)-6-oxopiperidin-1-yl]-1,4-dimethylquinolin-2(1H)-one* 6**
In a Paar pressure bottle, 7-[*rac*-3-nitro-2-(4-chlorophenyl)-6-oxopiperidin-1-yl]-1,4-dimethylquinolin-2(1*H*)-one (**5**; 400 mg g, 0.94 mmol) was dissolved in tetrahydrofuran (50 mL). Raney® Nickel (suspended in water; ~55 mg, 0.94 mmol) was added to the solution and the vessel was set on a Paar Hydrogenation Apparatus under an atmosphere of H_2 gas (40 psi). After 7 h, the catalyst was filtered through a bed of Celite and washed with MeOH (200 mL). The combined organic layer was evaporated to provide a yellow gummy solid. The crude compound was purified by FCC (MeOH/ dichloromethane with 0.1% NH_4OH) to yield the desired compound **6** (0.175 g, 41% in over three steps) as gummy solid.

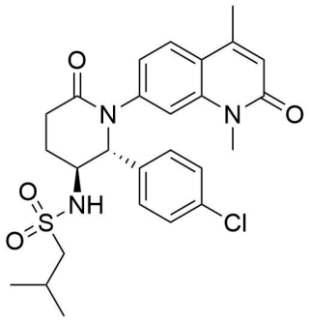
^{1}H NMR: (400 MHz, $CDCl_3$) δ 7.54 (d, *J* = 8.5 Hz, 1H), 7.24 (dd, *J* = 26.8, 8.4 Hz, 4H), 7.07 (d, *J* = 1.7 Hz, 1H), 6.98 (dd, *J* = 8.5, 1.7 Hz, 1H), 6.48 (s, 1H), 4.91 (d, *J* = 5.0 Hz, 0.2H), 4.70 (d, *J* = 6.0 Hz, 0.8H), 3.51 (s, 3H), 3.43 (s, 1H), 3.33 –3.3.41 (m, 1H), 2.74 –2.85 (m, 2H), 2.34 (s, 3H), 2.20 – 1.83 (m, 5H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 170.44, 162.05, 146.10, 143.52, 140.07, 137.72, 134.02, 129.04 (2C), 128.85 (2C), 125.79, 121.32, 120.95, 120.03, 113.77, 72.22, 52.76, 29.81, 29.21, 26.33, 18.87.
LC-MS: m/z = 395.2 (M-1).



***N-[rac-2-(4-chlorophenyl)-1-(1,4-dimethyl-2-oxo-1,2-dihydroquinolin-7-yl)-6-oxopiperidin-3-yl]-2-methylpropane-1-sulfonamide* rac-8**
In a round bottom flask fitted with a stir bar and nitrogen inlet, 7-[*rac*-3-amino-2-(4-chlorophenyl)-6-oxopiperidin-1-yl]-1,4-dimethylquinolin-2(1*H*)-one (**6**; 0.175 g, 0.44 mmol) was dissolved in dichloromethane (8 mL). Triethylamine (0.19 mL, 1.3 mmol) was added to the solution at room temperature at once, followed by isobutanesulfonyl chloride **7** (0.14 mL, 1.1 mmol), added dropwise.

After 3 h, the reaction was diluted with DCM (50 mL) and washed with brine (15 mL). The organic layer was dried over $MgSO_4$, filtered, and evaporated to provide a brown gummy residue. The crude material was purified by FCC (0-18% MeOH/ethyl acetate with 0.1% NH_4OH) to yield compound **rac-8** 0.160 g, 69%) as a light yellow solid.

^{1}H NMR: (400 MHz, $CDCl_3$, mixture) δ 7.60 (d, *J* = 8.5 Hz, 1H), 7.49 – 7.20 (m, 5H), 7.16 – 7.02 (m, 1H), 6.54 (s, 1H), 5.72 (d, *J* = 8.6 Hz, 1H), 5.27 (d, *J* = 3.0 Hz, 1H), 3.90-3.80 (m, 1H), 3.56 (s, 2.5H), 3.50 (s, 0.5H), 3.10-2.64 (m, 4H), 2.38 (s, 3H), 2.22 (d, *J* = 6.7 Hz, 2H), 1.88 – 1.73 (m, 1H), 1.12 (dd, *J* = 6.7, 2.9 Hz, 1H), 1.07 (dd, *J* = 6.7, 2.4 Hz, 5H).
LC-MS: m/z = 516.41, 518.41, 519.41



***N-[(2R,3S)-2-(4-chlorophenyl)-1-(1,4-dimethyl-2-oxo-1,2-dihydroquinolin-7-yl)-6-oxopiperidin-3-yl]-2-methylpropane-1-sulfonamide* 8**
The enantiomers from racemic **8** (0.133g) were separated by chiral SFC using ChiralPak AD-H HPLC column (4.6 × 100 mm; 35% EtOH in CO_2, 70 mL/min flow rate, 2.5 mL injection volume, 133 mg/ 10 mL methanol injection concentration; detection at λ 220 nm absorbance). Yield of (2*R*, 3*S*)-LP99 was 76 mg (55% yield) and, yield of (2*S*, 3*R*)-LP99 was 56 mg (42%).
50 mg of the desired isomer [(2*R*, 3*S*)-LP99] was purified again by Chromatotron flash chromatography using 4% MeOH in DCM to yield 33 mg (67% recovery, which is equal to ~40% final yield compared to 133 mg of crude product) of the pure (2*R*,3*S*)-LP99 **8** as off white solid.
The purity of the enantio-enriched sample was determined by analytical HPLC analysis (ChiralPak AD 4.6 x 250 mm, 45% Isopropanol in Hexanes, 0.5 mL/min flow rate, 1 mg/mL injection concentration, 5 µL injection volume, detection at λ = 220 nm absorbance): t_r (2*R*,3*S*)-LP99 = 13.75 min, 99.33% ee.

^{1}H NMR: (400 MHz, $CDCl_3$) δ 7.55 (d, *J* = 8.5 Hz, 1H), 7.31 (dt, *J* = 15.7, 8.5 Hz, 5H), 7.10 (dd, *J* = 8.5, 1.8 Hz, 1H), 6.56 (d, *J* = 8.0 Hz, 1H), 6.52 (d, *J* = 4.0 Hz, 1H), 5.27 (d, *J* = 2.8 Hz, 1H), 3.85 (dd, *J* = 7.8, 3.8 Hz, 1H), 3.52 (s, 3H), 2.95-2.71 (m, 4H), 2.35 (s, 3H), 2.23 (dt, *J* = 13.3, 6.7 Hz, 1H), 2.16 – 2.05 (m, 1H), 1.89-1.75 (m, 1H), 1.07 (d, *J* = 6.7 Hz, 6H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 169.91, 162.02, 146.00, 143.73, 140.32, 137.01, 134.34, 129.23 (2C), 128.15 (2C), 125.78, 121.20, 121.02, 120.24, 113.47, 70.77, 61.72, 53.79, 29.34, 27.89, 24.94, 22.54, 22.52, 22.37, 18.85.
LC-MS: m/z = 516.41, 518.41, 519.41

S10.3. Raw spectroscopic and chromatographic data.

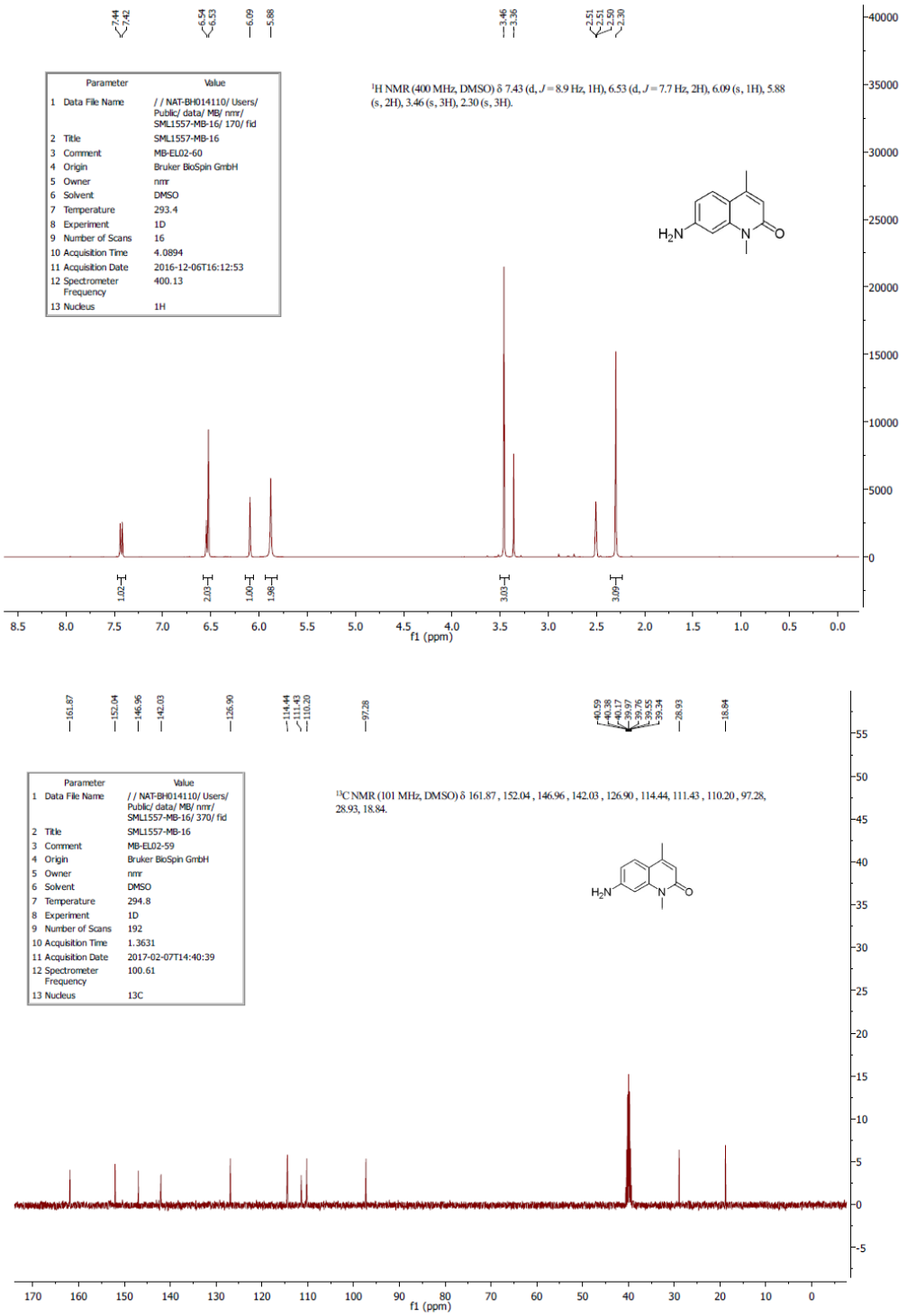


Figure S33. ^{1}H (top) and ^{13}C (bottom) NMR spectra of compound **1**.

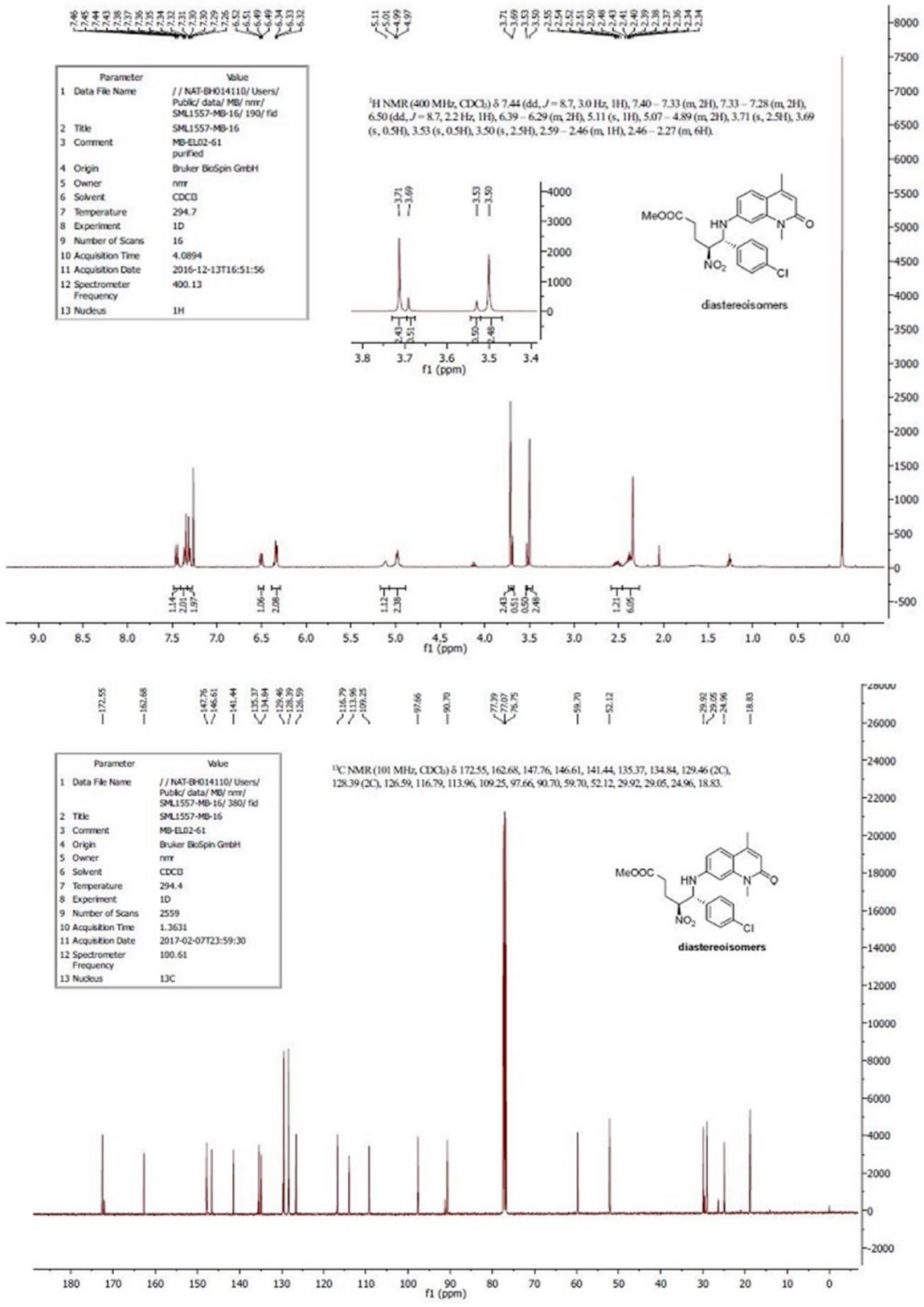


Figure S34. ^{1}H (top) and ^{13}C (bottom) NMR spectra of compound **4**.

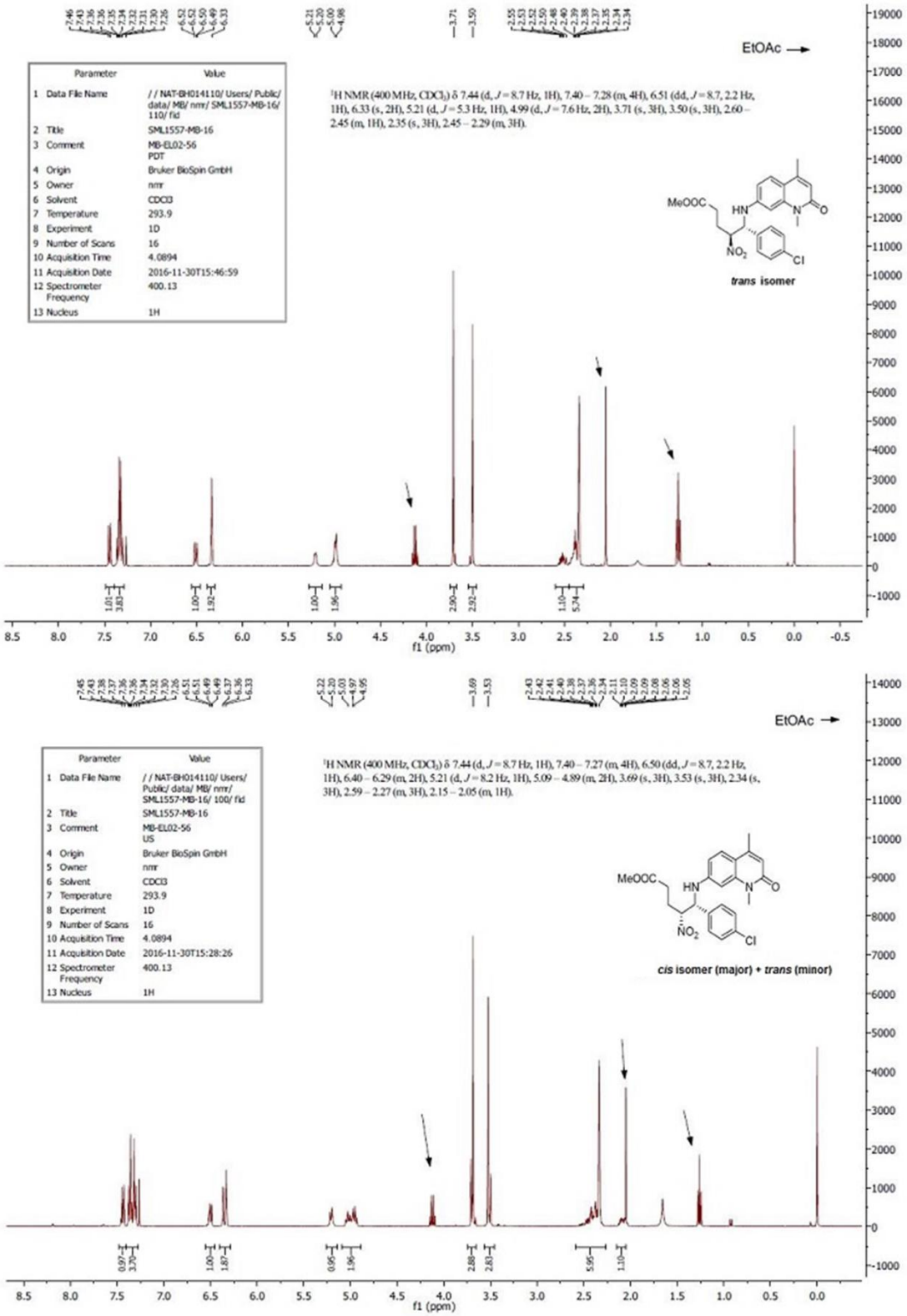


Figure S35. ^{1}H NMR spectra of the *trans* (top) and *cis* (bottom) isomers of compound **4**.

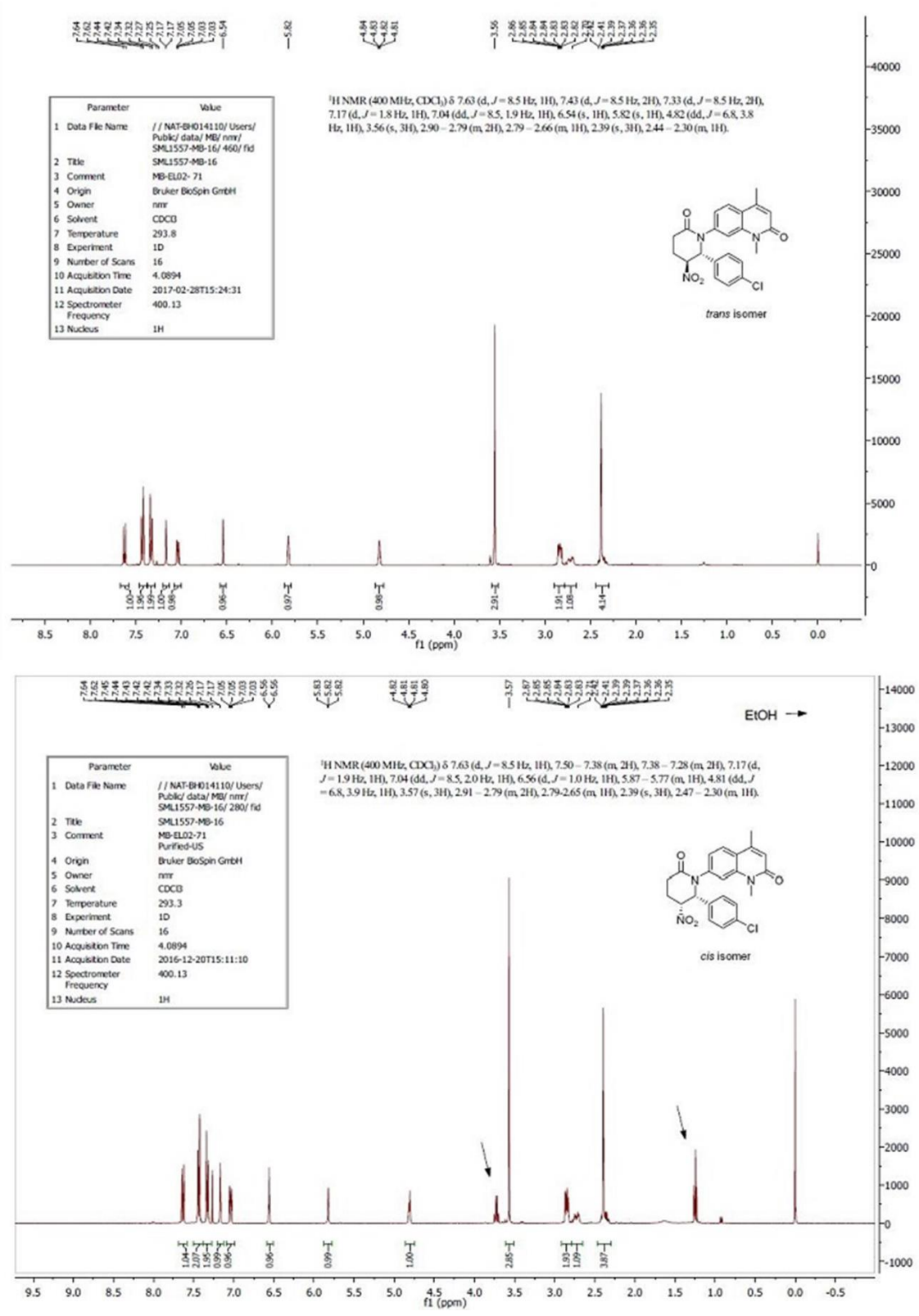


Figure S36. ^{1}H NMR spectra of the *trans* (top) and *cis* (bottom) isomers of compound **5**.

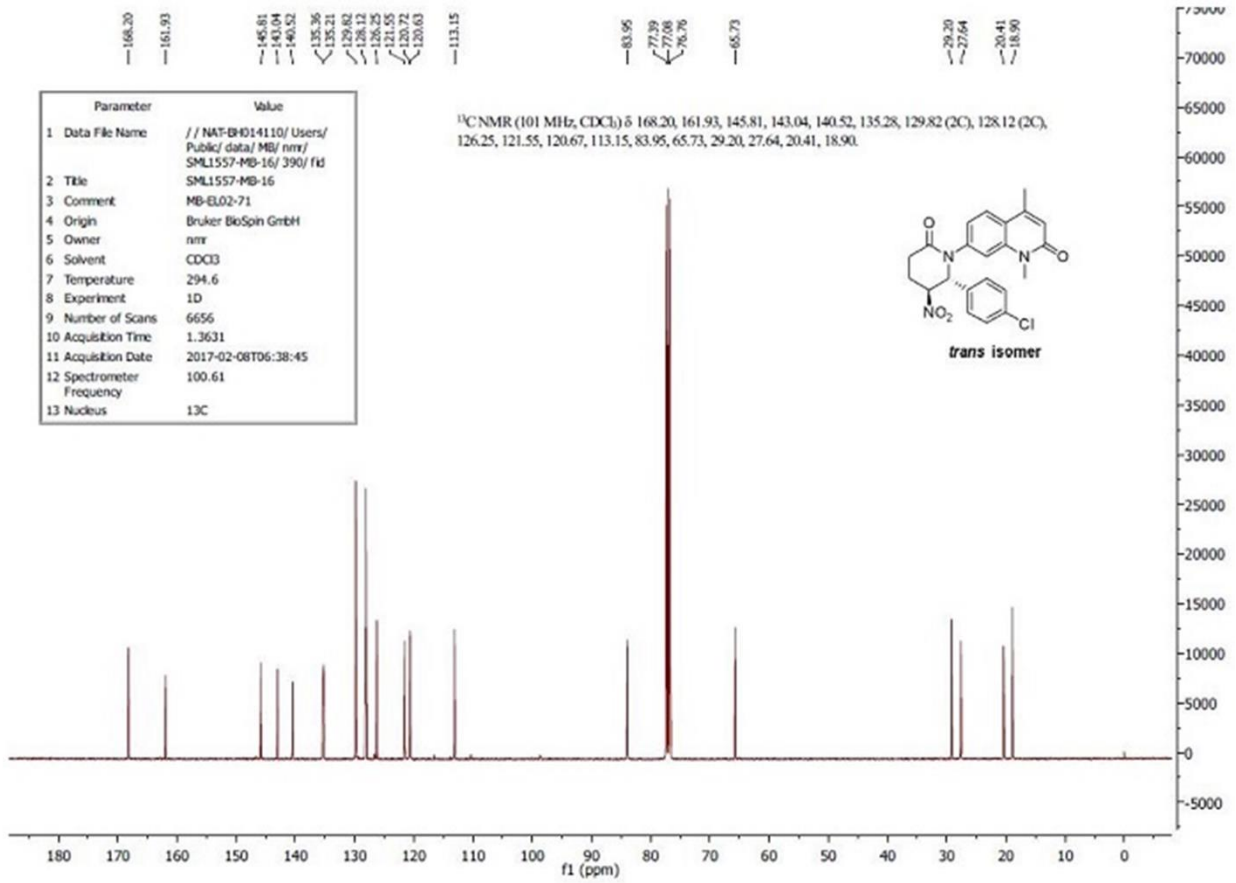


Figure S37. ^{13}C NMR spectrum of the *trans* isomer of compound **5**.

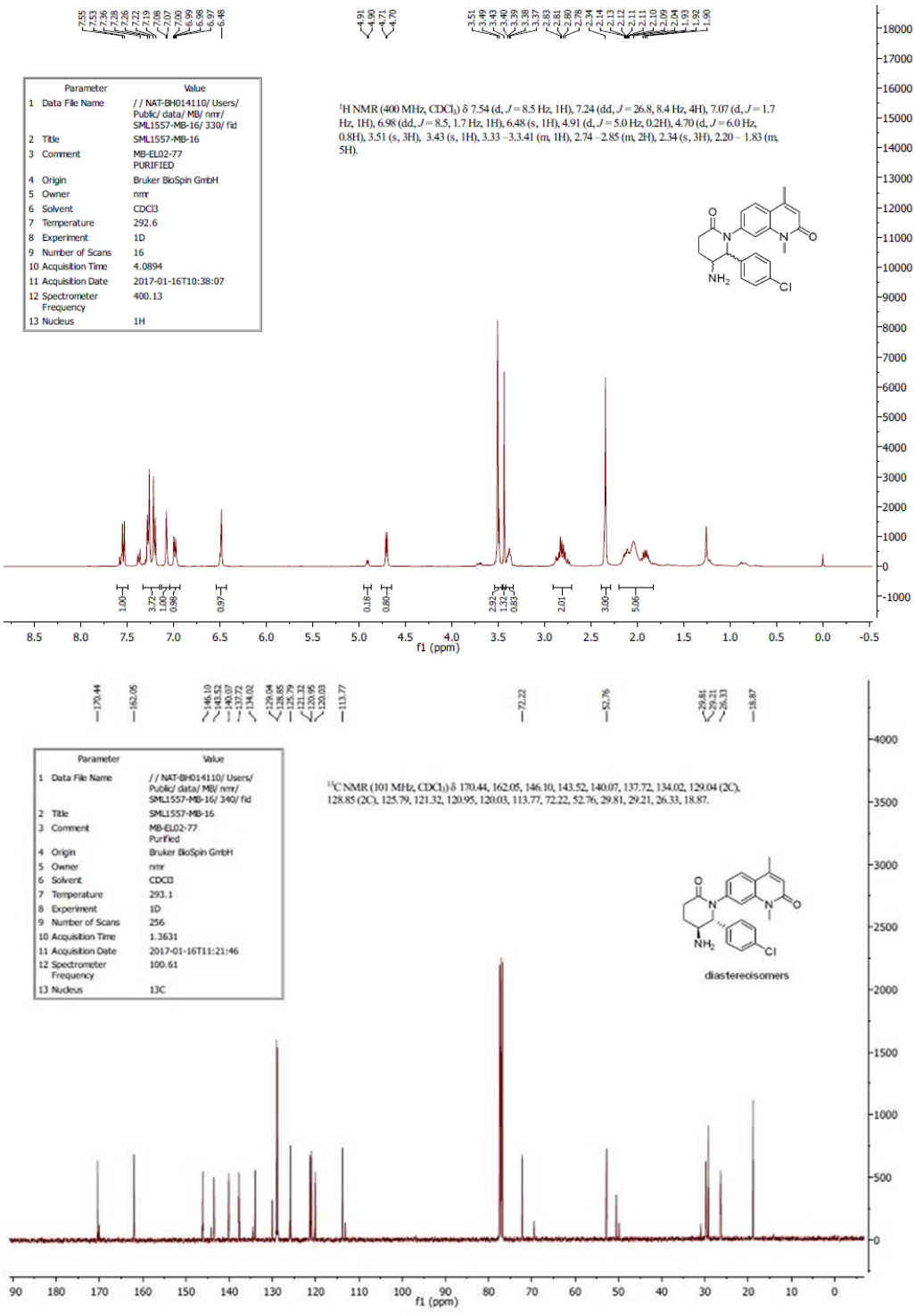


Figure S38. ^{1}H (top) and ^{13}C (bottom) NMR spectra of the mixture of diastereoisomers of compound **6**.

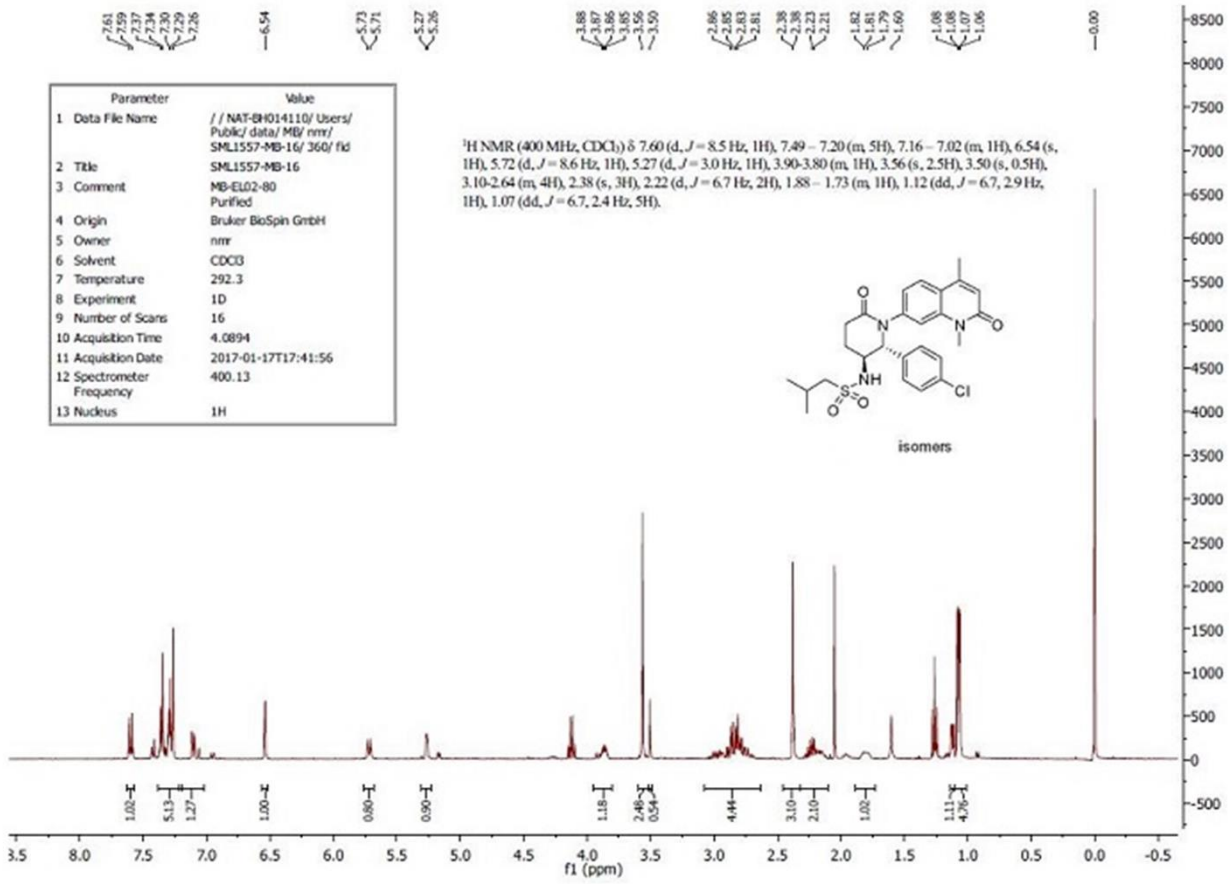


Figure S39. ^{1}H NMR spectra of compound **rac-8**.

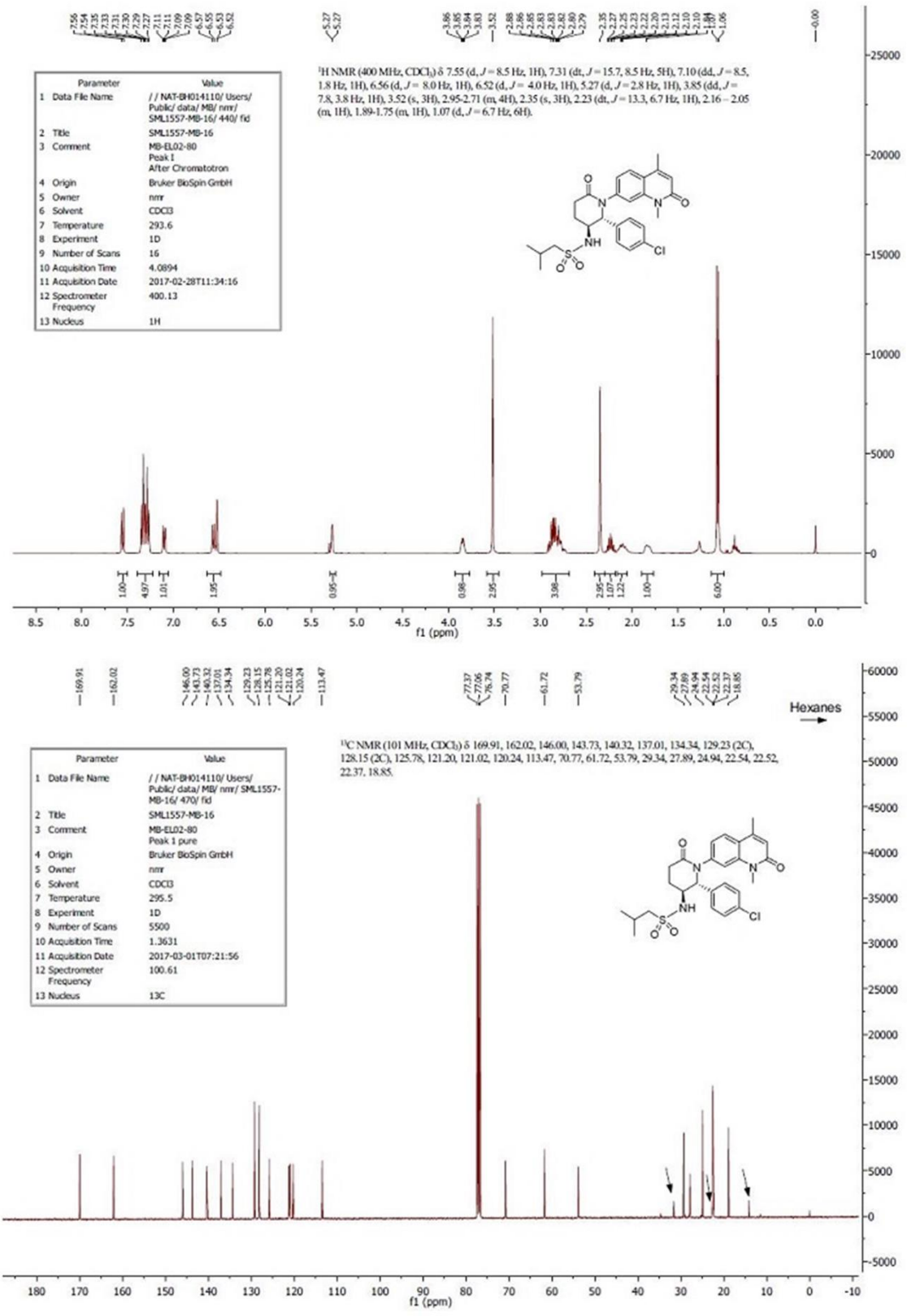
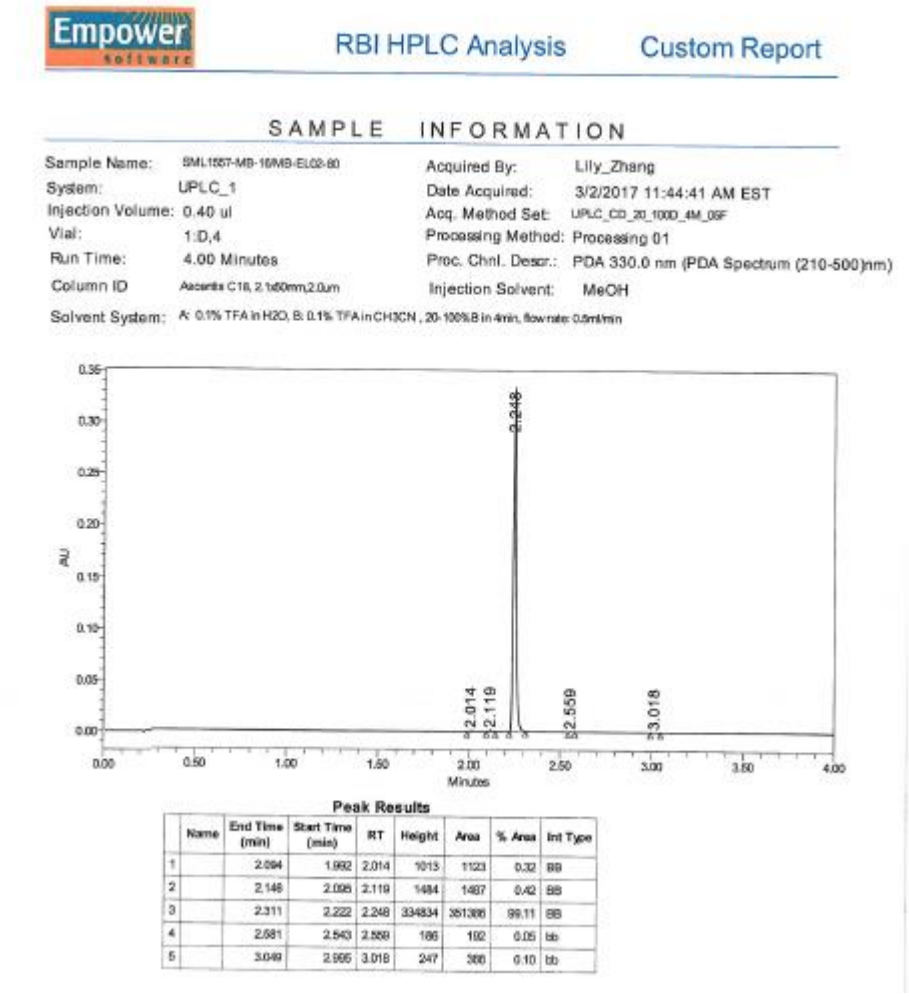


Figure S40. ^{1}H (top) and ^{13}C (bottom) NMR spectra of enantiopure compound **8**.



	Name	End Time (min)	Start Time (min)	RT	Height	Area	% Area	Int Type
1		2.094	1.992	2.014	1015	1123	0.32	BB
2		2.146	2.095	2.119	1484	1487	0.42	BB
3		2.311	2.222	2.248	334834	351386	99.11	BB
4		2.681	2.543	2.559	186	192	0.05	bb
5		3.049	2.995	3.018	247	366	0.10	bb

Figure S41. 99.11%, HPLC purity of compound **8**.

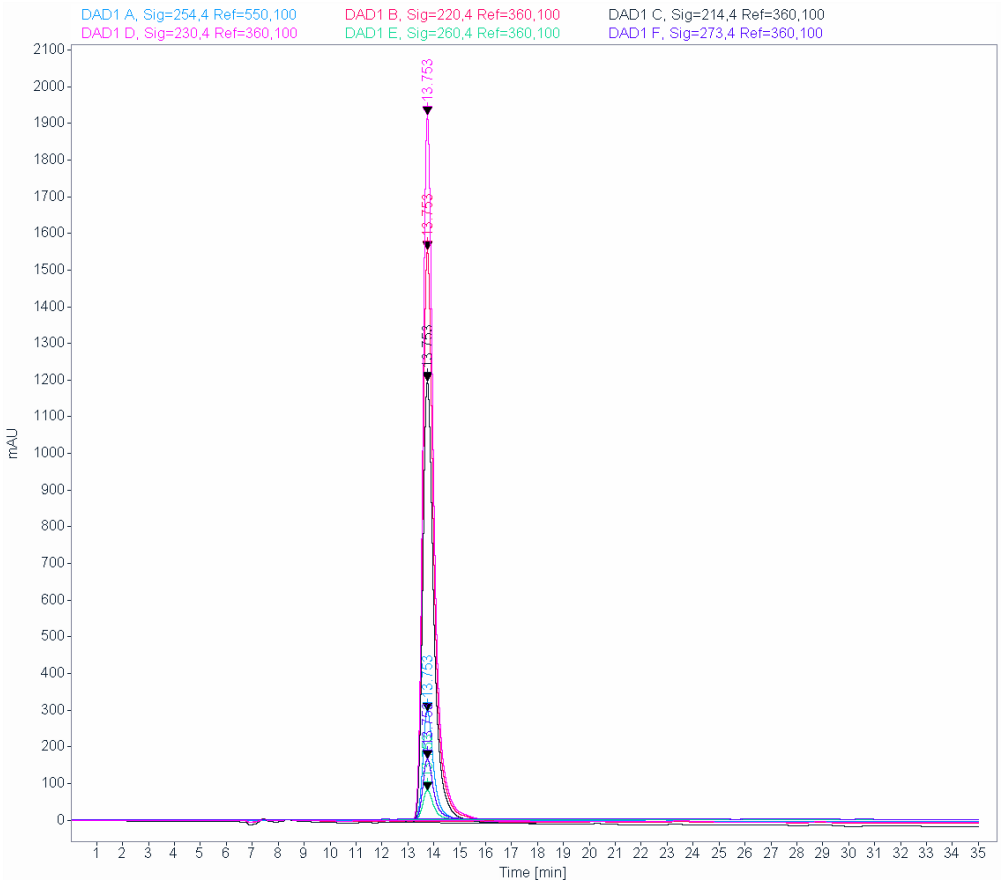
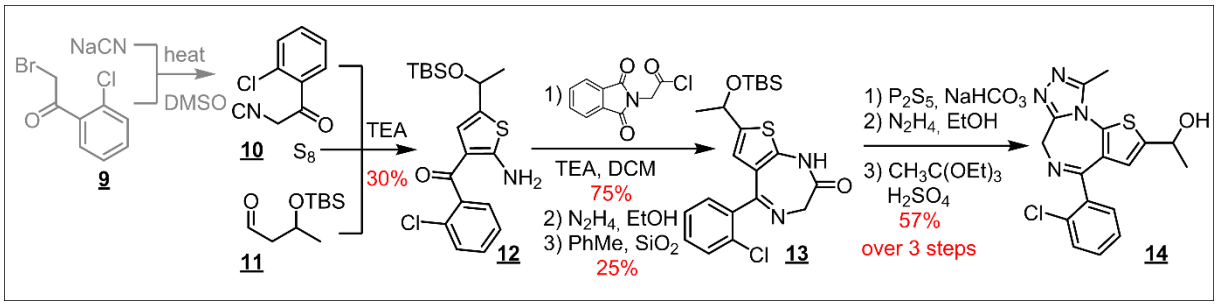


Figure S42. Chiral HPLC trace for ee purity of final product **8**.

Section S11. Synthesis of α-hydroxyetizolam, 14.

S11.1. The current synthetic route.

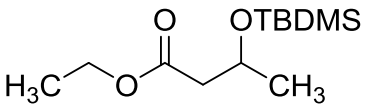


Scheme S2. Chematica-planned synthesis of α-Hydroxyetizolam (**14**); same as **Figure 2b**. Note there is no prior reported route to this compound.

S11.2. Synthetic details.

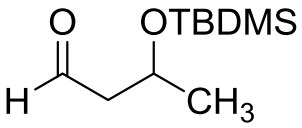
All reagents were used as received from vendors with no additional purification. Solvents were used as received in either Sigma Aldrich® Pure-Pac® II or Sure/Seal™ systems and were dispensed immediately prior to use. Intermediates and final products were purified using a CombiFlash RF system (Teledyne Isco).

Proton and 13-carbon NMR spectra were recorded on a JEOL ECS 400 (400 MHz) spectrometer at MilliporeSigma Round Rock. Chemical shifts are recorded in PPM on the δ scale and referenced to the sample solvent ($CHCl_3$ δ 7.26 and DMSO-d_6 δ 2.50) for ^{1}H-NMR and ($CHCl_3$ δ 77.2 and DMSO-d_6 δ 39.5) for ^{13}C-NMR. High resolution mass spectra (HRMS) were recorded on a Waters Xevo G2 QTof spectrometer using electrospray ion source (ESI) coupled with a Waters Acquity UPLC.



***Ethyl 3-[(tert-butyldimethylsilyl)oxy]butanoate* SI-2**
To a solution of ethyl 3-hydroxybutyrate (13.22 g, 0.100 mol, 1.0 equiv.) in DCM (240 mL) was added imidazole (13.60 g, 0.200 mol, 2.0 equiv.) and the solution was cooled to 0 °C. *tert*-Butyldimethylsilyl chloride (18.22 g, 0.120 mol, 1.2 equiv.) was added and the reaction was allowed to come to room temperature and stir for 12 h. Analysis by GC/FID confirmed the consumption of starting material. The reaction mixture was diluted with water (180 mL) and the aqueous phase was extracted with DCM (2 x 180 mL). The combined organic phases were washed with brine (100 mL), dried over Na_2SO_4, filtered and concentrated under reduced pressure. Purification of the crude product by MPLC (silica, 5-10% EtOAc/hexanes) afforded the silyl ether **SI-2** (25.58 g, quant.) as a clear, colorless liquid. Spectroscopic and physical data matched reported literature.

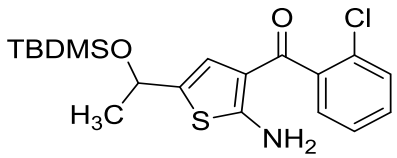
^{1}H-NMR (400 MHz, $CDCl_3$): 4.30-4.22 (m, 1H), 4.16-4.04 (m, 2H), 2.45 (dd, J = 14.2 Hz and J = 7.8 Hz, 1H), 2.34 (dd, J = 14.7 Hz and J = 5.5 Hz, 1H), 1.24 (t, J = 7.3 Hz, 3H), 1.18 (d, J = 6.0 Hz, 3H), 0.85 (s, 9H), 0.05 (s, 3H), 0.03 (s, 3H)
^{13}C-NMR (100 MHz, $CDCl_3$): 171.8, 66.0, 60.4, 45.1, 25.8, 24.0, 18.1, 14.3, -4.4, -5.0



***3-[(tert-butyldimethylsilyl)oxy]butanal* 11**

A solution of ethyl ester **SI-2** (25.58 g, 0.104 mol, 1.0 equiv.) in DCM (700 mL) was cooled to - 78 °C under N_2. DIBAL-H (110 mL, 1 M in hexanes, 0.109 mol, 1.05 equiv.) was added dropwise to the solution. The reaction was stirred at -78 °C for 1 h. Analysis by GC/FID confirmed the consumption of starting material. The reaction was quenched by adding methanol (35 mL) dropwise to the solution at -78 °C. The reaction mixture was then slowly poured into a stirring saturated solution of Rochelle's salt (700 mL) at 0 °C. The aqueous mixture was then allowed to warm to room temperature and stir for 2 h. The aqueous phase was separated and extracted with DCM (2 x 200 mL). The combined organic phases were then washed with brine (200 mL), dried over Na_2SO_4, filtered and concentrated under reduced pressure to afford aldehyde **11** (21.31 g, quant.) which was used directly in the next step. Spectroscopic and physical data matched reported literature.

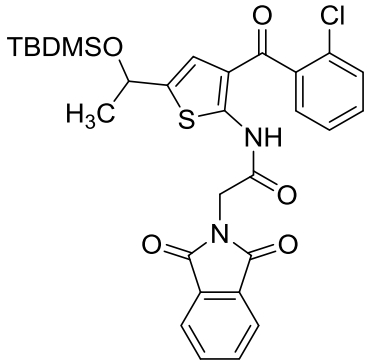
^{1}H-NMR (400 MHz, $CDCl_3$): 9.78 (br s, 1H), 4.38-4.30 (m, 1H), 2.54 (ddd, J = 16.0 Hz, J = 7.3 Hz, J = 2.8 Hz, 1H), 2.48 (ddd, J = 15.6 Hz, J = 7.3 Hz, J = 2.3 Hz, 1H), 1.22 (d, J = 6.4 Hz, 3H), 0.86 (s, 9H), 0.06 (s, 3H), 0.05 (s, 3H)
^{13}C-NMR (100 MHz, $CDCl_3$): 202.4, 64.6, 53.1, 25.8, 24.3, 18.0, -4.3, -4.9



***5-{1-[(tert-butyldimethylsilyl)oxy]ethyl}-3-(2-chlorobenzoyl)thiophen-2-amine* 12**

To a solution of aldehyde **11** (6.30 g, 0.0311 mol, 1.0 equiv.) in DMF (60 mL) was added 2-chlorobenzoylacetonitrile **10** (11.10 g, 0.0623 mol, 2.0 equiv.), sulfur (1.09 g, 0.0311 mol, 1.0 equiv.) and Et_3N (4.2 mL, 0.0311 mol, 1.0 equiv). The reaction was allowed to stir at room temperature for 4 h. Reaction progress was monitored by GC/FID (for consumption of aldehyde) and TLC (silica, 8:2 DCM/Hexanes, UV). An additional 2.75 g of 2-chlorobenzoylacetonitrile was added and stirred for an additional 3 h. The reaction was diluted with EtOAc (500 mL) and the organic phase was washed with water (2 x 200 mL), followed by brine (200 mL). The organic phase was then dried over Na_2SO_4, filtered and concentrated under reduced pressure. Purification of the crude product by MPLC (silica, 10-30% EtOAc/hexanes) afforded the thiophene **12** (3.64 g, 30%) as a light brown, viscous oil that eventually solidified into a yellow, waxy solid.

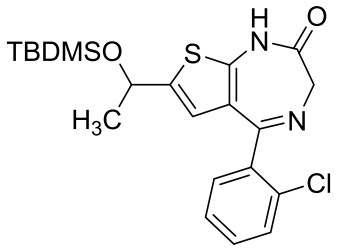
^{1}H-NMR (400 MHz, DMSO-d_6): 8.36 (s, 2H), 7.48 (dd, J = 7.8 Hz, J = 1.4 HZ, 1H), 7.44-7.35 (m, 2H), 7.30 (dd, J= 7.3 Hz, J = 1.8 Hz, 1H), 6.00 (d, J = 0.9 Hz, 1H), 4.81-4.77 (m, 1H), 1.22 (d, J = 6.4 Hz, 3H), 0.77 (s, 9H), -0.03 (s, 3H), -0.07(s, 3H)
^{13}C-NMR (100 MHz, DMSO-d_6): 187.4, 167.1, 140.9, 131.0, 130.4, 130.1, 129.6, 128.7, 127.7, 121.0, 113.2, 66.8, 26.4, 26.1, 18.3, -4.3, -4.6
HRMS (*m/z*): Calcd for $C_{19}H_{26}ClNO_2SSi$, $[M+H]^+$, 396.1220; found, 396.1216



***N*-(5-{1-[(tert-butyldimethylsilyl)oxy]ethyl}-3-(2-chlorobenzoyl)thiophen-2-yl)-2-(1,3-dioxo-2,3-dihydro-1H-isoindol-2-yl)acetamide SI-3**

To a solution of thiophene **12** (3.64 g, 0.0092 mol, 1.0 equiv) in DCM (75 mL) was added Et_3N (4.0 mL, 0.0276 mol, 3.0 eq) and the solution was cooled to 0 °C. *N*-phthaloylglycyl chloride (2.97 g, 0.0119 mol, 1.3 equiv) was added in portions. The reaction was then allowed to warm to room temperature and stir for 2 h. Reaction progress was monitored by TLC (silica, 3:7 EtOAc/Hexanes, UV). Volatiles were then concentrated under reduced pressure and the crude product was purified by MPLC (silica, 10-30% EtOAc/hexanes) to afford the *N*-phthaloylglycyl adduct **SI-3** (4.04 g, 75%) as a yellow solid.

^{1}H-NMR (400 MHz, $CDCl_3$): 12.22 (s, 1H), 7.91 (dd, J = 5.5 Hz and J = 2.8 Hz, 2H), 7.75 (dd, J = 6.0 Hz and J = 3.2 Hz, 2H), 7.46-7.31 (m, 4H), 6.43 (d, J = 0.9 Hz, 1H), 4.91-4.86 (m, 1H), 4.69 (s, 2H), 1.39 (d, J = 6.4 Hz, 3H), 0.83 (s, 9H), 0.01 (s, 3H), -0.03 (s, 3H)
^{13}C-NMR (100 MHz, $CDCl_3$): 191.6, 167.6, 164.2, 150.0, 141.6, 139.2, 134.5, 132.1, 131.1, 130.7, 130.2, 128.5, 126.8, 123.9, 120.8, 120.0, 66.9, 40.9, 26.8, 25.8, 18.2, -4.8, -4.9
HRMS (*m/z*): Calcd for $C_{29}H_{31}ClN_2O_5SSi$, $[M+H]^+$, 583.1490; found, 583.1487



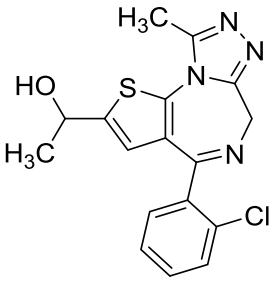
***7*-{1-[(tert-butyldimethylsilyl)oxy]ethyl}-5-(2-chlorophenyl)-1H,2H,3H-thieno[2,3-e][1,4]diazepin-2-one 13**

To a solution of **SI-3** (4.04 g, 0.00693 mol, 1.0 equiv) in EtOH (60 mL) was added hydrazine monohydrate (0.65 mL, 0.0118 mol, 1.7 equiv.) and the reaction was heated at reflux for 2.5 h. The reaction was monitored by TLC (silica, 1:1 EtOAc/Hexanes, UV) and LC/MS. The reaction was then allowed to cool to room temperature and the precipitated solids were removed by vacuum filtration and washed with EtOH. The filtrate was concentrated under reduced pressure to afford an amine intermediate with ~11% diazepine **13** as determined by LC/MS. The amine intermediate was suspended in toluene (80 mL) and silica (6.7 g) was added. The reaction was then heated at reflux for 30 h. The reaction was monitored by TLC (silica, 1:1 EtOAc/Hexanes, UV) and LC/MS. The reaction was allowed to cool and the silica was removed by vacuum filtration and washed with EtOAc. The filtrate was concentrated under reduced pressure. Purification of the crude product by MPLC (silica, 20-60% EtOAc/hexanes) afforded the diazepine **13** (0.76 g, 25%) as a tan solid.

^{1}H-NMR (400 MHz, $CDCl_3$): 7.45-7.30 (m, 4H), 6.24 (d, J = 0.9 Hz, 1H), 4.91-4.86 (m, 1H), 4.46 (s, 2H), 1.40 (d, J = 6.0 Hz, 3H), 0.86 (s, 9H), 0.04 (s, 3H), 0.00 (s, 3H)

^{13}C-NMR (100 MHz, $CDCl_3$): 167.5, 166.7, 144.0, 143.0, 138.2, 133.1, 130.8, 130.7, 130.0, 128.3, 127.0, 126.6, 120.1, 66.9, 57.5, 26.8, 25.8, 18.2, -4.7, -4.9
HRMS (*m/z*): Calcd for $C_{21}H_{27}ClN_2O_2SSi$, $[M+H]^+$, 435.1329; found, 435.1330

Note: Multiple conditions are reported for this cyclization step. Heating in toluene in the presence of silica described above proved to be the mildest reaction condition and resulted in the most reproducible isolated yields



***1-[7-(2-chlorophenyl)-13-methyl-3-thia-1,8,11,12-tetraazatricyclo[8.3.0.0², ⁶]trideca-2(6),4,7,10,12-pentaen-4-yl]ethan-1-ol* 14**
To a solution of diazepine **13** (0.76 g, 0.00175 mol, 1.0 equiv.) in diglyme (8 mL) was added $NaHCO_3$ (0.35 g, 0.00349 mol, 2.0 equiv.) and P_2S_5 (0.43 g, 0.00175 mol, 1.0 equiv.). The reaction was heated at 80 °C for 2 h. The reaction was monitored by TLC (silica, 1:1 EtOAc/Hexanes, UV) and LC/MS. The reaction was allowed to cool to room temperature and water (50 mL) was added. The aqueous layer was extracted with EtOAc (2 x 40 mL). The combined organic layer was washed with brine (15 mL), dried over Na_2SO_4, filtered and concentrated under reduced pressure. The crude product was taken on to the next step directly.

^{1}H-NMR (400 MHz, $CDCl_3$): 7.45-7.42 (m, 1H), 7.37-7.30 (m, 2H), 6.25 (d, J = 1.1 Hz, 1H), 4.93-4.86 (m, 3H), 1.40 (d, J = 6.1 Hz, 3H), 0.86 (s, 9H), 0.04 (s, 3H), 0.01 (s, 3H)
HRMS (*m/z*): Calcd for $C_{21}H_{27}ClN_2OS_2Si$, $[M+H]^+$, 451.1101; found, 451.1097

To a solution of crude material isolated from the procedure above in EtOH (20 mL) was added hydrazine monohydrate (0.40 mL, 00699 mol, 4.0 equiv.) and the reaction was stirred at room temperature for 1 h. The reaction was monitored by TLC (silica, 1:1 EtOAc/Hexanes, UV) and LC/MS. The solvent was concentrated under reduced pressure and the residue was dissolved in 1:1 toluene/EtOH (20 mL). Triethyl orthoacetate (1.0 mL, 0.00524 mol, 3.0 equiv) was added followed by the dropwise addition of H_2SO_4 (0.4 mL). The reaction was stirred at room temperature for 18 h. The reaction was monitored by TLC (silica, 1:1 EtOAc/Hexanes, UV) and LC/MS. 10% Na_2CO_3 was added (60 mL) and the aqueous layer was extracted with EtOAc (2 x 40 mL). The combined organic layer was washed with brine (20 mL), dried over Na_2SO_4, filtered and concentrated under reduced pressure. Purification of the crude product by MPLC (silica, 0-10% MeOH/EtOAc) afforded α-hydroxyetizolam **14** (0.355 g, 57%) as an off-white solid.

^{1}H-NMR (400 MHz, $CDCl_3$): 7.44 (m, 1H), 7.38-7.31 (m, 3H), 6.49 (s, 1H), 5.06 (dd, J = 9.4 Hz and J = 6.0 Hz, 1H), 4.91 (br d, J = 12.8 Hz, 2H), 2.70 (s, 3H), 1.56 (d, J = 6.4 Hz, 3H)
^{13}C-NMR (100 MHz, $CDCl_3$): 165.7, 153.1, 149.8, 146.4, 138.1, 134.7, 132.7, 131.1, 130.7, 130.2, 129.7, 127.2, 121.7, 66.1, 47.1, 25.4, 12.3
HRMS (*m/z*): Calcd for $C_{17}H_{15}ClN_4OS$, $[M+H]^+$, 359.0733; found, 359.0734

S11.3. Raw spectroscopic and chromatographic data.

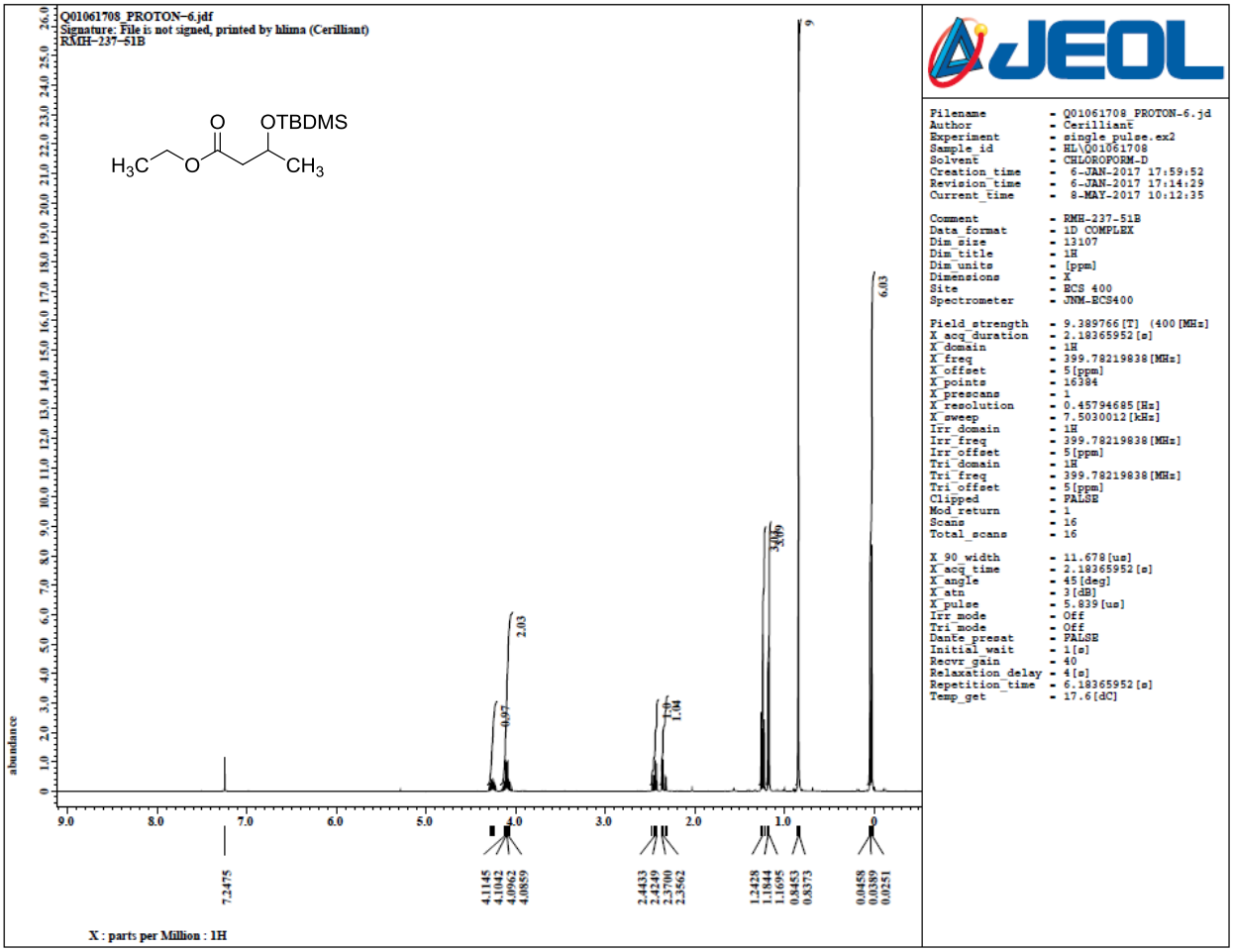


Figure S43. ^{1}H NMR of Ethyl 3-[(tert-butyldimethylsilyl)oxy]butanoate **SI-2**.

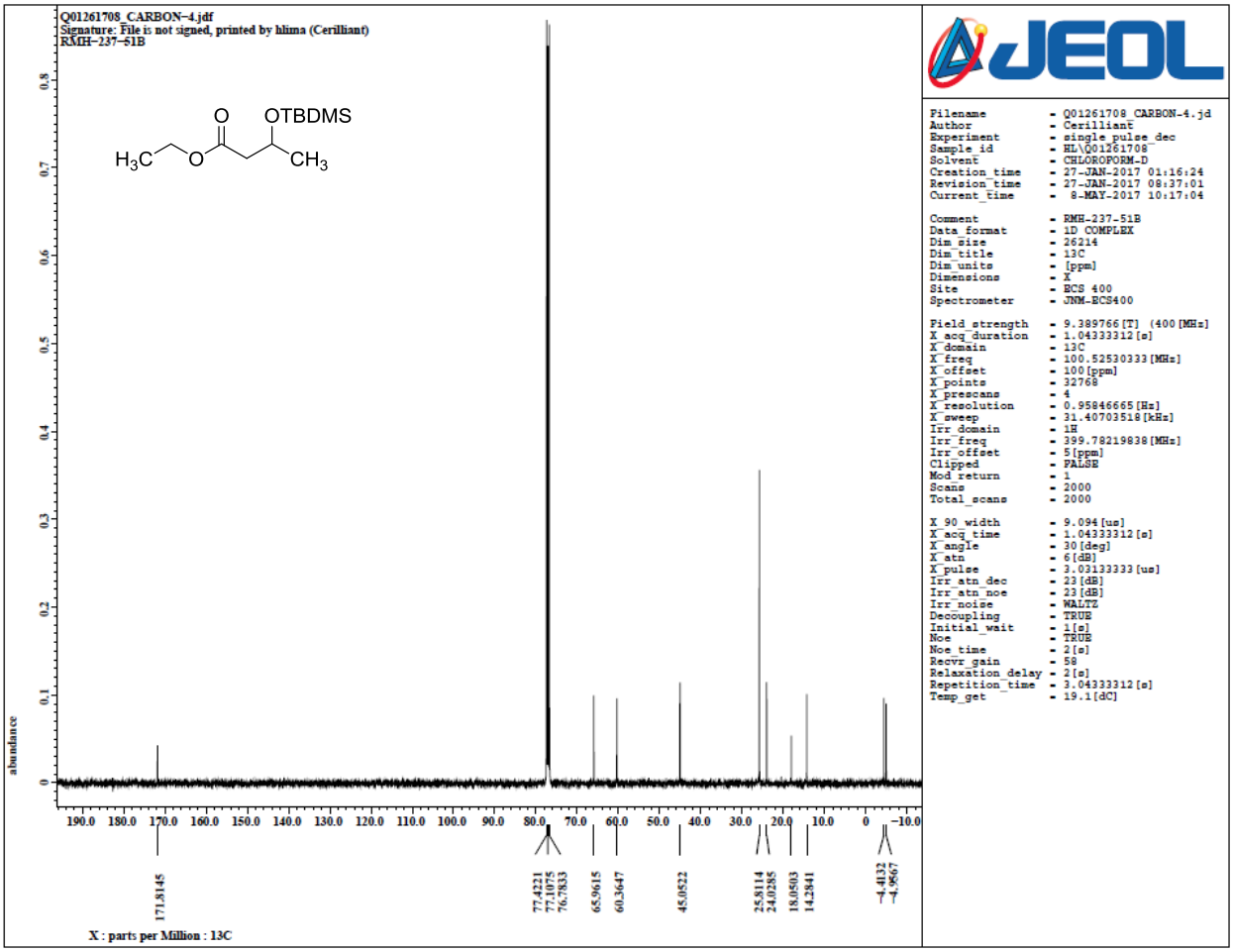


Figure S44. ^{13}C NMR of Ethyl 3-[(tert-butyldimethylsilyl)oxy]butanoate **SI-2**.

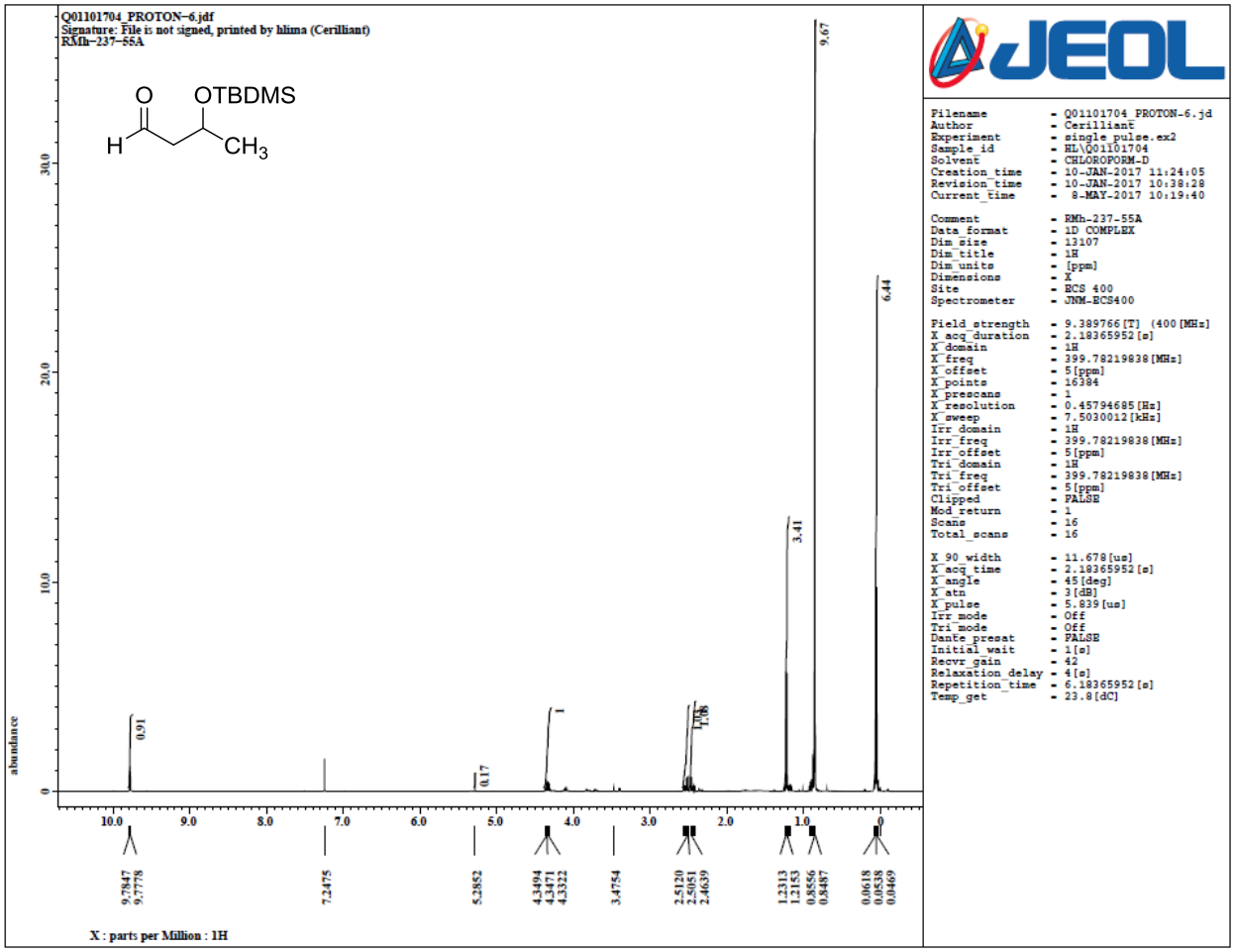


Figure S45. ^{1}H NMR of 3-[(tert-butyldimethylsilyl)oxy]butanal **11**.

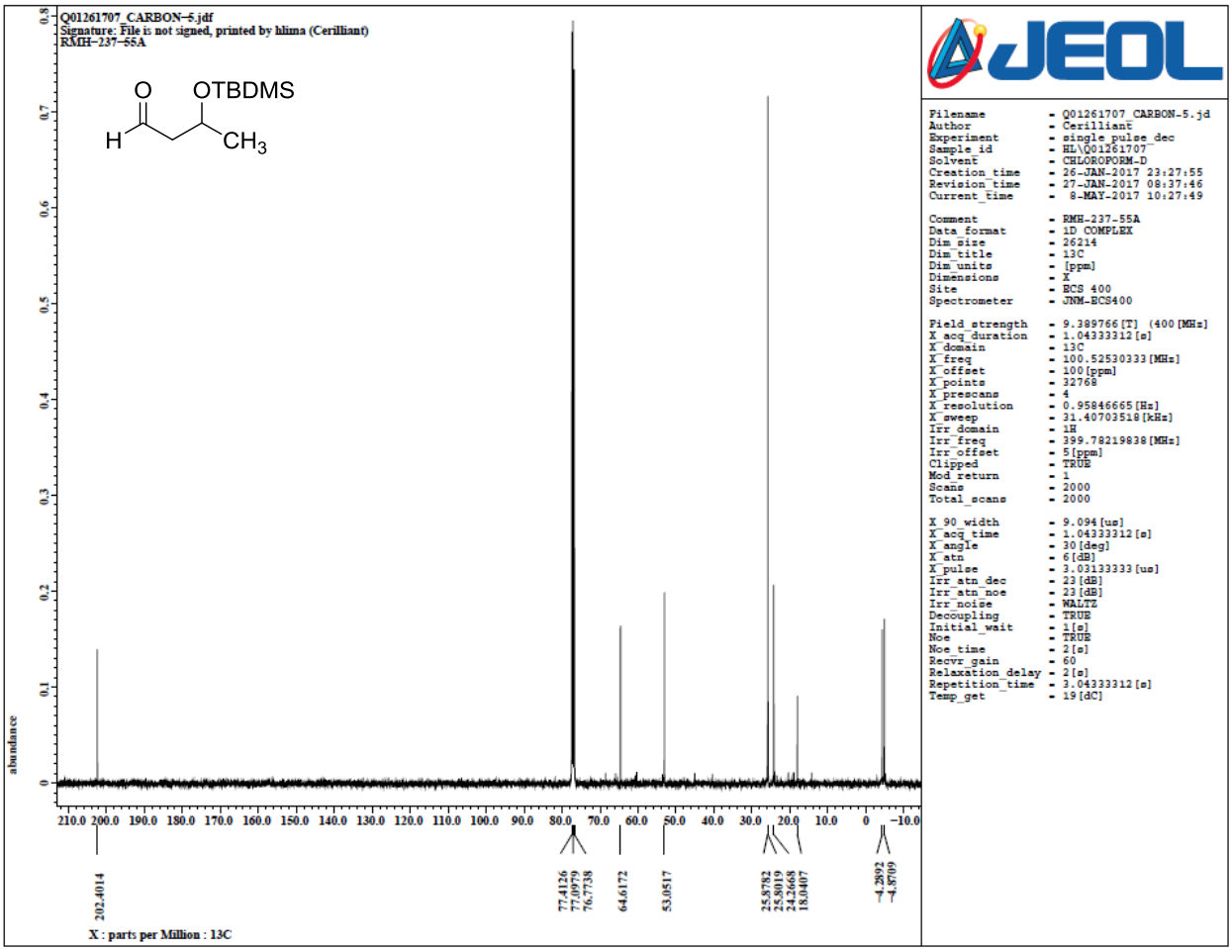


Figure S46. ^{13}C NMR of 3-[(tert-butyldimethylsilyl)oxy]butanal **11**.

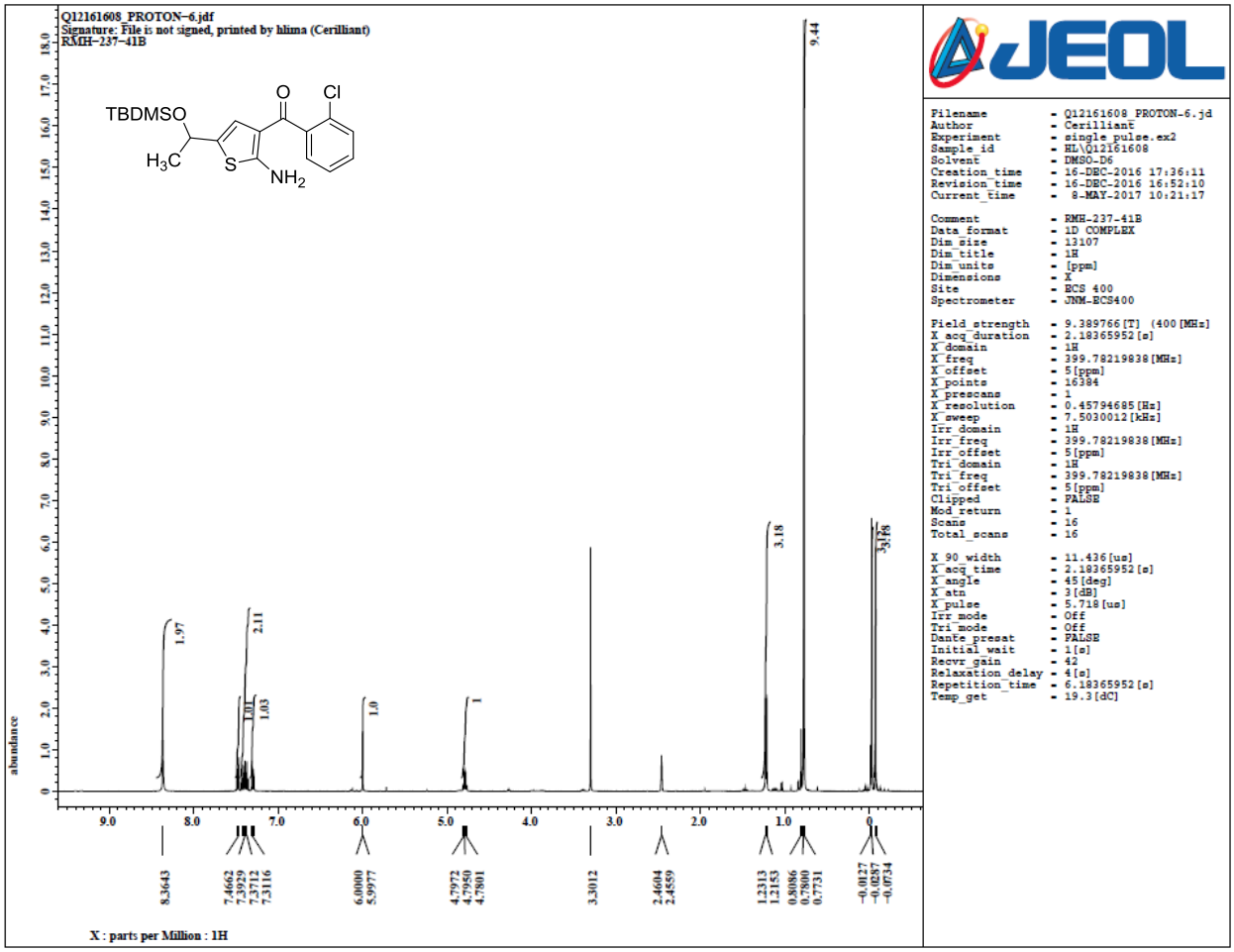


Figure S47. ^{1}H NMR of 5-{1-[(tert-butyldimethylsilyl)oxy]ethyl}-3-(2-chlorobenzoyl)thiophen-2-amine **12**.

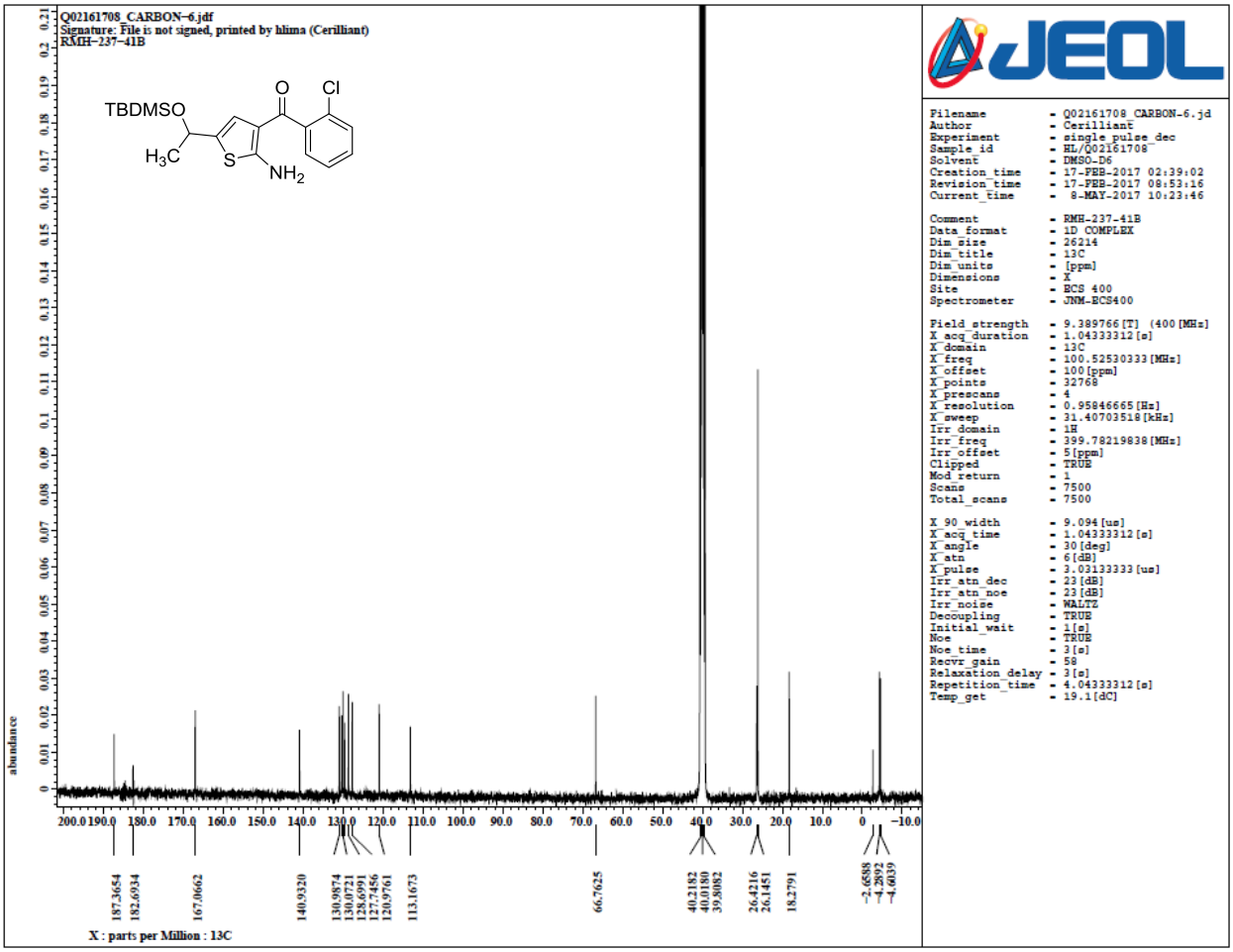


Figure S48. ^{13}C NMR of 5-{1-[(tert-butyldimethylsilyl)oxy]ethyl}-3-(2-chlorobenzoyl)thiophen-2-amine **12**.

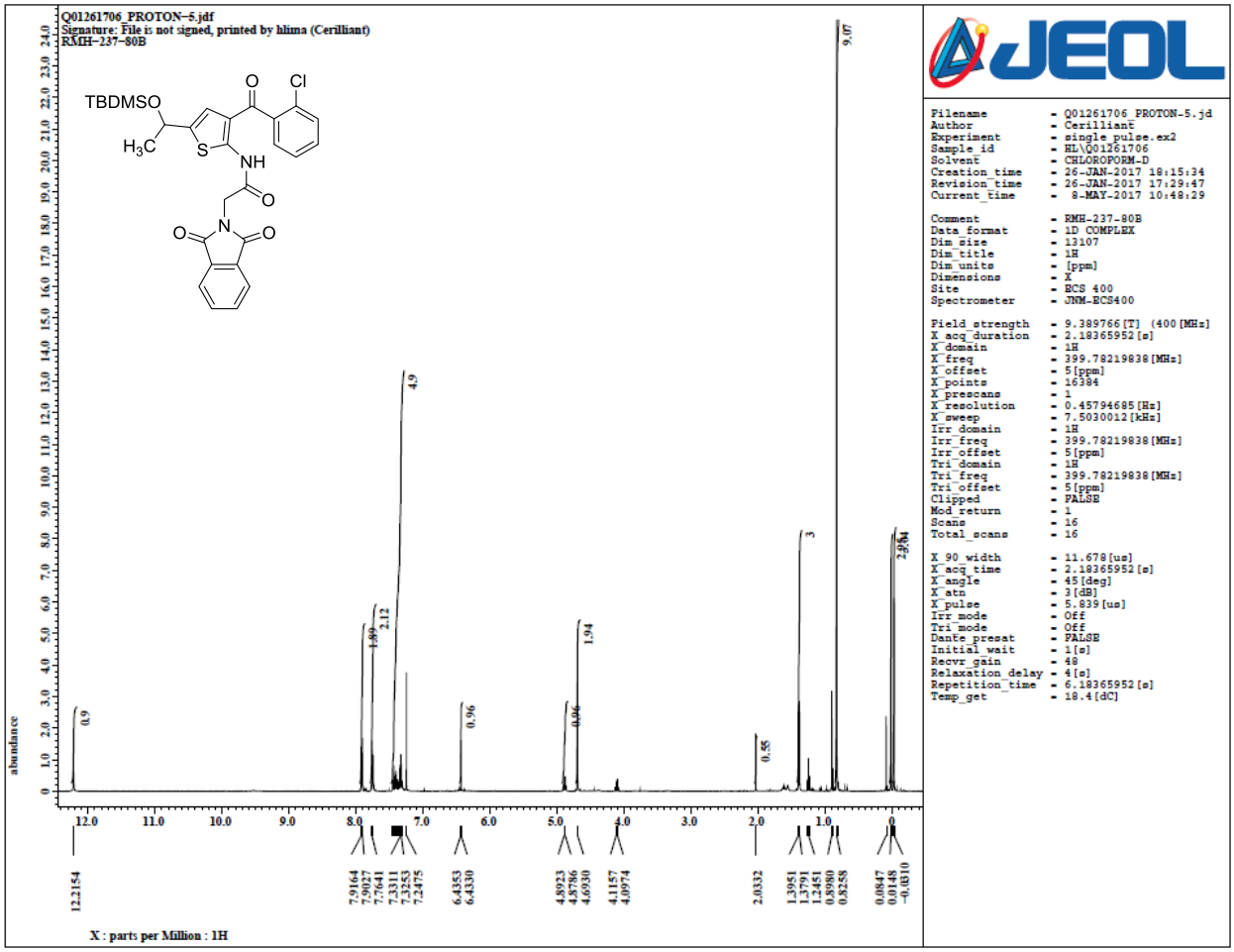


Figure S49. ^{1}H NMR of *N*-(5-{1-[(tert-butyldimethylsilyl)oxy]ethyl}-3-(2-chlorobenzoyl)thiophen-2-yl)-2-(1,3-dioxo-2,3-dihydro-1*H*-isoindol-2-yl)acetamide **SI-3**.

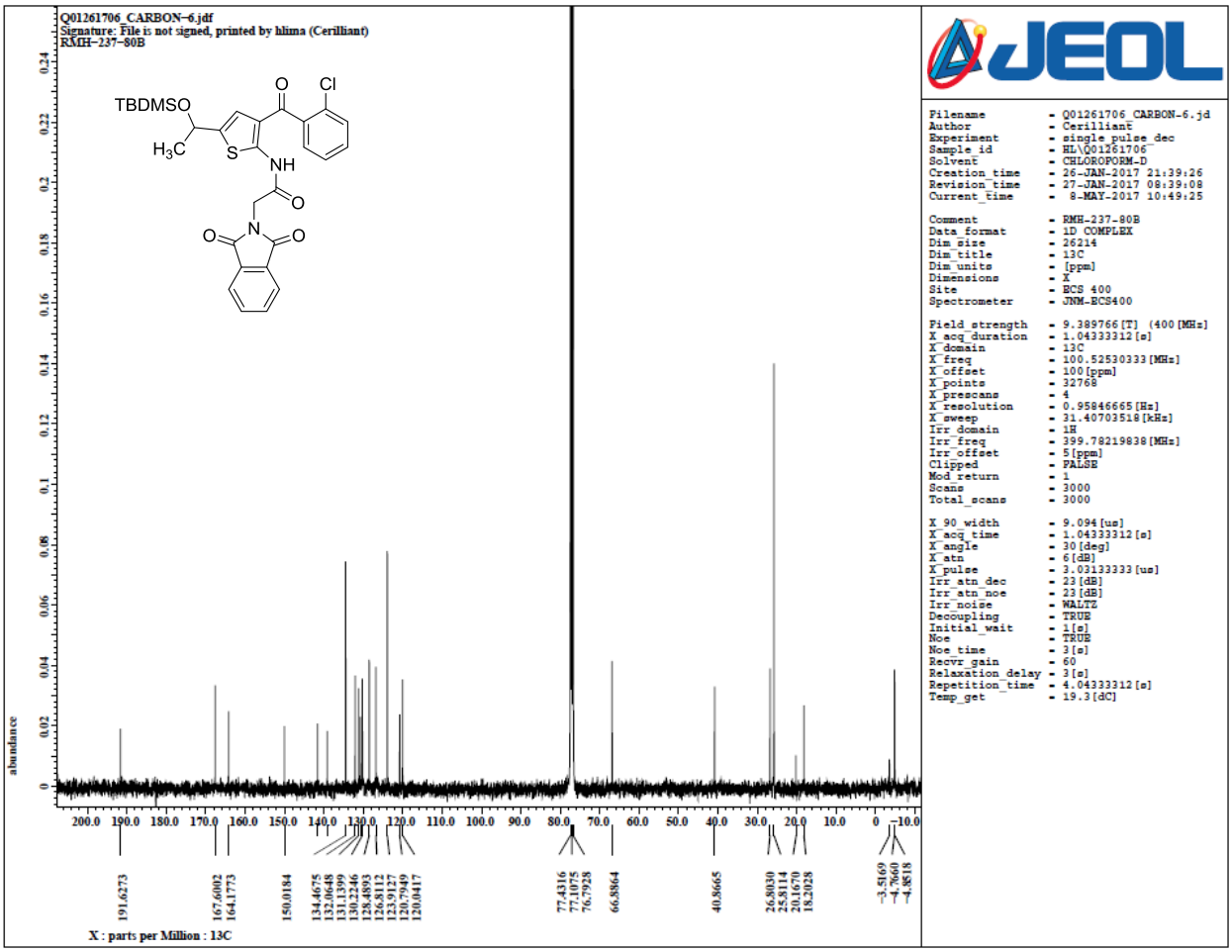


Figure S50. ^{13}C NMR of *N*-(5-{1-[(tert-butyldimethylsilyl)oxy]ethyl}-3-(2-chlorobenzoyl)thiophen-2-yl)-2-(1,3-dioxo-2,3-dihydro-1*H*-isoindol-2-yl)acetamide **SI-3**.

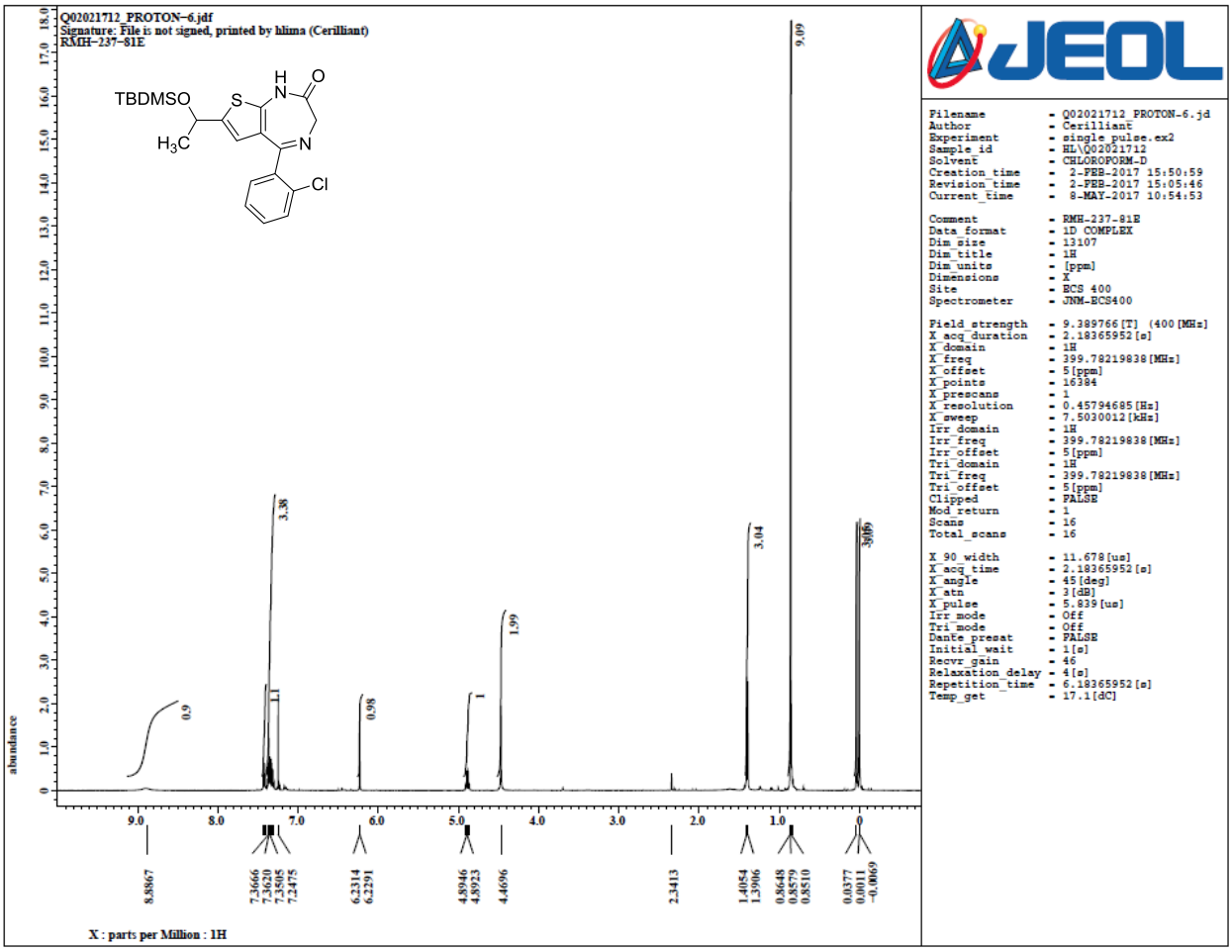


Figure S51. ^{1}H NMR of 7-{1-[(tert-butyldimethylsilyl)oxy]ethyl}-5-(2-chlorophenyl)-1*H*,2*H*,3*H*-thieno[2,3-*e*][1,4]diazepin-2-one **13**.

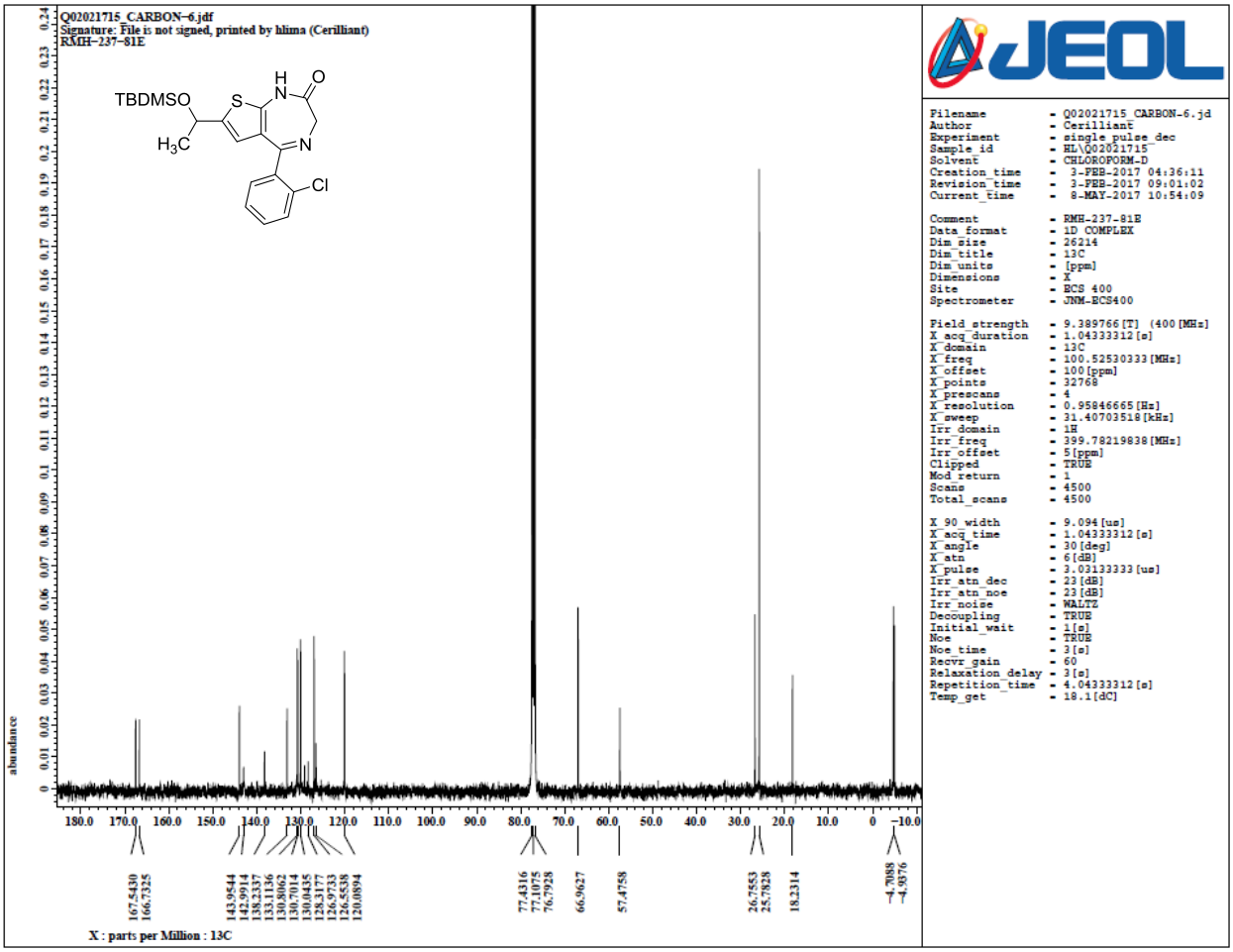


Figure S52. ^{13}C NMR 7-{1-[(tert-butyldimethylsilyl)oxy]ethyl}-5-(2-chlorophenyl)-1*H*,2*H*,3*H*-thieno[2,3-*e*][1,4]diazepin-2-one **13**.

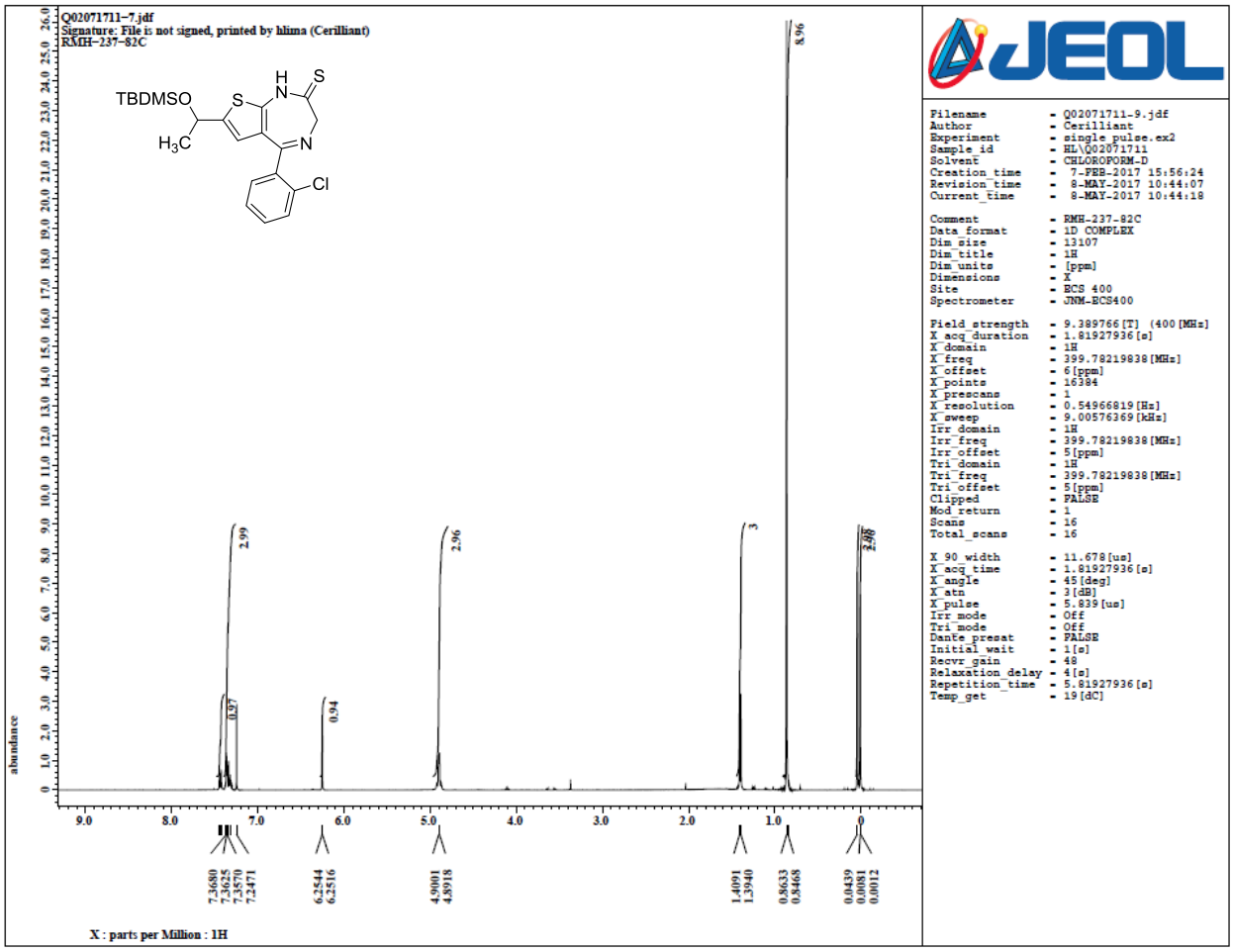


Figure S53. ^{1}H NMR of intermediate from synthesis of **14**.

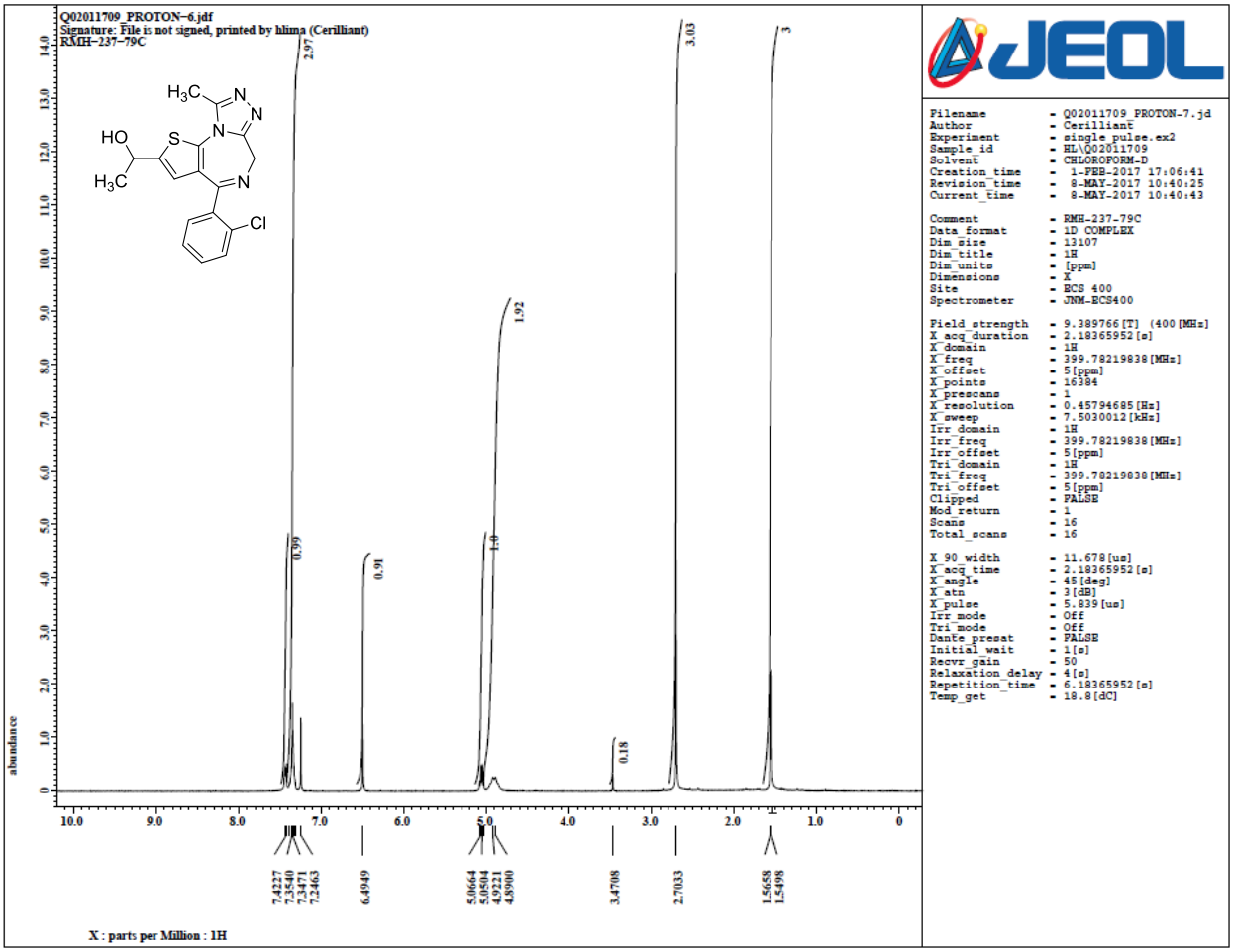


Figure S54. ^{1}H NMR of 1-[7-(2-chlorophenyl)-13-methyl-3-thia-1,8,11,12-tetraazatricyclo[8.3.0.0²,6] trideca-2(6),4,7,10,12-pentaen-4-yl]ethan-1-ol **14**.

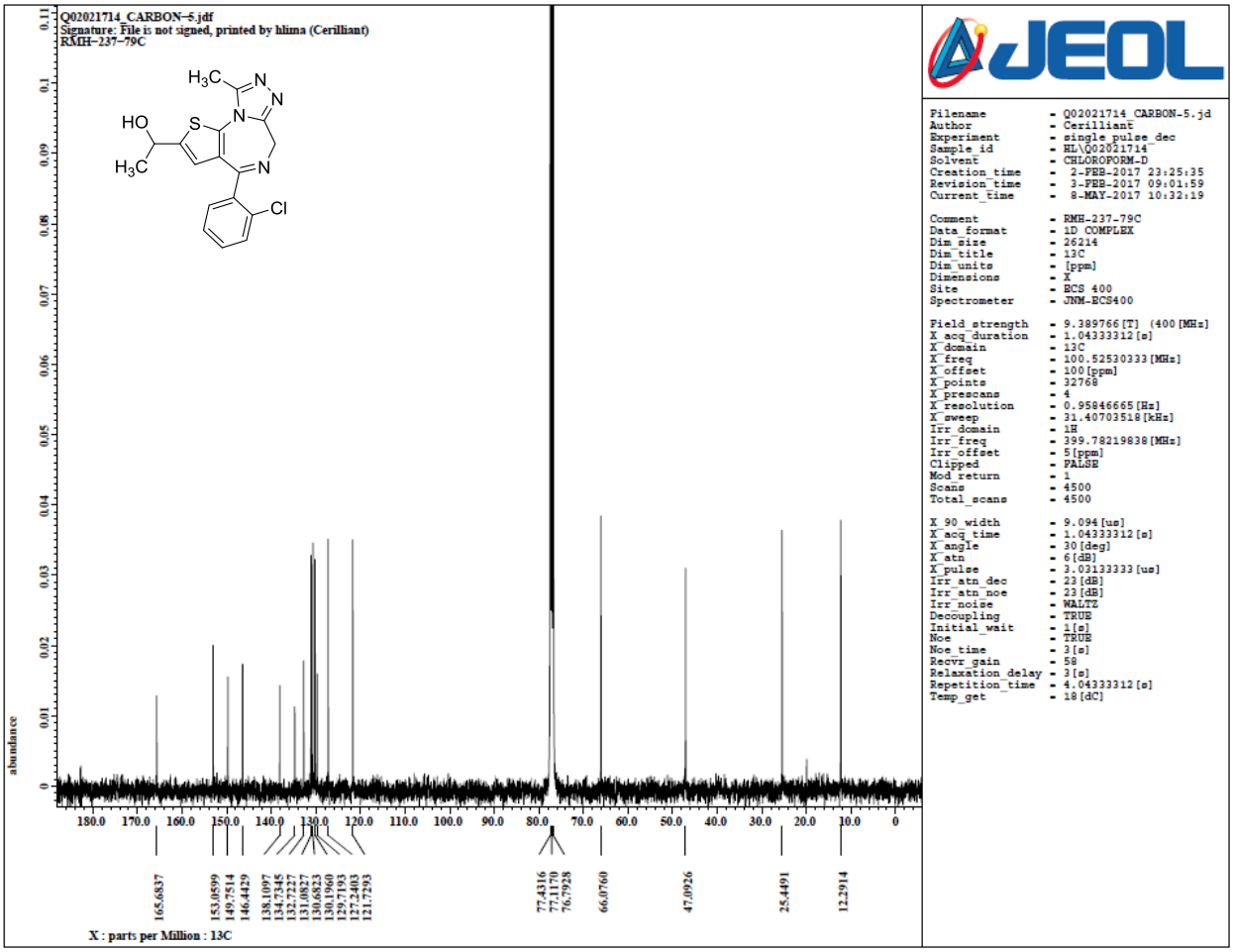
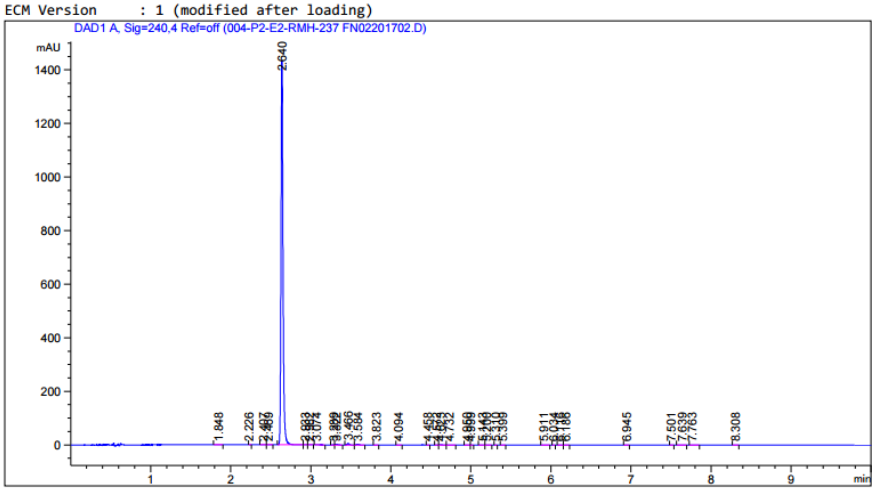


Figure S55. ^{13}C NMR of 1-[7-(2-chlorophenyl)-13-methyl-3-thia-1,8,11,12-tetraazatricyclo[8.3.0.0²,6] trideca-2(6),4,7,10,12-pentaen-4-yl]ethan-1-ol **14**.

ECM Path : \LAB\HPLC\LC 14\2017\RMH-237 P LC 14 2017-03-30 16-59-50.SC.SSIzip
ECM Version : 1 (modified after loading)



Area Percent Report

Sorted By : Signal
Multiplier : 1.0000
Dilution : 1.0000
Do not use Multiplier & Dilution Factor with ISTDs

Signal 1: DAD1 A, Sig=240,4 Ref=off

Peak #	RetTime [min]	Type	Width [min]	Area [mAU*s]	Height [mAU]	Area %
1	1.848	BB	0.0331	4.03446e-1	1.51260e-1	0.0159
2	2.226	BB	0.0125	3.47907e-2	3.68257e-2	1.368e-3
3	2.407	BV	0.0270	6.91236e-1	3.57882e-1	0.0272
4	2.469	VB	0.0326	2.08897e-1	8.51036e-2	8.216e-3
5	2.640	BB	0.0265	2518.30225	1432.73157	99.0466
6	2.933	BV	0.0232	8.59623e-1	5.67399e-1	0.0338
7	2.982	VB	0.0251	8.82799e-1	5.38925e-1	0.0347
8	3.074	BB	0.0274	9.22968e-1	5.27539e-1	0.0363
9	3.289	BV	0.0161	4.32168e-1	3.71792e-1	0.0170
10	3.322	VB	0.0250	4.12635	2.47327	0.1623
11	3.466	BB	0.0231	6.77998	4.50418	0.2667
12	3.584	BV	0.0236	3.26067	2.10749	0.1282
13	3.823	BB	0.0292	1.03139e-1	4.58635e-2	4.057e-3
14	4.094	BB	0.0304	1.24648e-1	5.02794e-2	4.902e-3
15	4.458	BB	0.0201	1.56450e-1	1.09783e-1	6.153e-3
16	4.572	BV	0.0230	3.83405e-1	2.44873e-1	0.0151
17	4.623	VB	0.0255	1.51624	8.84017e-1	0.0596
18	4.732	BB	0.0250	5.44802e-1	3.09298e-1	0.0214
19	4.950	BV	0.0286	1.32703e-1	5.60643e-2	5.219e-3
20	4.999	VB	0.0171	7.41958e-2	5.47713e-2	2.918e-3
21	5.143	BV	0.0286	2.92530e-1	1.32239e-1	0.0115
22	5.200	VB	0.0230	4.44234e-1	2.40549e-1	0.0175
23	5.310	BB	0.0151	4.01367e-2	3.32908e-2	1.579e-3
24	5.399	BB	0.0239	7.54686e-2	3.92732e-2	2.968e-3
25	5.911	BB	0.0436	1.94116e-1	5.50433e-2	7.635e-3
26	6.034	BB	0.0184	4.42144e-2	3.05085e-2	1.739e-3
27	6.116	BV	0.0258	1.85818e-1	8.65086e-2	7.308e-3
28	6.186	VB	0.0259	1.92480e-1	9.00482e-2	7.570e-3
29	6.945	BB	0.0257	1.08586e-1	5.29223e-2	4.271e-3
30	7.501	BB	0.0190	6.36186e-2	4.23301e-2	2.502e-3
31	7.639	VB	0.0353	1.92993e-1	6.60863e-2	7.591e-3
32	7.763	BB	0.0310	2.58797e-1	1.06005e-1	0.0102
33	8.308	BB	0.0255	5.08318e-1	2.42079e-1	0.0200
Totals :				2542.54206	1447.42506	

*** End of Report ***

Column: Ascentis Phenyl C8, 2.7 µm, 3.0 x 100 mm
Mobile Phase: A Acetonitrile
B 0.1% Ammonium acetate

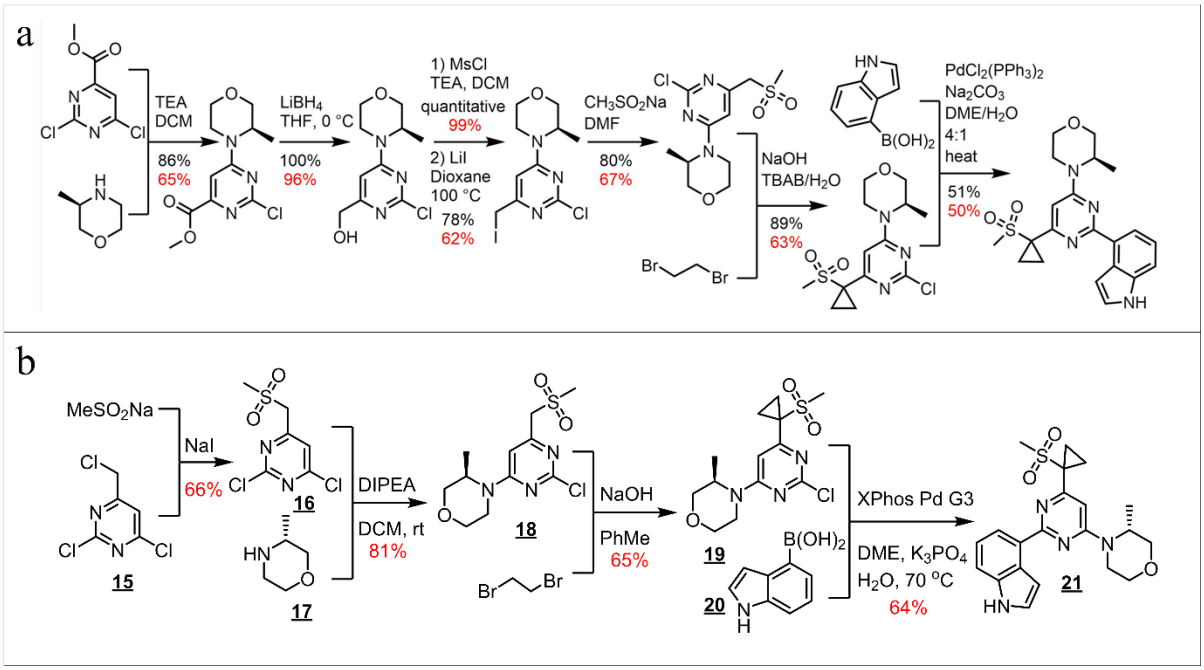
Gradient: Time (min)	%A	%B
0.0	20	80
8.0	85	15
9.0	85	15
9.1	20	80

Flow Rate: 0.7 mL/min
Wavelength: 240 nm
Temperature: 35 °C

Figure S56. HPLC of 1-[7-(2-chlorophenyl)-13-methyl-3-thia-1,8,11,12-tetraazatricyclo[8.3.0.0²,6] trideca-2(6),4,7,10,12-pentaen-4-yl]ethan-1-ol **14**.

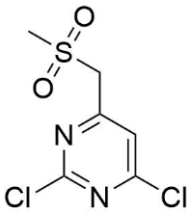
Section S12. Synthesis of ATR kinase inhibitor, 21.

S12.1. Previous vs. current synthetic routes.



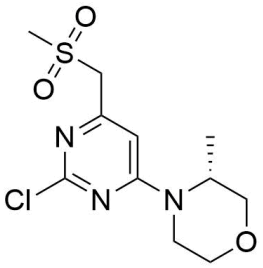
Scheme S3. (a) The original, seven-step preparation of **21** from the main-text reference[25]. Yields in black fonts are from the original reference which involved; yields in red fonts are from previous numerous attempts at Sigma-Aldrich. For comparison, **(b)** shows the Chematica route (same as in the main-text **Figure 2c**).

S12.2. Synthetic details.



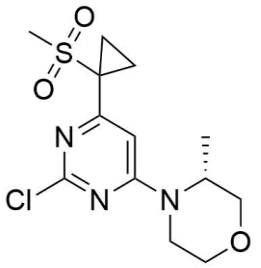
***2,4-dichloro-6-[(methylsulfonyl)methyl]pyrimidine* 16**

Solution of 2,4-dichloro-6-(chloromethyl)pyrimidine **15**, (2.00 g, 10.1 mmol), sodium methanesulfinate (3.10 g, 30.4 mmol), and sodium iodide (1.52 g, 10.1 mmol) in 20 mL of DMF was stirred at 55 °C for 4 h. At this time ^{1}H NMR analysis of reaction sample indicated complete conversion. The reaction mixture was cooled to room temperature and concentrated *in vacuo*. The residue was triturated with 10 mL of 50% aqueous methanol to give white solid, which was collected by filtration and dried under vacuum to afford 2,4-dichloro-6-[(methylsulfonyl)methyl]pyrimidine **16** (1.61g, 66%). ^{1}H NMR (400 MHz, DMSO-d_6) δ 3.14 (s, 3H), 4.80 (s, 2H), 7.88 (s, 1H). ^{13}C NMR (101 MHz, DMSO-d_6) δ 41.41, 60.39, 122.86, 159.60, 162.39, 164.01. MS-ESI *m/z* 239.03 $[M-H]^-$.



***(R)-4-(2-Chloro-6-methanesulfonylmethyl-pyrimidin-4-yl)-3-methylmorpholine* 18.**

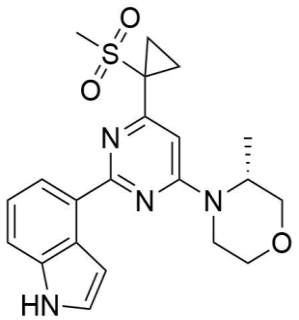
(*R*)-3-Methylmorpholine **17** (0.671 g, 6.64 mmol) was added to the solution of 2,4-dichloro-6-[(methylsulfonyl)methyl]pyrimidine, **16** (1.50g, 6.22 mmol), and N,N-diisopropylethylamine (1.73 mL, 9.95 mmol) in dichloromethane (30 mL). The resulting mixture was stirred at room temperature for 18 h. The reaction mixture was diluted with water (100 mL). The layers were separated and extracted with dichloromethane (100 mL). The combined organics were dried over magnesium sulfate, and concentrated *in vacuo*. The residue was purified by chromatography on silica (Biotage system, 40 g cartridge), eluting with 5:1 EtOAc/EtOH. The product after chromatography was triturated with 20 mL of methanol to afford a white solid which was collected by filtration and dried under vacuum to afford **18** (1.64 g, 81%). ^{1}H NMR (400 MHz, DMSO-d_6) *δ* 1.21 (d, J = 6.8 Hz, 3H), 3.11 (s, 3H), 3.21 (t, J = 12.6 Hz, 1H), 3.44 (td, J = 11.9, 3.0 Hz, 1H), 3.59 (dd, J = 11.7, 3.2 Hz, 1H), 3.73 (d, J = 11.6 Hz, 1H), 3.94 (dd, J = 11.5, 3.8 Hz, 1H), 4.00 (bs, 1H), 4.31 (bs, 1H), 4.46 (s, 2H), 6.93 (s, 1H). ^{13}C NMR (101 MHz, DMSO-d_6) *δ* 14.08, 41.72, 47.42, 60.98, 66.21, 70.29, 104.07, 159.42, 159.77, 163.03. MS-ESI m/z 306.15 $[MH]^+$.



***(3R)-4-[2-chloro-6-(1-methanesulfonylcyclopropyl)pyrimidin-4-yl]-3-methylmorpholine* 19**

NaOH (50% w/w aqueous solution, 125 mmol) was added to a mixture of (*R*)-4-(2-Chloro-6-methanesulfonylmethyl-pyrimidin-4-yl)-3-methylmorpholine **18** (1.30 g, 4.25 mmol), 1,2-dibromoethane (2.40 g, 12.8 mmol) and TBAF (0.5 g, 2.0 mmol) in toluene (40 mL) and the resulting suspension was stirred at 60°C for 18 h. Saturated aqueous sodium bicarbonate (200 mL) was added to the mixture. The phases were separated and extracted with toluene (2x75 mL). The combined organics were washed with water (100 mL), dried over magnesium sulfate and concentrated *in vacuo*.

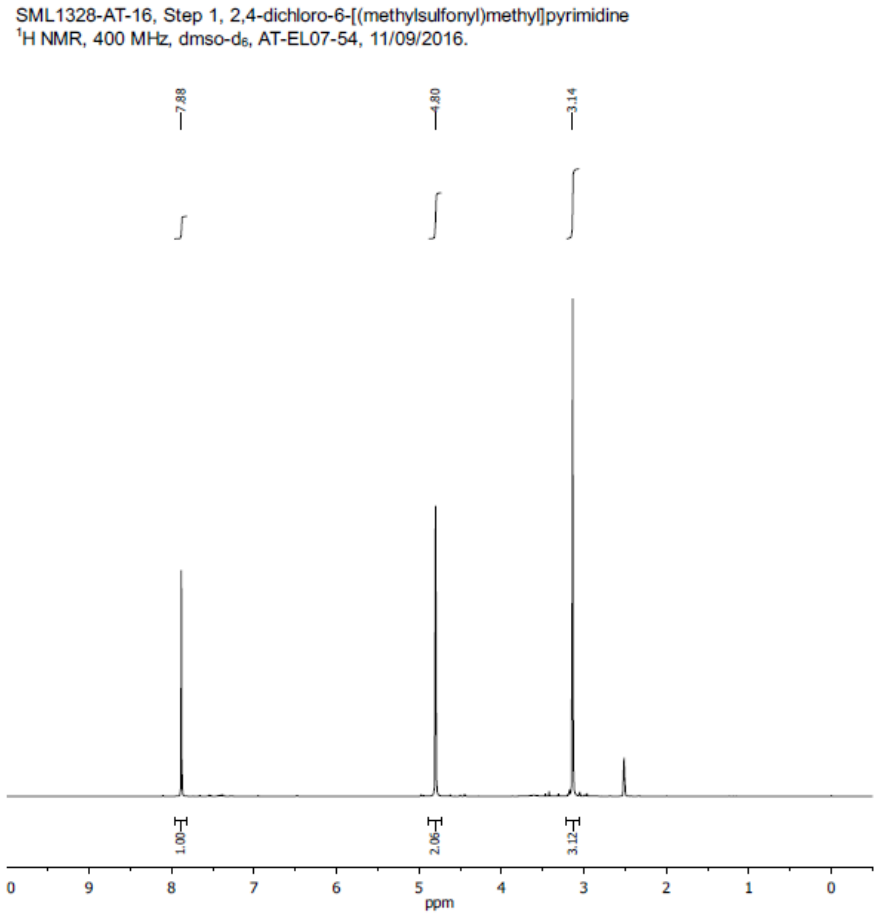
The residue was purified by chromatography on silica (Biotage system, 40 g cartridge) eluting with a gradient of 0-100% EtOAc in hexanes. Fractions containing product were combined, evaporated, and dried under vacuum to afford **19** as off-white solid (0.915 g, 65%). ^{1}H NMR (400 MHz, DMSO-d$_6$) δ 1.21 (d, J = 6.7 Hz, 3H), 1.52 (t, J = 3.8 Hz, 2H), 1.58 – 1.77 (m, 2H), 3.20 (s, 4H), 3.43 (td, J = 11.9, 3.0 Hz, 1H), 3.57 (dd, J = 11.7, 3.2 Hz, 1H), 3.72 (d, J = 11.6 Hz, 1H), 3.93 (dd, J = 11.6, 3.7 Hz, 1H), 4.05 (s, 1H), 4.42 (s, 1H), 6.94 (s, 1H). ^{13}C NMR (101 MHz, DMSO-d$_6$) δ 12.75, 14.12, 40.72, 46.11, 47.27, 66.26, 70.34, 102.98, 159.49, 163.17, 163.67. MS-ESI m/z 332.25 [MH]$^+$.



4-[4-[(3R)-3-methyl-4-morpholinyl]-6-[1-(methylsulfonyl)cyclopropyl]-2-pyrimidinyl]-1H-indole 21.

A mixture of (3*R*)-4-[2-chloro-6-(1-methanesulfonylcyclopropyl)pyrimidin-4-yl]-3-methylmorpholine **19** (0.672 g, 2.02 mmol), (1*H*-indo-4-yl)boronic acid, **20** (1.14 g, 7.09 mmol), 1.2 M aqueous solution of potassium phosphate (6.08 mmol), and 1,2-dimethoxyethane (15 mL) in 25 mL pressure flask was deoxygenated by passing a slow stream of nitrogen (about 50 mL/min) for 20 min. Then, XPhos Pd G3 (0.10 g, 0.12 mol) was added in one portion. The flask was closed and the mixture was heated at 70 °C with stirring for 20 h. The reaction mixture was diluted with 50 mL of EtOAc. The mixture was washed with water (1x50 mL). The organics were separated; the aqueous solution was extracted with ethyl acetate (3x50 mL). Combined organics were dried over magnesium sulfate, filtered and concentrated *in vacuo* to afford crude product (about 1.0 g) as brown oil. The crude was purified on silica (Biotage system, 40 g cartridge) with 0-30% gradient of EtOAc/EtOH (4:1) in hexanes. Fractions containing product were combined, evaporated, and dried under vacuum to afford **21** as white solid (0.531 g, 64%). ^{1}H NMR (400 MHz, DMSO-d$_6$) δ 1.26 (d, J = 6.7 Hz, 3H), 1.60 (td, J = 5.7, 5.0, 3.5 Hz, 2H), 1.71 (qd, J = 4.1, 3.4, 1.3 Hz, 2H), 3.22 (d, J = 3.9 Hz, 1H), 3.28 (s, 3H), 3.50 (td, J = 11.8, 3.0 Hz, 1H), 3.65 (dd, J = 11.5, 3.2 Hz, 1H), 3.78 (d, J = 11.4 Hz, 1H), 3.99 (dd, J = 11.4, 3.6 Hz, 1H), 4.20 (d, J = 13.5 Hz, 1H), 4.60 (s, 1H), 6.83 (s, 1H), 7.20 (t, J = 7.7 Hz, 1H), 7.31 (t, J = 2.6 Hz, 1H), 7.46 (t, J = 2.8 Hz, 1H), 7.54 (dd, J = 8.0, 1.0 Hz, 1H), 8.05 (dd, J = 7.5, 1.0 Hz, 1H), 11.28 (t, J = 2.3 Hz, 1H). ^{13}C NMR (101 MHz, DMSO-d$_6$) δ 12.78, 13.87, 39.34, 40.81, 46.62, 46.87, 66.54, 70.71, 101.01, 103.33, 114.29, 120.92, 121.06, 126.75, 126.80, 129.84, 137.48, 161.99, 162.45, 164.94. MS-ESI m/z 413.26 [MH]+. Anal. Found (% w/w): C, 61.27; H, 6.02; N, 13.42. $C_{21}H_{24}N_4O_3S$ requires C, 61.15; H, 5.86; N, 13.58. $[\alpha]^{23}_D$ -88.3 (c 1.03, DMSO).

S12.3. Raw spectroscopic and chromatographic data.



	Parameter	Value
1	Title	SML1328-1-FP.10.fid
2	Author	
3	Solvent	DMSO
4	Temperature	292.6
5	Pulse Sequence	zg30
6	Experiment	1D
7	Number of Scans	64
8	Acquisition Date	2016-11-09T10:52:54
9	Spectrometer Frequency	400.13
10	Nucleus	1H
11	Spectral Size	65536

^{1}H NMR (400 MHz, DMSO-d_6) δ 3.14 (s, 3H), 4.80 (s, 2H), 7.88 (s, 1H).

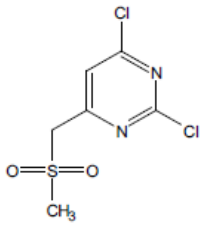


Figure S57. ^{1}H-NMR of 2,4-dichloro-6-[(methylsulfonyl)methyl]pyrimidine **16**.

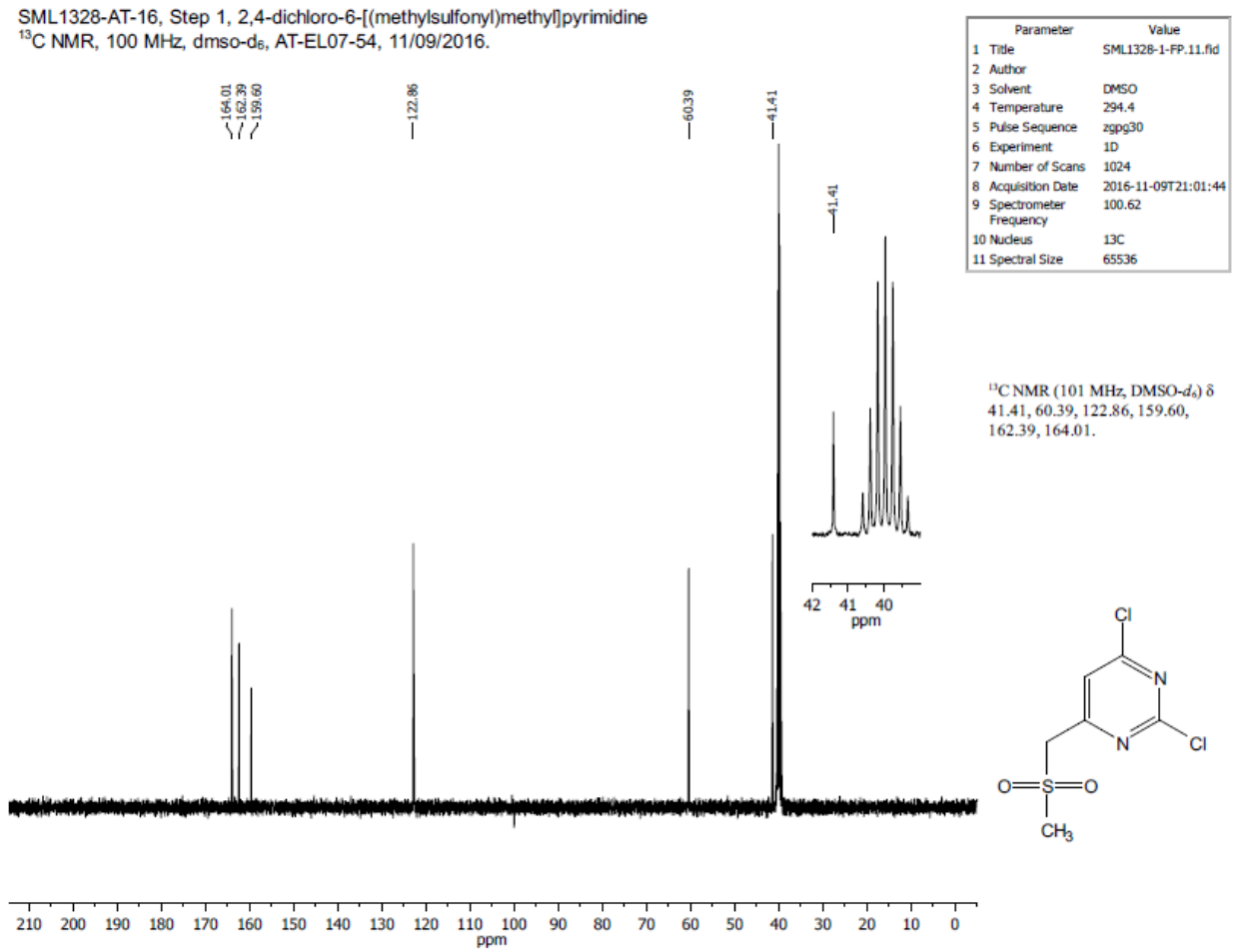


Figure S58. ^{13}C-NMR of 2,4-dichloro-6-[(methylsulfonyl)methyl]pyrimidine **16**.

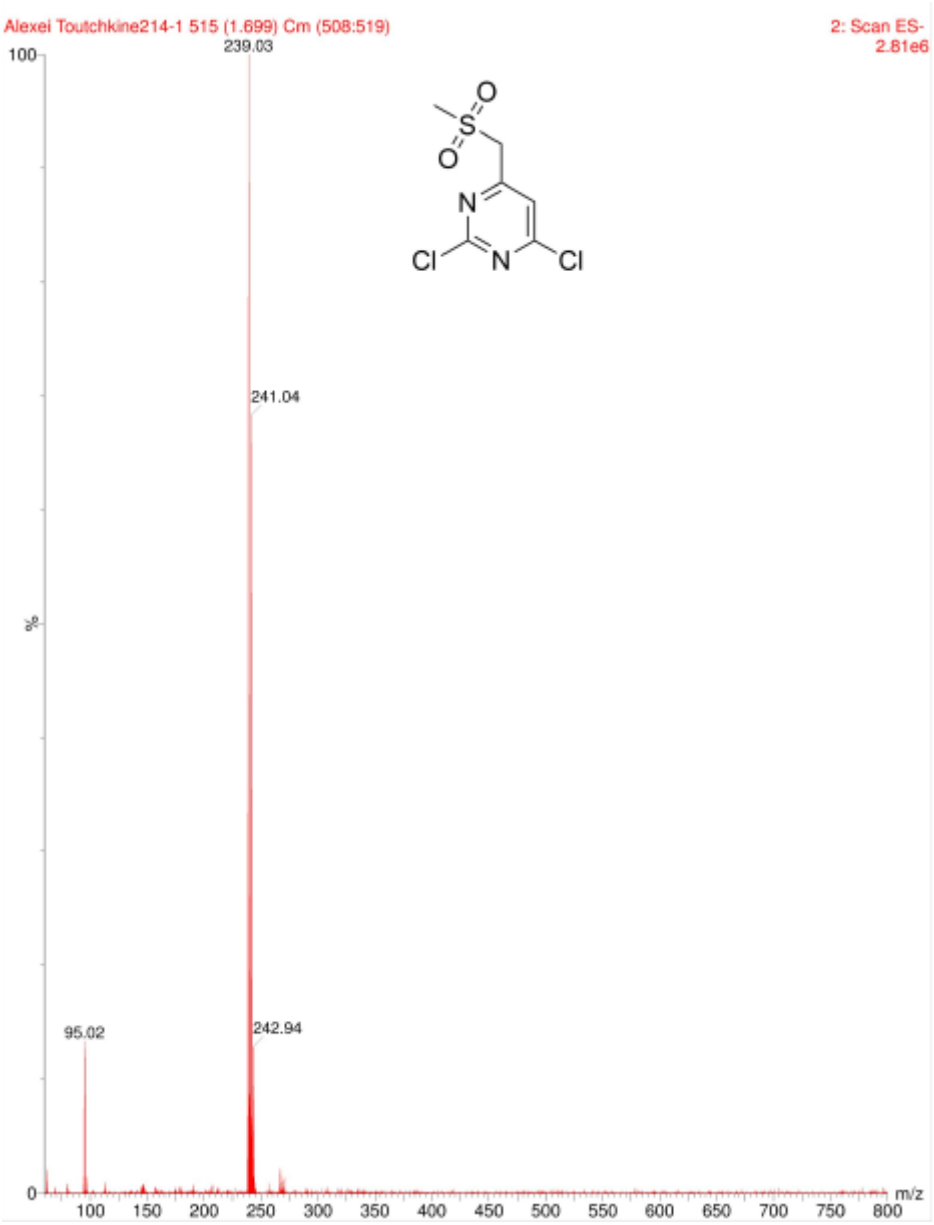


Figure S59. MS of 2,4-dichloro-6-[(methylsulfonyl)methyl]pyrimidine **16**.

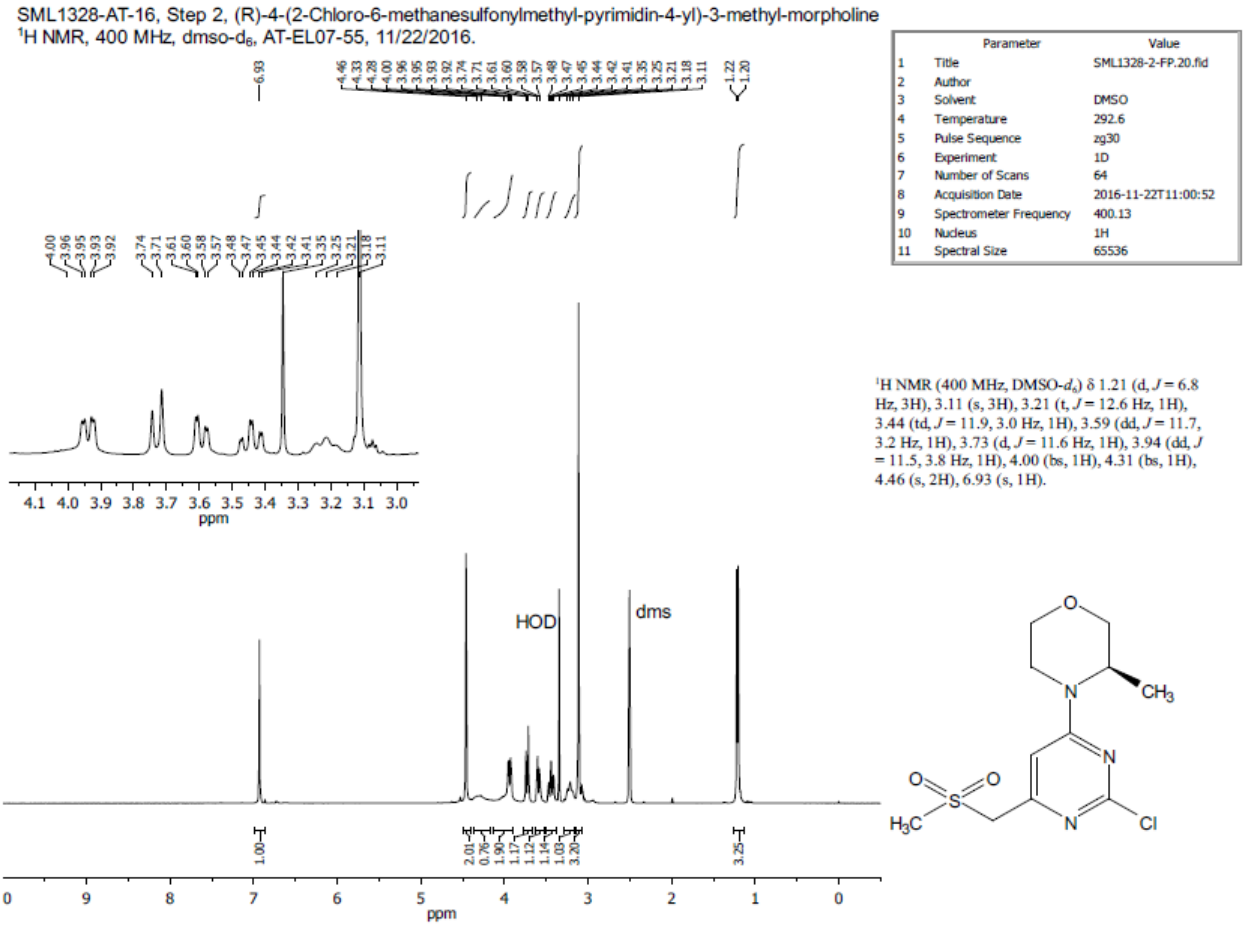


Figure S60. [1]H-NMR of (*R*)-4-(2-chloro-6-methanesulfonylmethyl-pyrimidin-4-yl)-3-methylmorpholine **18**.

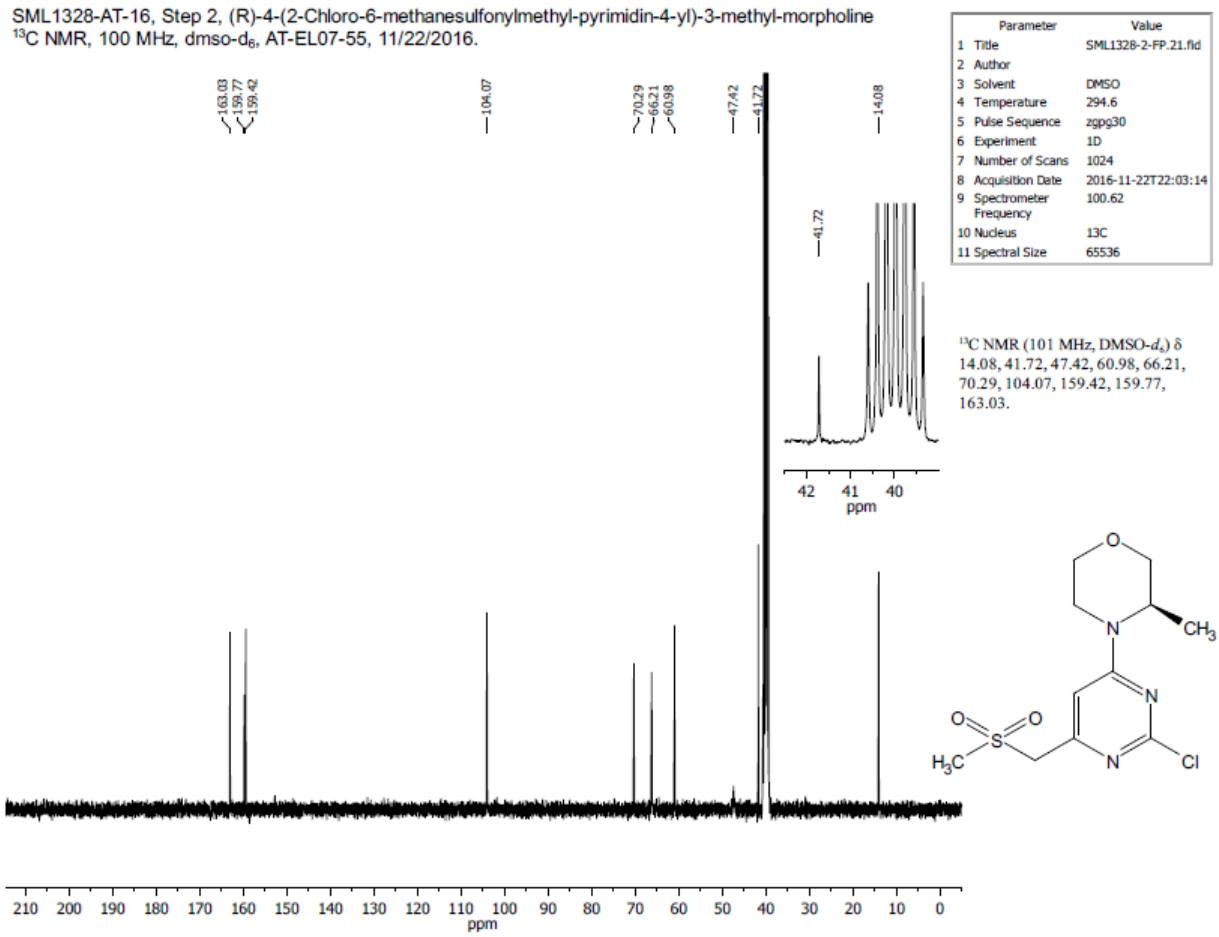


Figure S61. ^{13}C-NMR of (*R*)-4-(2-chloro-6-methanesulfonylmethyl-pyrimidin-4-yl)-3-methylmorpholine **18**.

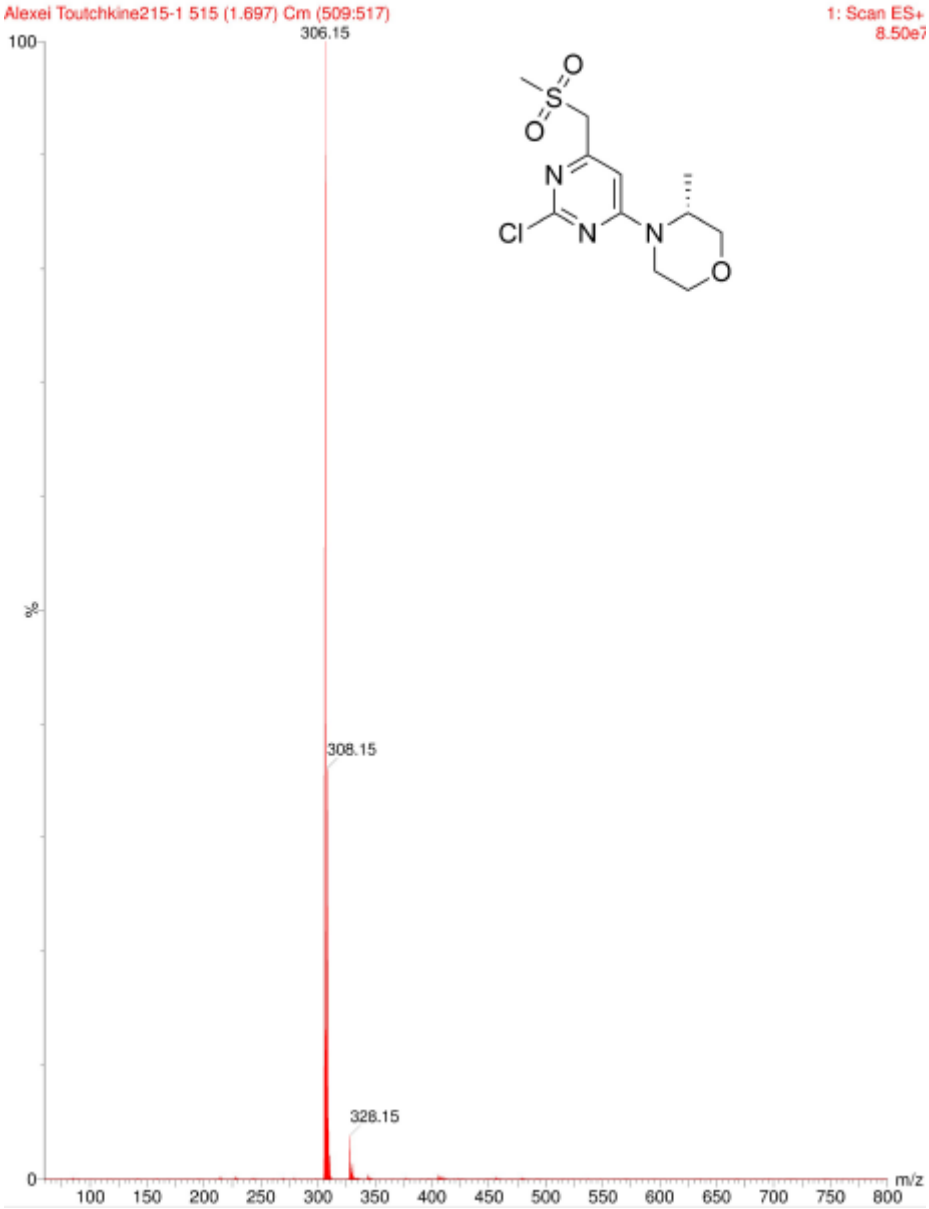


Figure S62. MS of *(R)*-4-(2-chloro-6-methanesulfonylmethyl-pyrimidin-4-yl)-3-methylmorpholine **18**.

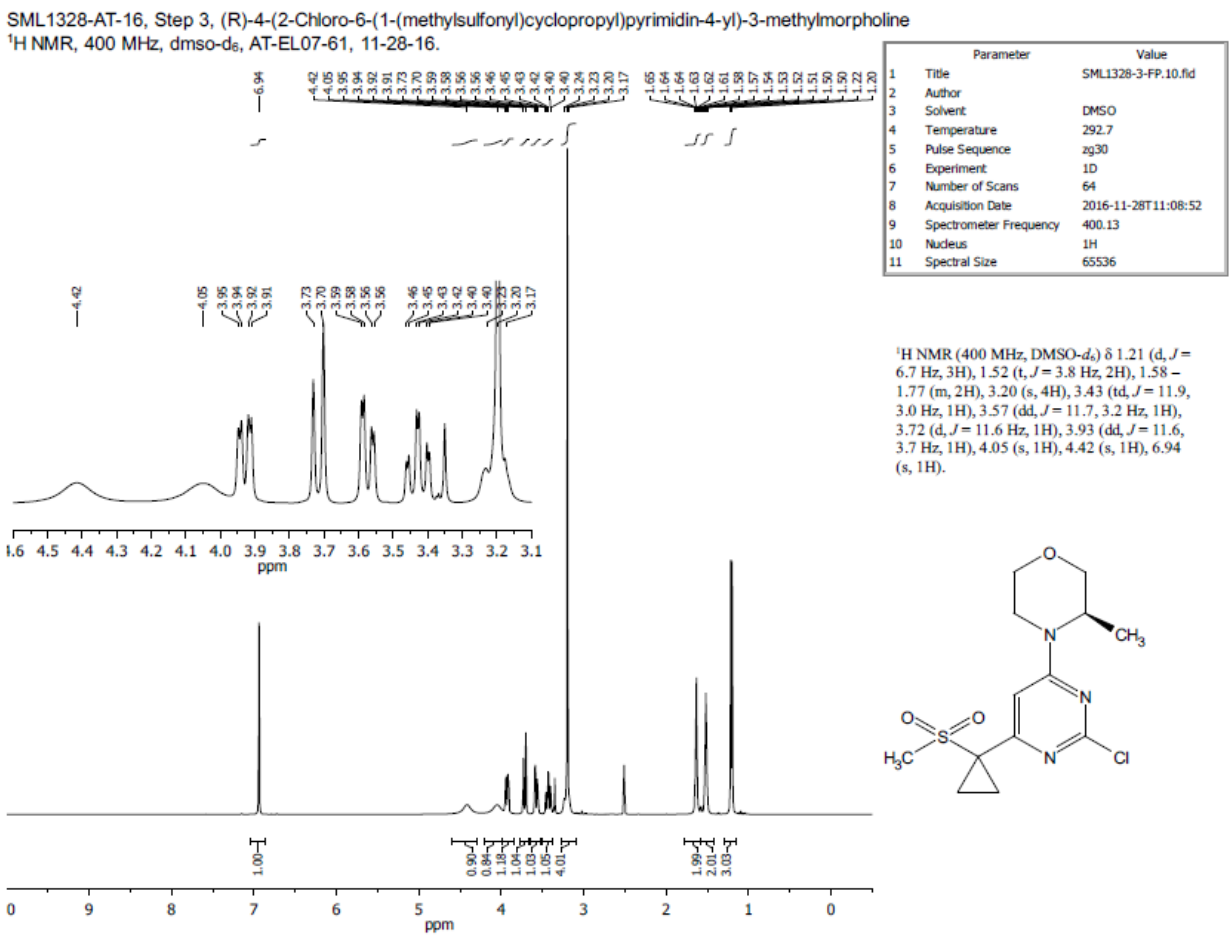


Figure S63. ^{1}H-NMR of (3*R*)-4-[2-chloro-6-(1-methanesulfonylcyclopropyl)pyrimidin-4-yl]-3-methylmorpholine **19**.

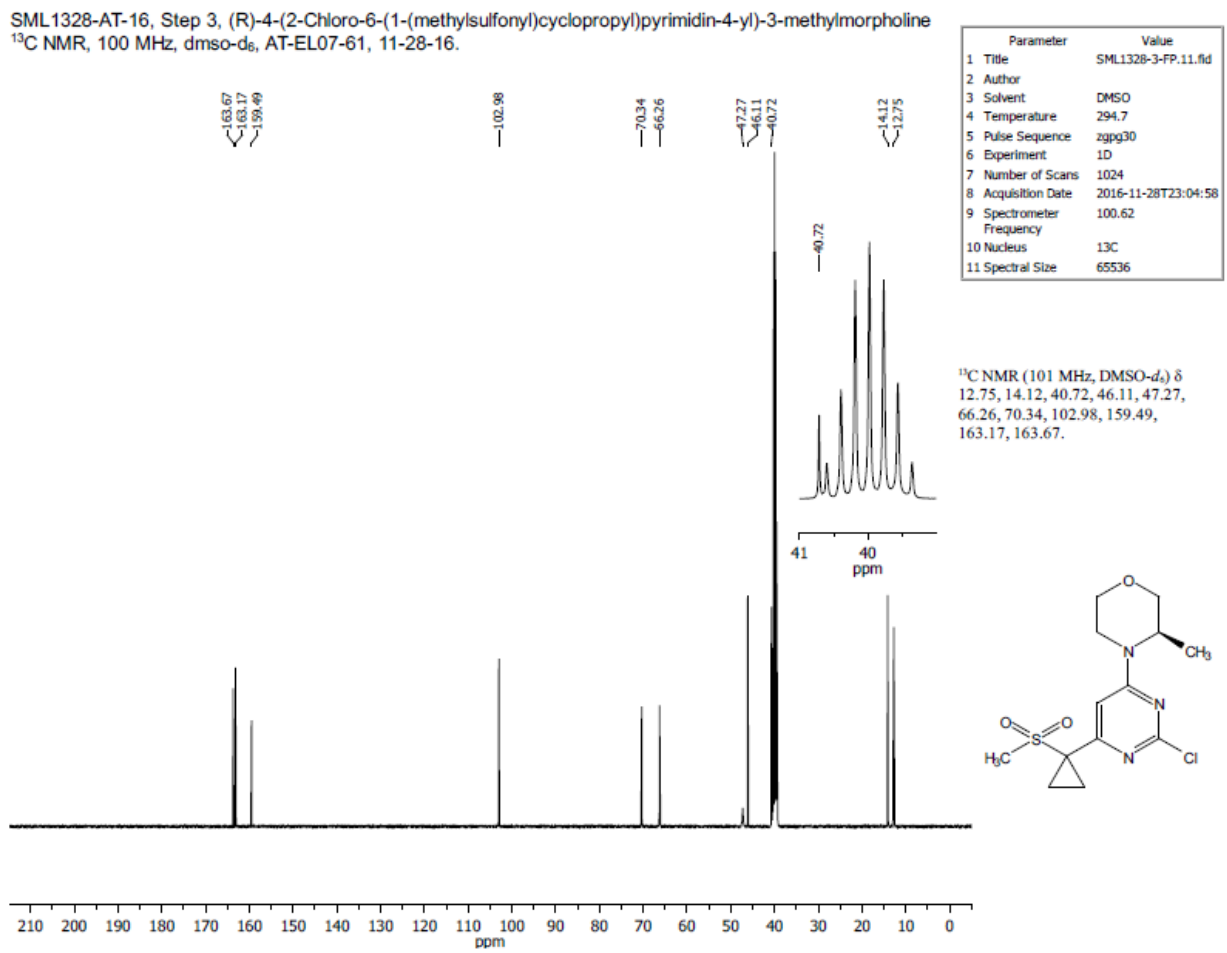


Figure S64. ^{13}C-NMR of (3*R*)-4-[2-chloro-6-(1-methanesulfonylcyclopropyl)pyrimidin-4-yl]-3-methylmorpholine **19**.

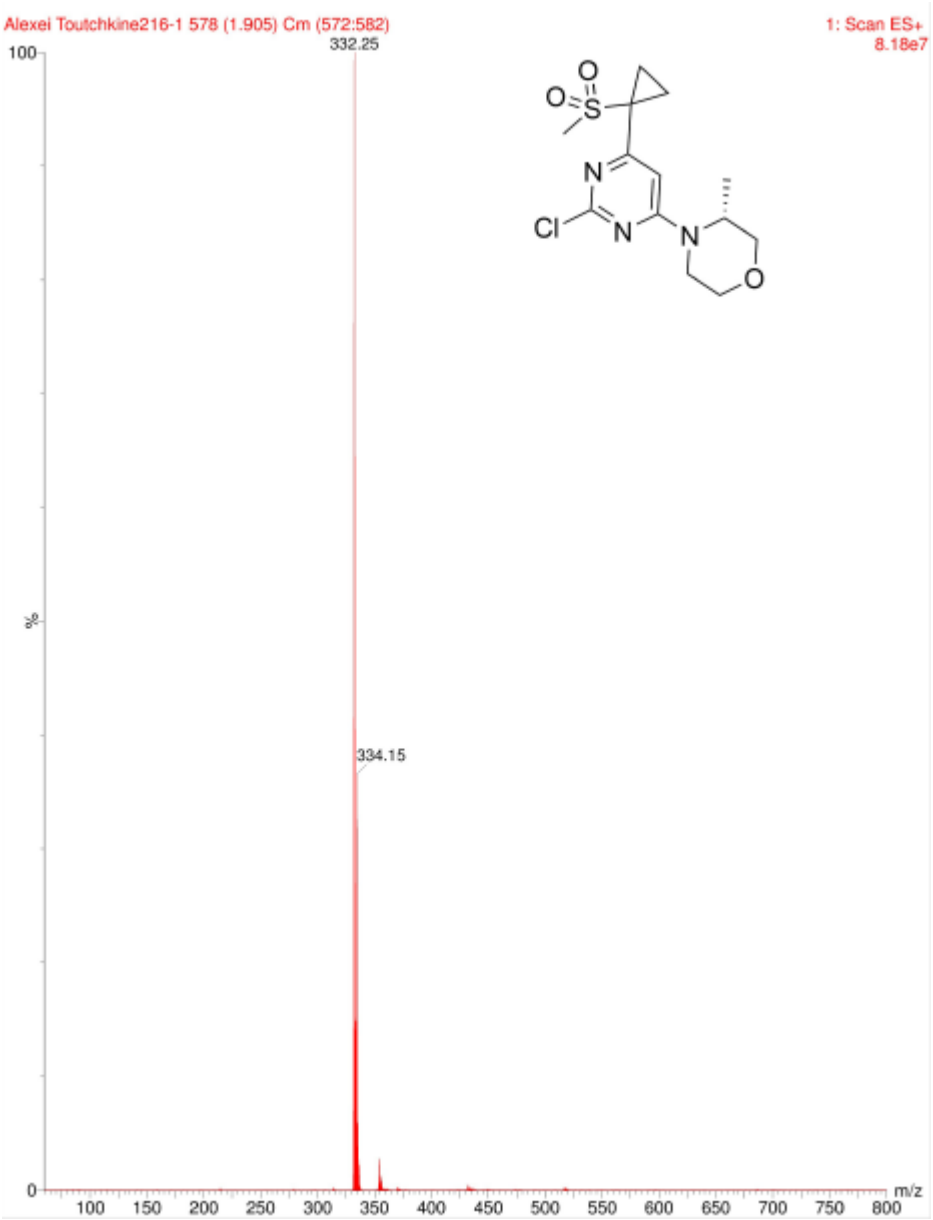


Figure S65. MS of (3*R*)-4-[2-chloro-6-(1-methanesulfonylcyclopropyl)pyrimidin-4-yl]-3-methylmorpholine **19**.

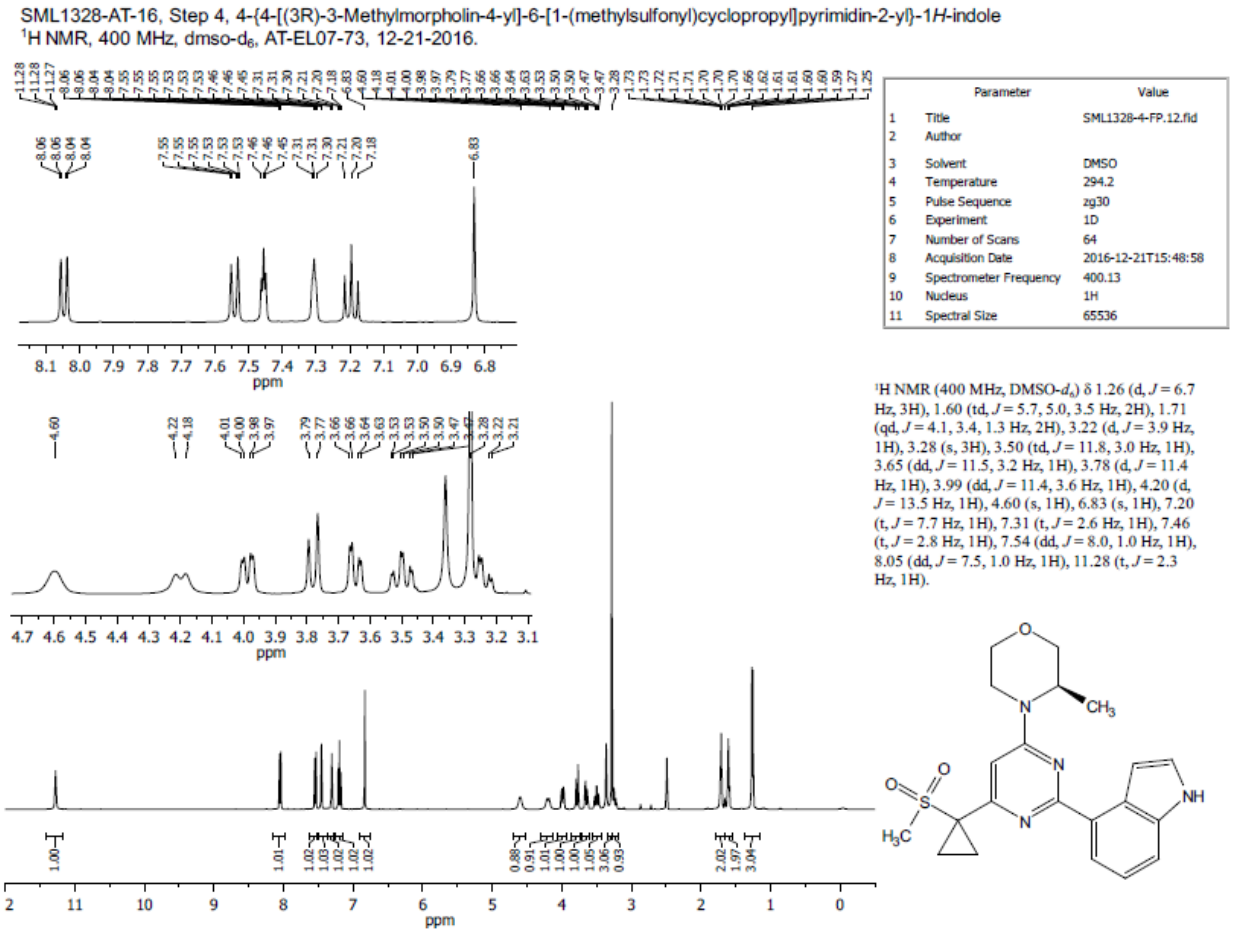


Figure S66. ^{1}H-NMR of 4-[4-[(3*R*)-3-methyl-4-morpholinyl]-6-[1-(methylsulfonyl)cyclopropyl]-2-pyrimidinyl]-1*H*-indole **21**.

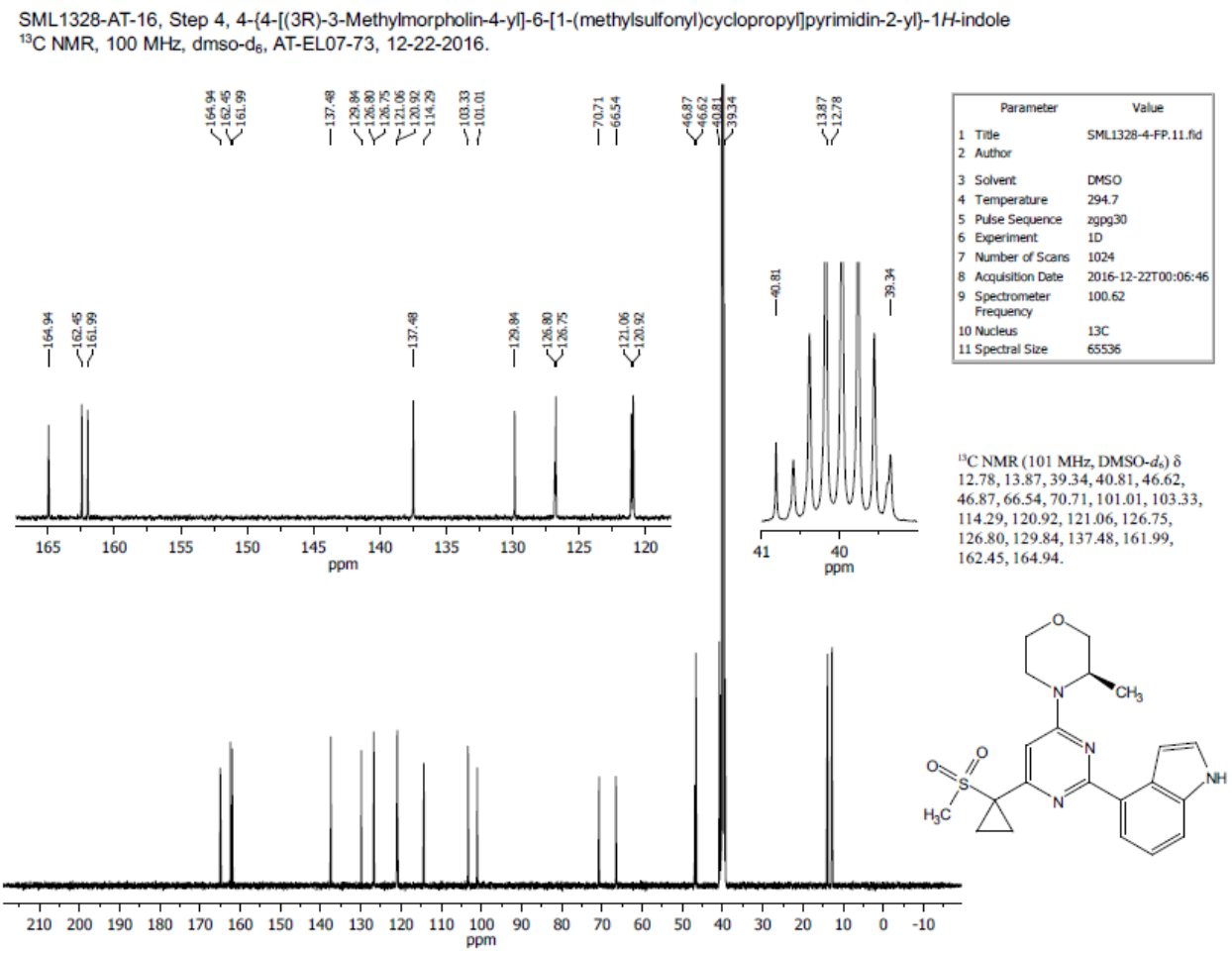


Figure S67. ^{13}C-NMR of 4-[4-[(3*R*)-3-methyl-4-morpholinyl]-6-[1-(methylsulfonyl)cyclopropyl]-2-pyrimidinyl]-1*H*-indole **21**.

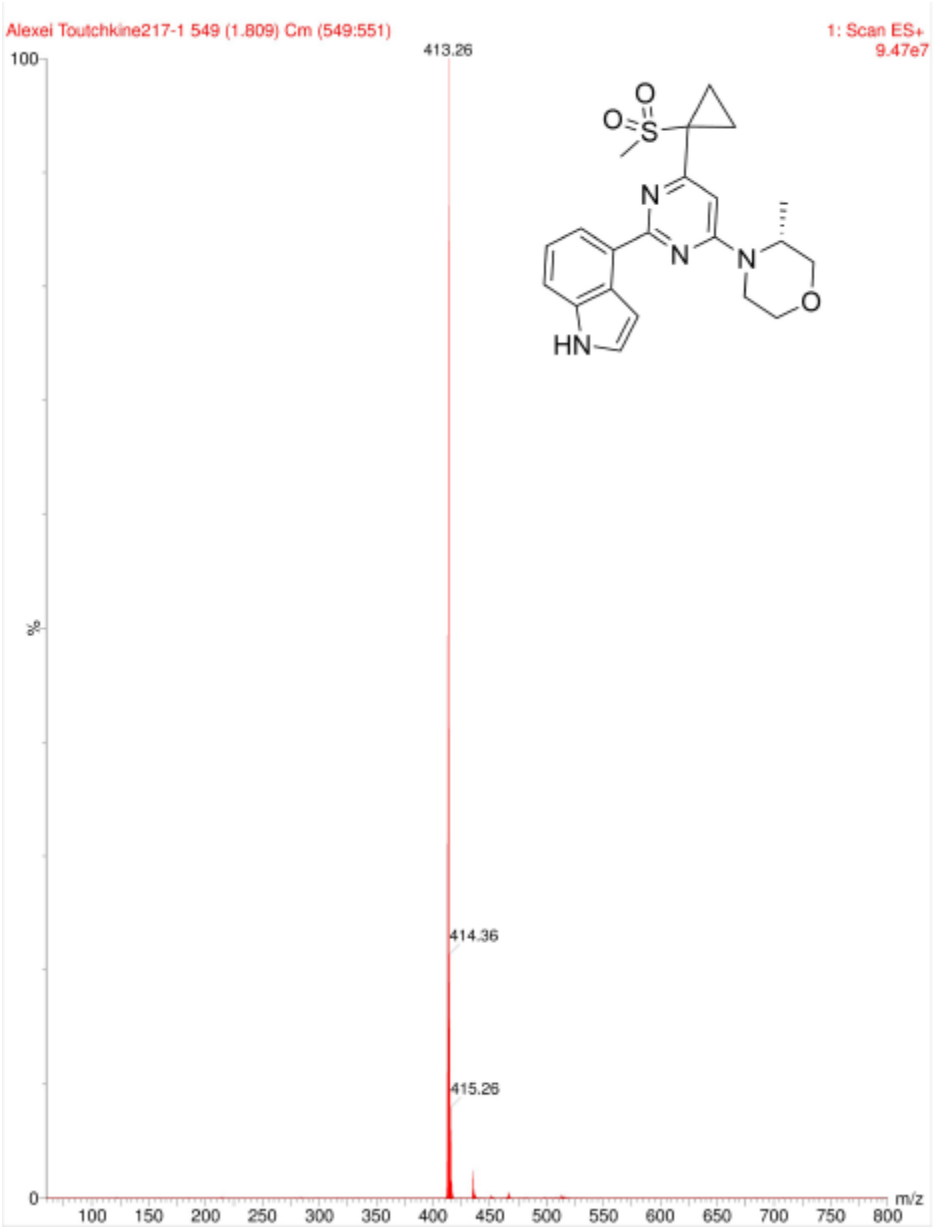


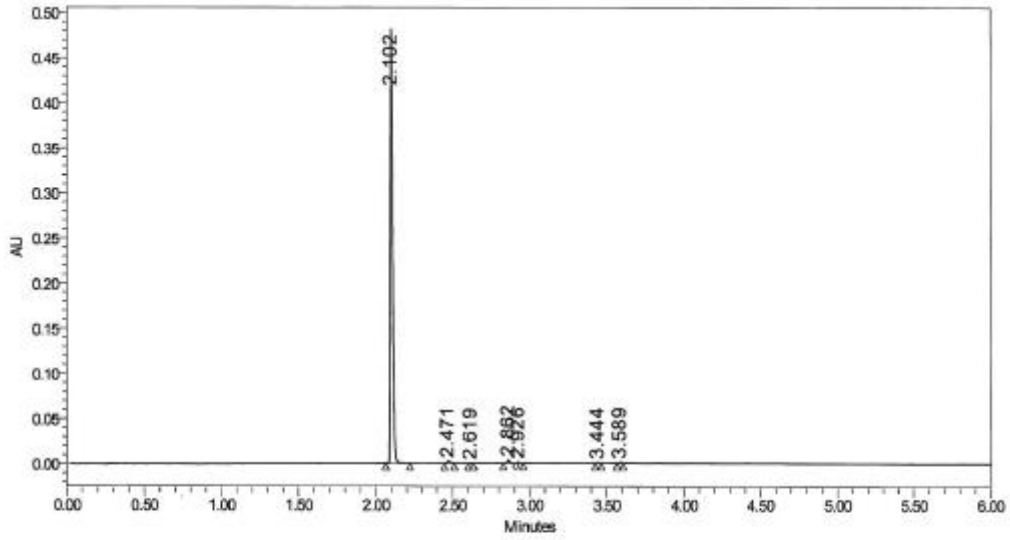
Figure S68. MS of 4-[4-[(3*R*)-3-methyl-4-morpholinyl]-6-[1-(methylsulfonyl)cyclopropyl]-2-pyrimidinyl]-1*H*-indole **21**.

Empower software

RBI HPLC Analysis Custom Report

SAMPLE INFORMATION

Sample Name:	SML1328-AT-16/AT-EL07-73	Acquired By:	Lily_Zhang
System:	UPLC_1	Date Acquired:	12/20/2016 2:24:11 PM EST
Injection Volume:	0.80 ul	Acq. Method Set:	UPLC_CD_0_100D_4M_H2M_05F
Vial:	2:E,8	Processing Method:	Processing 01
Run Time:	6.00 Minutes	Proc. Chnl. Descr.:	PDA 285.0 nm (PDA Spectrum (210-500)nm
Column ID	Ascentis C18, 2.1x50mm,2.1um	Injection Solvent:	MeOH
Solvent System:	A: 0.1% TFA in H2O, B: 0.1% TFA in CH3CN , 0-100%B in 4min, hold 2min, flow rate: 0.5ml/min		



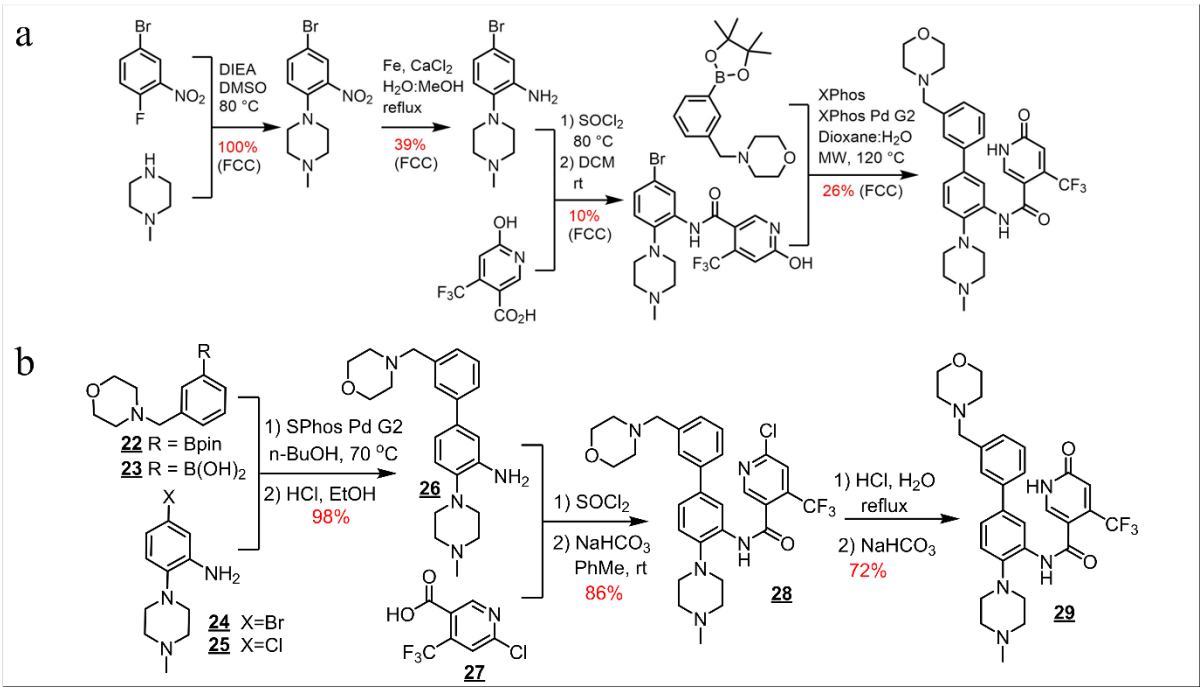
Peak Results

	Name	End Time (min)	Start Time (min)	RT	Height	Area	% Area	Int Type
1		2.222	2.065	2.102	481428	498231	97.74	BB
2		2.512	2.448	2.471	2442	4067	0.80	Bb
3		2.637	2.602	2.619	659	698	0.14	bb
4		2.909	2.828	2.862	3652	3509	0.69	bV
5		2.958	2.909	2.926	1351	1351	0.26	Vb
6		3.486	3.428	3.444	822	794	0.16	bb
7		3.607	3.569	3.589	1218	1107	0.22	bb

Figure S69. UPLC of 4-[4-[(3*R*)-3-methyl-4-morpholinyl]-6-[1-(methylsulfonyl)cyclopropyl]-2-pyrimidinyl]-1*H*-indole **21**.

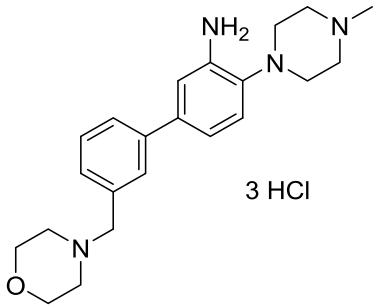
Section S13. Synthesis of anti-leukemia drug candidate, 29.

S13.1. Previous vs. current synthetic routes.



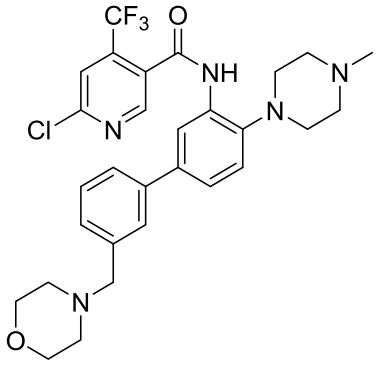
Scheme S4. (a) The original, low-yielding preparation of **29** from the main-text reference[26]. For comparison, **(b)** shows the Chematica route (same as in the main-text **Figure 2d**).

S13.2. Synthetic details.



***4-(4-Methyl-piperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-ylamine hydrochloride* 26**

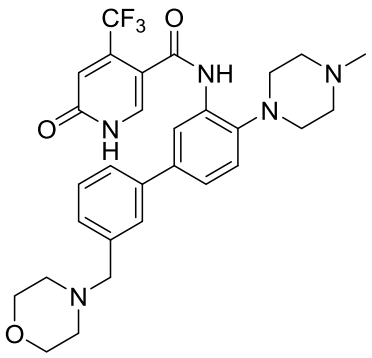
A stirring solution of 4-[3-(4,4,5,5-tetramethyl-[1,3,2]dioxaborolan-2-yl)-benzyl]-morpholine **22** (5.37 g, 17.7 mmol), 5-bromo-2-(4-methyl-piperazin-1-yl)-phenylamine **24** (3.97 g, 14.7 mmol), potassium phosphate (9.43 g, 44.4 mmol), water (4.12 mL) and 1-butanol (68.0 mL) in a 250 mL round bottom sealed tube pressure vessel was purged with nitrogen, then SPhos Pd G2 (0.750 g, 1.04 mmol) was added in one portion, the mixture was purged for an additional 5 min, then the mixture was sealed and heated at 70 °C in an oil bath overnight. The crude reaction mixture was allowed to cool to room temperature, was filtered through celite with warm 95% EtOH:H_2O (3 x 50 mL) and concentrated to afford a brown solid. The crude solid was diluted with EtOH (200 mL), cooled to 0 °C in an ice bath and concentrated HCl (20 mL) was added in one portion. The mixture was removed from the ice bath and stirred for 1h at room temperature. The resulting suspension was filtered, washed with cold EtOH (2 x 50 mL) and dried under vacuum to afford 6.85 g (98%) of 4-(4-methyl-piperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-ylamine hydrochloride **26**. ^{1}H NMR (400 MHz, DMSO-d_6) δ 7.97 (s, 1H), 7.72 – 7.60 (m, 4H), 7.59 – 7.51 (m, 1H), 7.38 (d, J = 8.3 Hz, 1H), 4.44 (s, 2H), 3.95 – 3.79 (m, 4H), 3.49 (d, J = 11.5 Hz, 2H), 3.42 – 3.08 (m, 11H), 2.82 (s, 3H). ^{13}C NMR (400 MHz, DMSO-d_6) δ 143.12, 139.45, 136.94, 131.22, 130.83, 130.11, 130.01, 129.53, 127.55, 125.13, 122.75, 120.57, 63.05, 58.78, 52.64, 50.63, 48.37, 42.15. LRMS m/z calcd. for $C_{22}H_{30}N_4O$ ([M+H]+) 367.25, found 367.46.



***6-Chloro-N-[4-(4-methylpiperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-yl]-4-trifluoromethylnicotinamide* 28**

To a stirring solution of 6-chloro-4-trifluoromethyl-nicotinoyl chloride hydrochloride **27** (3.83 g, 13.6 mmol) and 4-(4-methyl-piperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-ylamine trihydrochloride **26** (5.00 g, 10.5 mmol) in toluene (50.0 mL) was added a solution of sodium bicarbonate (4 g, 50 mmol) in water (50 mL), and the mixture was stirred at room temperature for 15h. The crude reaction mixture was poured into a 1L separatory funnel, diluted with EtOAc (500 mL), washed with saturated sodium bicarbonate (3 x 100 mL), brine (2 x 100 mL), dried over sodium sulfate, filtered, concentrated and dried under vacuum to afford 5.4g (86%) of an off-white solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 9.07 (s, 1H), 8.81 – 8.72 (m, 2H), 7.73 (s, 1H), 7.58 (s, 1H), 7.53 (d, J = 7.6 Hz, 1H), 7.43 – 7.30 (m, 4H), 3.77

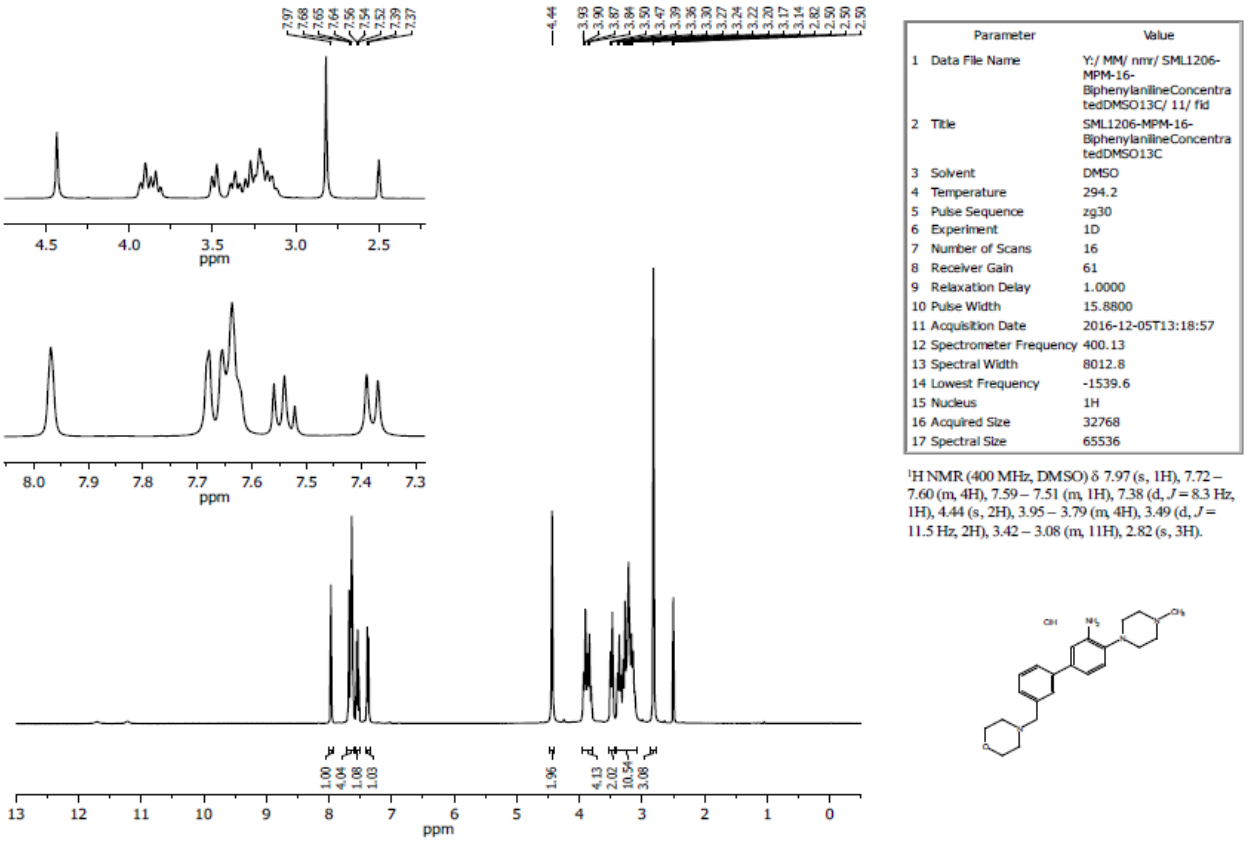
– 3.68 (m, 4H), 3.57 (s, 2H), 3.02 – 2.90 (m, 4H), 2.70 – 2.44 (m, 8H), 2.35 (s, 3H). ^{13}C NMR (400 MHz, $CDCl_3$) δ 161.58, 153.92, 149.72, 140.52, 140.39, 139.08, 138.37, 138.22, 138.03, 137.69, 137.35, 133.34, 129.36, 128.69, 128.32, 127.97, 126.14, 125.68, 123.79, 122.94, 121.55, 121.52, 121.47, 121.42, 120.20, 118.31, 117.47, 66.95, 63.36, 55.58, 53.58, 52.40, 45.92. LRMS m/z calcd. for $C_{29}H_{32}F_3N_5O_3$ ([M+H]+) 574.21, found 574.36.



***6-Oxo-4-trifluoromethyl-1,6-dihydro-pyridine-3-carboxylic acid [4-(4-methyl-piperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-yl]-amide* 29**

A stirring solution of 6-chloro-N-[4-(4-methylpiperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-yl]-4-trifluoromethylnicotinamide **28** (5.4 g, 9.4 mmol) in 12M HCl (170 mL) was refluxed at 120 °C for 2h. After 1h, additional 12M HCl was added (50 mL). After 2h, the crude reaction mixture was concentrated under reduced pressure, dissolved by refluxing in EtOH (250 mL) for 1h and the resulting solution was filtered rapidly through celite. The filter was washed with warm EtOH (2 x 20 mL) and the liquor was brought back to reflux for another 30 min before cooling slowly to room temperature with stirring. The resulting suspension was filtered, washed with EtOH (2 x 50 mL) and the solid was dried under vacuum to afford 3.5 g of hydrochloride salt. The hydrochloride salt was taken up in sat. sodium bicarbonate solution (300 mL) and extracted with EtOAc (4 x 100 mL), washed with brine, dried over sodium sulfate, filtered, concentrated and dried under vacuum to afford 2.7 g (52%) of the final product **29**. The mother liquor from the recrystallized HCl salt was concentrated, and the resulting solid was treated with sat. sodium bicarbonate solution (100 mL), extracted with EtOAc (4 x 100 mL). The combined organic phase was washed with brine, dried over sodium sulfate, filtered and concentrated to afford an off-white solid. The solid was further purified by silica gel chromatography using 5-20% MeOH:DCM. Purified fractions were combined, concentrated and dried under vacuum to afford an additional 1.1 g (20%) of **29** as a white powder for a combined yield 3.8 g (72%). Matches reported data.

S13.3. Raw spectroscopic and chromatographic data.



	Parameter	Value
1	Data File Name	Y:/ MM/ nmr/ SML1206-MPM-16-BiphenylanilineConcentratedDMSO13C/ 11/ fid
2	Title	SML1206-MPM-16-BiphenylanilineConcentratedDMSO13C
3	Solvent	DMSO
4	Temperature	294.2
5	Pulse Sequence	zg30
6	Experiment	1D
7	Number of Scans	16
8	Receiver Gain	61
9	Relaxation Delay	1.0000
10	Pulse Width	15.8800
11	Acquisition Date	2016-12-05T13:18:57
12	Spectrometer Frequency	400.13
13	Spectral Width	8012.8
14	Lowest Frequency	-1539.6
15	Nucleus	1H
16	Acquired Size	32768
17	Spectral Size	65536

^{1}H NMR (400 MHz, DMSO) δ 7.97 (s, 1H), 7.72 – 7.60 (m, 4H), 7.59 – 7.51 (m, 1H), 7.38 (d, J = 8.3 Hz, 1H), 4.44 (s, 2H), 3.95 – 3.79 (m, 4H), 3.49 (d, J = 11.5 Hz, 2H), 3.42 – 3.08 (m, 11H), 2.82 (s, 3H).

Figure S70. ^{1}H-NMR of 4-(4-Methyl-piperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-ylamine hydrochloride **26**.

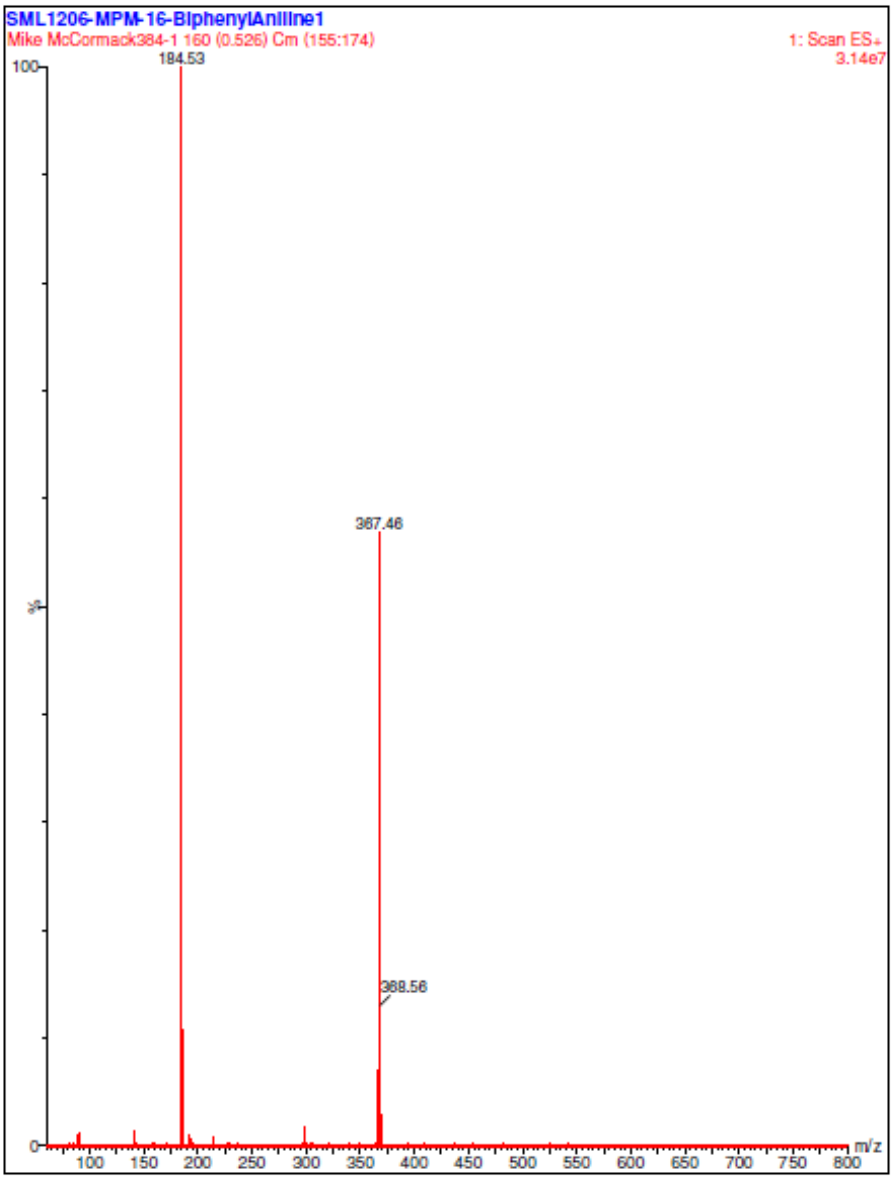


Figure S71. Mass Spectrum of 4-(4-Methyl-piperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-ylamine hydrochloride **26**.

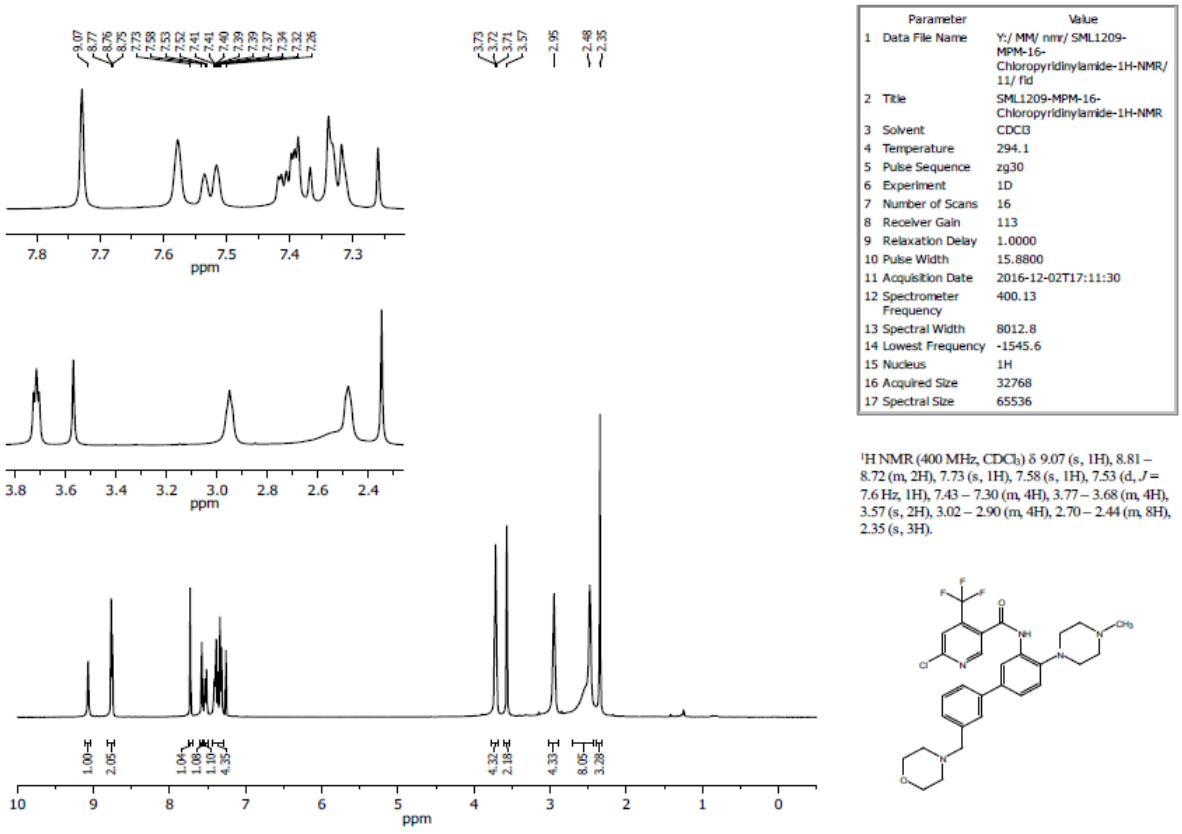
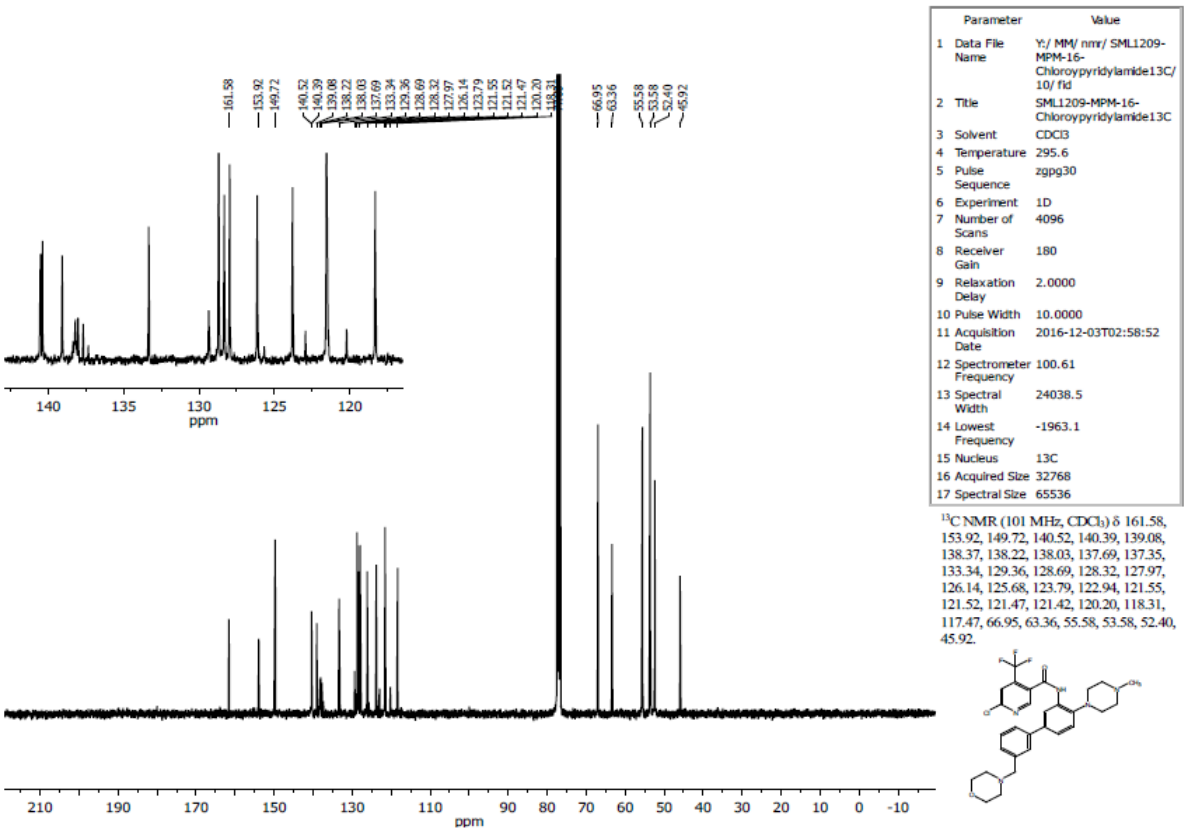


Figure S72. ^{1}H-NMR of 6-Chloro-*N*-[4-(4-methylpiperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-yl]-4-trifluoromethylnicotinamide **28**.



	Parameter	Value
1	Data File Name	Y:/ MM/ nmr/ SML1209-MPM-16-Chloroypyridylamide13C/ 10/ fid
2	Title	SML1209-MPM-16-Chloroypyridylamide13C
3	Solvent	CDCl3
4	Temperature	295.6
5	Pulse Sequence	zgpg30
6	Experiment	1D
7	Number of Scans	4096
8	Receiver Gain	180
9	Relaxation Delay	2.0000
10	Pulse Width	10.0000
11	Acquisition Date	2016-12-03T02:58:52
12	Spectrometer Frequency	100.61
13	Spectral Width	24038.5
14	Lowest Frequency	-1963.1
15	Nucleus	13C
16	Acquired Size	32768
17	Spectral Size	65536

^{13}C NMR (101 MHz, $CDCl_3$) δ 161.58, 153.92, 149.72, 140.52, 140.39, 139.08, 138.37, 138.22, 138.03, 137.69, 137.35, 133.34, 129.36, 128.69, 128.32, 127.97, 126.14, 125.68, 123.79, 122.94, 121.55, 121.52, 121.47, 121.42, 120.20, 118.31, 117.47, 66.95, 63.36, 55.58, 53.58, 52.40, 45.92.

Figure S73. ^{13}C-NMR of 6-Chloro-*N*-[4-(4-methylpiperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-yl]-4-trifluoromethylnicotinamide **28**.

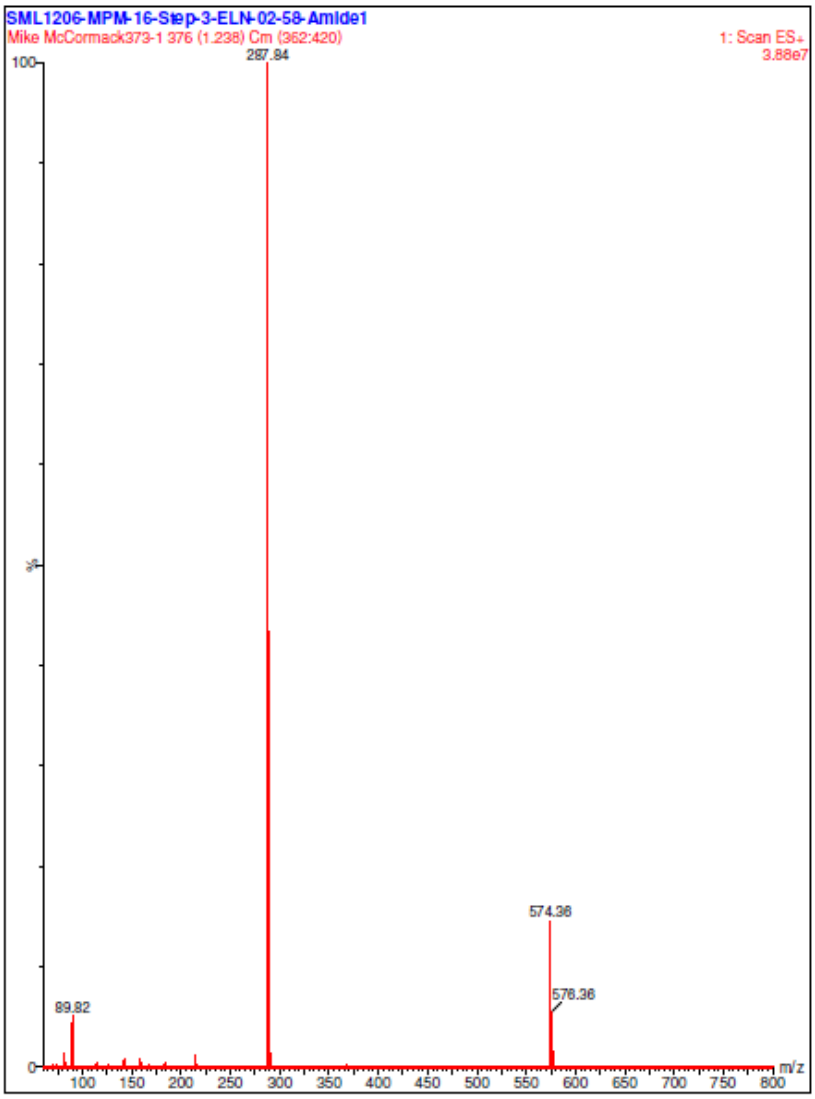
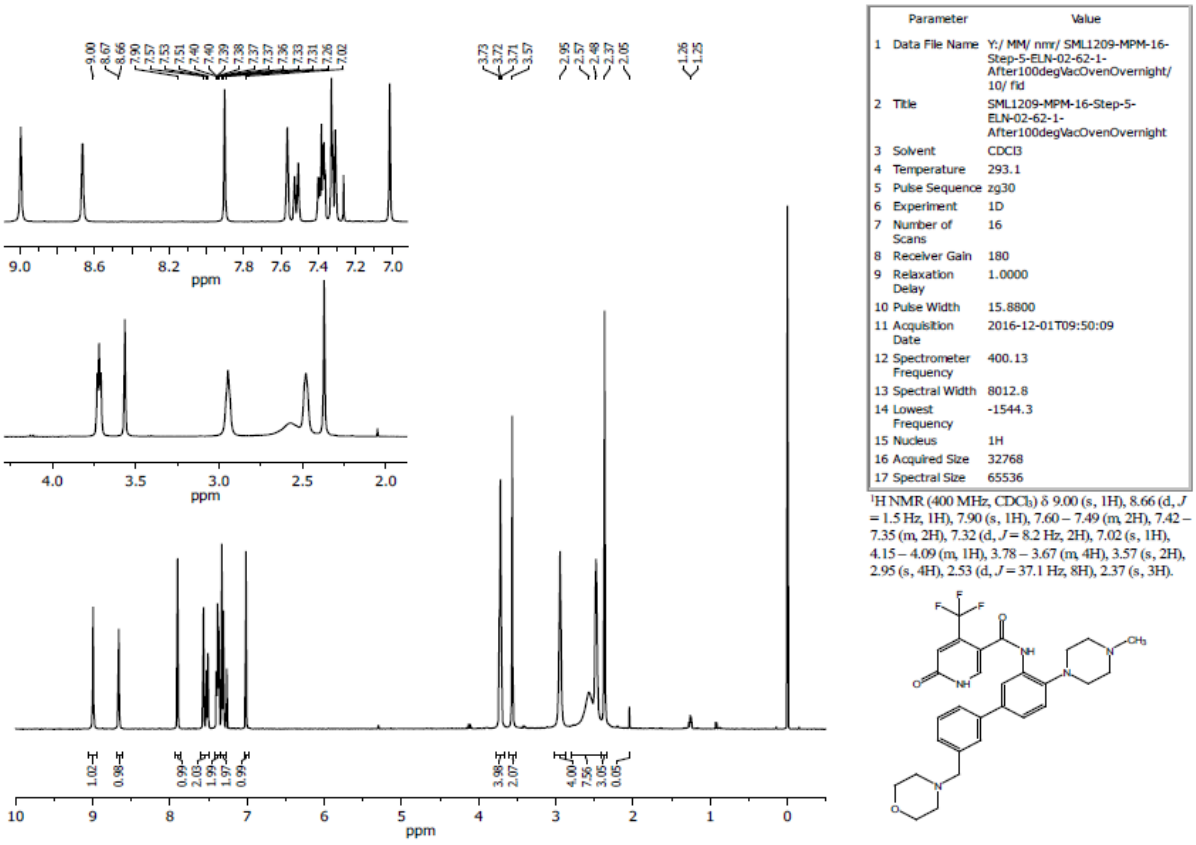


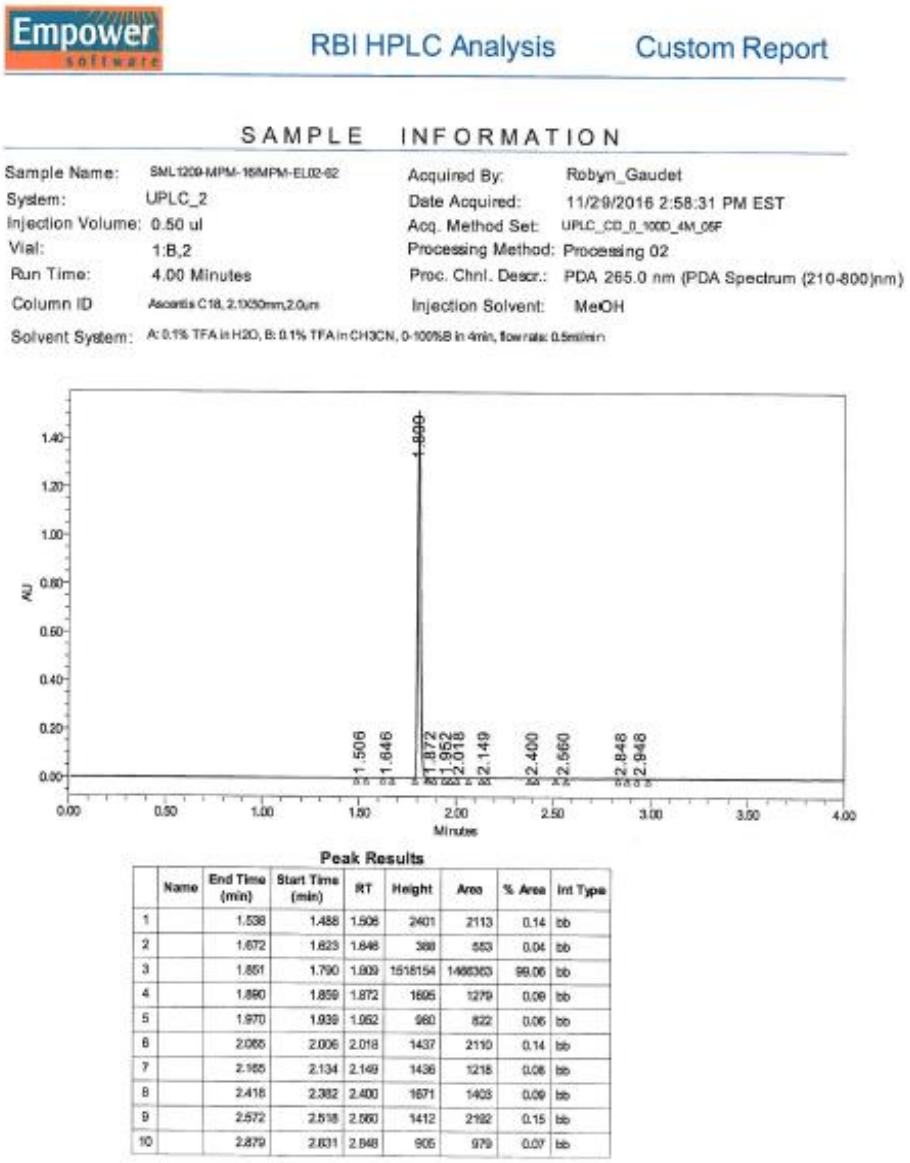
Figure S74. Mass Spectrum of 6-Chloro-*N*-[4-(4-methylpiperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-yl]-4-trifluoromethylnicotinamide **28**.



	Parameter	Value
1	Data File Name	Y:/ MM/ nmr/ SML1209-MPM-16-Step-5-ELN-02-62-1-After100degVacOvenOvernight/ 10/ fid
2	Title	SML1209-MPM-16-Step-5-ELN-02-62-1-After100degVacOvenOvernight
3	Solvent	CDCl3
4	Temperature	293.1
5	Pulse Sequence	zg30
6	Experiment	1D
7	Number of Scans	16
8	Receiver Gain	180
9	Relaxation Delay	1.0000
10	Pulse Width	15.8800
11	Acquisition Date	2016-12-01T09:50:09
12	Spectrometer Frequency	400.13
13	Spectral Width	8012.8
14	Lowest Frequency	-1544.3
15	Nucleus	1H
16	Acquired Size	32768
17	Spectral Size	65536

^{1}H NMR (400 MHz, $CDCl_3$) δ 9.00 (s, 1H), 8.66 (d, J = 1.5 Hz, 1H), 7.90 (s, 1H), 7.60 – 7.49 (m, 2H), 7.42 – 7.35 (m, 2H), 7.32 (d, J = 8.2 Hz, 2H), 7.02 (s, 1H), 4.15 – 4.09 (m, 1H), 3.78 – 3.67 (m, 4H), 3.57 (s, 2H), 2.95 (s, 4H), 2.53 (d, J = 37.1 Hz, 8H), 2.37 (s, 3H).

Figure S75. ^{1}H-NMR of 6-Oxo-4-trifluoromethyl-1,6-dihydro-pyridine-3-carboxylic acid [4-(4-methyl-piperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-yl]-amide **29**.



Peak Results

	Name	End Time (min)	Start Time (min)	RT	Height	Area	% Area	Int Type
1		1.538	1.488	1.506	2401	2113	0.14	bb
2		1.672	1.623	1.646	388	553	0.04	bb
3		1.851	1.790	1.909	1516154	1466363	99.06	bb
4		1.890	1.859	1.872	1695	1279	0.09	bb
5		1.970	1.939	1.952	980	822	0.06	bb
6		2.055	2.006	2.018	1437	2110	0.14	bb
7		2.165	2.134	2.149	1436	1218	0.08	bb
8		2.418	2.382	2.400	1671	1403	0.09	bb
9		2.572	2.518	2.560	1412	2192	0.15	bb
10		2.879	2.831	2.848	905	979	0.07	bb

Figure S76. HPLC of 6-Oxo-4-trifluoromethyl-1,6-dihydro-pyridine-3-carboxylic acid [4-(4-methyl-piperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-yl]-amide **29**.

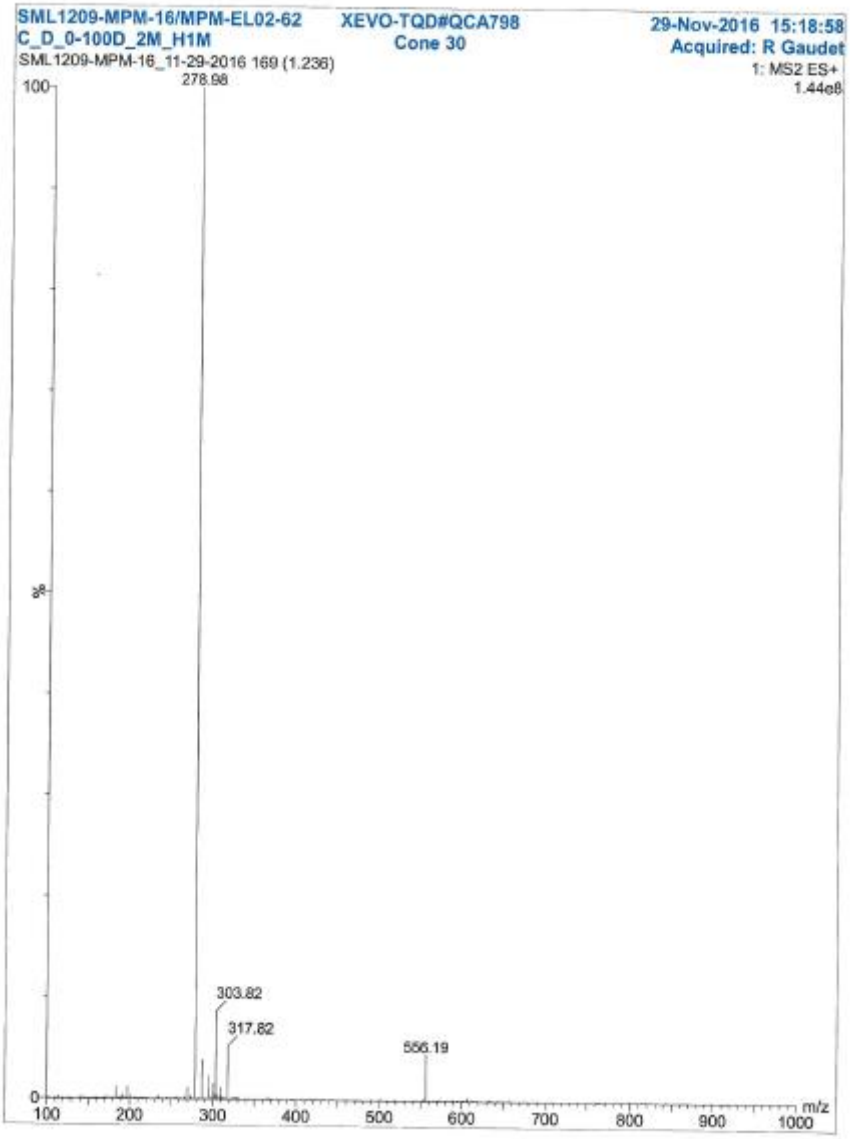
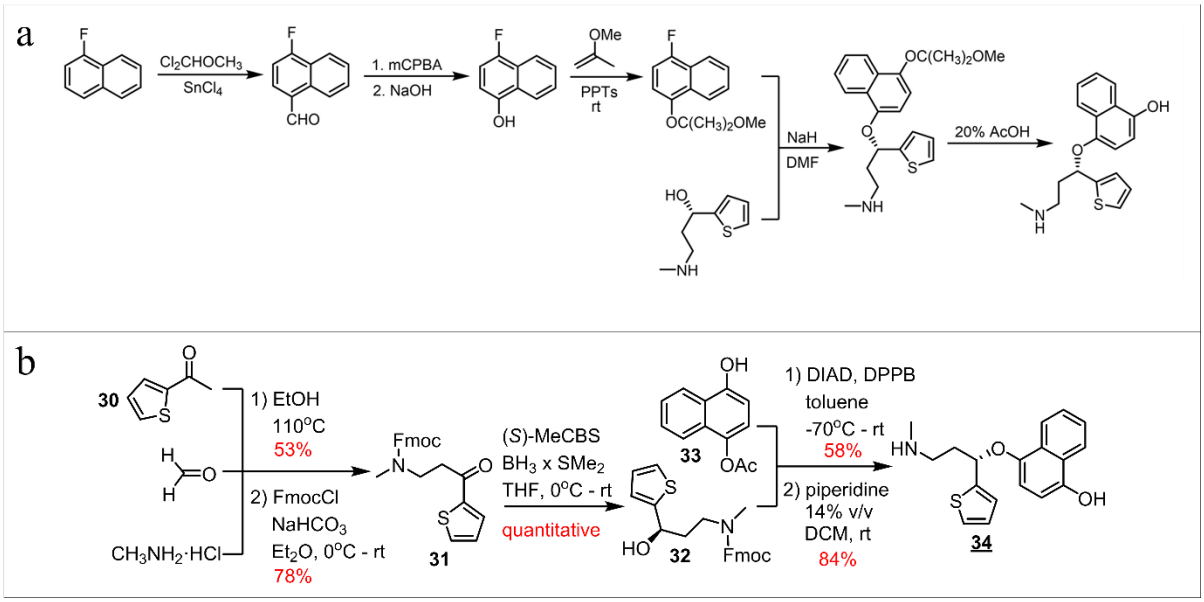


Figure S77. Mass Spectrum of 6-Oxo-4-trifluoromethyl-1,6-dihydro-pyridine-3-carboxylic acid [4-(4-methyl-piperazin-1-yl)-3'-morpholin-4-ylmethyl-biphenyl-3-yl]-amide **29**.

Section S14. Synthesis of (*S*)-hydroxyduloxetine (34)

S14.1. Previous vs. current synthetic routes.

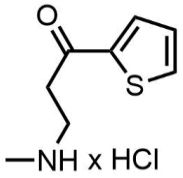


Scheme S5. (a) The original preparation of **34** from the main-text reference[27]. Authors did not provide experimental procedures, spectral data and syntheses yields. The scheme could not be reproduced by Sigma's chemists on multiple tries. For comparison, **(b)** shows the *Chematica* route (same as in the main-text **Figure 3a**).

S14.2. Synthetic details.

Reagents and solvents were purchased from commercial sources (Aldrich, ABCR, POCH, Chempur). All reagents were used without further purification unless otherwise noted. Flash column chromatography was performed using Merck silica gel 60 (230-400 mesh, 40-63 μm). Reactions were monitored using Macherey-Nagel silica gel 60F254 aluminium plates. TLC's were visualized by UV fluorescence (254 nm) or iodine vapors.

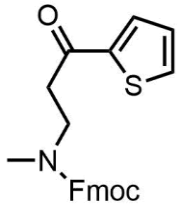
NMR spectra were recorded on a Bruker 400 MHz Avance III spectrometer at room temperature. Chemical shifts (δ) were reported in parts per million (ppm) relative to residual solvent peaks rounded to the nearest 0.01 (ref: $CHCl_3$ [^{1}H: 7.26, ^{13}C: 77.2]). Coupling constants (J) were reported in Hz to the nearest 0.1 Hz. Peak multiplicity was indicated as follows: s (singlet), d (doublet), t (triplet), q (quartet), qi (quintet), sx (sextet) and m (multiplet). HRMS spectra were recorded on AutoSpec Premier (Waters) or MaldiSYNAPT G2-S HDMS (Waters) spectrometers and are given in m/z. Enantiomeric excess of chiral compounds was measured using HPLC Merck HITACHI, pump L-7100, UV detector L-7400.



***3-(Methylamino)-1-(thiophen-2-yl)propan-1-one hydrochloride* SI-4**

Methylamine hydrochloride (5.30 g, 78.46 mmol), paraformaldehyde (3.00 g, 100 mmol) and 2-acetylthiophene **30** (9.00 g, 71.32 mmol) were placed in a glass vial, dissolved in EtOH (30 mL) and sealed. The reaction was stirred for 48 hrs at 110°C and then cooled to the room temperature. During the cooling process precipitation of a pale yellow powder is observed. The reaction mixture was evaporated to half the initial volume, AcOEt (60 mL) was added and the mixture was left overnight. The resulting precipitate was filtered off, dissolved in *i*PrOH (200 mL) and left for another night to crystallization. The resulting crystals were filtered off, washed with *i*PrOH and Et_2O and dried to yield **SI-4** (7.68 g, 53%).

^{1}H NMR (400 MHz, DMSO-d_6) δ 8.94 (s, 2H), 8.07 (dd, *J* = 4.9, 1.1 Hz, 1H), 8.01 (dd, *J* = 3.8, 1.1 Hz, 1H), 7.29 (dd, *J* = 4.9, 3.8 Hz, 1H), 3.46 (t, *J* = 6.8 Hz, 2H), 3.22 (t, *J* = 6.8 Hz, 2H), 2.58 (s, 3H).
^{13}C NMR (101 MHz, DMSO-d_6) δ 190.39, 143.20, 135.92, 134.37, 129.36, 43.68, 35.13, 33.03.
HRMS: (*m*/*z*): calcd for $C_8H_{12}NOS$, $[M+H]^+$, 170.0640; found 170.0637

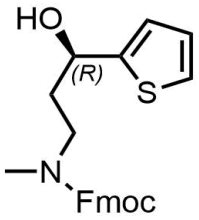


***3-(Methylamino)-3-(9-fluorenylmethoxycarbonyl)-1-(thiophen-2-yl)propan-1-one* 31**

A round-bottom flask was charged with **SI-4** (3.00 g, 14.67 mmol), FmocCl (4.61 g, 16.13 mmol) and $NaHCO_3$ (2.83g, 33.79 mmol) and Et_2O (80 mL) was added. The Ar atmosphere was established and the reaction mixture was cooled to 0°C. Then it was allowed to warm up to the room temperature and the reaction mixture was stirred for 20 hrs. Reaction was quenched by addition of water (110 mL) and water phase was extracted with AcOEt (3x60 mL). Combined organic phases were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by the flash column chromatography (hexane:AcOEt, 4:1) to yield **31** (4.48 g, 78%) as a white powder.
^{1}H NMR (400 MHz, $CDCl_3$) δ 7.67 (m, 6H), 7.37 (m, 4H), 7.14 (m, 1H), 4.57 (s, 1H), 4.43 (s, 1H), 4.25 (s, 1H), 3.71 (s, 1H), 3.43 (s, 1H), 3.22 (s, 1H), 2.95 (d, 3H), 2.79 (s, 1H).

^{13}C NMR (126 MHz, DMSO-d_6) δ 192.48, 144.50, 143.10, 139.91, 137.92, 134.94, 133.47, 129.33, 129.02, 127.68, 121.75, 120.39, 109.88, 52.55, 42.17, 40.59, 37.06.

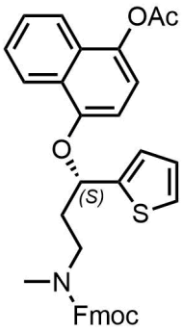
HRMS: (*m*/*z*): calcd for $C_{23}H_{21}NO_3SNa$, [M+Na]$^+$, 414.1140; found 414.1130



***(R)-3-(Methylamino)-3-(9-fluorenylmethoxycarbonyl)-1-(thiophen-2-yl)propan-1-ol* 32**

A round-bottom flask was charged with THF (8 mL) and Ar atmosphere was established. 1M solution of (*S*)-MeCBS in toluene (0.51 mL) was added dropwise, the reaction mixture was cooled to 0 ^{o}C and the 2M solution of $BH_3 \cdot SMe_2$ in THF (2.81 mL) was added. The reaction mixture was stirred for 10 min and then the solution of **31** in THF (40 mL) was added dropwise at the rate 1 mL/min. The reaction mixture was stirred for 4.5 hrs at room temperature. The reaction was quenched by addition of water (45 mL) and water phase was extracted with AcOEt (4x40 mL). Combined organic phases were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by the flash column chromatography (hexane:AcOEt, 3:1) to yield **32** (2.088 g, quant., 91% *ee*) as a colorless, thick oil.

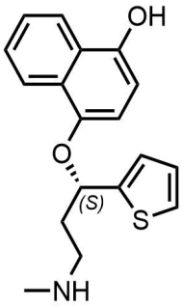
^{1}H NMR (400 MHz, $CDCl_3$) δ 7.78 (s, 2H), 7.62 (d, 2H), 7.46 – 7.30 (m, 4H), 7.25 (d, 1H), 6.98 (d, 2H), 4.76 (m, 1H), 4.64 – 4.40 (m, 2H), 4.25 (d, 1H), 3.87 (s, 1H), 3.19 (m, 1H), 2.89 (s, 3H), 1.98 (m, 2H).
^{13}C NMR (101 MHz, $CDCl_3$) δ 157.35, 148.07, 144.03, 141.38, 127.69, 127.10, 126.56, 124.87, 124.20, 123.08, 119.96, 67.49, 66.44, 47.46, 45.71, 36.81, 33.99.
HRMS: (*m*/*z*): calcd for $C_{23}H_{23}NO_3SNa$, [M+Na]$^+$, 416.1296; found 416.1289



(S)-3-(4-O-acylonaphthalen-1-yl)-N-(9-fluorenylmethoxycarbonyl)-N-methyl-3-(thiophen-2-yl)propan-1-amine SI-5

A round-bottom flask was charged with **32** (0.300 g, 0.762 mmol), **33** (0.231 g, 1.14 mmol) and 1,4-bis(diphenylphosphino)butane (0.423 g, 0.991 mmol). The flask was then capped with a rubber septum and the content was dried on high-vacuum pump for 10 min. Then, the N_2 atmosphere was established and dry toluene (4 mL) was added. The reaction mixture was cooled to -70 ^{o}C and a solution of DIAD (0.200 g, 0.990 mmol) in dry toluene (1 mL) was added dropwise. The resulting yellow suspension was stirred for 18 hrs, during that time the reaction mixture was allowed to warm up to room temperature. Next, the solvent was evaporated and the resulting brown oil was purified by the flash column chromatography ($CHCl_3$:AcOEt, 40:1) to yield **SI-5** (0.253 g, 58%) as white foam. The NMR analysis revealed traces of impurities, but attempts at purifying the product resulted in its partial decomposition – accordingly, it was used in the next reaction without further purification.

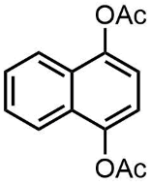
^{1}H NMR (400 MHz, $CDCl_3$) δ 8.36 (d, 1H), 7.76 (m,, 3H), 7.66 – 7.29 (m, 8H), 7.24 (m, 1H), 7.07 (m, 2H), 6.96 (m, 1H), 6.79 (s, 1H), 5.59 (d, 1H), 4.44 (d, 2H), 4.27 – 4.06 (m, 1H), 3.76 – 3.32 (m, 2H), 2.93 (s, 3H), 2.44 (s, 3H), 2.39 – 2.22 (m, 1H), 2.21 – 1.96 (m, 1H).
^{13}C NMR (101 MHz, $CDCl_3$) δ 169.71, 156.21, 151.16, 144.14, 144.01, 141.34, 140.36, 127.62, 127.55, 127.03, 126.93, 126.76, 126.66, 125.89, 125.02, 124.75, 122.45, 120.98, 119.92, 117.56, 106.05, 88.79, 74.21, 67.09, 47.37, 46.47, 37.19, 24.11, 20.95.
HRMS: (*m*/*z*): calcd for $C_{35}H_{31}NO_5SNa$, [M+Na]$^+$, 600.1821; found 600.1827



(S)-3-(4-hydroxynaphthalen-1-yl)-N-methyl-3-(thiophen-2-yl)propan-1-amine 34

A round-bottom flask was charged with **SI-5** (0.443 g, 0.767 mmol) and freshly distilled DCM (2.5 mL) was added. The flask was capped with a rubber septum, the N_2 atmosphere was established, and piperidine (0.4 mL) was added dropwise. The reaction mixture was stirred for 4 hrs at room temperature. Then, the solvents were removed *in vacuo* and the residue was purified by the flash column chromatography ($CHCl_3$:MeOH, 20:1 → $CHCl_3$:MeOH, 10:1→ $CHCl_3$:MeOH, 3:1) to yield **34** (0.203 g, 84%.) as a white glassy solid.

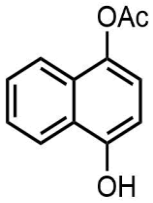
^{1}H NMR (400 MHz, Acetone-d_6) δ 8.27 (dd, 1H), 8.20 (dd, 1H), 7.52 – 7.44 (m, 2H), 7.34 (dd, 1H), 7.12 (d, 1H), 6.94 (dd, 1H), 6.82 (d, 1H), 6.76 – 6.69 (d, 1H), 5.86 (dd, 1H), 4.75 (s, 2H), 2.89 – 2.76 (m, 2H), 2.53 – 2.44 (m, 1H), 2.41 (s, 3H), 2.22 (m, 1H).
^{13}C NMR (101 MHz, Acetone-d_6) δ 147.27, 146.41, 145.88, 127.29, 126.35, 125.82, 125.35, 125.06, 124.93, 124.71, 122.07, 121.86, 108.45, 107.03, 74.88, 47.82, 38.37, 35.45.
HRMS: (*m*/*z*): calcd for $C_{18}H_{20}NO_2S$, $[M+H]^+$, 314.1215; found 314.1219



1,4-di-O-acylonaphtalene SI-6

A round-bottom flask was charged with 1,4-naphtoquinone (7.25 g, 45.84 mmol), which was dissolved in AcOEt (150 mL). The flask was capped with a rubber septum and the Ar atmosphere was established. The septum was removed, 10% Pd/C (0.17 g) was added and the flask was again capped with a septum. The reaction mixture was flushed with H_2 and was then stirred at room temperature under H_2 atmosphere for 18 hrs. After completion of the reaction, Pd/C was filtered off, and the reaction mixture was filtered through silica pad with AcOEt. Concentration *in vacuo* afforded 6.98 g (98% crude) 1,4-dihydroxynaphtalene as a brown solid which was used in the next step without further purification.
1,4-dihydroxynaphtalene (6.97 g, 43.55 mmol) was dissolved in pyridine (69.95 mL) and acetic anhydride (62.57 mL) was added. The reaction was carried at room temperature for 5 hrs and then the reaction mixture was evaporated to dryness. Purification by column chromatography (Hexane:AcOEt = 3:1) afforded 9.33 g of 1,4-di-*O*-acylonaphtalene **SI-6** as white powder (83% after two steps).

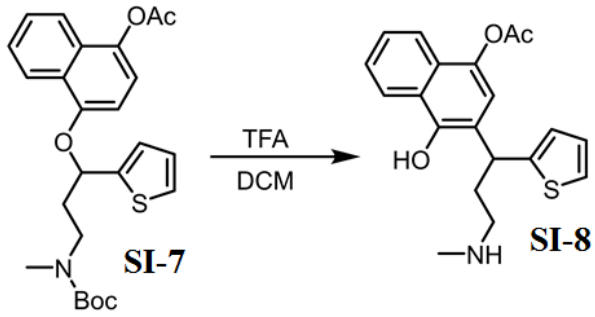
^{1}H NMR (400 MHz, $CDCl_3$) δ 7.92 (dd, *J* = 6.5, 3.2 Hz, 2H), 7.58 (dd, *J* = 6.5, 3.2 Hz, 2H), 7.29 (s, 2H), 2.48 (s, 6H).
^{13}C NMR (101 MHz, $CDCl_3$) δ 169.29, 144.35, 127.67, 126.97, 121.63, 117.66, 20.97.
HRMS: (*m*/*z*): calcd for $C_{14}H_{12}O_4Na$, $[M+Na]^+$, 267.0633; found 267.0629



***1-O-acylo-4-hydroxynaphtalene* 33**

A round-bottom flask was charged with 1,4-di-*O*-acylonaphtalene (**SI-6**) (1.34 g, 5.50 mmol), which was suspended in EtOH (54 mL). Then, $NaBH_4$ (0.11 g, 3.02 mmol) was added, the Ar atmosphere was established and the reaction was stirred for 2.5 hrs at room temperature. Reaction was quenched by addition of water (50 mL). The reaction mixture was cooled to 0°C, acidified to pH = 2 and extracted with DCM (3x50 mL). Combined organic phases were washed with $NaHCO_3$ (50 mL), water (50 mL) and brine (50 mL), dried over anhydrous $MgSO_4$ and evaporated to dryness. Purification by the column chromatography (hexane : AcOEt = 7:1) afforded 0.878 g of 1-*O*-acylo-4-hydroxynaphtalene (**33**) (79%).

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.13 (d, 1H), 7.79 (d, 1H), 7.52 (dt, 2H), 7.01 (d, 1H), 6.59 (d, 1H), 5.74 (s, 1H), 2.49 (s, 4H).
^{13}C NMR (101 MHz, $CDCl_3$) δ 170.53, 149.66, 139.92, 127.44, 126.92, 125.63, 125.15, 122.24, 120.90, 117.84, 107.76, 21.00.
HRMS: (*m/z*): calcd for $C_{12}H_{10}O_3Na$, $[M+Na]^+$, 225.0526; found 225.0528



A round-bottom flask was charged with Boc-protected amine (**SI-7**) (0.092 g, 0.202 mmol), which was suspended in DCM (1.00 mL). Ar atmosphere was established, TFA (0.165 mL) was added and the reaction was stirred for 80 min at room temperature. Reaction was quenched by addition of 10% water solution of NaOH and extracted with Et_2O. Combined organic phases were dried over anhydrous $MgSO_4$ and evaporated to dryness. NMR of crude reaction mixture revealed that an ether rearrangement has occurred.

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.43 (dd, 1H), 7.71 – 7.66 (m, 1H), 7.54 – 7.45 (m, 2H), 7.23 (d, 1H), 7.00 (dd, 2H), 6.78 (s, 1H), 5.04 (dd, 1H), 2.85 – 2.72 (m, 1H), 2.59 – 2.49 (m, 4H), 2.36 (s, 3H), 2.22 (dd, 1H).
ESI(+): (m/z): for $C_{18}H_{20}NO_2S$, $[M+H]^+$, 356.3

S14.3. Raw spectroscopic and chromatographic data.

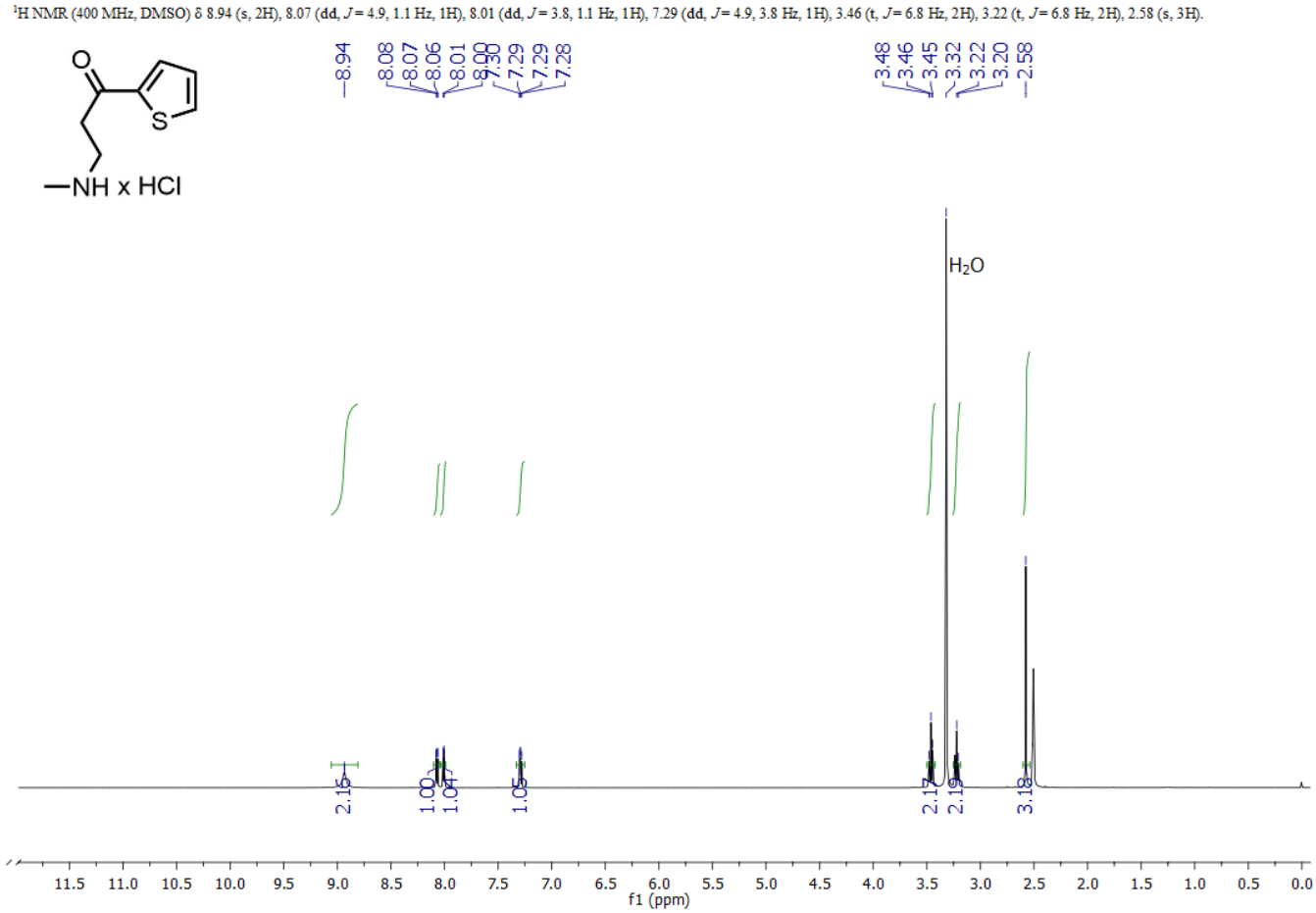


Figure S78. ^{1}H NMR spectrum of compound **SI-4**.

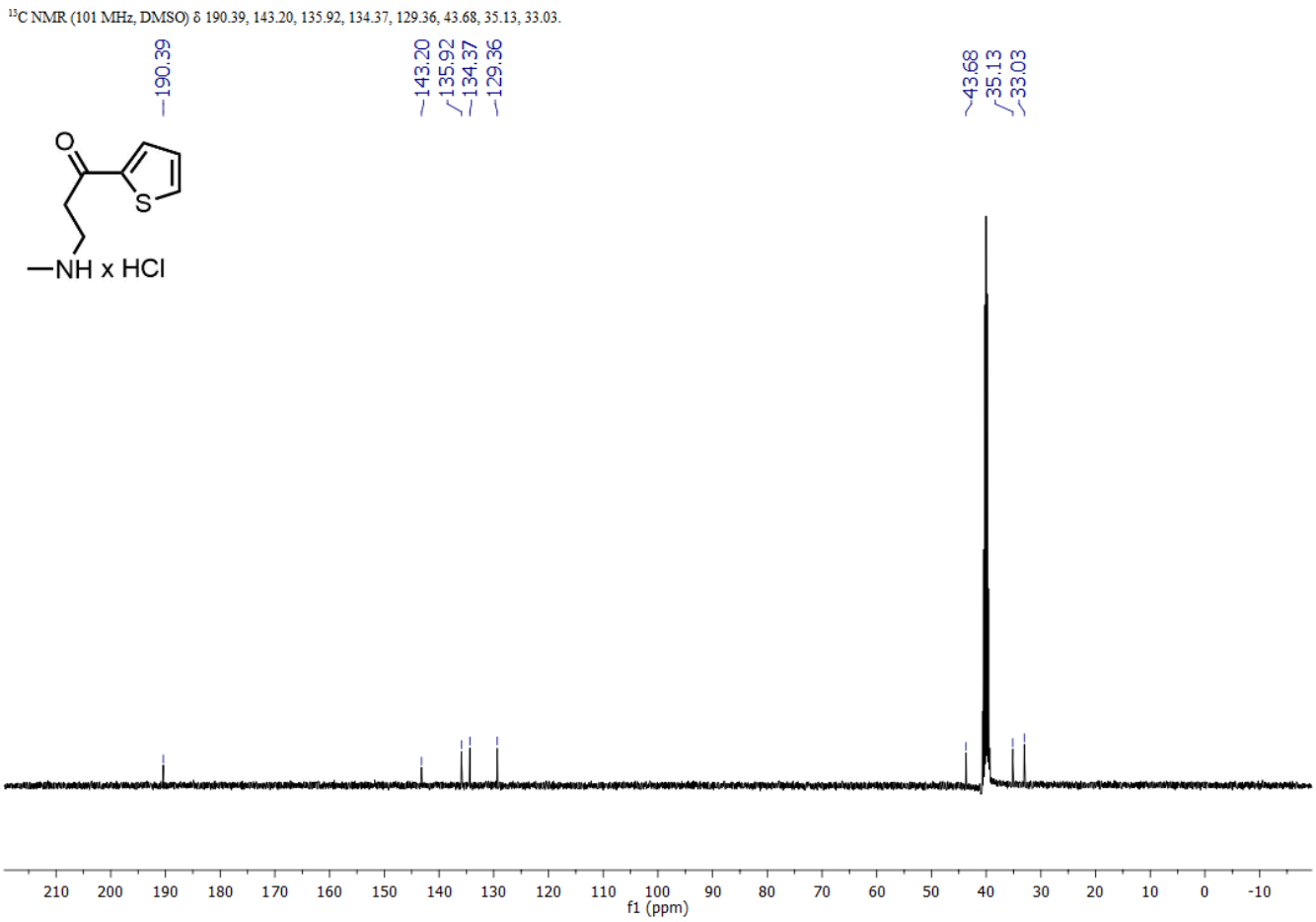


Figure S79. ^{13}C NMR spectrum of compound **SI-4**.

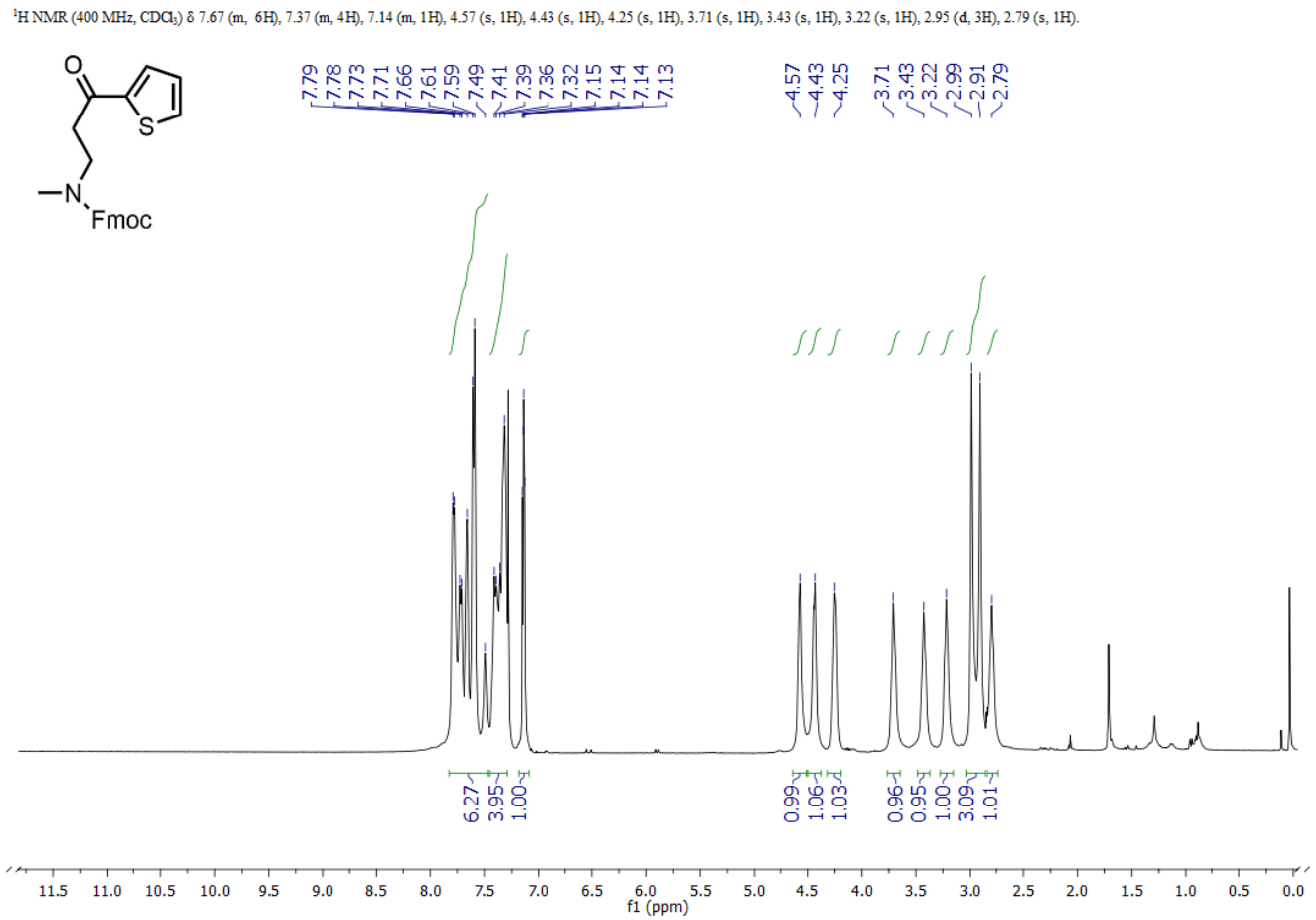


Figure S80. ^{1}H NMR spectrum of compound **31**.

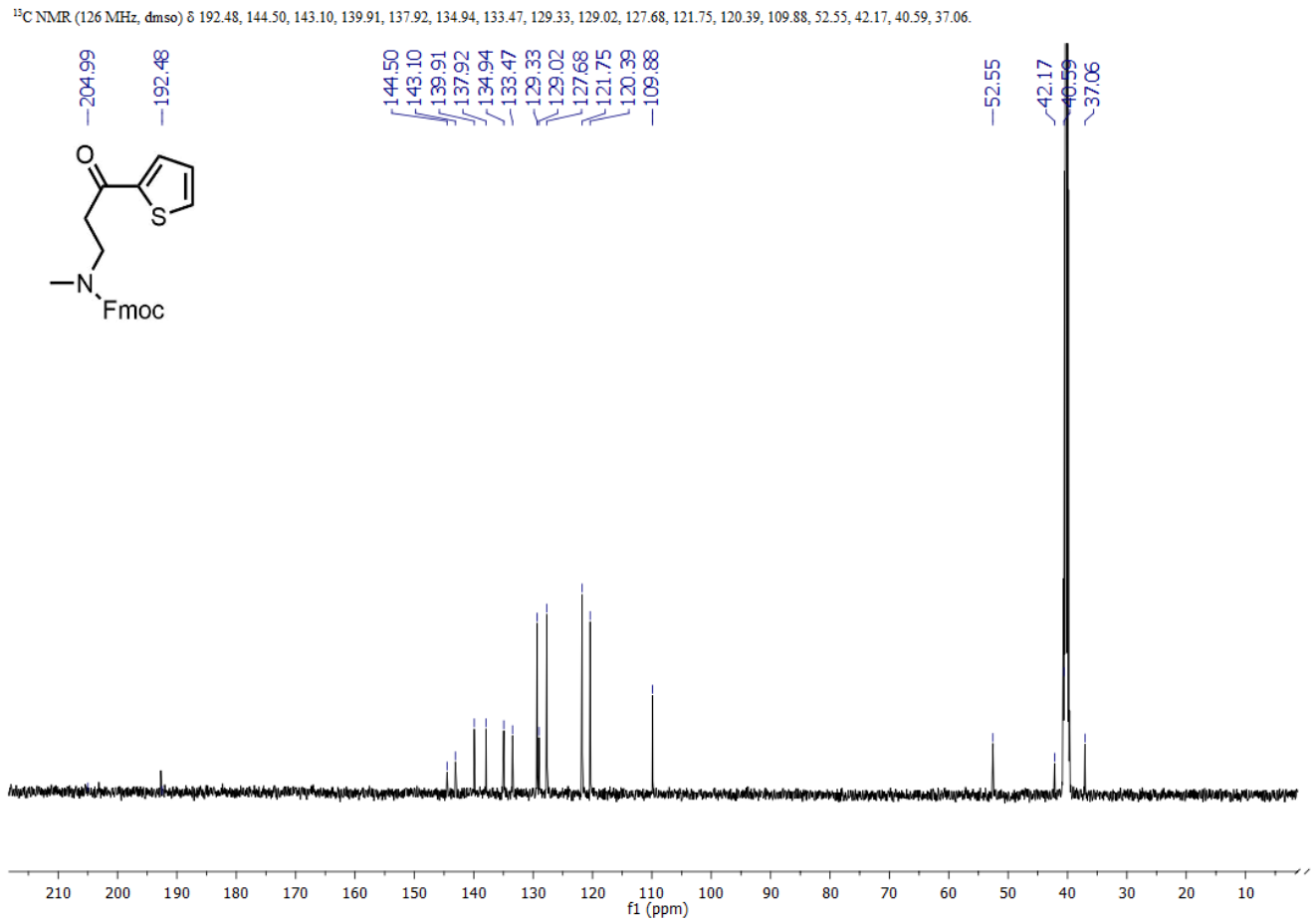


Figure S81. ^{13}C NMR spectrum of compound **31**.

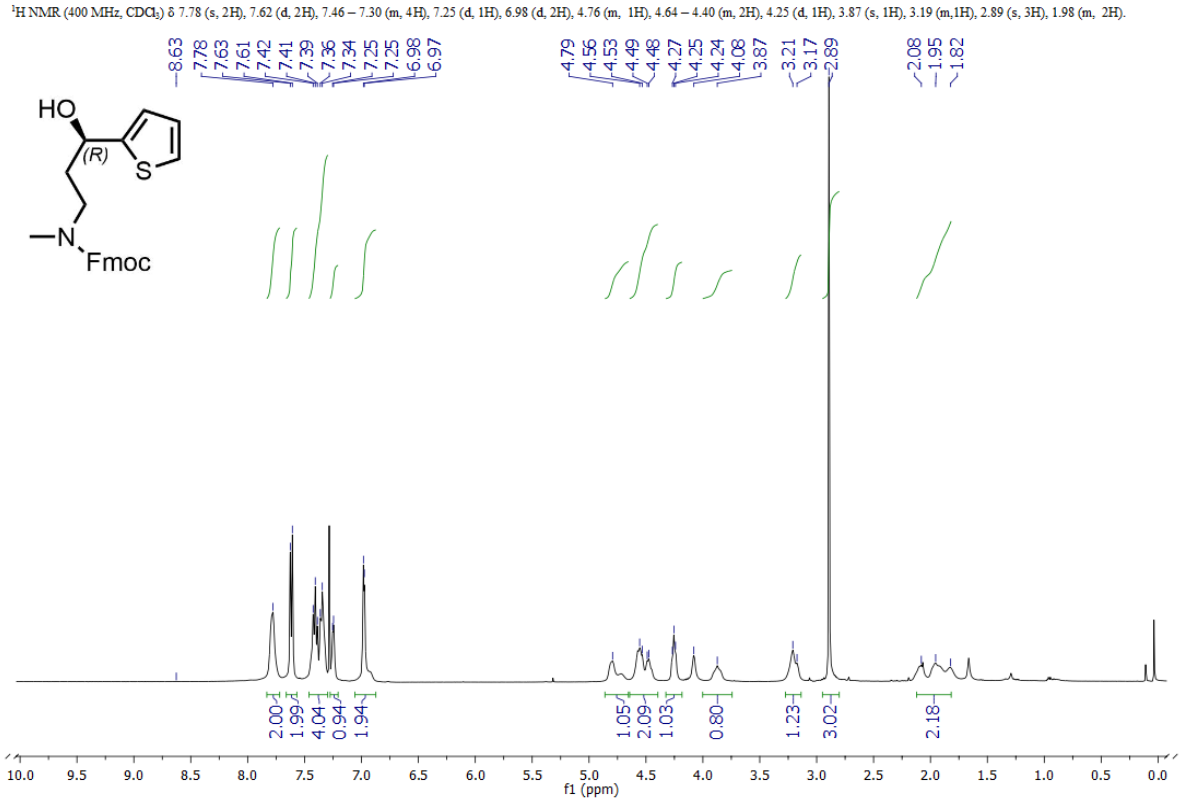


Figure S82. ^{1}H NMR spectrum of compound **32**.

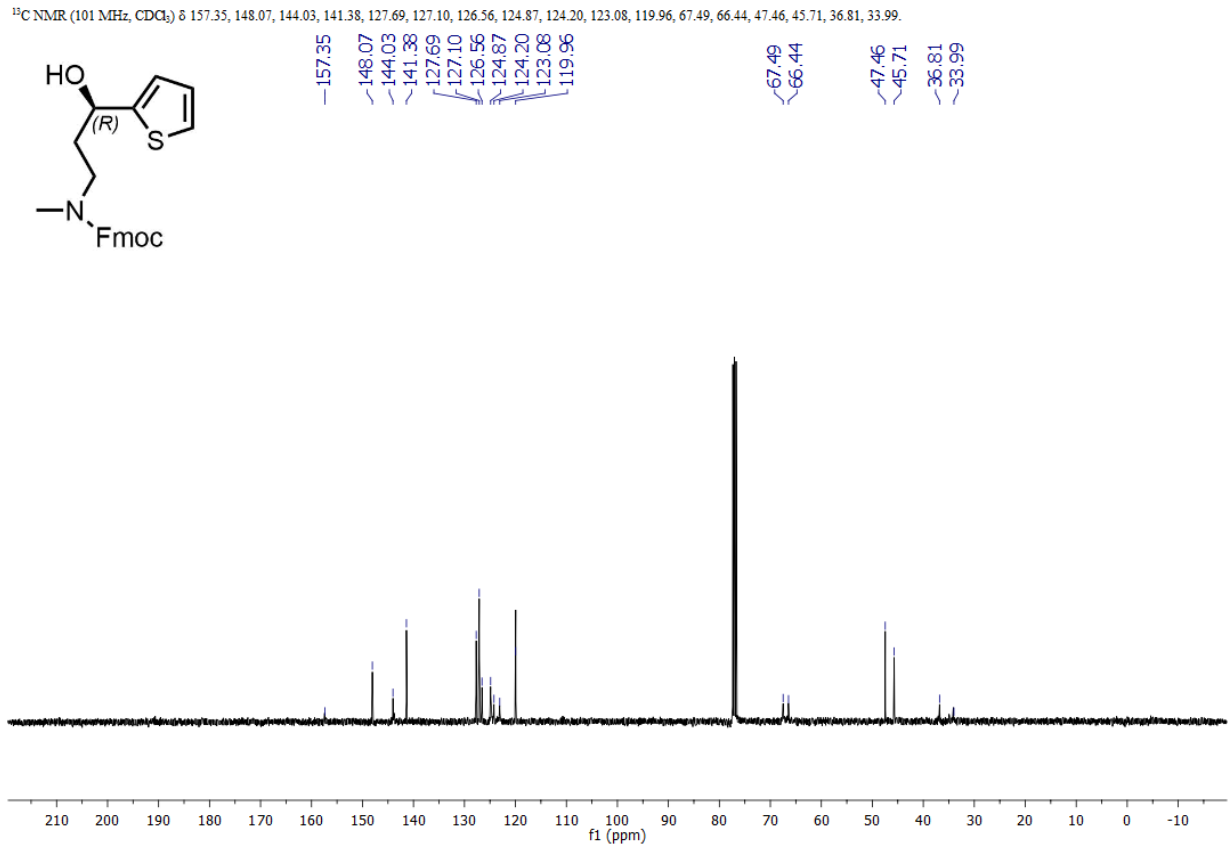


Figure S83. ^{13}C NMR spectrum of compound **32**.

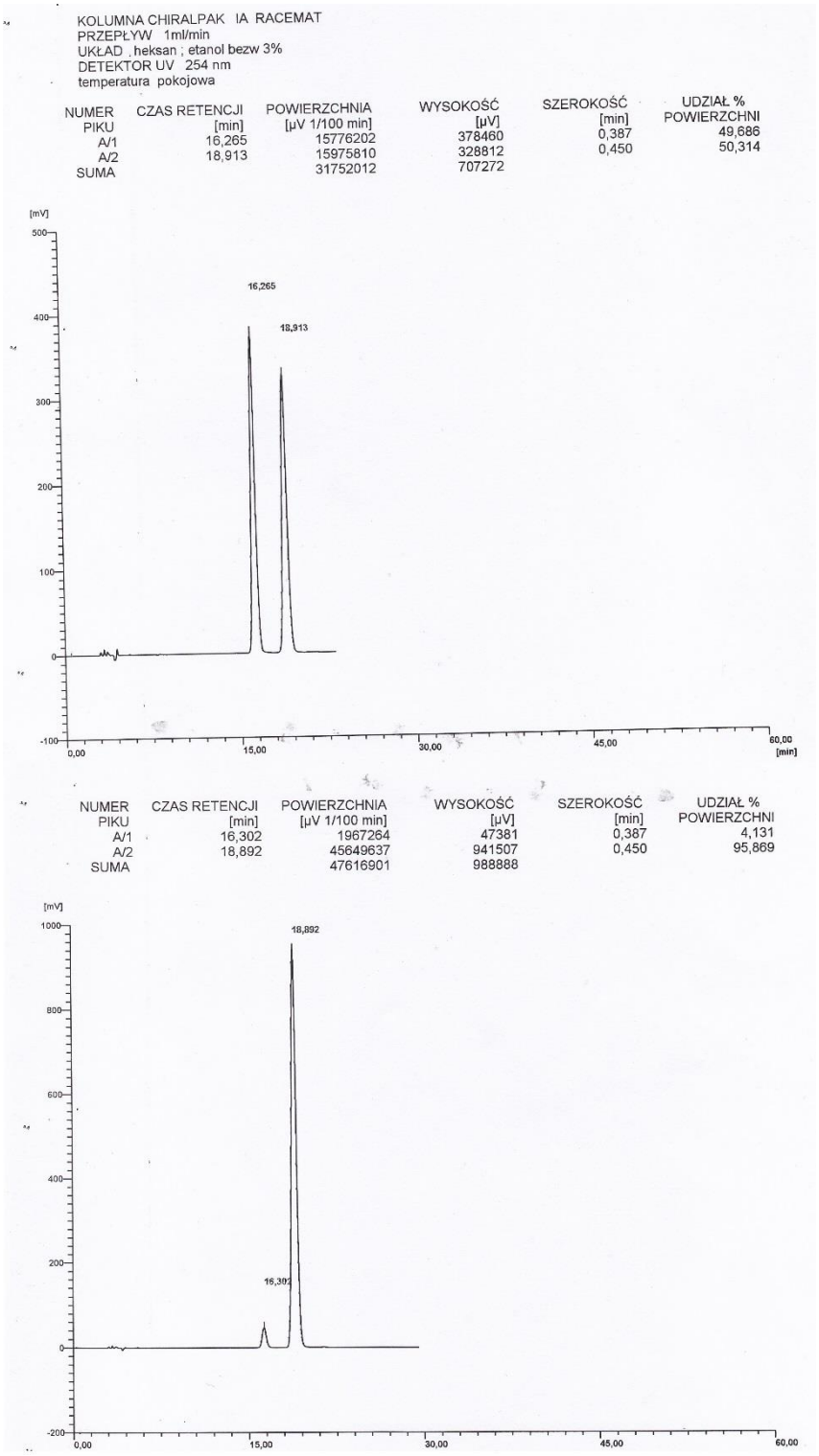


Figure S84. HPLC of compound **32.** Enantiomeric excess was measured using ChiralPak IA HPLC column (3% EtOH in hexane, 1 mL/min flow rate, detection at λ = 254 nm absorbance).

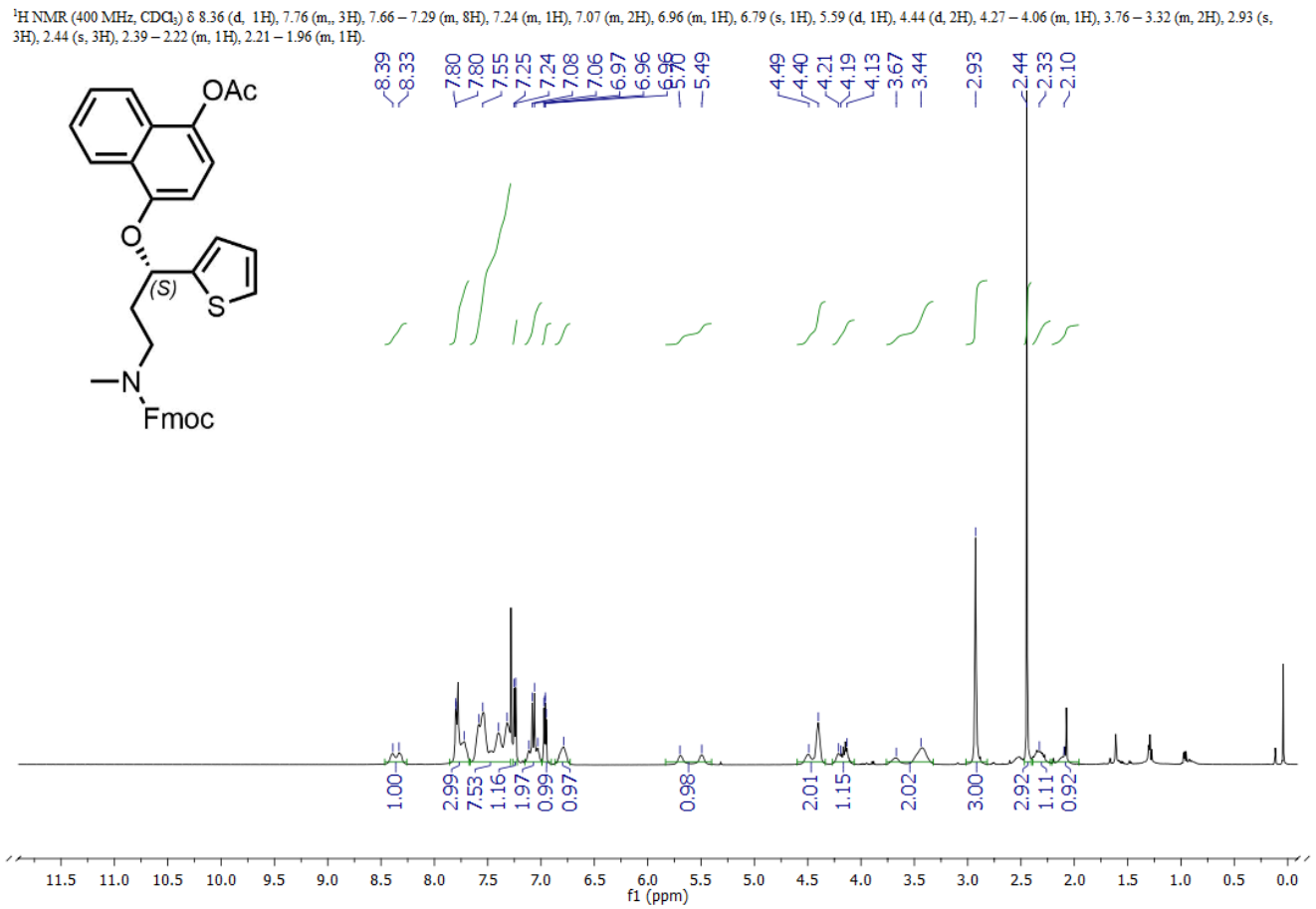


Figure S85. ^{1}H NMR spectrum of compound **SI-5**.

^{13}C NMR (101 MHz, $CDCl_3$) δ 169.71, 156.21, 151.16, 144.14, 144.01, 141.34, 140.36, 127.62, 127.55, 127.03, 126.93, 126.76, 126.66, 125.89, 125.02, 124.75, 122.45, 120.98, 119.92, 117.56, 106.05, 88.79, 74.21, 67.09, 47.37, 46.47, 37.19, 24.11, 20.95.

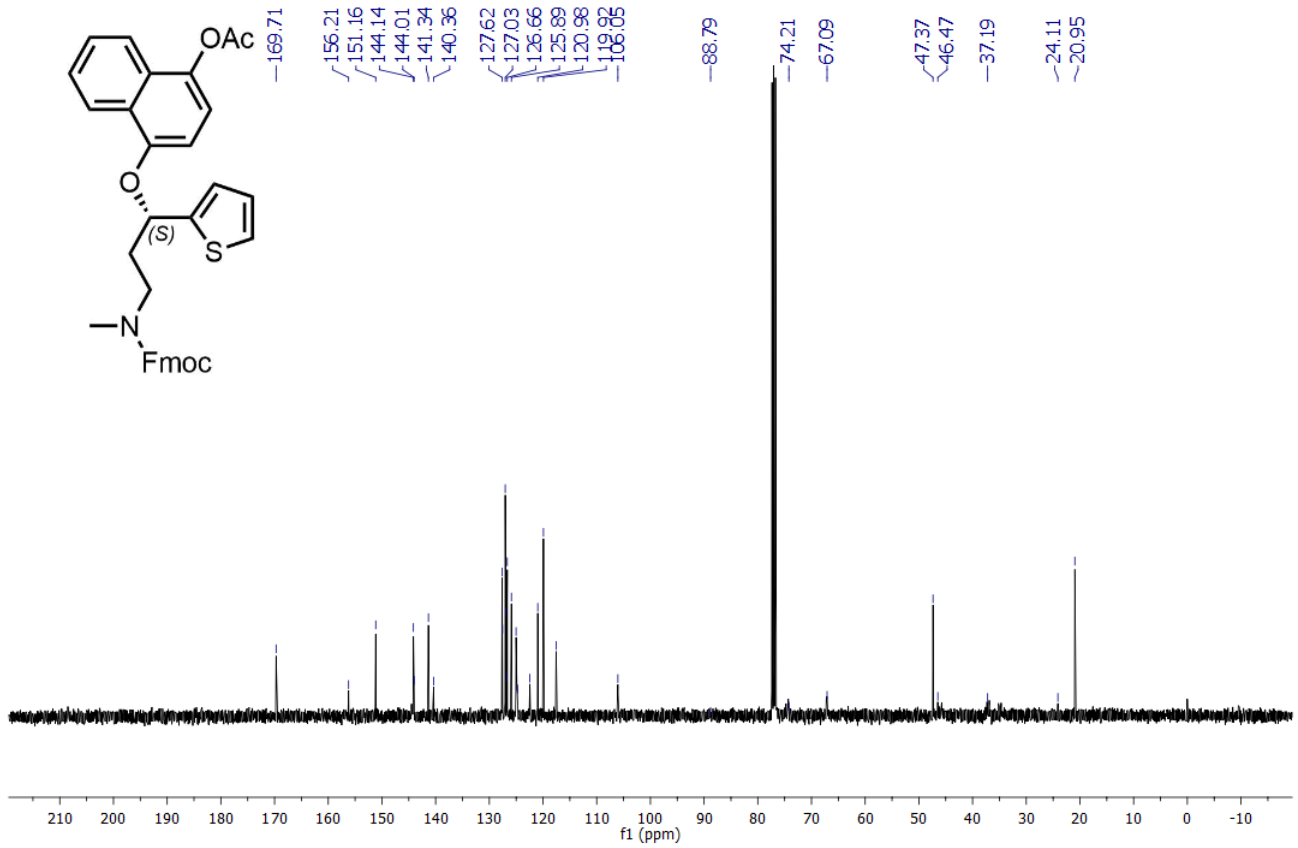


Figure S86. ^{13}C NMR spectrum of compound **SI-5**.

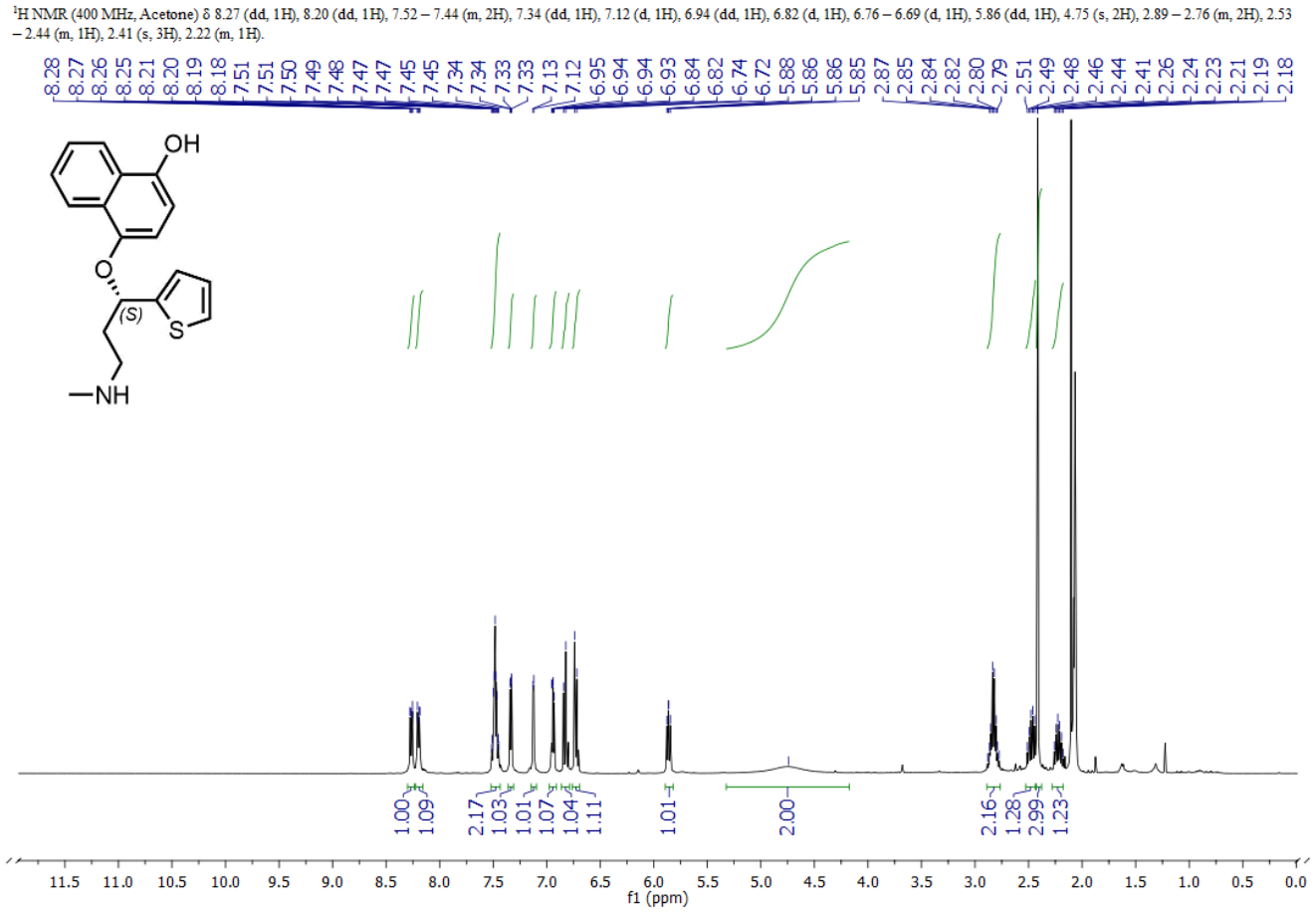


Figure S87. ^{1}H NMR spectrum of compound **34**.

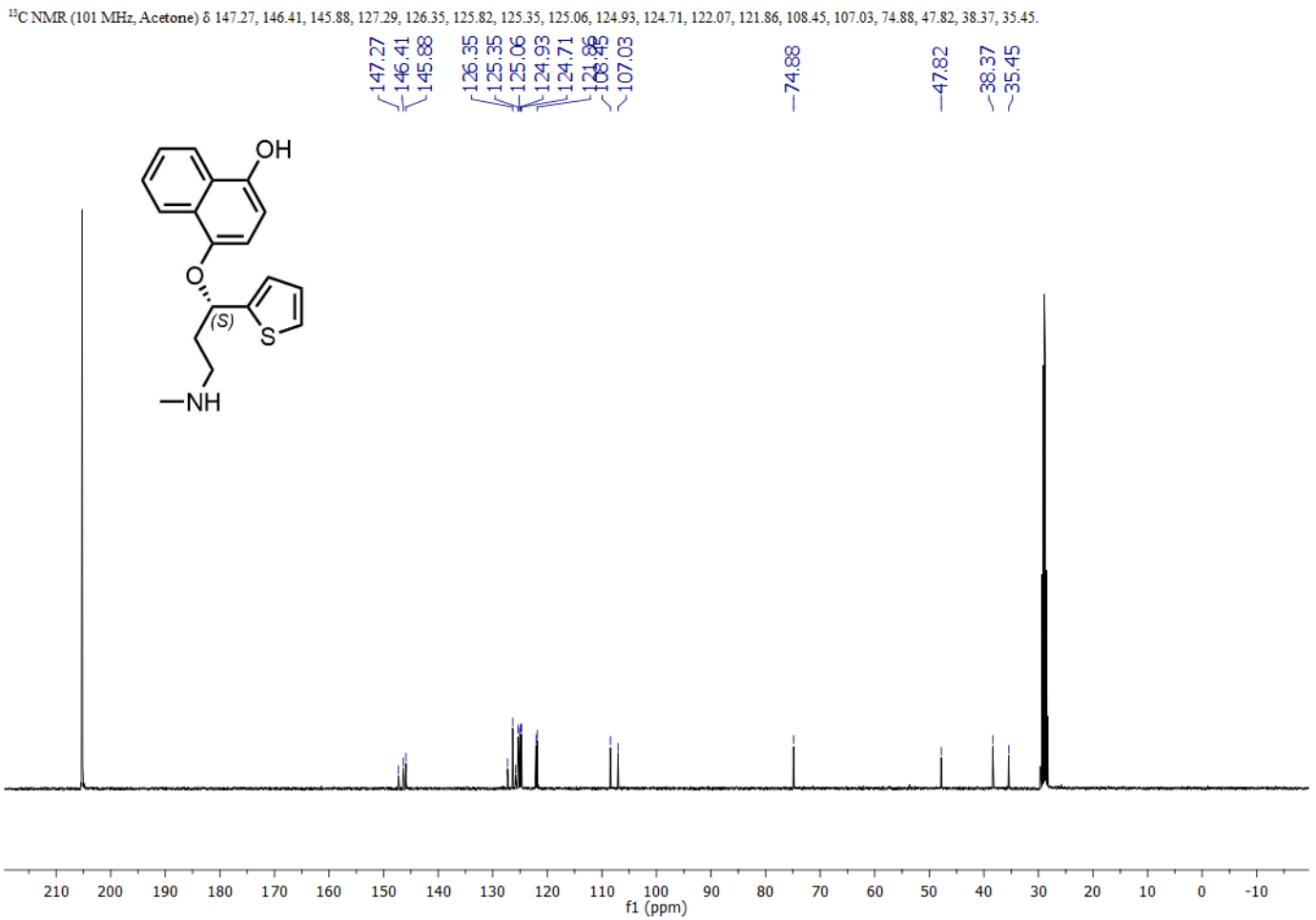


Figure S88. ^{13}C NMR spectrum of compound **34.**

Nr of peaks	Retention time	Area	Height	Area %
NUMER PIKU	CZAS RETENCJI [min]	POWIERZCHNIA [µV 1/100 min]	WYSOKOŚĆ [µV]	UDZIAŁ % POWIERZCHNI
A/1	3,500	8019401	263749	94,404
A/2	5,005	7696	641	0,091
A/3	5,637	8115	535	0,096
A/4	5,907	185834	5295	2,188
A/5	8,015	177123	8746	2,085
A/6*	8,627	45270	1742	0,533
A/7*	9,133	26091	1085	0,307
A/8	10,108	25245	1261	0,297
SUMA		8494775	283054	

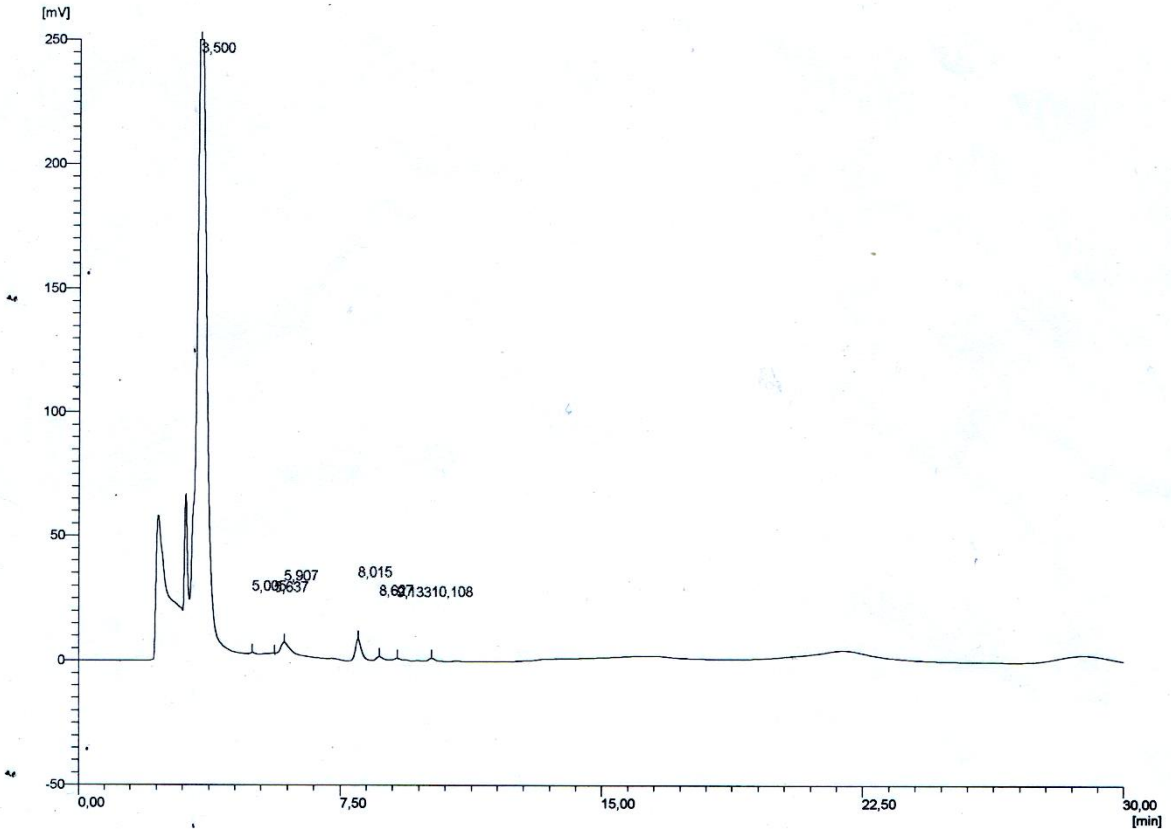


Figure S89. HPLC purity of compound 34 (column ID: KROMASIL SI60 C18 5µL, column size: 250 x 4.6mm, Solvent system: MeOH : H_2O : HCO_2H (55:45:0.025), Flow rate: 0.9 mL/min, Detector UV: 254nm absorbance).

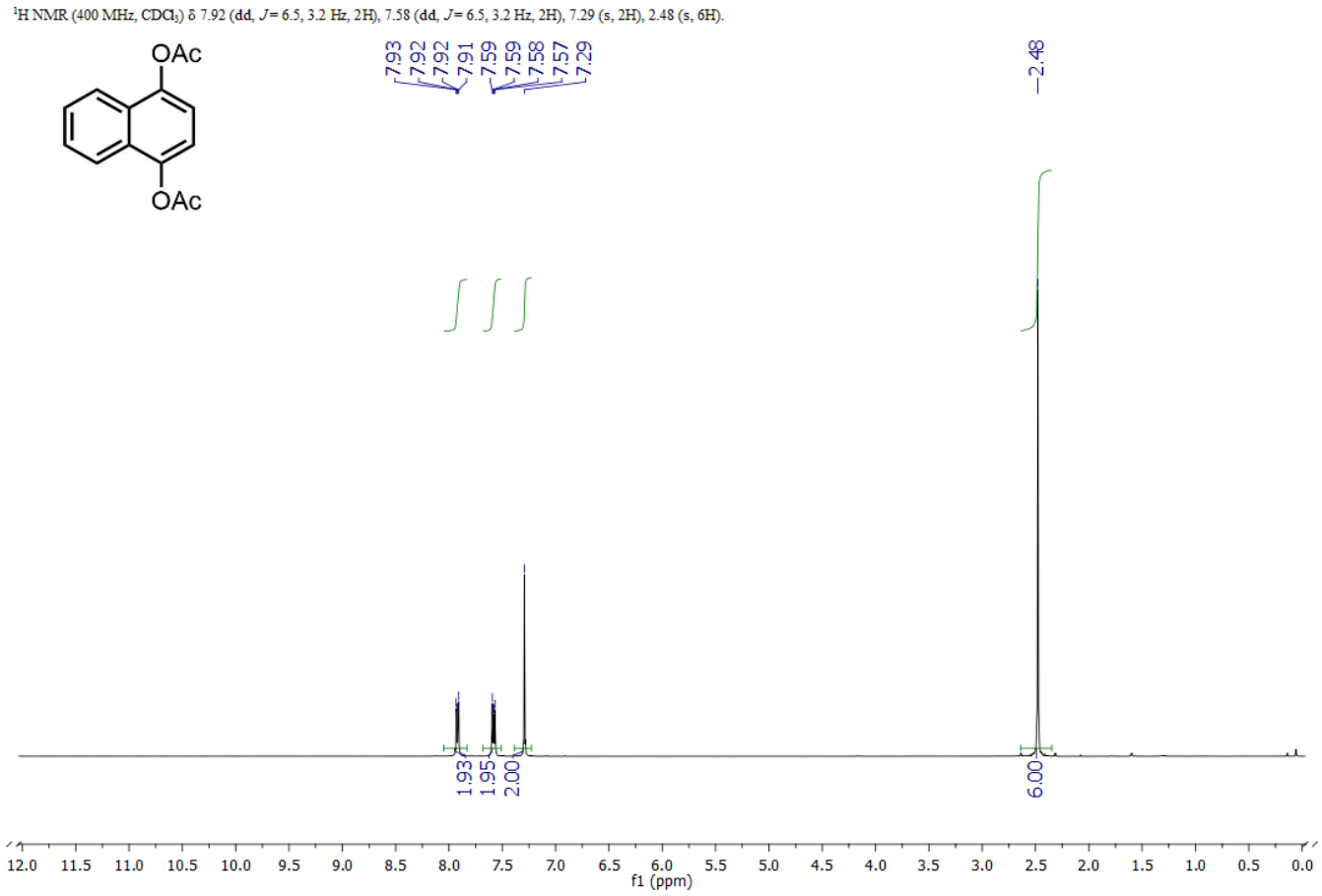


Figure S90. ^{1}H NMR spectrum of compound **SI-6**.

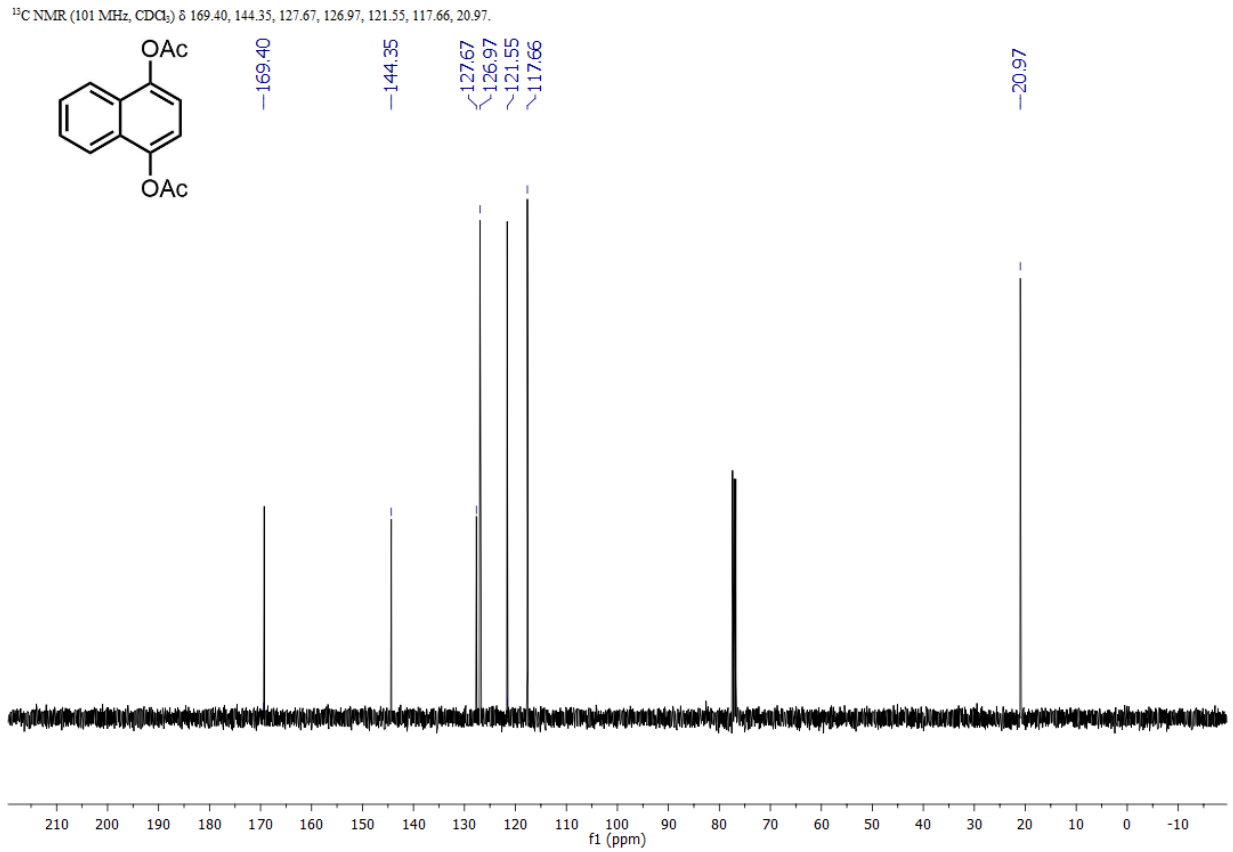


Figure S91. ^{13}C NMR spectrum of compound **SI-6**.

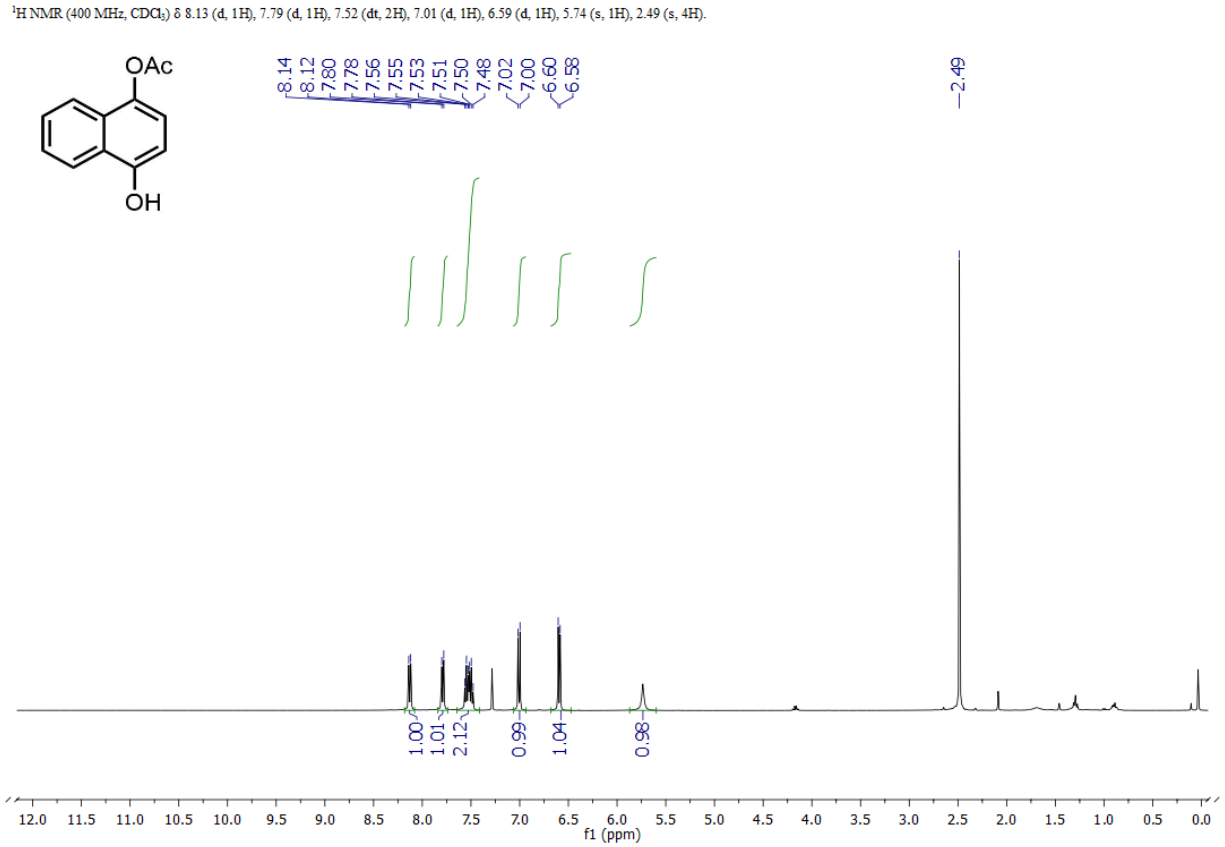


Figure S92. ^{1}H NMR spectrum of compound **33**.

^{13}C NMR (101 MHz, $CDCl_3$) δ 170.53, 149.66, 139.92, 127.44, 126.92, 125.63, 125.15, 122.24, 120.90, 117.84, 107.76, 21.00.

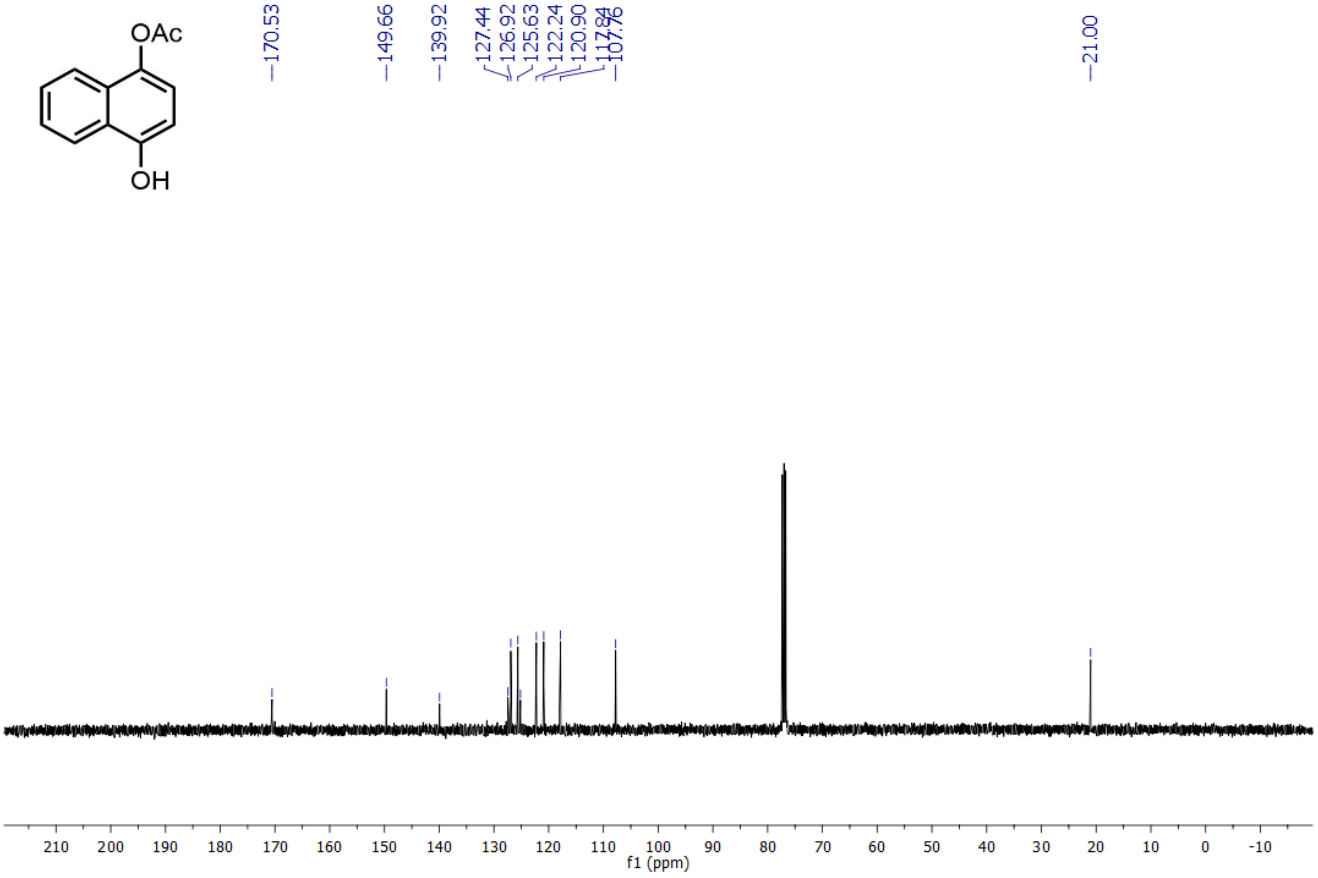


Figure S93. ^{13}C NMR spectrum of compound **33**.

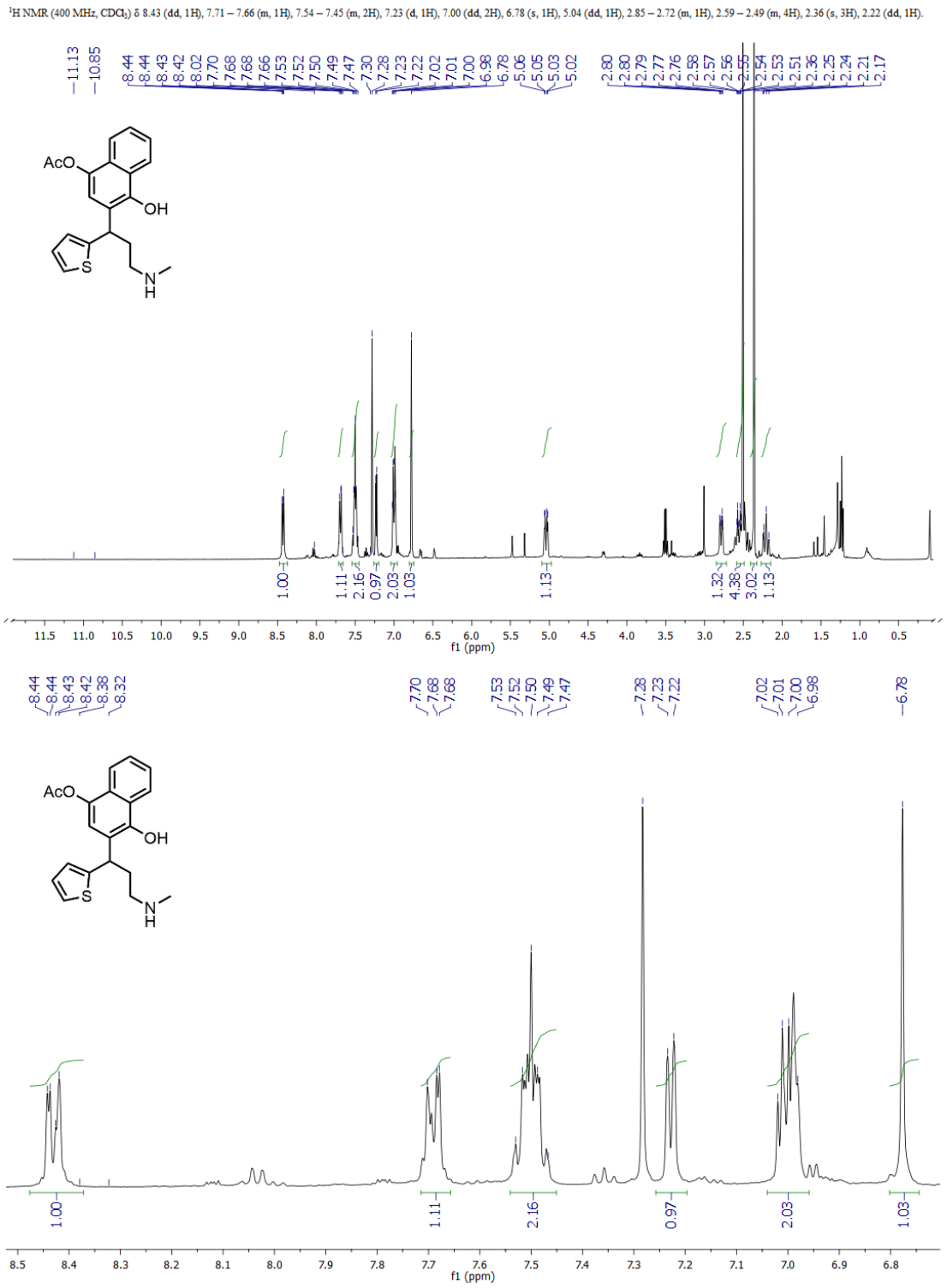


Figure S94. ^{1}H NMR spectrum of compound **SI-8** (crude reaction mixture).

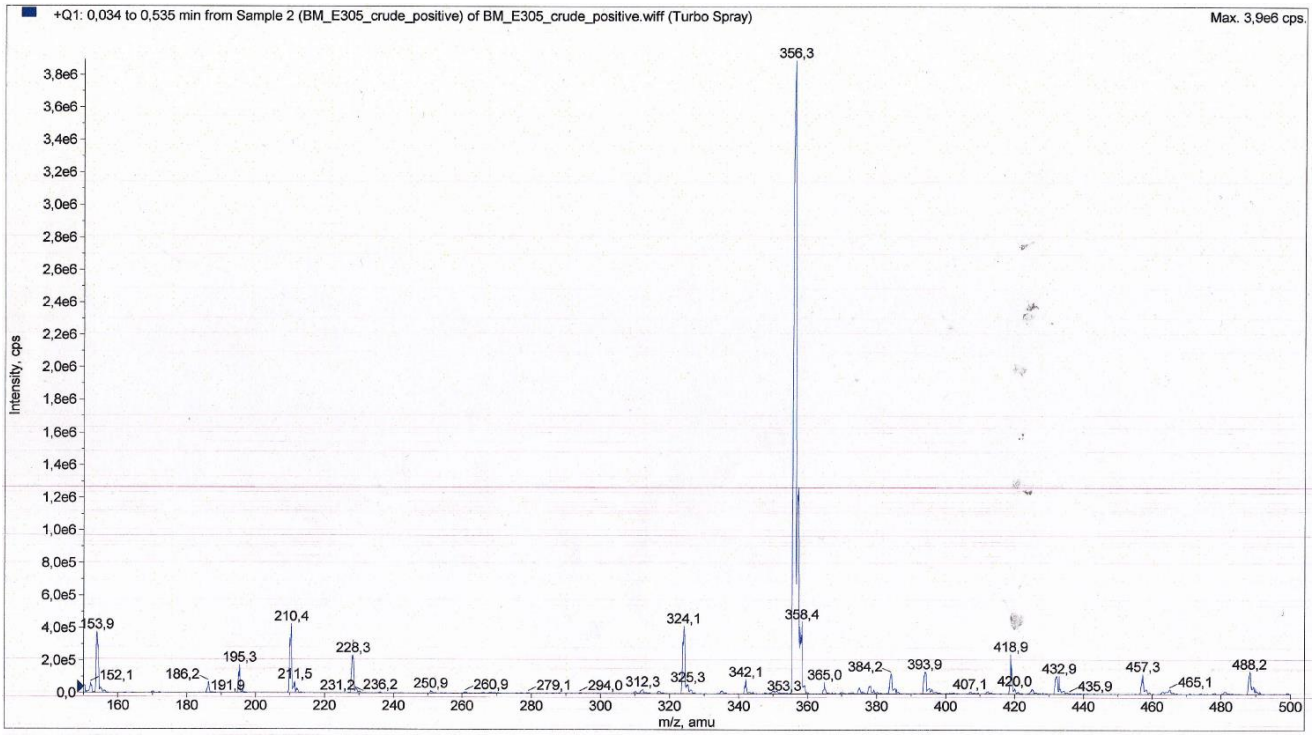
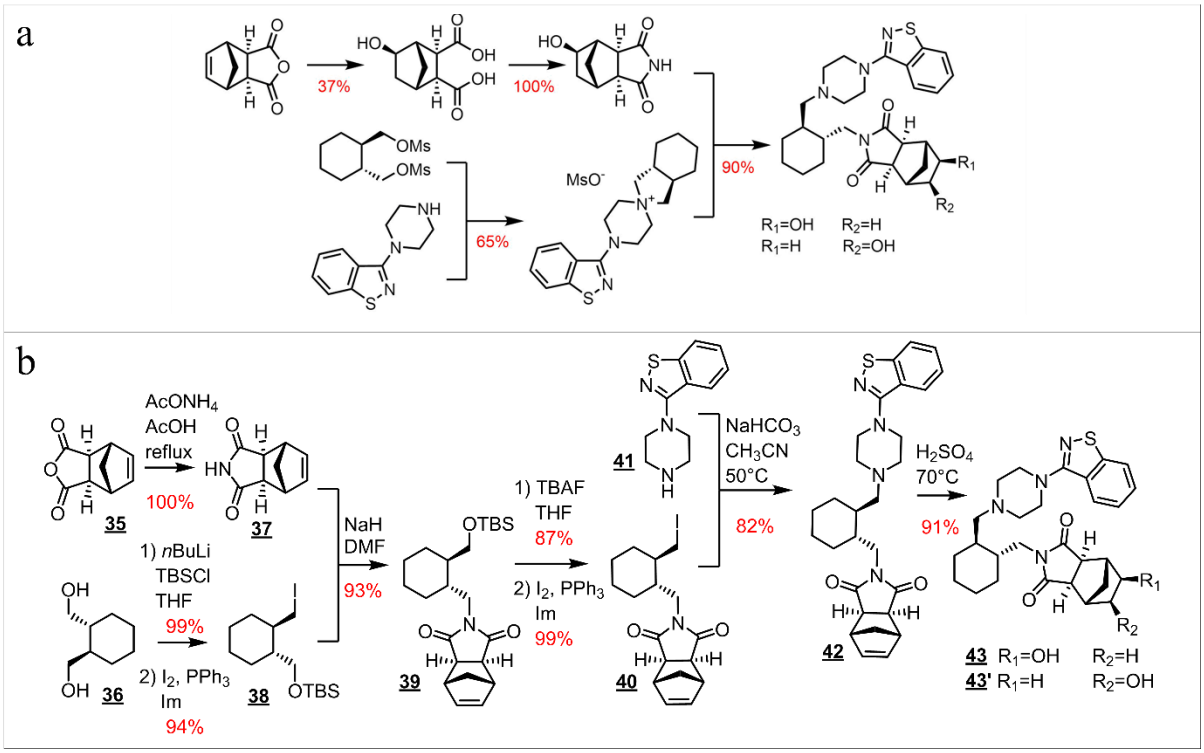


Figure S95. ESI(+) spectrum of compound **SI-8** (crude reaction mixture).

Section S15. Synthesis of 5β/6β-hydroxylurasidone, 43,43'.

S15.1. Previous vs. current synthetic routes.

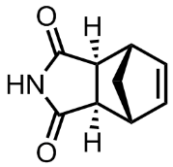


Scheme S6. (a) The original and patented preparation of **43**,**43'** from the main-text reference [29]. For comparison, **(b)** shows the Chematica route (same as in the main-text **Figure 3b**).

S15.2. Synthetic details.

Reagents and solvents were purchased from commercial sources (Aldrich, ABCR, POCH, Chempur). All reagents were used without further purification unless otherwise noted. Flash column chromatography was performed using Merck silica gel 60 (230-400 mesh, 40-63 μm). Reactions were monitored using Macherey-Nagel silica gel 60F254 aluminium plates. TLC's were visualized by UV fluorescence (254 nm) or iodine vapors.

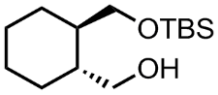
NMR spectra were recorded on a Bruker 400 MHz Avance III spectrometer at room temperature. Chemical shifts (δ) were reported in parts per million (ppm) relative to residual solvent peaks rounded to the nearest 0.01 (ref: $CHCl_3$ [^{1}H: 7.26, ^{13}C: 77.2]). Coupling constants (J) were reported in Hz to the nearest 0.1 Hz. Peak multiplicity was indicated as follows: s (singlet), d (doublet), t (triplet), q (quartet), qi (quintet), sx (sextet) and m (multiplet). HRMS spectra were recorded on AutoSpec Premier (Waters) or MaldiSYNAPT G2-S HDMS (Waters) spectrometers and are given in m/z.



***(3aR,4R,7S,7aS)-3a,4,7,7a-tetrahydro-1H-4,7-methanoisoindole-1,3(2H)-dione* 37**

To a stirring solution of cis-5-norbornene-exo-2,3-dicarboxylic anhydride **35** (4.76 g, 29.0 mmol) in acetic acid (75 ml), $AcONH_4$ (6.71 g, 87.1 mmol) was added. The reaction mixture was stirred under reflux for 4 days. After completion of the reaction, the crude reaction mixture was evaporated to dryness and the resulting oil was redissolved in DCM (100 mL), washed with the saturated aqueous solution of $NaHCO_3$, dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* to yield **37** (4.74 g, 100%) as a white powder.

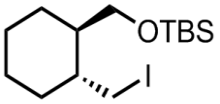
^{1}H NMR: (400 MHz, CDCl3): δ 8.84 (s, 1H), 6.29 (t, 2H), 3.48 – 3.12 (m, 2H), 2.74 (d, 2H), 1.63 – 1.54 (m, 1H), 1.47 (d, 1H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 178.52, 137.75, 49.19, 45.14, 42.90.
HRMS: (*m*/*z*): calcd for $C_9H_9NO_2$ [$M^{+\bullet}$], 163.0633; found 163.0629



***((1R,2R)-2-(((tert-butyldimethylsilyl)oxy)methyl)cyclohexyl)methanol* SI-9**

A round-bottom flask was charged with a solution of ((1*R*,2*R*)-cyclohexane-1,2-diyl)dimethanol **36** (5.36 g, 37.17 mmol) in THF (150 mL). The flask was then capped with a rubber septum and Ar atmosphere was established. Reaction mixture was cooled to 0 °C and the solution of *n*BuLi in cyclohexane (3.64 mL, 2M) was added dropwise. The reaction was stirred at 0 °C for 15 min and at room temperature for 3 h. Then, TBSCl (5.60 g, 37.17 mmol) was added and the reaction was stirred at room temperature for another 1 h. The reaction was quenched by addition of saturated water solution of $NaHCO_3$ (80 mL). Next, AcOEt (55 mL) was added, the phases were separated, and the water phase was extracted with DCM (3x80 mL). Combined organic phases were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 5:4) to yield **SI-9** (9.47 g, 99%) as a colorless oil.

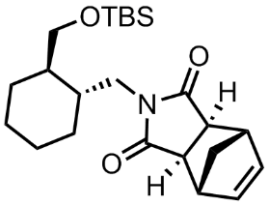
^{1}H NMR: (400 MHz, $CDCl_3$) δ 3.61 (s, 1H), 3.57 (d, 2H), 3.52 – 3.45 (m, 1H), 1.79 – 1.69 (m, 2H), 1.70 - 1.62 (m, 2H), 1.38 – 0.97 (m, 6H), 0.92 (s, 9H), 0.10 (t, 6H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 68.59, 67.43, 45.52, 44.07, 30.07, 29.86, 26.16, 26.15, 25.81, 18.17, -5.45, -5.57.
HRMS: (*m*/*z*): calcd for $C_{14}H_{30}O_2Si$, $[M+Na]^+$, 281.1913; found 281.1907



***tert-butyl(((1R,2R)-2-(iodomethyl)cyclohexyl)methoxy)dimethylsilane* 38**

To a stirring solution of **SI-9** (0.42 g, 1.63 mmol) in DCM (15 mL), PPh_3 (0.857 g, 3.27 mmol) and imidazole (0.28 g, 4.08 mmol) were added sequentially. The reaction mixture was stirred at room temperature for 20 min., then iodine (0.83 g, 3.27 mmol) was added. The reaction was stirred for another 1 h at room temperature, and was then quenched by addition of saturated aqueous solution of $Na_2S_2O_3$ (20 mL) and DCM (40 mL). Phases were separated and the aqueous phase was extracted with DCM (3x20 mL). Combined organic phases were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 10:1) to yield **38** (0.57 g, 94%) as a colorless oil.

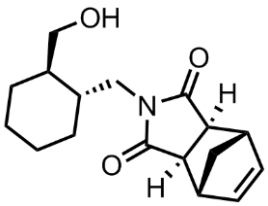
^{1}H NMR: (400 MHz, $CDCl_3$) δ 3.62 (dd, 1H), 3.52 (dd, 1H), 3.43 (dd, 1H), 3.32 (dd, 1H), 1.80 – 1.66 (m, 4H), 1.37 – 1.24 (m, 4H), 1.23 – 1.06 (m, 2H), 0.92 (s, 9H), 0.06 (s, 6H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 65.18, 43.85, 39.20, 33.21, 29.37, 26.06, 25.92, 25.71, 18.27, 16.74, -5.45, -5.47.
HRMS: Under ionization conditions decomposition of compound **38** is observed.
Elemental analysis: (%C; %H; %J): calcd for $C_{14}H_{29}IOSi$: 45.65%C, 7.94%H, 34.45%I; found: 45.57%C, 7.92%H, 34.25%I



***(3aR,4R,7S,7aS)-2-(((1R,2R)-2-(((tert-butyldimethylsilyl)oxy)methyl)cyclohexyl)methyl)-3a,4,7,7a-tetrahydro-1H-4,7-methanoisoindole-1,3(2H)-dione* 39**

37 (2.73 g, 16.72 mmol) was added to a stirred solution of NaH (0.44 g, 18.39 mmol) in DMF (40 mL). The flask was then capped with a rubber septum and Ar atmosphere was established. The reaction was stirred at room temperature for 30 min, and then the solution of **38** (6.16 g, 16.72 mmol) in DMF (40 mL) was added dropwise. The reaction mixture was stirred for 3 days at room temperature. The reaction was quenched by addition of saturated water solution of NH_4Cl (80 mL). Phases were separated and the aqueous phase was extracted with AcOEt (3x100 mL). Combined organic phases were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 5:1) to yield **39** (6.29 g, 93%) as a white powder.

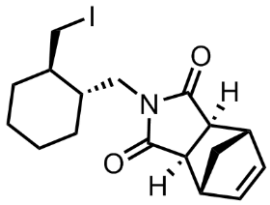
^{1}H NMR: (400 MHz, $CDCl_3$) δ 6.29 (s, 2H), 3.79 – 3.64 (m, 2H), 3.57 (dd, 1H), 3.33 (dd, 1H), 3.29 (s, 2H), 2.68 (s, 2H), 1.85 – 1.73 (m, 1H), 1.72 – 1.61 (m, 3H), 1.57 – 1.48 (m, 2H), 1.33 – 0.98 (m, 6H), 0.91 (s, 9H), 0.06 (d, 6H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 178.34, 178.29, 137.83, 137.78, 66.00, 47.81, 47.76, 45.12, 43.04, 42.93, 42.57, 37.80, 29.64, 29.04, 25.93, 25.37, 25.16, 18.26, -5.40, -5.47.
HRMS: (*m/z*): calcd for $C_{23}H_{37}O_3NSi$, $[M+Na]^+$, 426.2440; found 426.2428



***(3aR,4R,7S,7aS)-2-(((1R,2R)-2-(hydroxymethyl)cyclohexyl)methyl)-3a,4,7,7a-tetrahydro-1H-4,7-methanoisoindole-1,3(2H)-dione* SI-10**

A round-bottomed flask was charged with a solution of **39** (5.87 g, 15.22 mmol) in THF (70 mL). The flask was then capped with a rubber septum and Ar atmosphere was established. To the stirring reaction mixture a solution of TBAF in THF (1 M, 18.9 mL) was added dropwise and the reaction was stirred overnight at room temperature. The reaction was quenched by addition of saturated water solution of NH_4Cl (90 mL). Phases were separated and the water phase was extracted with AcOEt (5 x 100 mL). Combined organic phases were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 1:1) to yield **SI-10** (3.65 g, 87%) as a colorless oil.

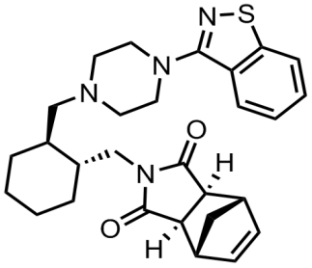
^{1}H NMR: (400 MHz, $CDCl_3$) δ 6.31 (s, 2H), 3.96 – 3.80 (m, 1H), 3.71 – 3.55 (m, 2H), 3.39 (dd,1H), 3.29 (s, 2H), 2.71 (s, 2H), 2.45 – 2.25 (s, 1H), 1.80 – 1.62 (m, 4H), 1.56 (m, 2H), 1.37 – 1.20 (m, 4H), 1.08 (m, 2H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 178.66, 178.56, 137.84, 137.80, 65.41, 47.85, 47.78, 45.19, 45.15, 42.79, 42.36, 41.96, 37.92, 30.84, 29.48, 25.64, 25.45.
HRMS: (*m*/*z*): calcd for $C_{17}H_{23}O_3N$, $[M+Na]^+$, 312.1576; found 312.1570



***(3aR,4R,7S,7aS)-2-(((1R,2R)-2-(iodomethyl)cyclohexyl)methyl)-3a,4,7,7a-tetrahydro-1H-4,7-methanoisoindole-1,3(2H)-dione* 40**

To a stirring solution of **SI-10** (3.49 g, 12.06 mmol) in DCM (140 mL), PPh_3 (6.33 g, 24.12 mmol) and imidazole (2.05 g, 30.15 mmol) were added sequentially. The reaction mixture was cooled to 0 °C and stirred for 30 min. then iodine (6.12 g, 26.12 mmol) was added in small portions. The reaction was stirred for another 1.5 h at room temperature. The reaction was quenched by addition of aqueous solution of $Na_2S_2O_3$ (100 mL). Phases were separated and the water phase was extracted with DCM (4x100 mL). Combined organic phases were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 1:1) to yield **40** (4.77 g, 99%) as a colorless oil.

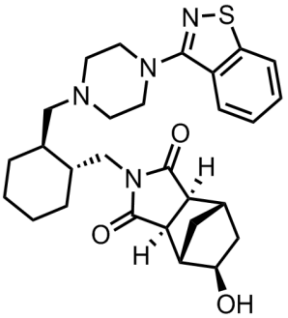
^{1}H NMR: (400 MHz, $CDCl_3$) δ 6.31 (s, 2H), 3.61 (dd, 1H), 3.48 – 3.39 (m, 2H), 3.38 – 3.31 (m, 1H), 3.30 (s, 2H), 2.72 (s, 2H), 1.80 (d, 1H), 1.75 – 1.59 (m, 4H), 1.55 (d, 1H), 1.37 – 1.17 (m, 4H), 1.13 – 0.95 (m, 2H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 178.38, 178.25, 137.80, 137.78, 47.84, 47.80, 45.15, 45.13, 43.03, 42.05, 41.24, 40.25, 32.78, 29.63, 25.28, 25.06, 15.03.
HRMS: (*m*/*z*): calcd for $C_{17}H_{22}O_2NI$, $[M+Na]^+$, 422.0593; found 422.0582



(3aR,4R,7S,7aS)-2-(((1R,2R)-2-((4-(benzo[d]isothiazol-3-yl)piperazin-1-yl)methyl)cyclohexyl)methyl)-3a,4,7,7a-tetrahydro-1H-4,7-methanoisoindole-1,3(2H)-dione 42

A round-bottom flask was charged with **40** (0.509 g, 1.27 mmol), 3-(1-piperazinyl)-1,2-benzisothiazole **41** (0.475 g, 2.16 mmol), $NaHCO_3$ (0.139 g, 1.66 mmol), and CH_3CN (7 mL). The flask was then capped with a rubber septum and Ar atmosphere was established. The reaction was stirred for 22 h at room temperature. The reaction was quenched by addition of water (20 mL). Phases were separated and the water phase was extracted with AcOEt (4x20 mL). Combined organic phases were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 2:1) to yield **42** (0.514 g, 82%) as a white solid.

^{1}H NMR: (400 MHz, $CDCl_3$) δ 7.92 (d, 1H), 7.81 (d, 1H), 7.47 (t, 1H), 7.36 (t, 1H), 6.30 (s, 2H), 3.97 (dd, 1H), 3.54 (t, 4H), 3.34 (dd, 1H), 3.29 (s, 2H), 2.69 (s, 2H), 2.68 – 2.58 (m, 5H), 2.25 (dd, *J* = 12.5, 1H), 1.90 (d, 1H), 1.68 (d, 2H), 1.56 (dd, 3H), 1.40 (dd, 1H), 1.27 (t, 2H), 1.15 (d, 1H), 1.08 – 0.97 (m, 2H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 178.40, 164.07, 152.74, 137.84, 137.78, 128.11, 127.44, 123.96, 123.80, 120.51, 63.66, 53.50, 50.18, 47.83, 47.79, 45.13, 42.92, 42.66, 40.74, 37.56, 30.77, 29.91, 25.42, 25.02.
HRMS: (*m*/*z*): calcd for $C_{28}H_{34}N_4O_2S$, $[M+H]^+$, 491.2481; found 491.2472



(3aR,4S,5R,7S,7aS)-2-(((1R,2R)-2-((4-(benzo[d]isothiazol-3-yl)piperazin-1-yl)methyl)cyclohexyl)methyl)-5-hydroxyhexahydro-1H-4,7-methanoisoindole-1,3(2H)-dione 43,43'

A round-bottom flask was charged with **42** (0.502 g, 1.02 mmol) and the solution of H_2SO_4 (2 mL) in water (4 mL). The reaction was stirred overnight at 70 °C. The reaction was stirred for another 1.5 h at room temperature. The reaction was quenched by addition of saturated water solution of K_2CO_3 (200 mL), AcOEt (100 mL) and water (50 mL). The water phase was extracted with AcOEt (3x100 mL), then the organic phases were combined, dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (AcOEt, 100%) to yield **43**,**43'** (0.475 g, 91%) as a white solid.
^{1}H NMR: (400 MHz, $CDCl_3$) δ 7.92 (d, 1H), 7.82 (d, 1H), 7.47 (t, 1H), 7.36 (t, 1H), 3.97 (m, 2H), 3.54 (m, 4H), 3.34 (dd, 1H), 2.75 (d, 1H), 2.64 (m, 6H), 2.52 (s, 2H), 2.24 (dd, 1H), 1.99 – 1.77 (m, 3H), 1.75 - 1.63 (m, 3H), 1.60 - 1.45 (m, 3H), 1.45 - 1.35 (m, 1H), 1.32 - 0.95 (m, 6H).
^{13}C NMR: (101 MHz, $CDCl_3$) δ 178.68, 178.51, 164.09, 152.72, 128.10, 127.48, 123.97, 120.53, 77.23, 72.97, 63.67, 53.49, 50.18, 47.91, 47.45, 44.66, 42.89, 40.67, 40.54, 38.68, 37.57, 30.76, 29.84, 29.54, 25.40, 25.00.

HRMS: (*m*/*z*): calcd for $C_{28}H_{36}N_4O_3S$, $[M+H]^+$, 509.2586; found 509.2583

S15.3. Raw spectroscopic and chromatographic data.

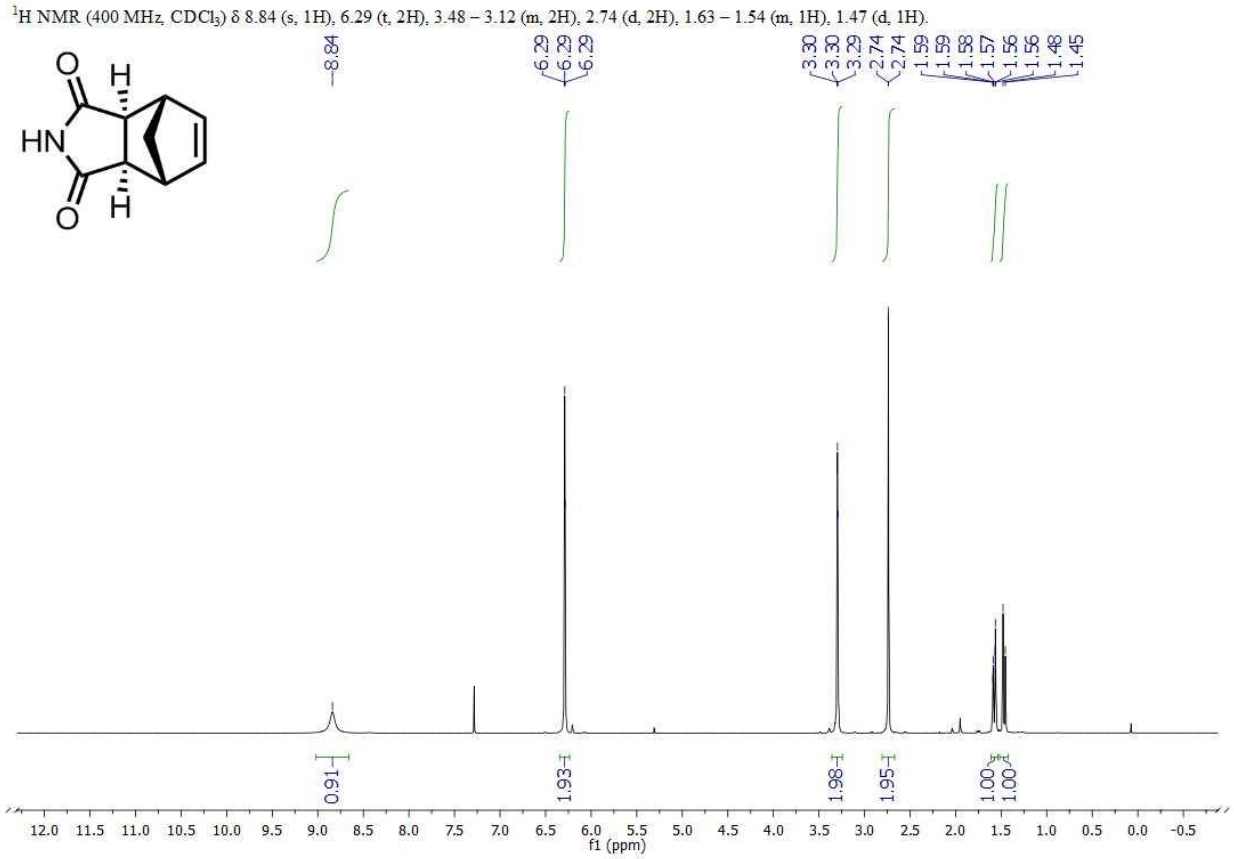


Figure S96. ^{1}H NMR spectrum of compound **37**.

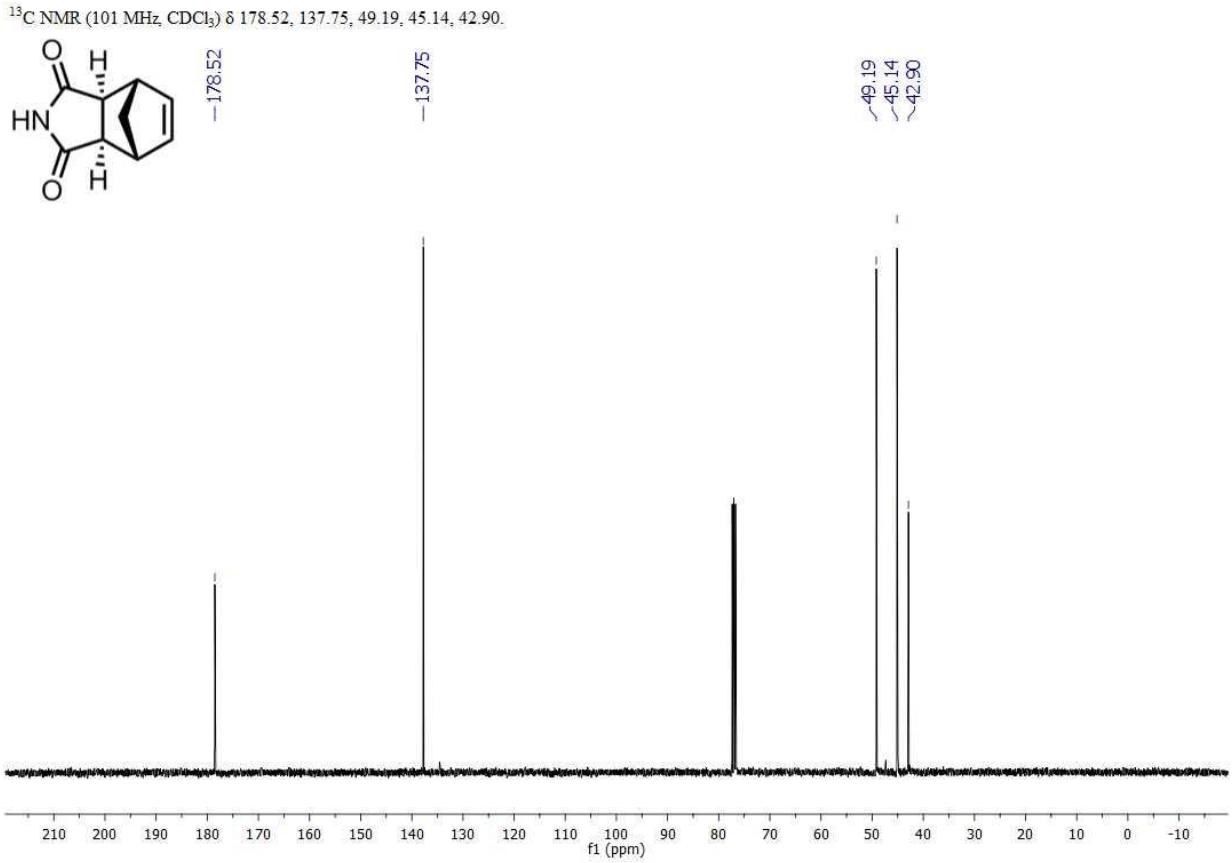


Figure S97. ^{13}C NMR spectrum of compound **37.**

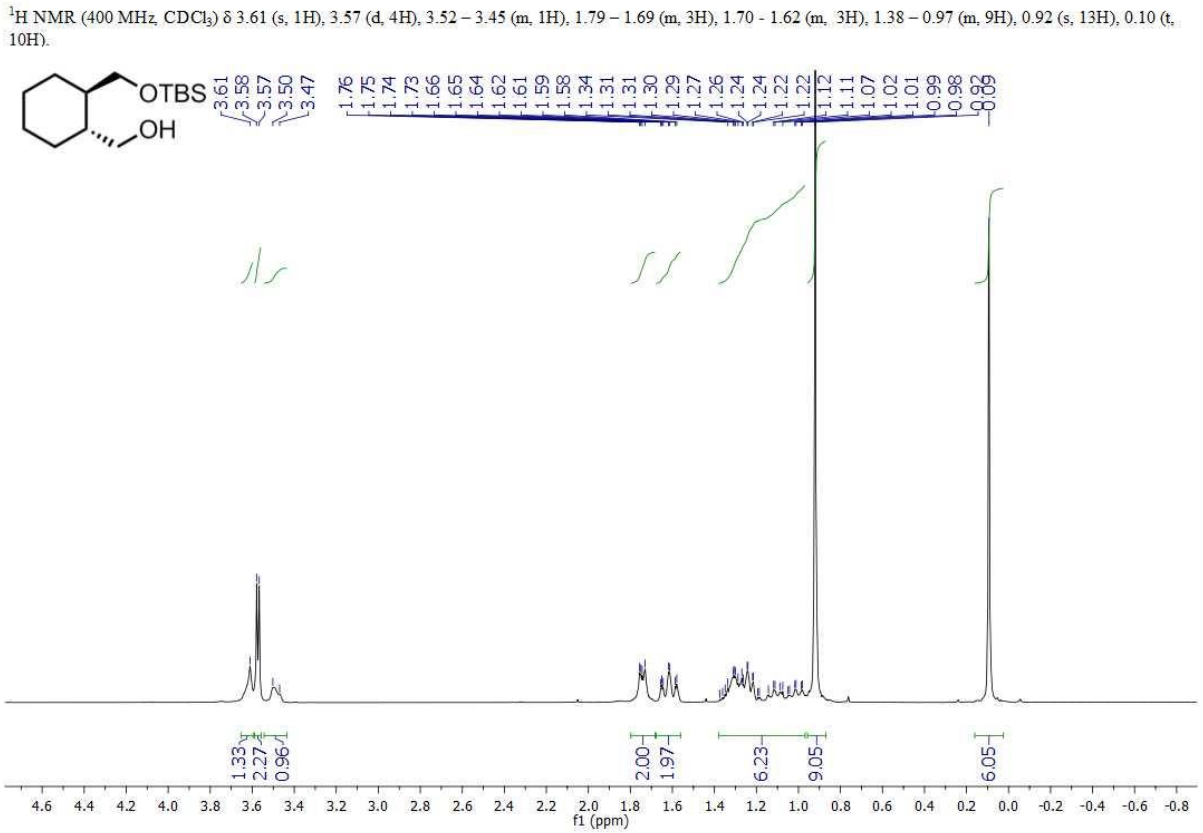


Figure S98. ^{1}H NMR spectrum of compound **SI-9**.

^{13}C NMR (101 MHz, $CDCl_3$) δ 68.59, 67.43, 45.52, 44.07, 30.07, 29.86, 26.16, 26.15, 25.81, 18.17, -5.45, -5.57.

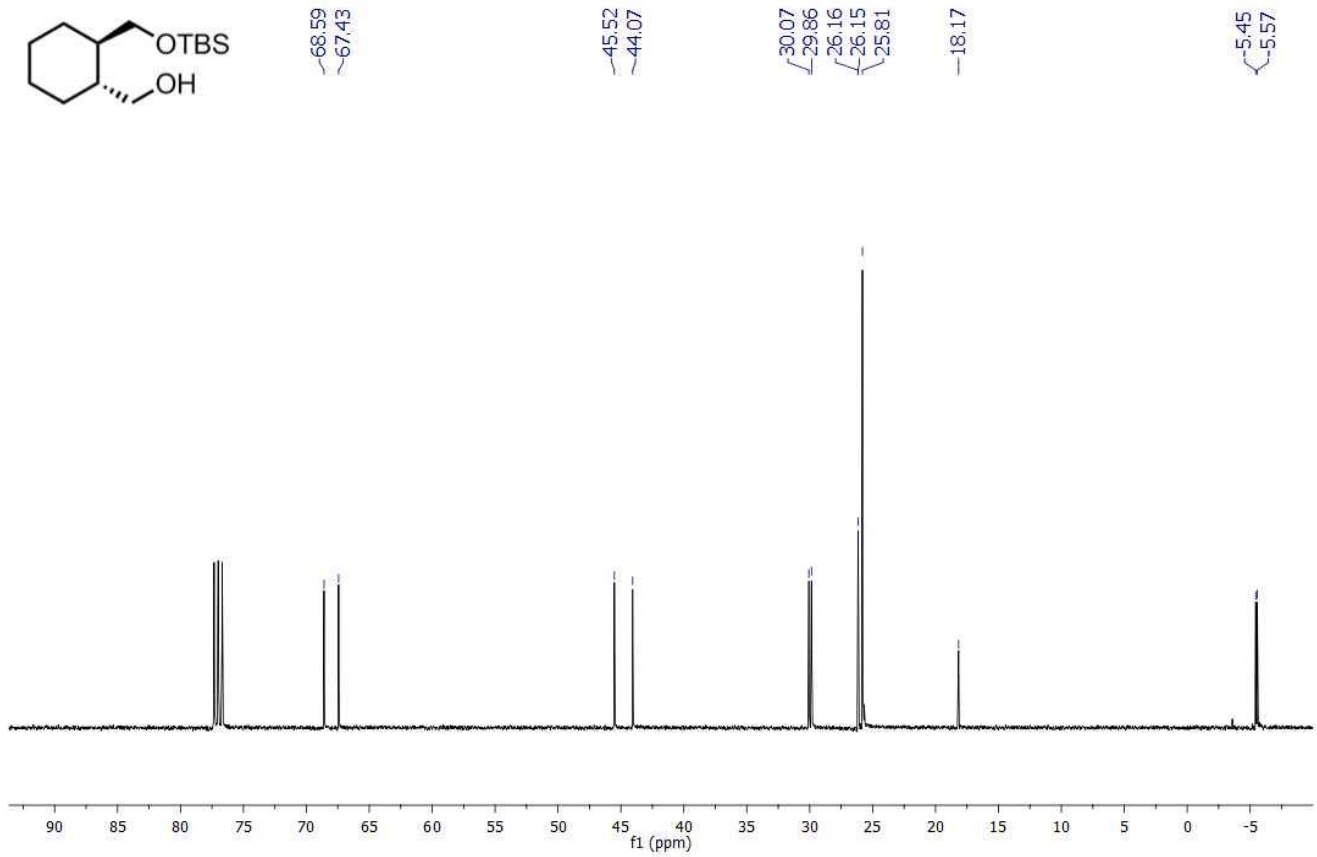


Figure S99. ^{13}C NMR spectrum of compound **SI-9**.

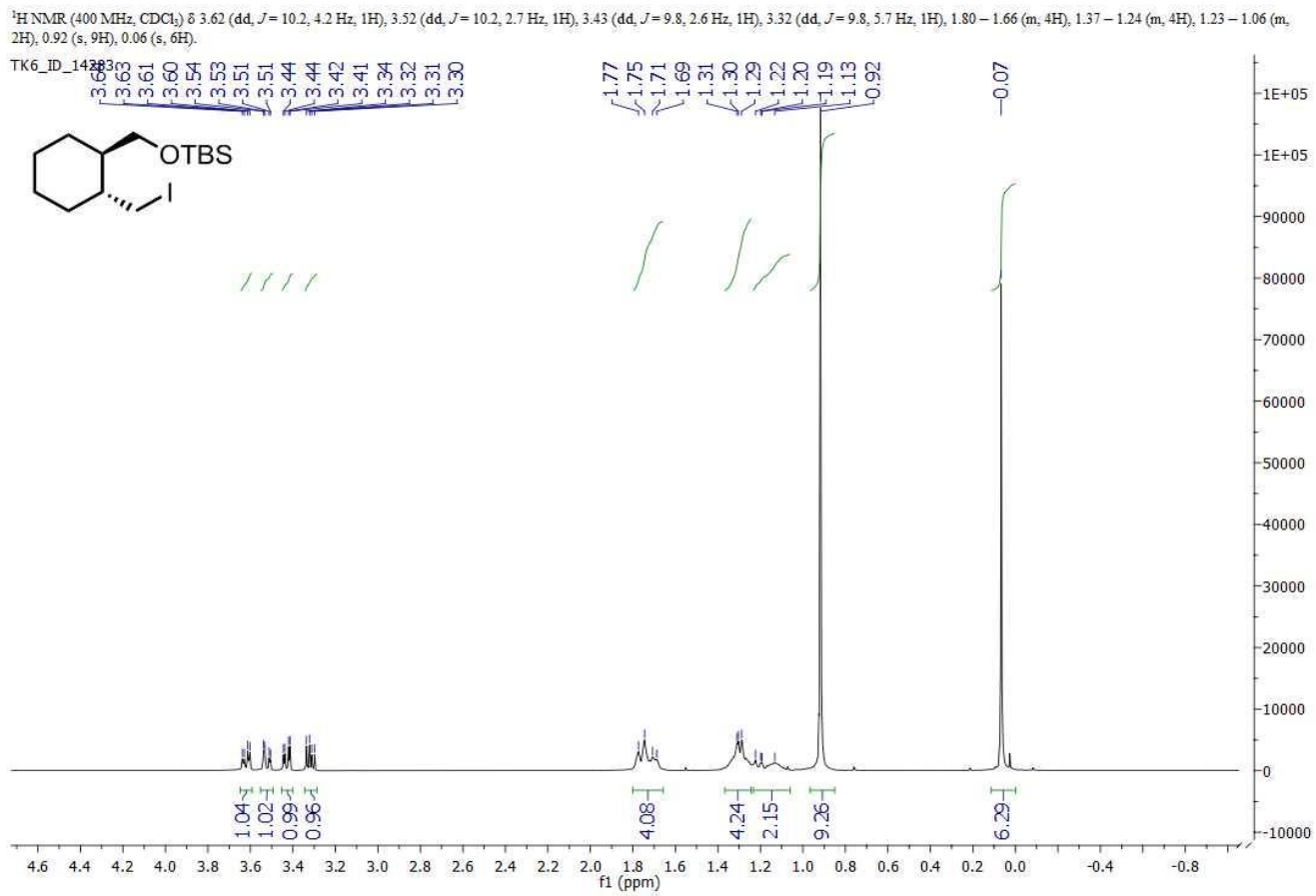


Figure S100. ^{1}H NMR spectrum of compound **38**.

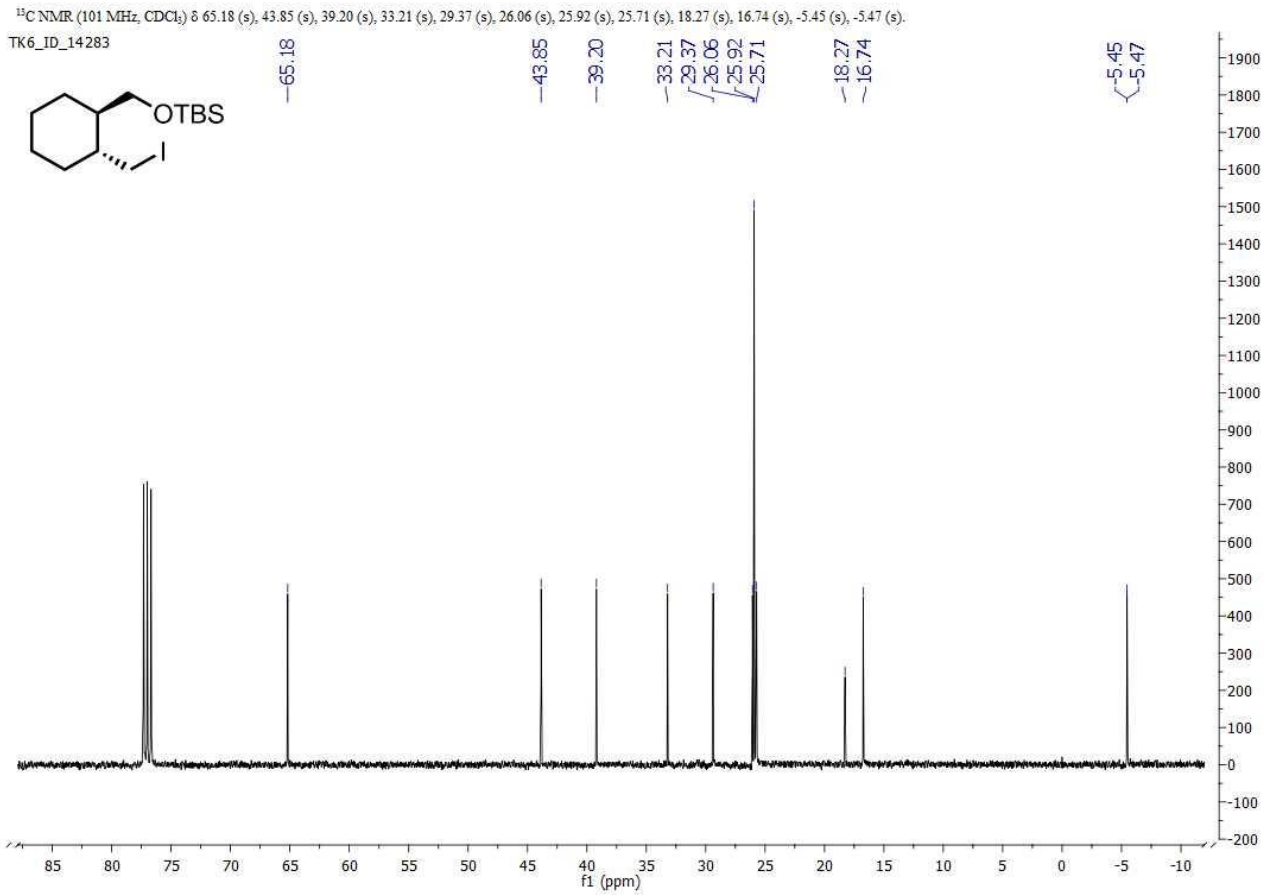


Figure S101. ^{13}C NMR spectrum of compound **38**.

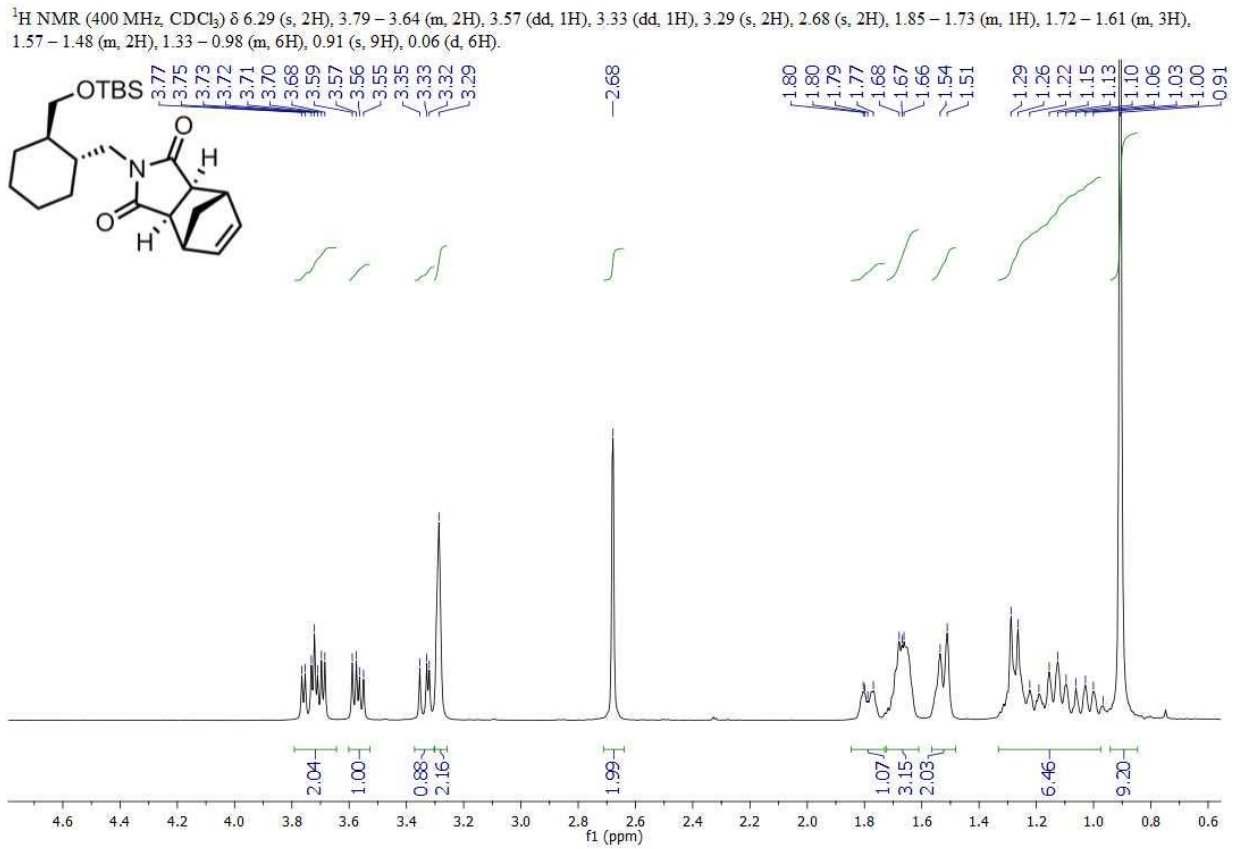


Figure S102. ^{1}H NMR spectrum of compound **39**.

^{13}C NMR (101 MHz, $CDCl_3$) δ 178.34, 178.29, 137.83, 137.78, 66.00, 47.81, 47.76, 45.12, 43.04, 42.93, 42.57, 37.80, 29.64, 29.04, 25.93, 25.37, 25.16, 18.26, -5.40, -5.47.

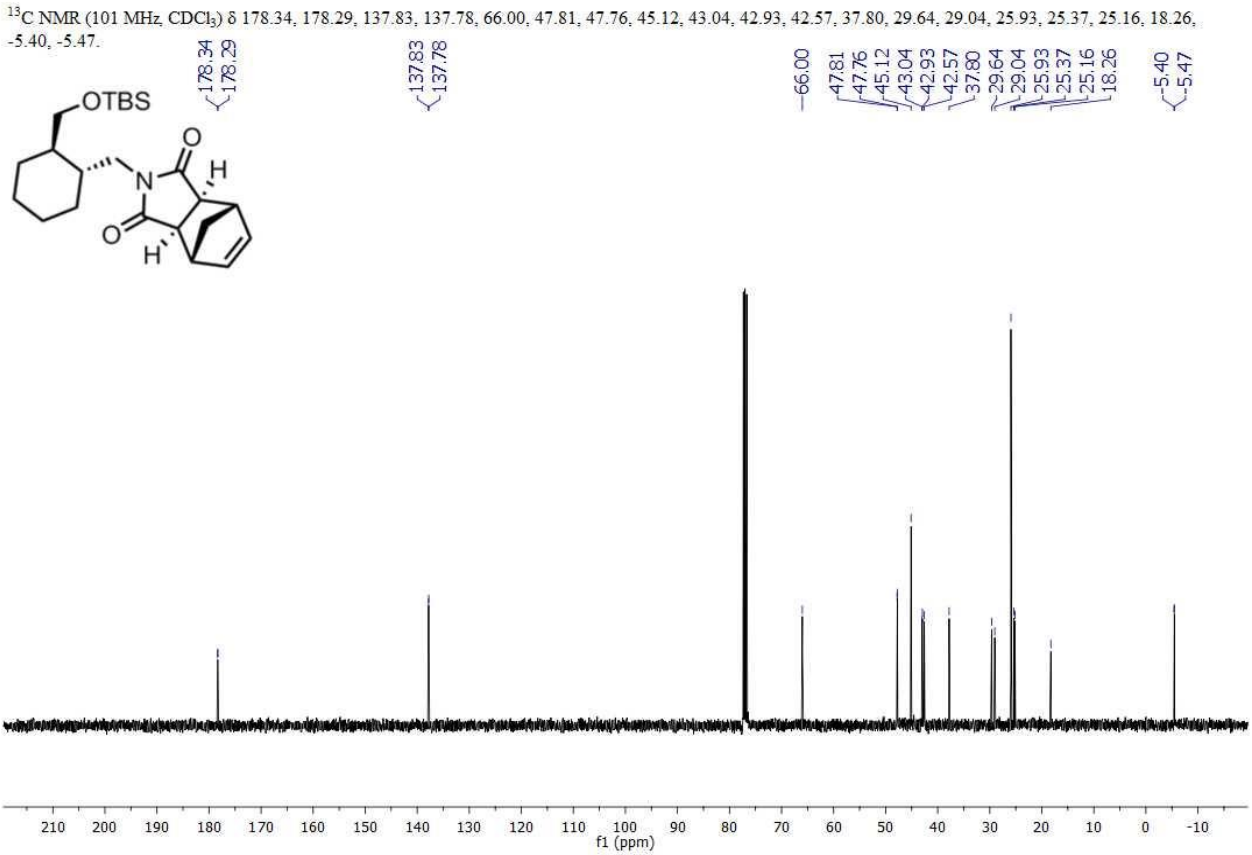


Figure S103. ^{13}C NMR spectrum of compound **39**.

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.31 (s, 2H), 3.96 – 3.80 (m, 1H), 3.71 – 3.55 (m, 2H), 3.39 (dd, 1H), 2.71 (s, 2H), 2.45 – 2.25 (m, 1H), 1.80 – 1.62 (m, 2H), 1.56 (m, 1H), 1.37 – 1.20 (m, 2H), 1.08 (m, 1H).

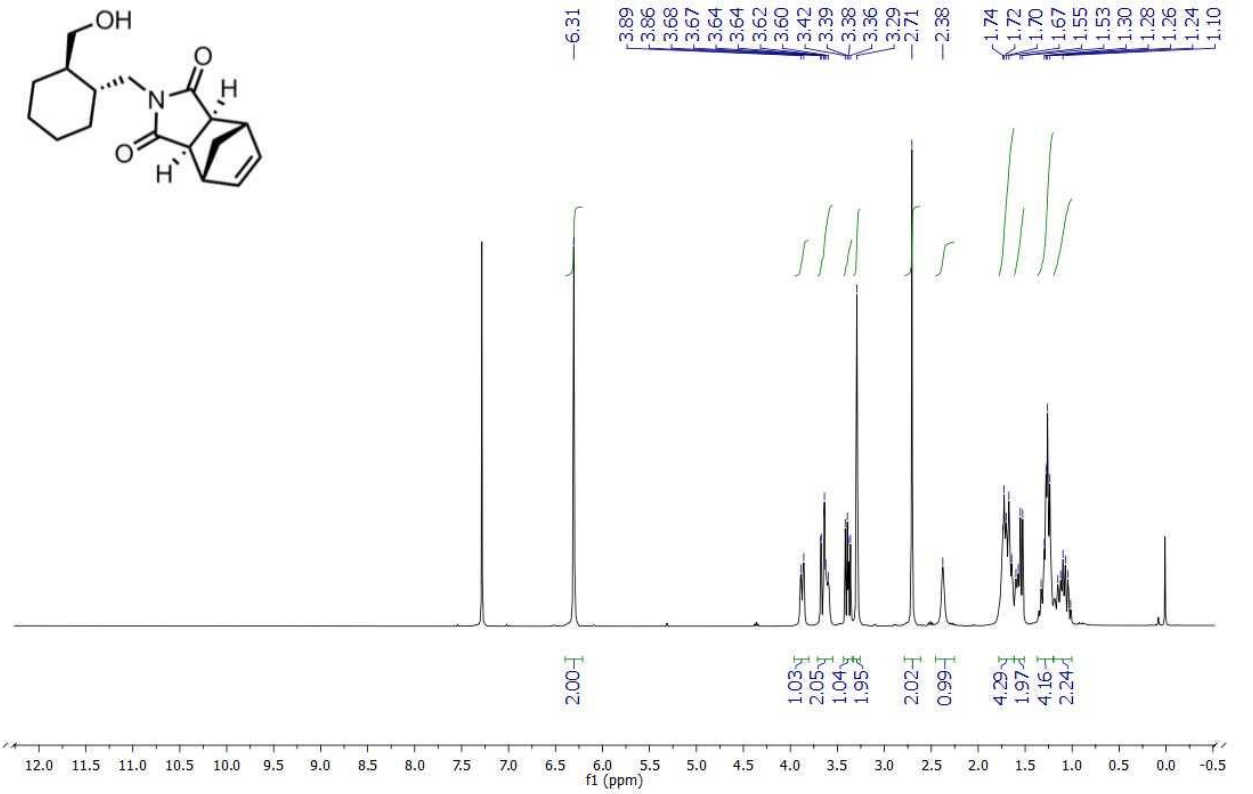


Figure S104. ^{1}H NMR spectrum of compound **SI-10**.

^{13}C NMR (101 MHz, $CDCl_3$) δ 178.66, 178.56, 137.84, 137.80, 65.41, 47.85, 47.78, 45.19, 45.15, 42.79, 42.36, 41.96, 37.92, 30.84, 29.48, 25.64, 25.45.

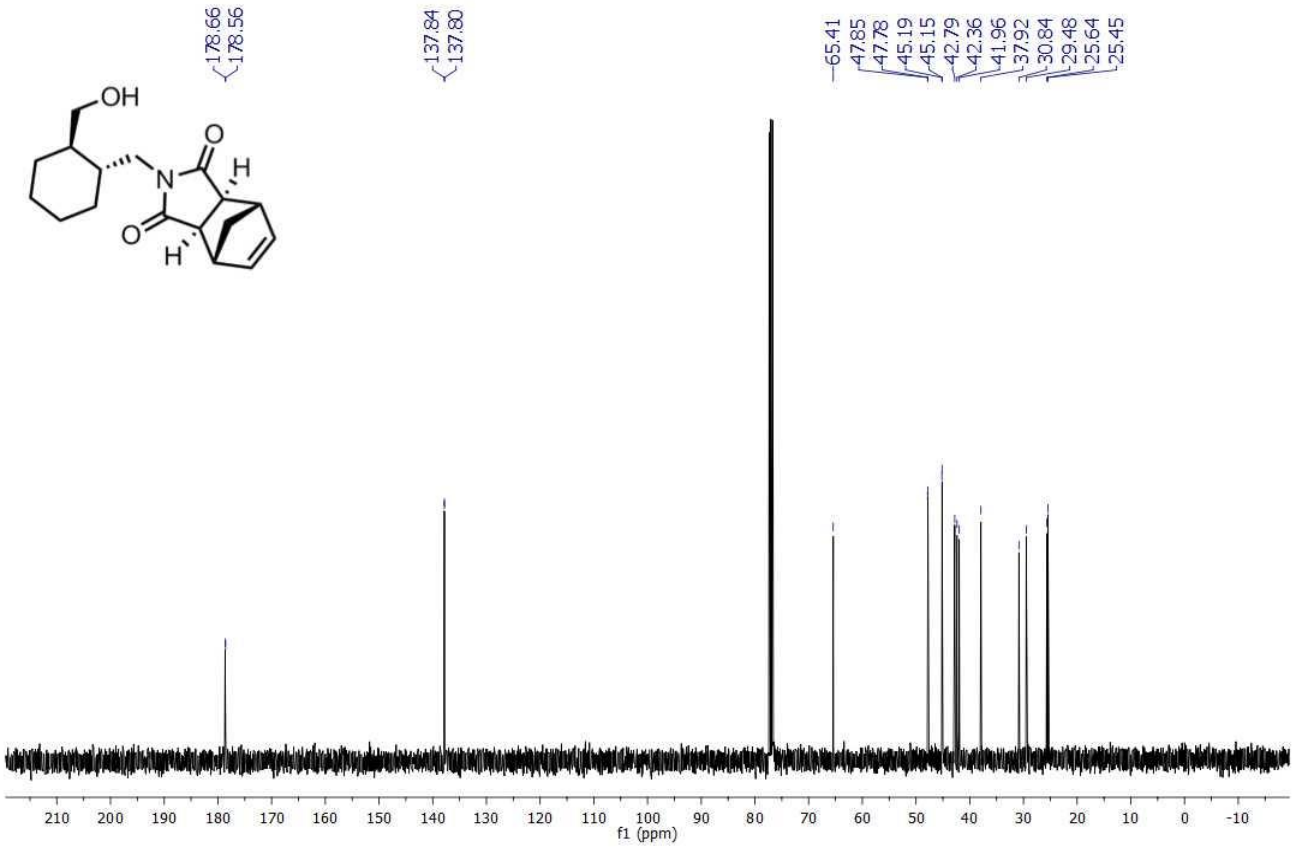


Figure S105. ^{13}C NMR spectrum of compound **SI-10**.

^{1}H NMR (400 MHz, $CDCl_3$) δ 6.31 (s, 1H), 3.61 (dd, J = 13.4, 4.6 Hz, 1H), 3.48 – 3.39 (m, 1H), 3.38 – 3.31 (m, 1H), 3.30 (s, 1H), 2.72 (s, J = 17.0 Hz, 1H), 1.80 (d, J = 12.3 Hz, 1H), 1.75 – 1.59 (m, 2H), 1.55 (d, J = 9.8 Hz, 1H), 1.37 – 1.17 (m, 2H), 1.13 – 0.95 (m, 1H).

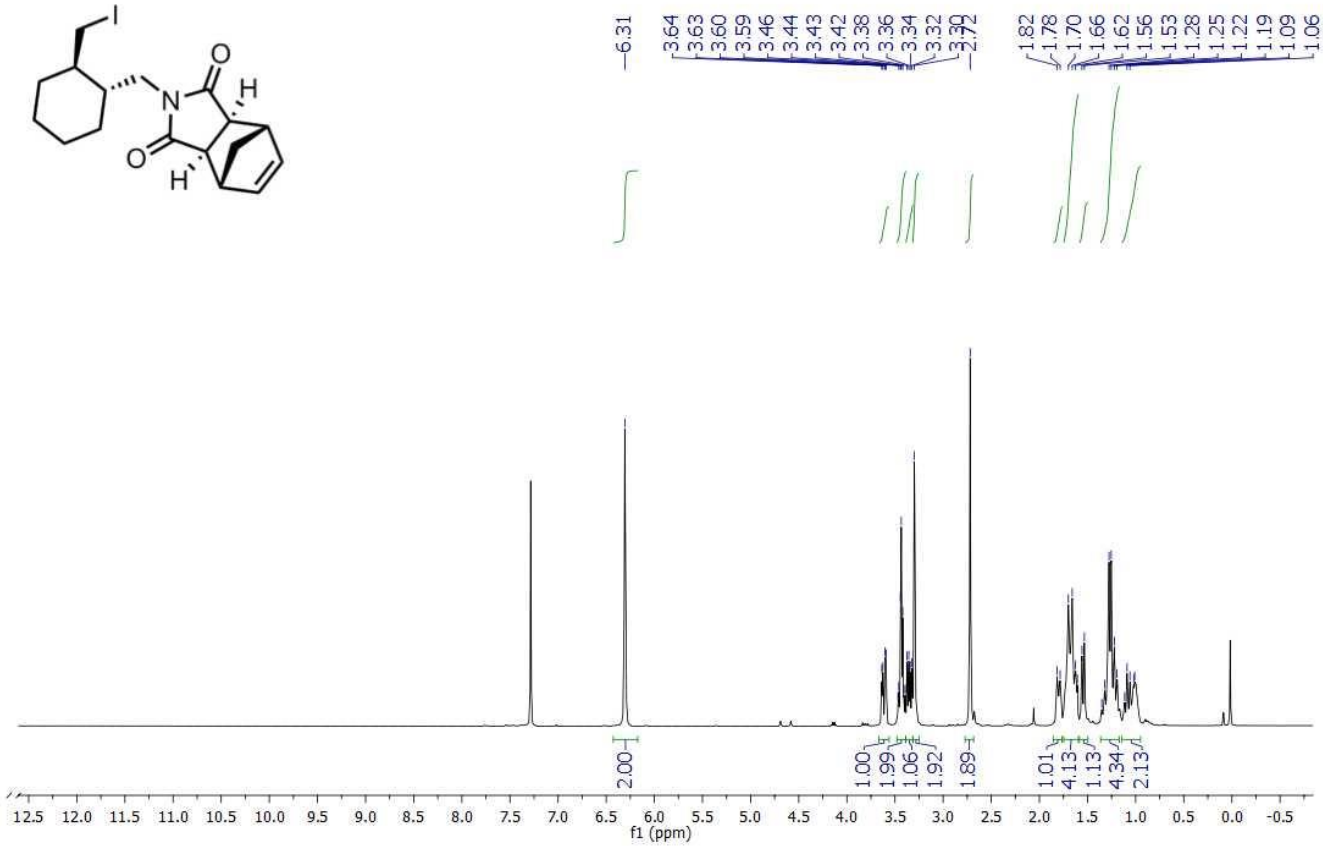


Figure S106. ^{1}H NMR spectrum of compound **40**.

^{13}C NMR (101 MHz, $CDCl_3$) δ 178.38, 178.25, 137.80, 137.78, 47.84, 47.80, 45.15, 45.13, 43.03, 42.05, 41.24, 40.25, 32.78, 29.63, 25.28, 25.06, 15.03.

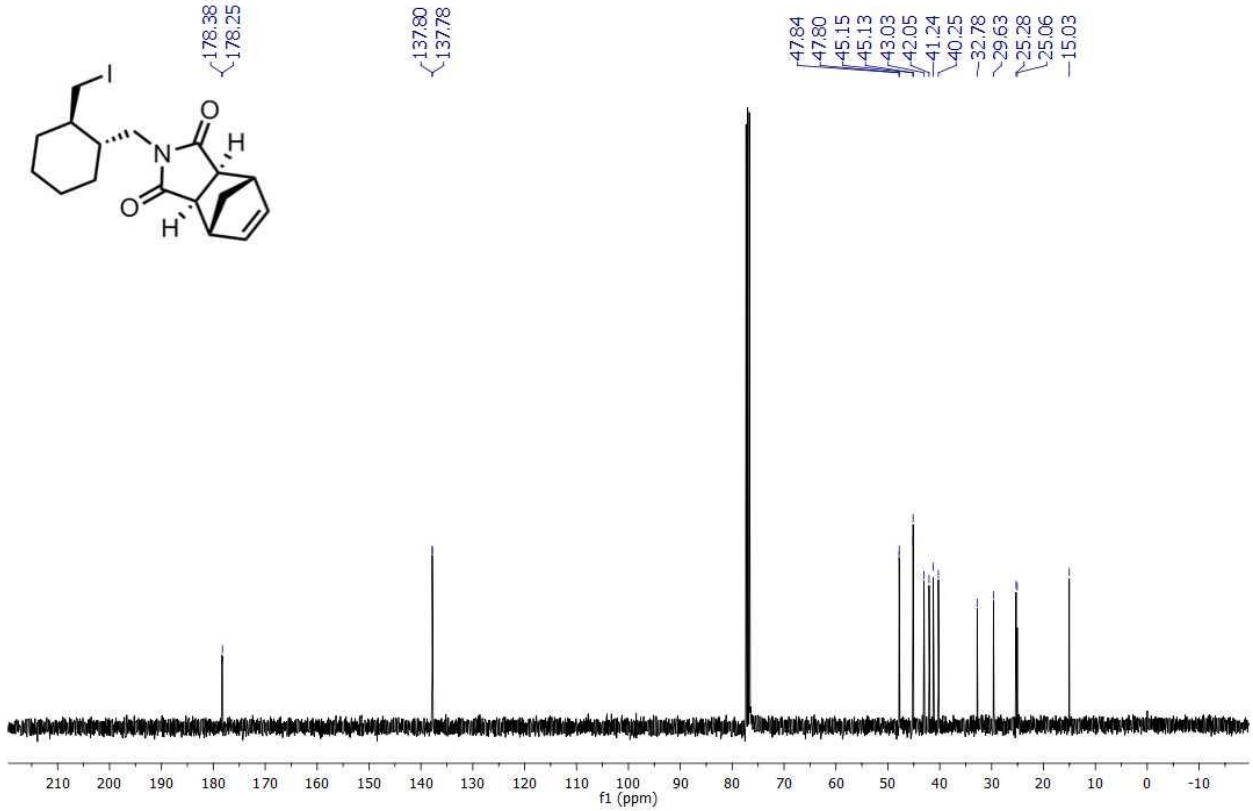


Figure S107. ^{13}C NMR spectrum of compound **40**.

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.92 (d, 1H), 7.81 (d, 1H), 7.47 (t, 1H), 7.36 (t, 1H), 6.30 (s, 2H), 3.97 (dd, 1H), 3.54 (t, 4H), 3.34 (dd, 1H), 3.29 (s, 2H), 2.69 (s, 2H), 2.68 – 2.58 (m, 5H), 2.25 (dd, J = 12.5, 1H), 1.90 (d, 1H), 1.68 (d, 2H), 1.56 (dd, 3H), 1.40 (dd, 1H), 1.27 (t, 2H), 1.15 (d, 1H), 1.08 – 0.97 (m, 2H).

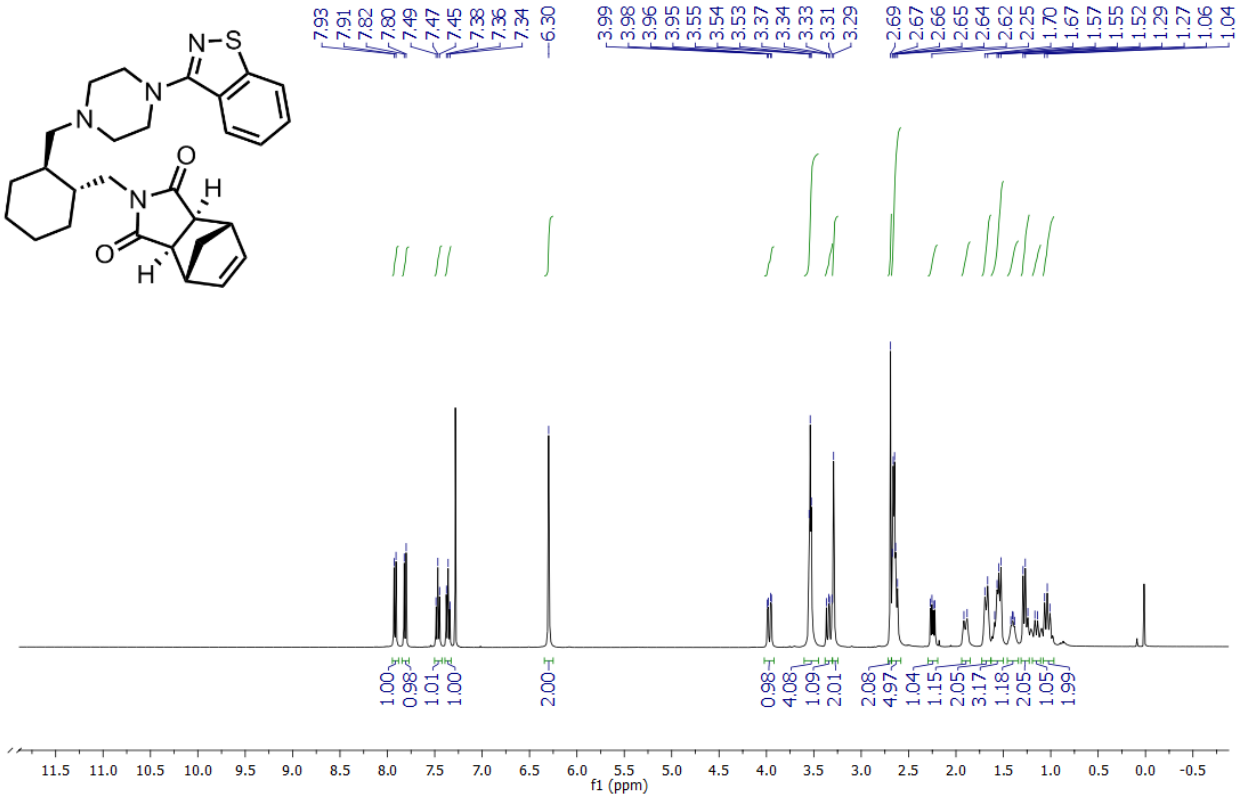


Figure S108. ^{1}H NMR spectrum of compound **42**.

^{13}C NMR (101 MHz, $CDCl_3$) δ 178.40, 164.07, 152.74, 137.84, 137.78, 128.11, 127.44, 123.96, 123.80, 120.51, 63.66, 53.50, 50.18, 47.83, 47.79, 45.13, 42.92, 42.66, 40.74, 37.56, 30.77, 29.91, 25.42, 25.02.

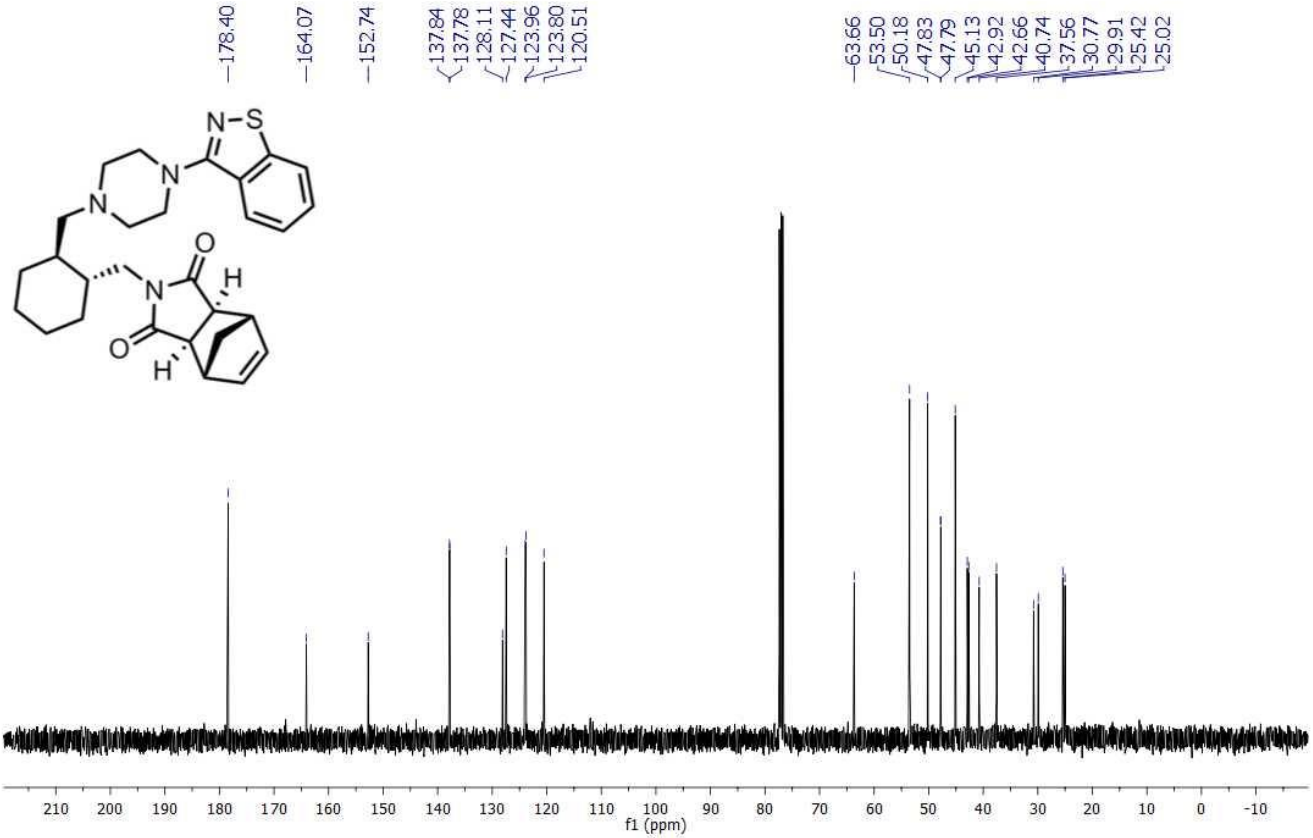


Figure S109. ^{13}C NMR spectrum of compound **42**.

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.92 (d, 1H), 7.82 (d, 1H), 7.47 (t, 1H), 7.36 (t, 1H), 3.97 (m, 2H), 3.54 (m, 4H), 3.34 (dd, 1H), 2.75 (d, 1H), 2.64 (m, 6H), 2.52 (s, 2H), 2.24 (dd, , 1H), 1.99 – 1.77 (m, 3H), 1.75 - 1.63 (m, 3H), 1.60 - 1.45 (m, 3H), 1.45 - 1.35 (m, 1H), 1.32 - 0.95 (m, 6H).

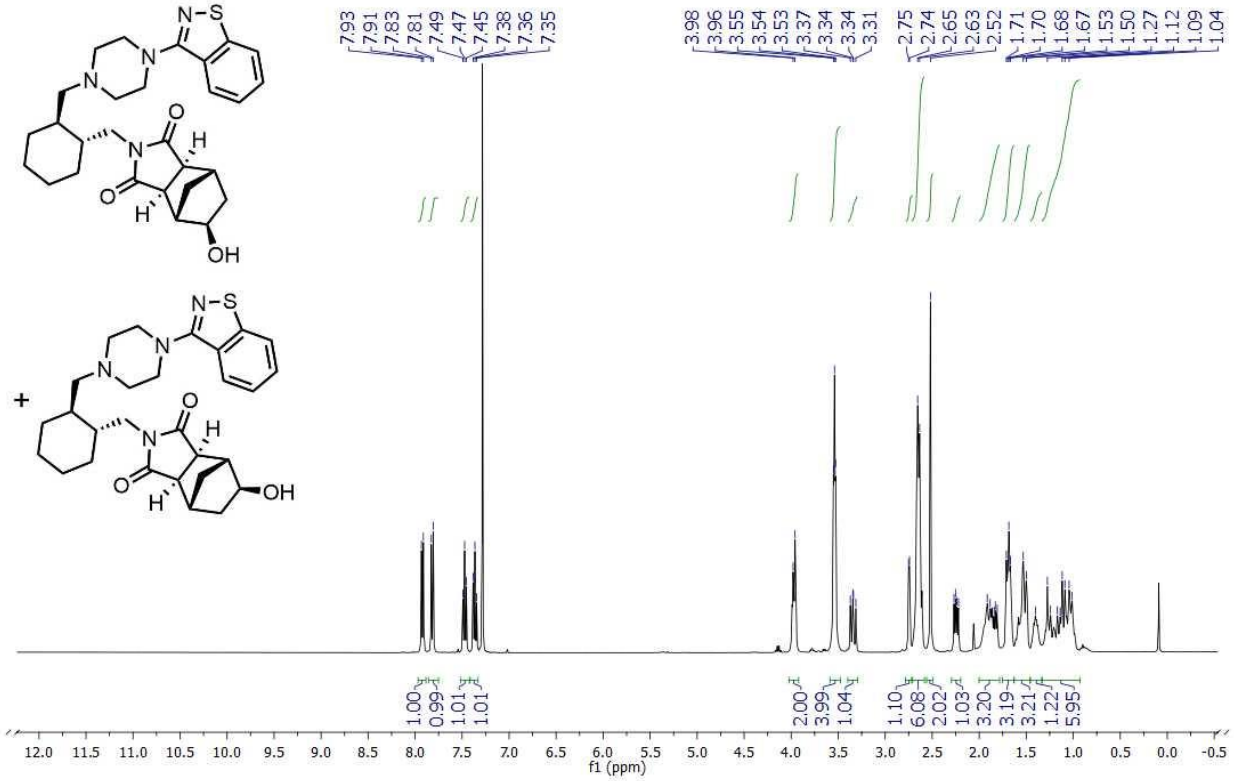


Figure S110. ^{1}H NMR spectrum of compound **43**,**43'**.

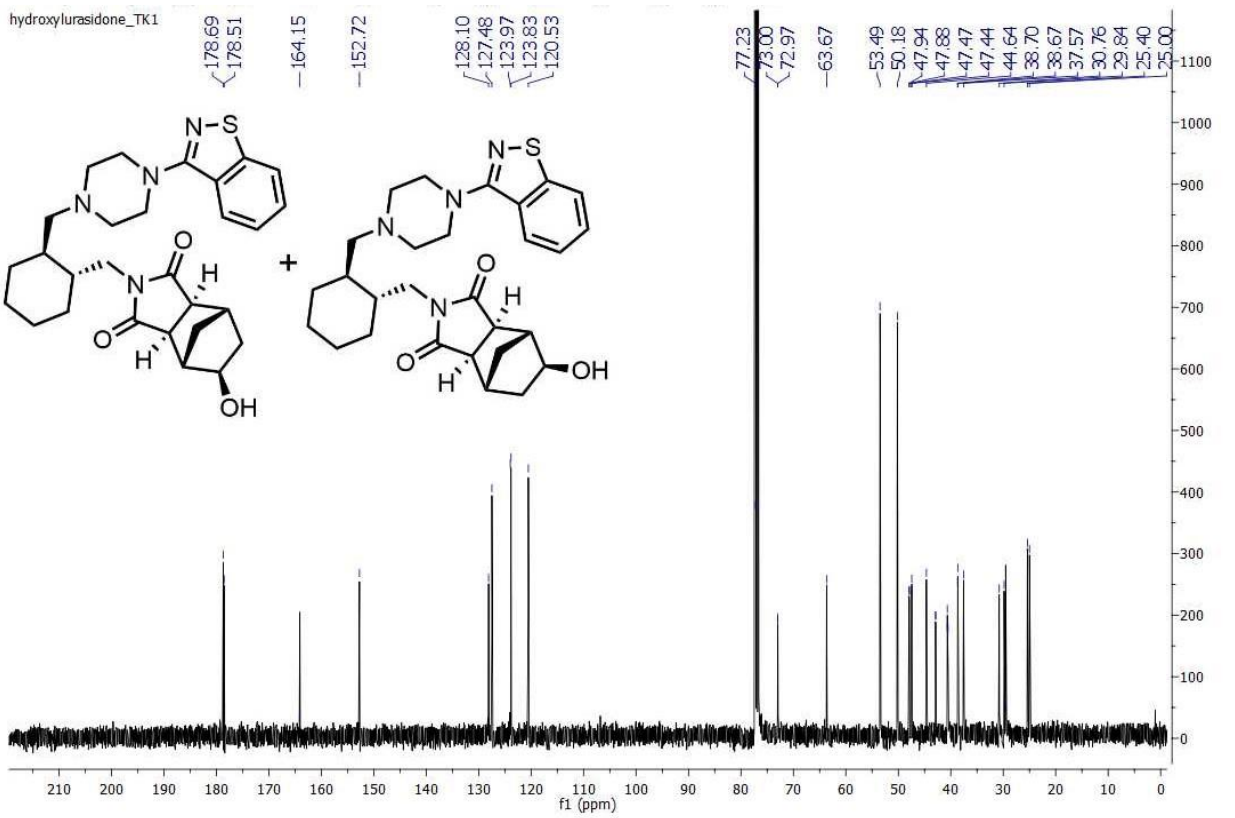


Figure S111. ^{13}C NMR spectrum of compound **43**,**43'**.

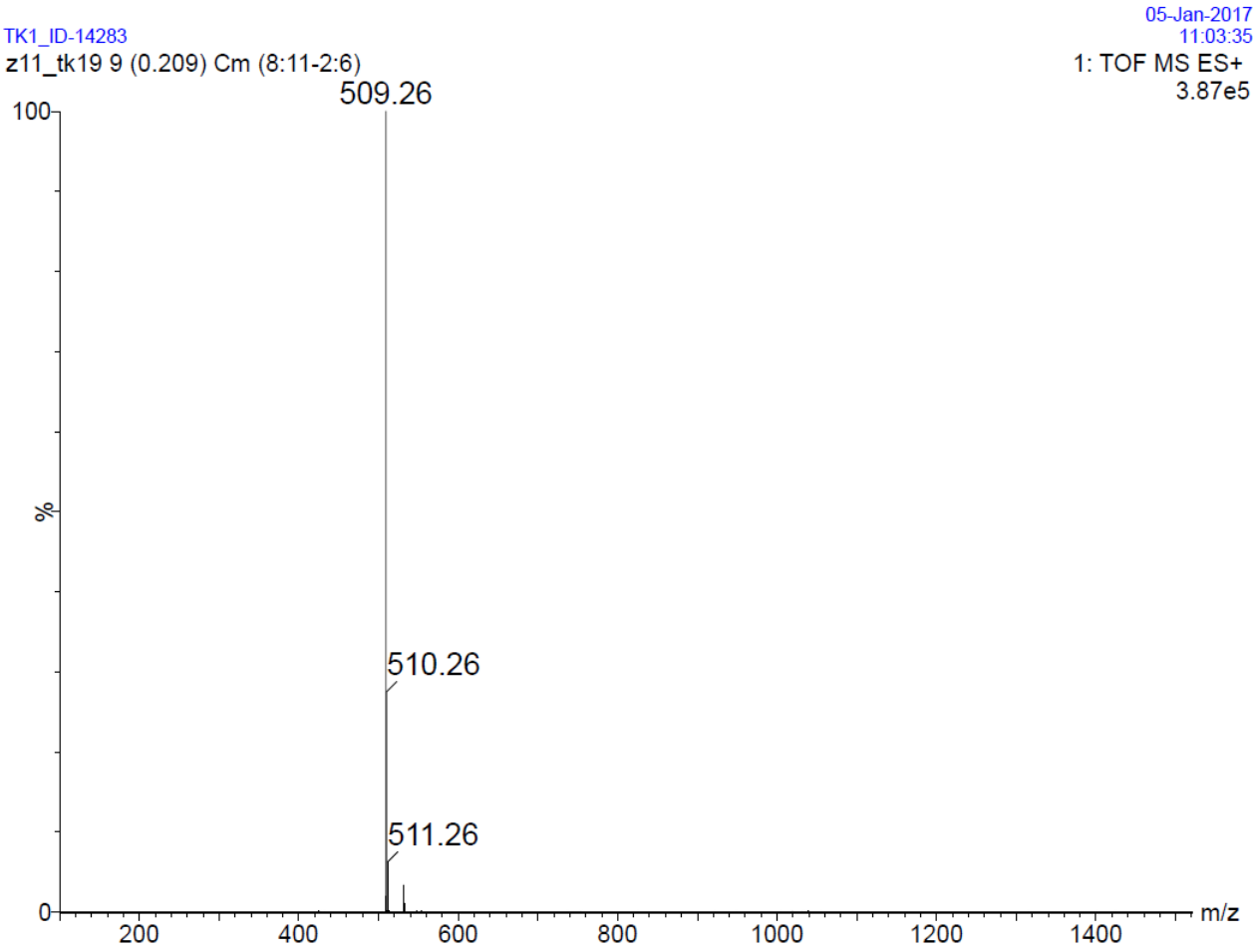


Figure S112. Mass Spectrum of compound **43**,**43'**.

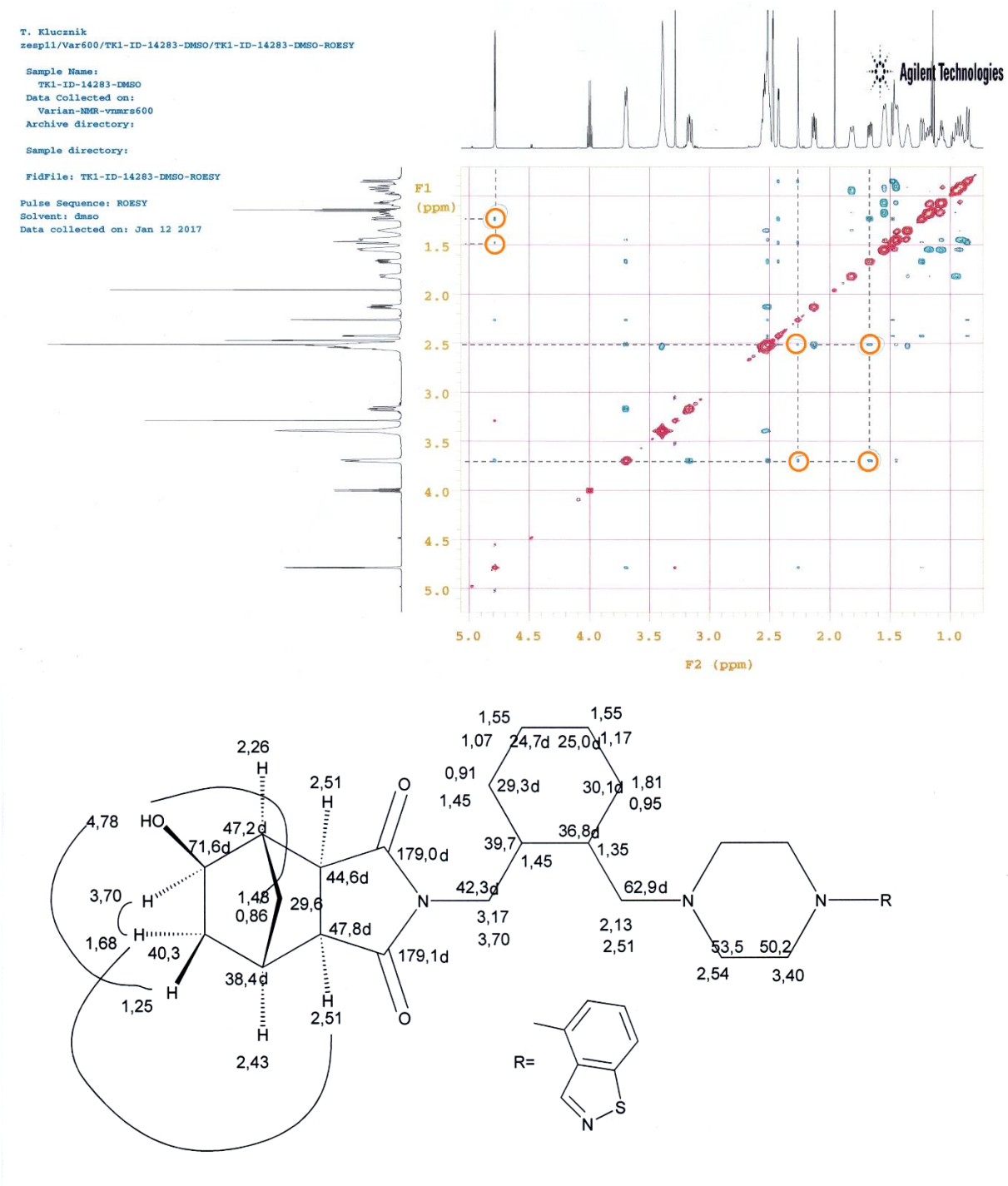
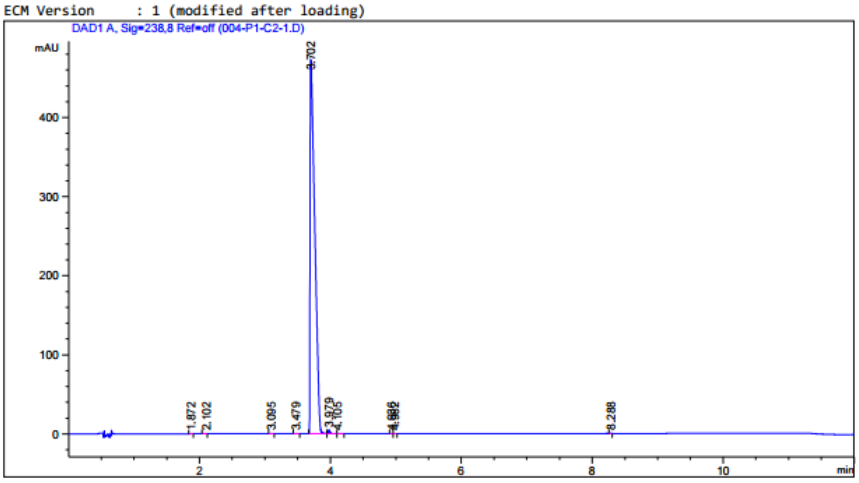


Figure S113. ROESY spectrum of compound **43**,**43'**.

ECM Path : \LAB\HPLC\LC 20\2017\RMH-240 P LC 20 2017-05-24 01-39-51.SC.SSIzip
ECM Version : 1 (modified after loading)



Peak #	RetTime [min]	Type	Width [min]	Area [mAU*s]	Height [mAU]	Area %
1	1.872	BV	0.0292	2.11181e-1	1.16434e-1	9.043e-3
2	2.102	BV	0.0343	6.32651e-2	2.28509e-2	2.709e-3
3	3.095	BV	0.0281	7.55542e-2	3.43683e-2	3.235e-3
4	3.479	VB	0.0349	5.95381e-2	2.07777e-2	2.550e-3
5	3.702	BV	0.0710	2326.64355	472.24716	99.6308
6	3.979	VV	0.0246	7.76837	4.62779	0.3327
7	4.105	VB	0.0330	1.48829e-1	5.78808e-2	6.373e-3
8	4.936	VV	0.0258	7.41235e-2	4.05160e-2	3.174e-3
9	4.982	VB	0.0290	6.78387e-2	3.37477e-2	2.905e-3
10	8.288	BBA	0.0220	1.54068e-1	1.10750e-1	6.597e-3
Totals :				2335.26633	477.31228	

*** End of Report ***

Area Percent Report

Sorted By : Signal
Multiplier : 1.0000
Dilution : 1.0000
Do not use Multiplier & Dilution Factor with ISTDs

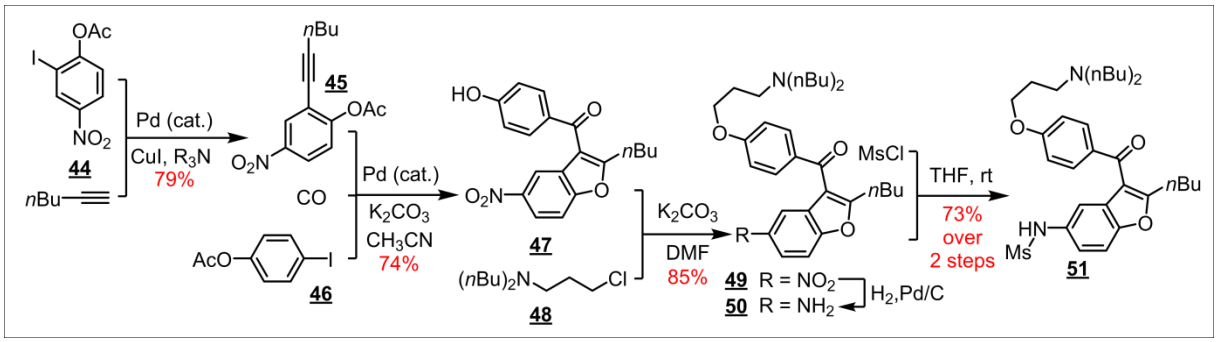
Signal 1: DAD1 A, Sig=238,8 Ref=off

Column:	Ascentis Express Phenyl Hexyl, 2.7 µm, 3.0 x 100 mm		
Mobile Phase:	A Acetonitrile		
	B 0.1% Ammonium acetate		
Gradient:	Time (min)	%A	%B
	0.0	40	60
	8.0	80	20
	10.0	80	20
	10.1	40	60
	15.0	40	60
Flow Rate:	0.7 mL/min		
Wavelength:	238 nm		
Temperature:	35 °C		

Fig. S114. 99.64% HPLC purity of compound **43**, **43'**.

Section S16. Synthesis of Dronedarone 51

S16.1. Previous vs. current synthetic routes.



Scheme S7. Chematica-planned synthesis of Dronedarone (**51**); same as **Figure 3c**.

List of patents protecting syntheses of Dronedarone

[1] J. Gubin, P. Chatelain, J. Lucchetti, G. Rosseells, H. Inion (Sanofi), US 5223510, **1993**

[2] A. Gutman, G. Nisneyich, L. Yudovich (Isp Investments Inc.), WO 2003/040120 A1, **2003**

[3] N. Fino, C. Leroy (Sanofi-Synthelabo), US 6828448 B2, **2004**

[4] M. Biard (Sanofi-Synthelabo), US 6846936 B2, **2005**

[5] A. Gutman, G. Nisnevich, L. Yudovitch (Isp Investments Inc.), US 7312345 B2, **2007**

[6] L. Eklund (Cambrex Karlskoga Ab), WO 2010/038029 A1, **2010**

[7] Q. Fuqing, H. Yang, M. Zhao (Eno Dubbo Chengdu Pharmaceutical Co., Ltd.), CN 101993427 A, **2011**

[8] M. Sada, A. Nardi, S. Maiorana (Laboratorio Chimico Internazionale S.P.A.), WO 2011/104591 A1, **2011**

[9] A. Friesz, Z. Dombrady, M. Csatarin Nagy (Sanofi-Aventis), WO 2011/70380 A1, **2011**

[10] A. Friesz, M. Csatarin Nagy (Sanofi), WO 2011/83346 A1, **2011**

[11] A. Friesz (Sanofi), WO 2011/158050 A1, **2011**

[12] E. Marom, M. Mizhiritskii, S. Rubnov (Mapi Pharma Holdings (Cyprus) Ltd.), WO 2011/99010 A1, **2011**

[13] J. Stohandl, J. Sindelarova (Ratiopharm GmbH), EP 2371824 A1, **2011**

[14] J. Stohandl, J. Sindelarova (Ratiopharm GmbH), EP 2371808 A1, **2011**

[15] X. Hou, Y. Chen (Jiangsu Hengrui Medicine Co., Ltd.), WO 2011153923 A1, **2011**

[16] 尚积金, 李丕永, 李岩, 林泉生 (New Materials Co., Ltd. Shandong Zouping Exhibition), CN 101948455 A, **2011**

[17] 张飞龙, 李建其, 王冠 (Shanghai Institute of Pharmaceutical Industry), CN 102070577 A, **2011**

[18] B. Baillon, M. Comte (Sanofi-Aventis), US 20110230553 A1**, 2011**

[19] S. D. Dwivedi, V. K. Patel, J. M. Pandya (Cadila Healthcare Ltd.), WO 2012/032545 A1, **2012**

[20] S. Srivastava, A. M. Castro, M. Gharpure, D. B. Deore, S. B. Narayanan (Glenmark Generics Ltd.), WO 2012/007959 A1, **2012**

[21] V. Jayaraman, R. Pillai, G. Bijukumar, J. Kevat, J. Nirmal, B. Dubadia (Alembic Pharmaceuticals Ltd.), WO 2012/153225 A1, **2012**

[22] R. Kumar, V. Tambe, S. Patil, K. Chavan, S. S. Naim, R. Dandala (Frichem Private Ltd.), WO 2012/4658 A2, **2012**

[23] A. Friesz, C. Huszar (Sanofi), WO 2012/10913 A1, **2012**

[24] A. Friesz, Z. Parkanyi, Z. Dombrady (Sanofi), WO 2012/131409 A1, **2012**

[25] F. Richter, E. Schreiner, S. Pirc, A. Copar (Lek Pharmaceuticals D.D.), WO 2012/62918 A1, **2012**

[26] R. Vishnu Newadkar, A. Changdeo Gaikwad, A. Madhukar Harad (Laboratories Lesvi, S.L.), WO 2012/52448 A1, **2012**

[27] S. Srivastava, A. M. Crasto, M. Gharpure, D. B. Deore, S. B. Narayanan (Glenmark Generics Ltd.), WO 2012/7959 A1, **2012**

[28] A. Friesz, C. Huszar (Sanofi), WO 2012/131408 A1, **2012**

[29] A. Friesz, C. Huszar (Sanofi), WO 2012/131410 A1, **2012**

[30] R. L. Giri, V. D. Mohite, S. Kambhampati, T. R. Chitturi, R. Thennati (Sun Pharmaceutical Industries Ltd.), WO 2012/120544 A2, **2012**

[31] A. Friesz, C. Huszar (Sanofi), WO 2012/10913 A1, **2012**

刘华全，宋俊松，朱海溪，欧阳平凯，王德才，韦萍 (Nanjing University of Technology), CN 102382087 A, **2012**

[31] 付清泉，杨海波，赵茂先 (Eno Dubbo Chengdu Pharmaceutical Co., Ltd.), CN 101993427 B, **2012**

[32] A. Copar, S. Pirc, F. Richter, E. Schreiner (Lek Pharmaceuticals D.D.), WO 2012062918 A1, **2012**

[33] A. Friesz, Z. Dombrady (Sanofi), WO 2013/14479 A1, **2013**

[34] A. Friesz, Z. Dombrady (Sanofi), WO 2013/121235 A2, **2013**

[35] C. Huszar, A. Hegedus, Z. Dombrady (Sanofi), WO 2013/121234 A1, **2013**

[36] A. Friesz, Z. Dombrady, M. Csatarine Nagy, C. Huszar (Sanofi), WO 2013/14480 A1, **2013**

[37] A. Friesz (Sanofi), EP 2617718 A1, **2013**

[38] A. Friesz, Z. Dombrady (Sanofi), WO 2013/14478 A1, **2013**

[39] C. Huszar, A. Hegedus, Z. Dombrady (Sanofi), WO 2013/178337 A1, **2013**

[40] F. Bailly, B. Grimaud, I. Malejonock, P. Vayron (Sanofi), US 2013/12729 A1**, 2013**

[41] C. Huszar, A. Hegedus, Z. Dombrady (Sanofi), WO 2013/124745 A1, **2013**

[42] F. Bailly, T. Priem, P. Vayron (Sanofi), US 2013/165673 A1, **2013**

[43] X. Bon, C. Biencourt, C. Leroy, J. Mateos-Caro, P. Vayron (Sanofi), US 2013/165675 A1, **2013**

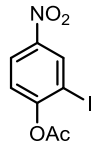
[44] B. Grimaud, P. J. Grossi (Sanofi), US 2014/18553 A1, **2014**

[45] X. Bon, J. L. Delepine, L. Jourdin, D. Largeau, P. Vayron (Sanofi), US 2015/31901 A1, **2015**

[46] J. Chan, J. Vitale (Gilead Sciences Inc.), WO 2015/31352 A1, **2015**

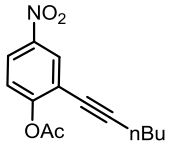
S16.2. Synthetic details

Reagents and solvents were purchased from commercial sources (Aldrich, ABCR, POCH, Chempur). More sensitive compounds were stored in a desiccator. Reagents were used without further purification unless otherwise noted. Carbon monoxide was purchased in gas cylinder and small portion was transferred to a balloon before each usage.
Flash column chromatography was performed using Merck silica gel 60 (230-400 mesh, 40-63 μm). Reactions were monitored using Macherey-Nagel silica gel 60F254 aluminium plates. TLC's were visualized by UV fluorescence (254 nm) or iodine vapors.
NMR spectra were recorded on a Bruker 400 MHz Avance III spectrometer at room temperature. Chemical shifts (δ) were reported in parts per million (ppm) relative to residual solvent peaks rounded to the nearest 0.01 (ref: $CHCl_3$ [^{1}H: 7.26, ^{13}C: 77.2]). Coupling constants (J) were reported in Hz to the nearest 0.1 Hz. Peak multiplicity was indicated as follows: s (singlet), d (doublet), t (triplet), q (quartet), qi (quintet), sx (sextet) and m (multiplet). HRMS spectra were recorded on AutoSpec Premier (Waters) or MaldiSYNAPT G2-S HDMS (Waters) spectrometers and are given in m/z.



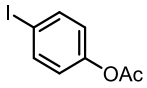
***2-iodo-4-nitrophenyl acetate* (<u>44</u>):** To a solution of 2-iodo-4-nitrophenol (0.303 g, 1.143 mmol) in THF (0.7 mL) and DCM (0.7 mL), TEA (0.350 mL, 1.372 mmol) was added dropwise at 0 ºC. Then, AcCl (0.098 mL, 1.372 mmol) was added dropwise at 0 ºC. The reaction mixture was stirred at 0 ºC for 30 min and then warmed to room temperature and stirred for another 3.5 hrs. After completion of the reaction, the reaction mixture was diluted with water and DCM. The aqueous layer was extracted with DCM three times and the combined organic layers were dried over anhydrous Na_2SO_4 and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 5:1) to yield **<u>44</u>** as a white powder (0.303 g, 94%).

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.71 (d, *J* = 2.6 Hz, 1H), 8.27 (dd, *J* = 8.9, 2.6 Hz, 1H), 7.30 (d, *J* = 8.9 Hz, 1H), 2.43 (s, 3H).
^{13}C NMR (101 MHz, $CDCl_3$): δ 167.57, 156.15, 145.73, 134.78, 124.74, 123.27, 90.44, 21.15.
HRMS (EI+): m/z calcd for $C_8H_6INO_4$: 306.9342 , found: 306.9343.



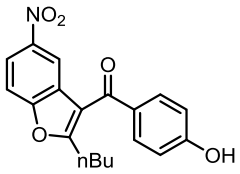
***2-(hex-1-yn-1-yl)-4-nitrophenyl acetate* (<u>45</u>):** A round-bottomed flask was charged with compound **<u>44</u>** (3.482 mmol, 1.069 g), $Pd(PPh_3)_4$ (0.052 mmol, 0.060 g) and CuI (0.108 mmol, 0.021 g). The flask was then capped with a rubber septum and Ar atmosphere was established. Degassed THF (25 mL), TEA (34.820 mmol, 4.853 mL) and 1-hexyne (6.963 mmol, 0.800 mL) were added *via* syringe. The reaction mixture was stirred at room temperature for 5.5 hrs. After completion of the reaction, the reaction mixture was diluted with water and AcOEt. The aqueous layer was extracted with AcOEt three times and the combined organic layers were dried over anhydrous Na_2SO_4 and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 20:1) to yield **<u>45</u>** as a yellow solid (0.761 g, 84%).

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.32 (d, *J* = 2.7 Hz, 1H), 8.16 (dd, *J* = 8.9, 2.7 Hz, 1H), 7.25 (d, *J* = 8.9 Hz, 1H), 2.47 (t, *J* = 7.0 Hz, 2H), 2.38 (s, 3H), 1.66 – 1.57 (m, 2H), 1.56 – 1.45 (m, 2H), 0.98 (t, *J* = 7.3 Hz, 3H).
^{13}C NMR (101 MHz, $CDCl_3$): δ 167.85, 156.06, 145.33, 128.43, 123.61, 123.12, 119.80, 98.59, 73.92, 30.44, 29.68, 21.90, 20.74, 19.18, 13.53.
HRMS (EI+): m/z calcd for $C_{14}H_{15}NO_4$: 261.1001 , found: 261.0997.



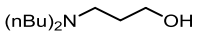
***4-iodophenyl acetate* (46):** 4-iodophenol (10.06 mmol, 2.214 g,) and DMAP (0.60 mmol, 0.074 g) were dissolved in DCM (25 mL) and the flask was capped with rubber septum. Then TEA (8.05 mmol, 1.12 mL) and Ac_2O (16.100 mmol, 1.52 mL) were added dropwise. The reaction mixture was stirred at room temperature for 1.5 hrs. Afterwards, the reaction mixture was diluted with water and DCM. The aqueous layer was extracted with DCM three times and the combined organic layers were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 7:1) to yield **46** as a white powder (2.589 g, 98%).

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.71 (d, *J* = 8.8 Hz, 1H), 6.89 (d, *J* = 8.8 Hz, 1H), 2.31 (s, 2H).
^{13}C NMR (101 MHz, $CDCl_3$) δ 168.99, 150.54, 138.47, 123.77, 89.81, 21.08.
HRMS (EI+): m/z calcd for $C_8H_7O_2$: 261.9491 , found: 261.9500.



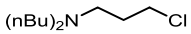
***(2-butyl-5-nitrobenzofuran-3-yl)(4-hydroxyphenyl)methanone* (47):** A round-bottomed flask was charged with compounds **45** (0.491 mmol, 0.128 g) and **46** (1.576 mmol, 0.913 g), $Pd(PPh_3)_4$ (0.030 mmol, 0.034 g) and anhydrous K_2CO_3 (2.357 mmol, 0.326 g). The flask was then capped with a rubber septum, evacuated and refilled with CO (three times) and then equipped with CO-filled balloon. Then, CH_3CN (1.2 mL), which was previously degassed and saturated with CO, was added *via* syringe. The reaction mixture was stirred at 55 ºC for 24 hrs. After completion of the reaction, the reaction mixture was diluted with water and DCM. The aqueous layer was extracted with DCM three times and the combined organic layers were dried over anhydrous Na_2SO_4 and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography (hexane:AcOEt, 15:1) to yield **47** as a white powder (0.126 g, 76%).

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.36 (d, *J* = 2.3 Hz, 1H), 8.24 (dd, *J* = 9.0, 2.4 Hz, 1H), 7.81 (d, *J* = 8.7 Hz, 2H), 7.59 (d, *J* = 9.0 Hz, 1H), 6.98 (d, *J* = 8.7 Hz, 2H), 6.42 (s, 1H), 2.95 (t, 2H), 1.79 (q, *J* = 7.6 Hz, 2H), 1.37 (sx, *J* = 7.4 Hz, 2H), 0.91 (t, *J* = 7.4 Hz, 3H).
^{13}C NMR (101 MHz, $CDCl_3$): δ 189.35, 168.60, 168.01, 156.30, 154.65, 144.80, 135.82, 130.70, 127.59, 122.02, 120.41, 117.73, 117.01, 111.44, 29.90, 29.68, 28.09, 22.30, 21.17, 13.59.
HRMS (EI+): m/z calcd for $C_{19}H_{17}NO_5$: 339.1107 , found: 339.1104.



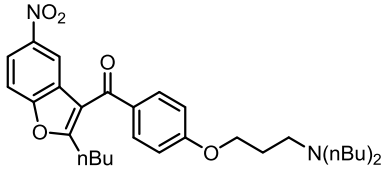
***3-(dibutylamino)propan-1-ol (SI-11)*:** Ethyl acrylate (9.9 mmol, 0.991 g) and dibutylamine (9.0 mmol, 1.151 g) were dissolved in methanol (12 mL) and the reaction mixture was stirred at room temperature for 8 hrs. Then, the reaction solvent was removed *in vacuo* to afford crude ethyl 3-(dibutylamino)propanoate as a transparent oil (2.095 g), which was used in the next reaction without further purification. The crude ethyl 3-(dibutylamino)propanoate (9.13 mmol, 2.095 g) was dissolved in THF (50 mL) and $LiAH_4$ (40.0 mmol, 1.518 g) was added. The reaction mixture was stirred at room temperature for 8 hrs. After completion of the reaction, the reaction mixture was diluted with water (10 mL) and dioxane (10 mL). The solvents were removed *in vacuo* to afford grey powder. The powder was then dissolved in DCM (200 mL), filtered through a celite pad and dried over anhydrous $MgSO_4$. $MgSO_4$ was filtered off and the solution was evaporated to dryness under vacuum to yield **SI-11** as a pale yellow oil (1.1861 g, 70% over two steps).

^{1}H NMR (400 MHz, $CDCl_3$) δ 5.62 (s, *J* = 135.8 Hz, 1H), 3.77 (t, 2H), 2.62 (t, 2H), 2.47 – 2.35 (m, 4H), 1.66 (qi, *J* = 5.4 Hz, 2H), 1.44 (qi, 4H), 1.29 (sx, *J* = 7.3 Hz, 4H), 0.90 (t, *J* = 7.3 Hz, 6H).
^{13}C NMR (101 MHz, $CDCl_3$) δ 64.70, 55.25, 53.92, 28.99, 27.82, 20.64, 13.99.
HRMS (EI+): m/z calcd for $C_{11}H_{25}NO$: 187.1936 , found: 187.1936.



N-butyl-N-(3-chloropropyl)butan-1-amine **(<u>48</u>):** Compound **<u>SI-11</u>** (2.230 mmol, 0.418 g) was dissolved in $CHCl_3$ (5 mL), and $SOCl_2$ (4.460 mmol, 0.328 mL) was added *via* syringe. The reaction mixture was stirred under reflux for 7 hrs. After completion of the reaction, the reaction mixture was diluted with water and DCM, and K_2CO_3 (1.2 g) was added. The organic layer was washed with saturated water solution of $NaHCO_3$ (5 mL), then washed again with water (5 mL). The organic layer was dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* to yield **<u>48</u>** as a yellow oil (0.430 g, 94%).

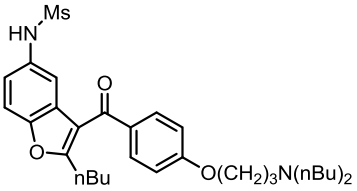
^{1}H NMR (400 MHz, CDCl3) δ 3.62 (t, J = 6.5 Hz, 1H), 2.55 (t, J = 6.8 Hz, 1H), 2.40 (t, 2H), 1.89 (qi, 1H), 1.48 – 1.36 (m, 2H), 1.36 – 1.26 (m, 2H), 0.93 (t, J = 7.3 Hz, 3H).
^{13}C NMR (101 MHz, CDCl3) δ 54.02, 51.06, 43.45, 30.61, 29.43, 20.67, 14.05.
HRMS (EI+): m/z calcd for $C_{11}H_{24}NCl$: 205.1597 , found: 205.1596.



(2-butyl-5-nitrobenzofuran-3-yl)(4-(3-(dibutylamino)propoxy)phenyl)methanone **(<u>49</u>):**

Compounds **<u>47</u>** (0.730 mmol, 0.248 g), **<u>48</u>** (0.803 mmol, 0.165 g) and K_2CO_3 (0.730 mmol, 0.101 g) were dissolved in DMF (4.5 mL) and the reaction mixture was stirred at 85 ºC for 4 hrs. After completion of the reaction, the reaction mixture was diluted with water and DCM. The aqueous layer was extracted with DCM three times and the combined organic layers were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography ($CHCl_3$:MeOH, 30:1) to yield **<u>49</u>** as a yellow oil (0.316 g, 85%).

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.36 (d, 1H), 8.23 (dd, *J* = 9.0, 2.3 Hz, 1H), 7.83 (d, *J* = 8.7 Hz, 2H), 7.57 (d, *J* = 9.0 Hz, 1H), 7.00 (d, *J* = 8.8 Hz, 2H), 4.14 (t, *J* = 6.3 Hz, 2H), 2.93 (t, *J* = 7.6 Hz, 2H), 2.64 (t, *J* = 6.9 Hz, 2H), 2.51 – 2.38 (m, 4H), 2.03 – 1.91 (m, 2H), 1.83 – 1.72 (m, 2H), 1.48 – 1.40 (m, 4H), 1.38 – 1.29 (m, 6H), 0.91 (t, 9H).
^{13}C NMR (101 MHz, $CDCl_3$) δ 189.02, 167.01, 163.64, 156.30, 144.64, 131.69, 130.75, 127.94, 120.16, 117.69, 117.34, 114.48, 111.33, 66.62, 53.95, 50.33, 29.90, 29.47, 29.06, 27.93, 27.00, 22.31, 20.68, 14.05, 13.62.
HRMS (ESI+): m/z [$M+H^+$] calcd for $C_{30}H_{41}N_2O_5$: 509.3015 , found: 509.3008.



N-(2-butyl-3-(4-(3-(dibutylamino)propoxy)benzoyl)benzofuran-5yl)methanesulfonamide **(<u>51</u>):**
Compound **<u>49</u>** (0.230 mmol, 0.117 g) was dissolved in THF (3 mL). The flask was capped with a rubber septum and Ar atmosphere was established . Then, Pd/C (10%) (0.028 g) was added, the flask was evacuated, refilled with H_2 and equipped with H_2-filled balloon. The reaction mixture was stirred overnight at room temperature. After completion of the reaction, the reaction mixture was

filtered through a celite pad, and the pad was washed with AcOEt three times. The solvents were removed *in vacuo* to obtain **50** as a yellow oil (0.112 g) which was used in the next step without further purification.
Crude **50** (0.112 g) was dissolved in DCM (3 mL), then Py (0.351 mmol, 0.025 mL) and MsCl (0.234 mmol, 0.018 mL) were added dropwise. The reaction mixture was stirred at room temperature for 1.5 hrs. After completion of the reaction, the reaction mixture was diluted with water and DCM. The aqueous layer was extracted with DCM three times and the combined organic layers were dried over anhydrous $MgSO_4$ and filtered. The solvents were removed *in vacuo* and the residue was then purified by flash column chromatography ($CHCl_3$:MeOH, 40:1) to yield **51** as a yellow oil (0.094 g, 73% over two steps).

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.83 (d, *J* = 8.8 Hz, 2H), 7.48 (d, *J* = 8.6 Hz, 1H), 7.32 – 7.26 (m, 2H), 6.98 (d, *J* = 8.8 Hz, 2H), 4.14 (t, *J* = 6.3 Hz, 2H), 2.95 (s, 3H), 2.89 (t, 2H), 2.65 (t, *J* = 6.8 Hz, 2H), 2.47 (t, 4H), 2.02 – 1.93 (m, 2H), 1.76 (qi, *J* = 7.6 Hz, 2H), 1.50 – 1.40 (m, 4H), 1.39 – 1.28 (m, 8H), 0.95 – 0.86 (m, 9H).
^{13}C NMR (101 MHz, $CDCl_3$) δ 190.20, 165.79, 163.25, 151.84, 132.38, 131.65, 131.30, 128.26, 120.14, 116.77, 115.55, 114.29, 111.78, 66.54, 53.88, 50.35, 39.05, 30.02, 29.69, 29.13, 28.00, 26.93, 22.31, 20.68, 14.05, 13.65.
HRMS (ESI+): m/z [M+H$^+$] calcd for $C_{31}H_{45}N_2O_5S$: 557.3049 , found: 557.3048.

S16.3. Raw spectroscopic and chromatographic data.

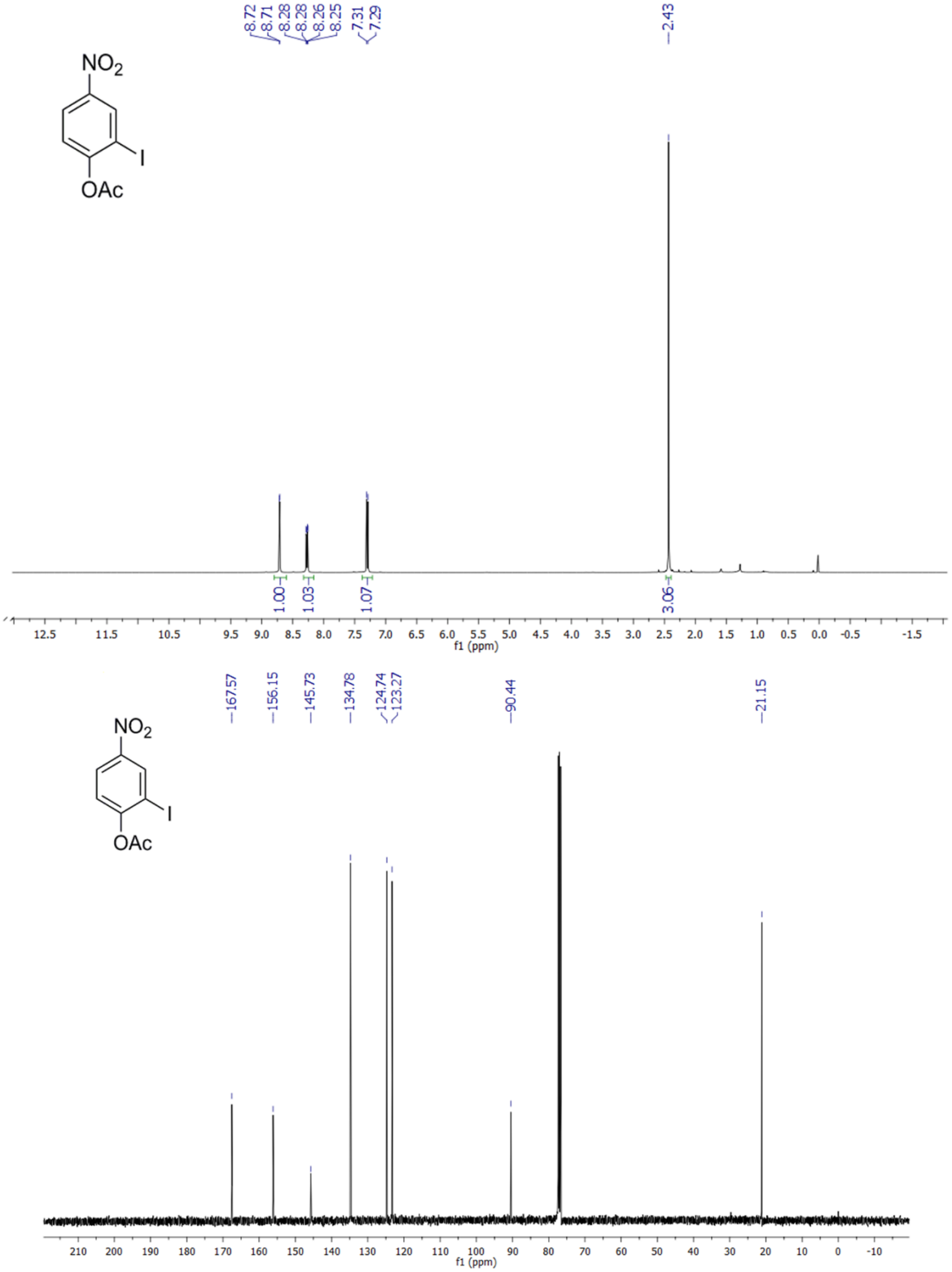


Figure S115. ^{1}H (top) an ^{13}C NMR (bottom) spectrum of compound **44**.

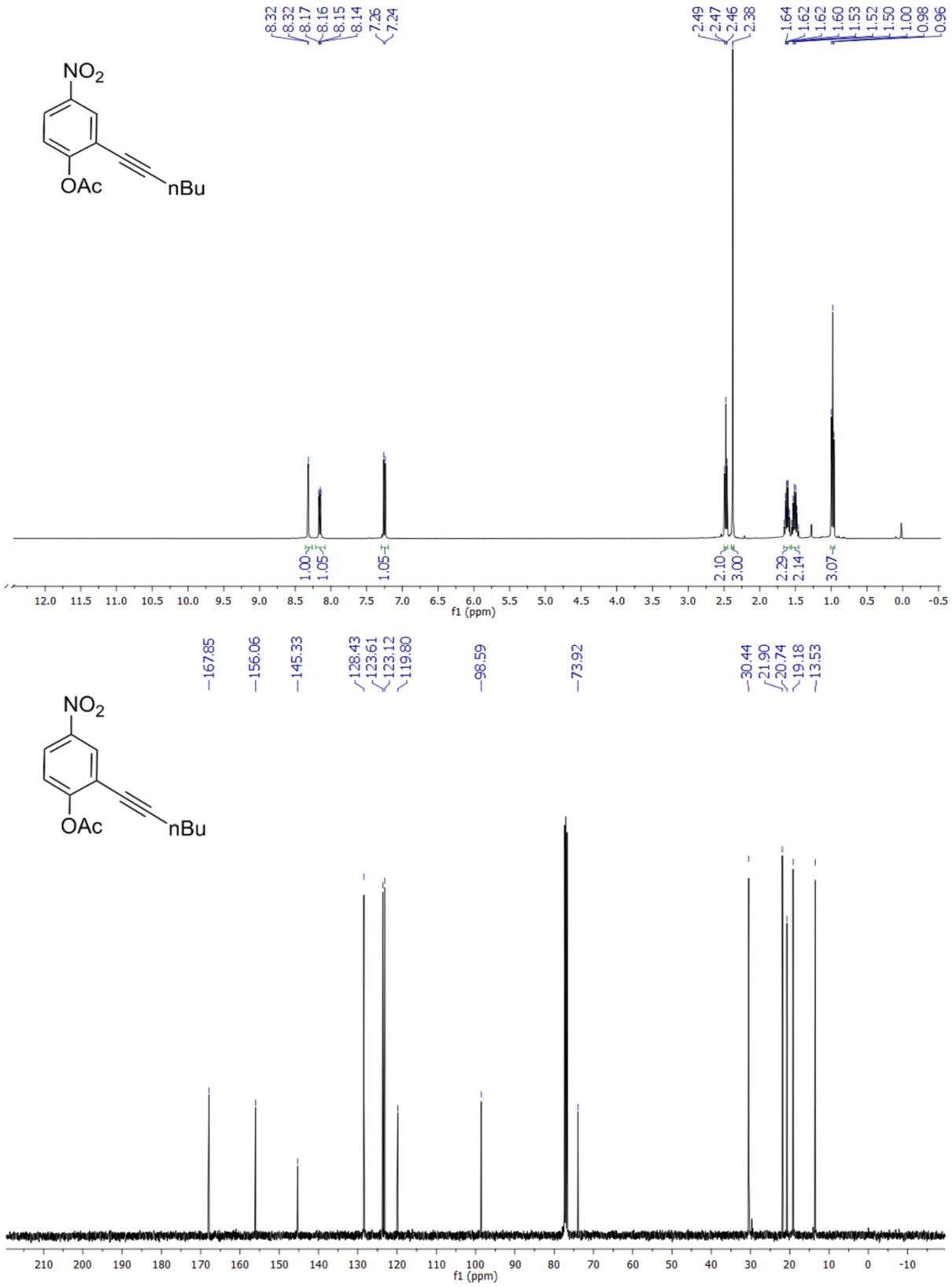


Figure S116. ^{1}H (top) an ^{13}C NMR (bottom) spectrum of compound **45**.

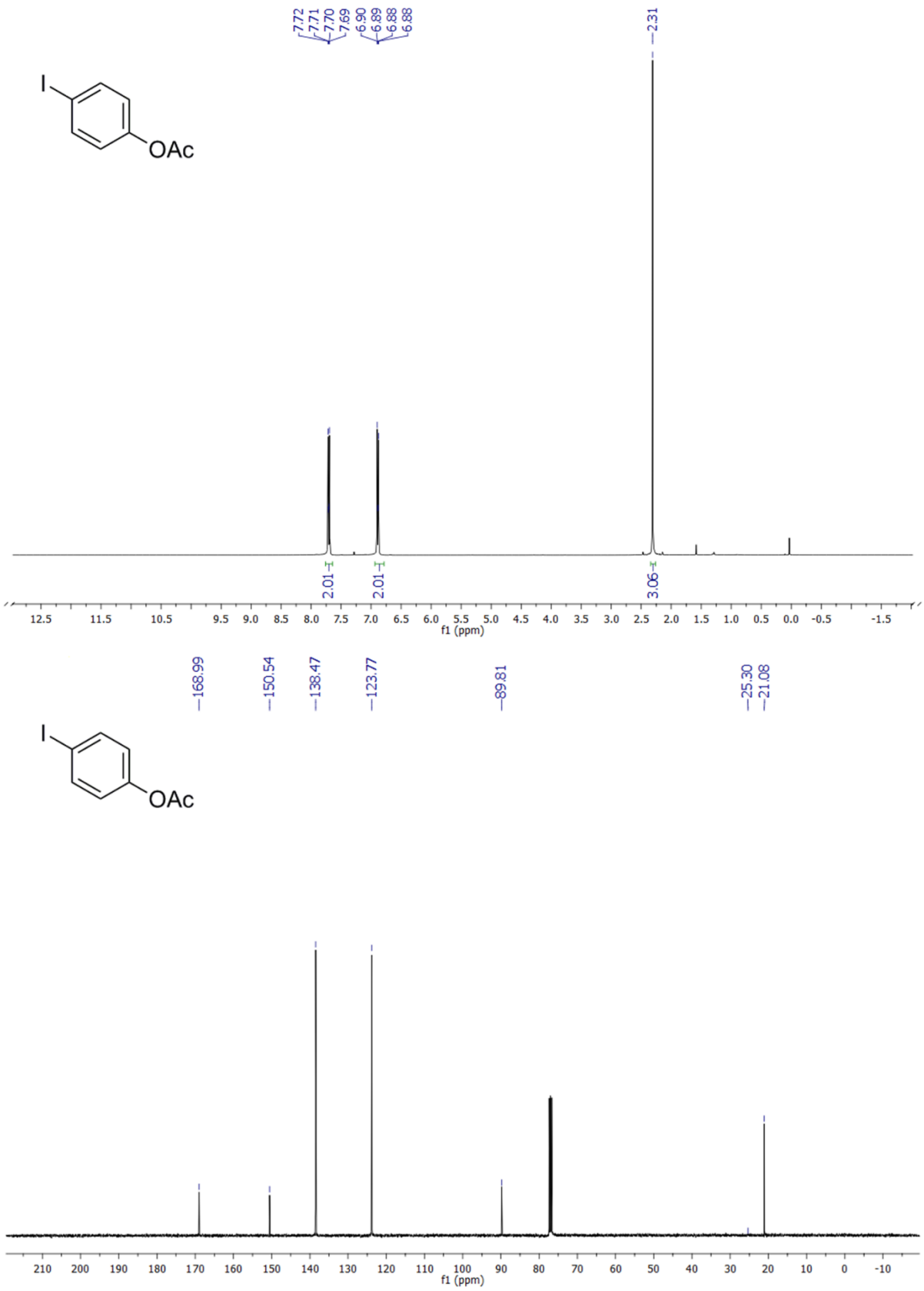


Figure S117. ^{1}H (top) an ^{13}C NMR (bottom) spectrum of compound **46**.

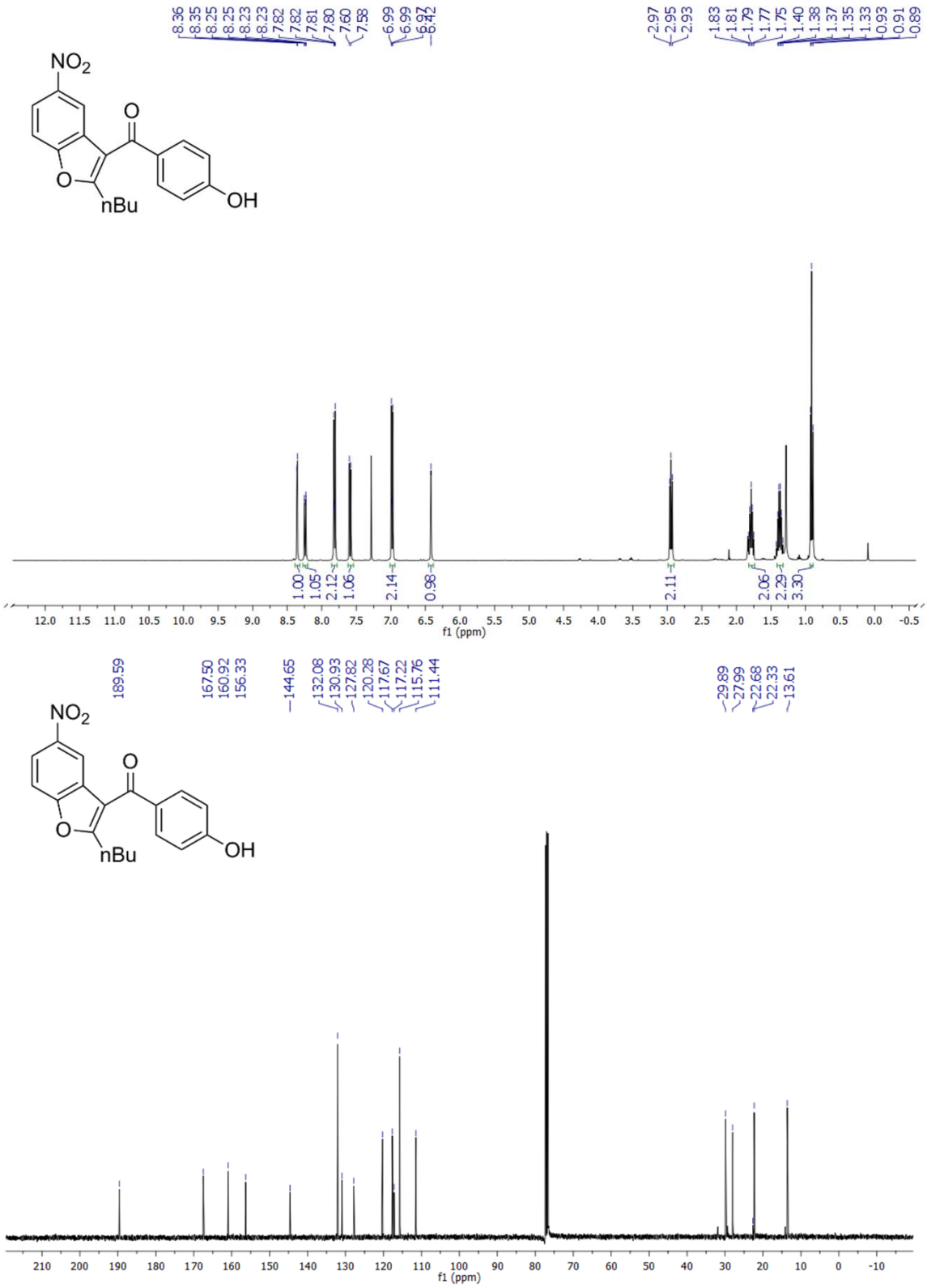


Figure S118. ^{1}H (top) an ^{13}C NMR (bottom) spectrum of compound **47**.

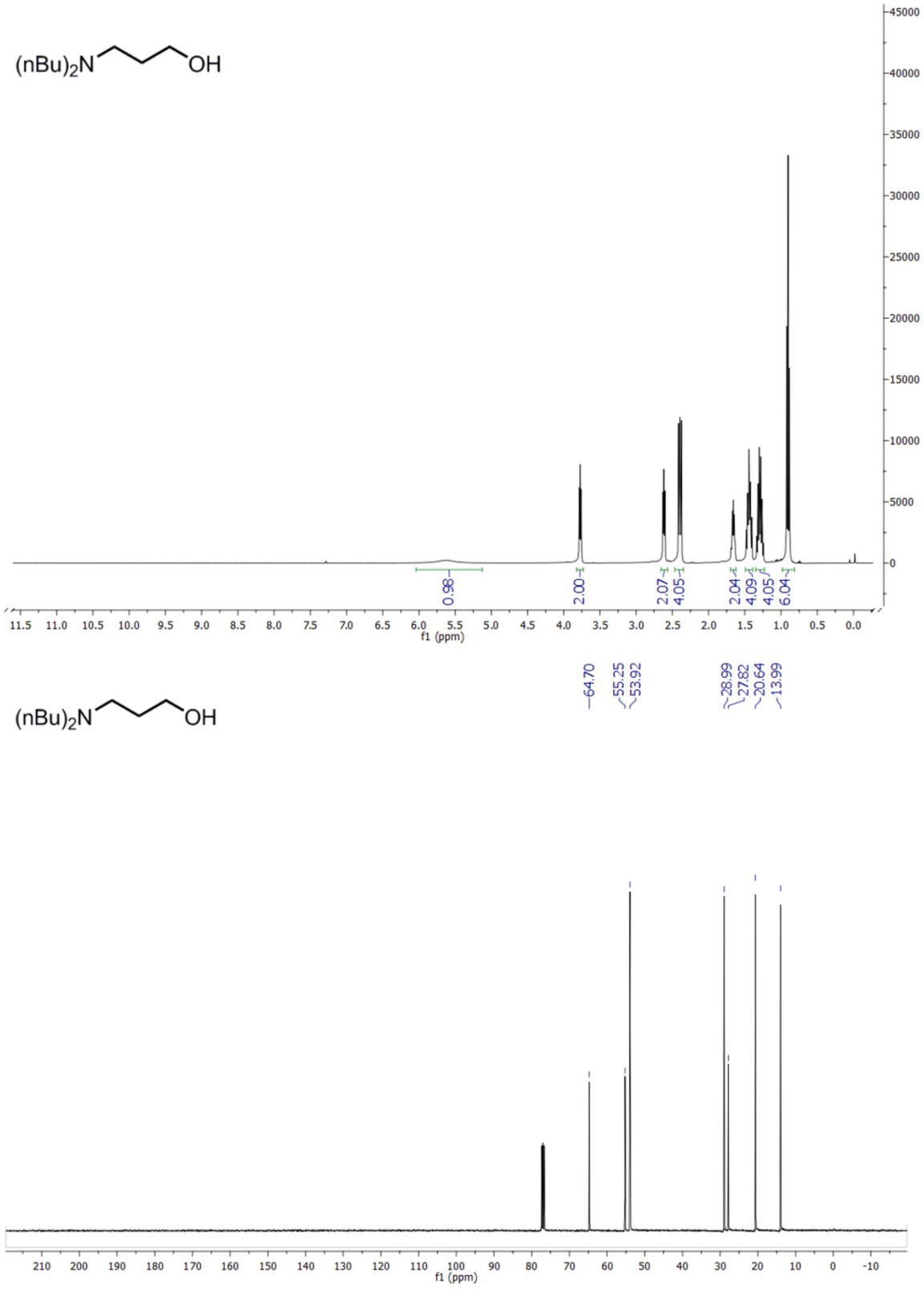


Figure S119. ^{1}H (top) an ^{13}C NMR (bottom) spectrum of compound **SI-11**.

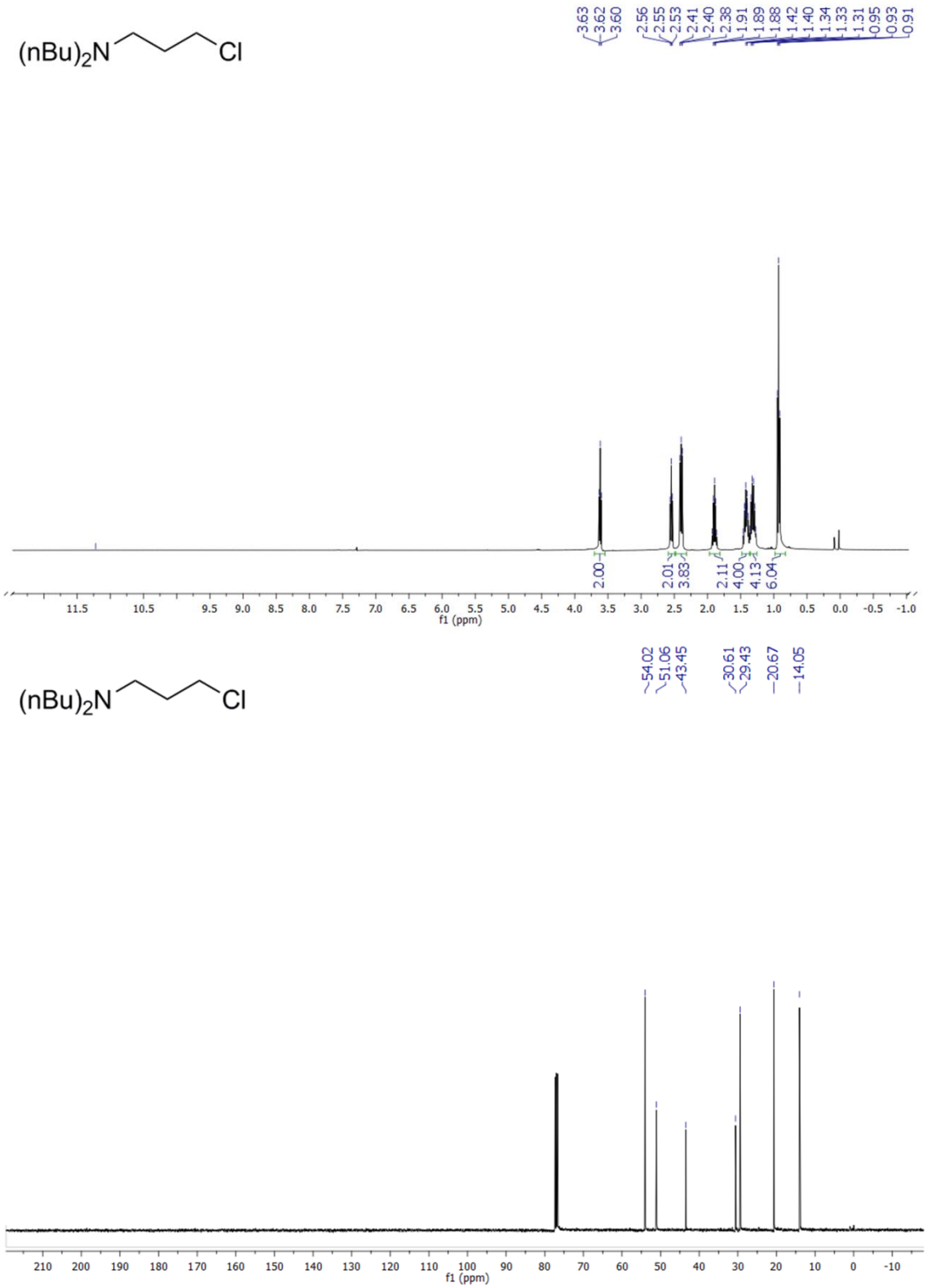


Figure S120. ^{1}H (top) an ^{13}C NMR (bottom) spectrum of compound **48**.

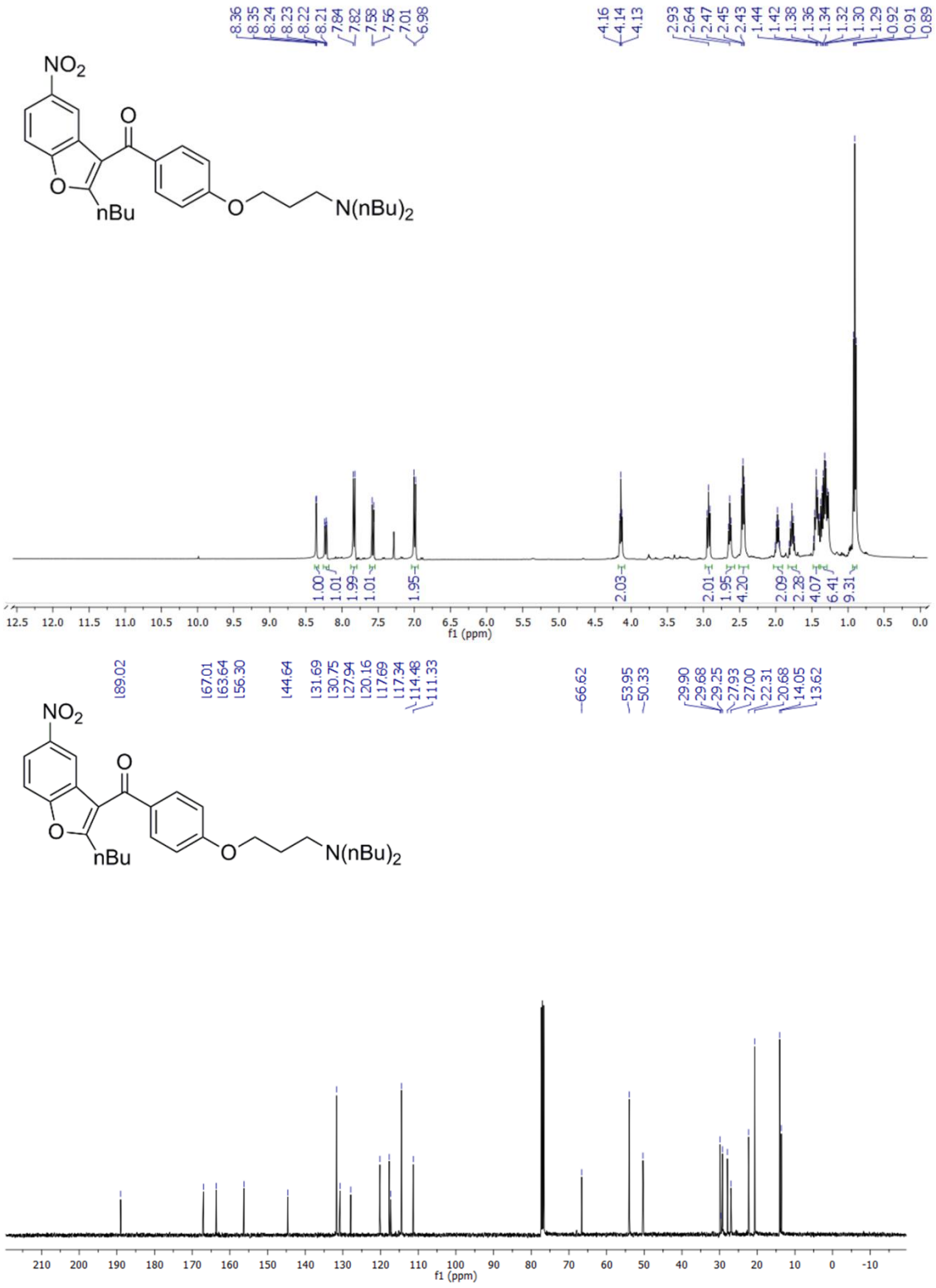


Figure S121. ^{1}H (top) an ^{13}C NMR (bottom) spectrum of compound **49**.

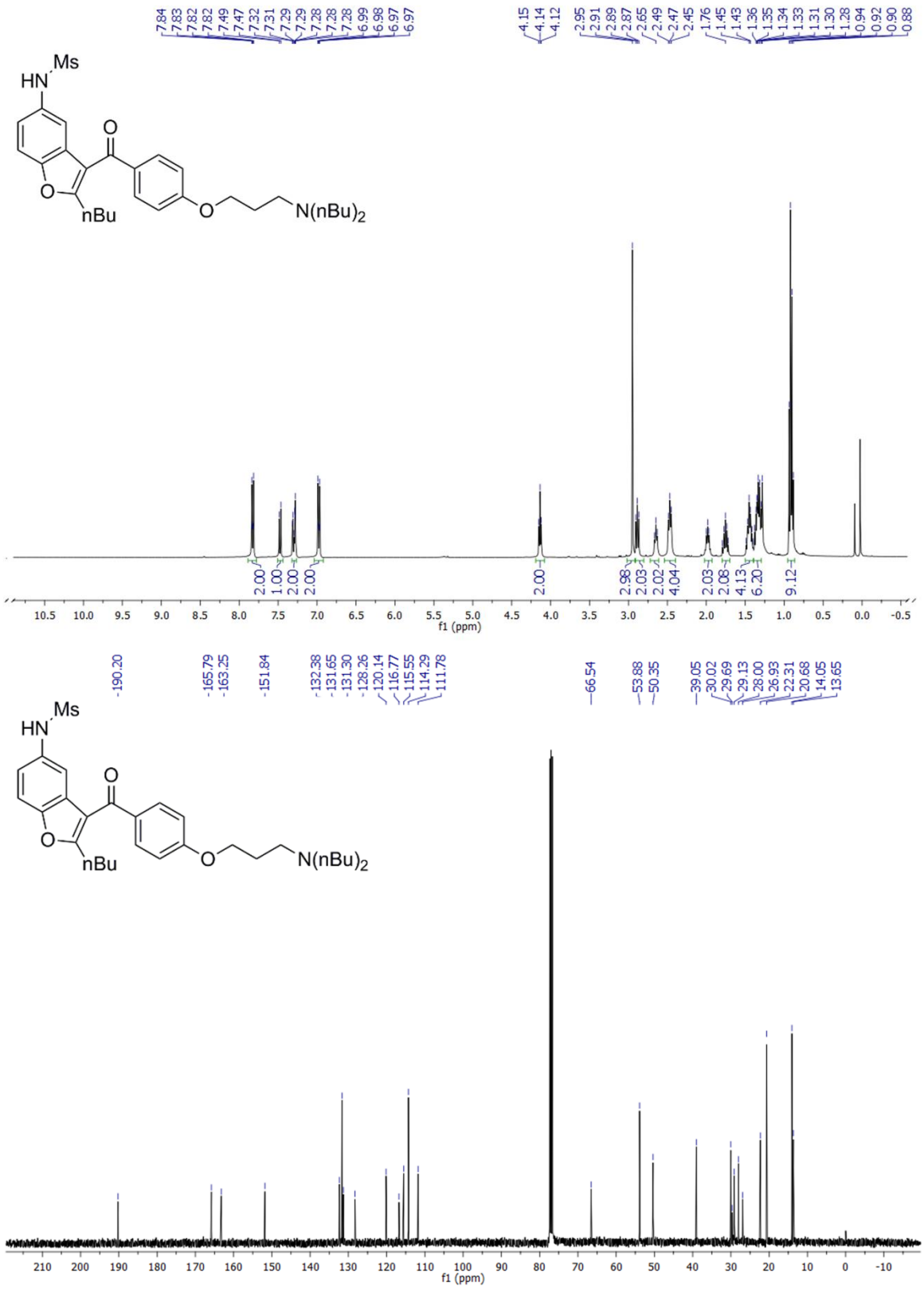
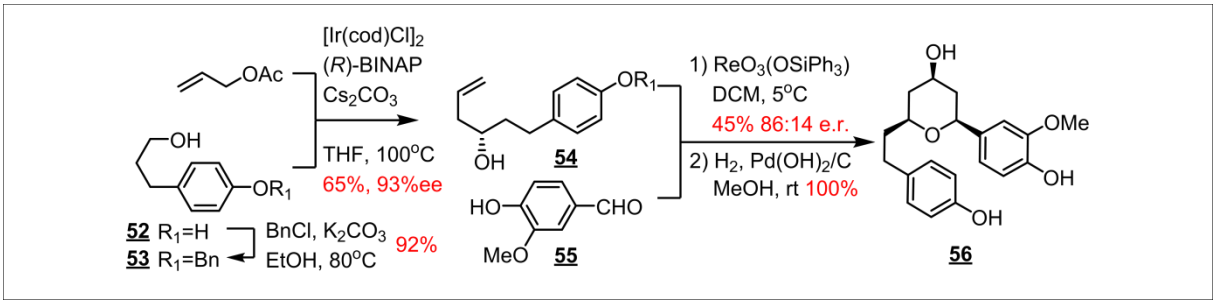


Figure S122. ^{1}H (top) an ^{13}C NMR (bottom) spectrum of compound **51**.

Section S17. Synthesis of Engelheptanoxide C 56

S17.1. The current synthetic route.

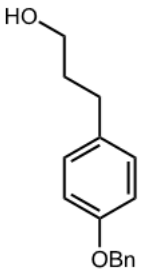


Scheme S8. Chematica-planned synthesis of Engelheptanoxide C (**56**); same as **Figure 3d**. Note there is no prior reported route to this compound.

S17.2 Synthetic details

General Information

All reactions were carried out under a nitrogen atmosphere in flame dried glassware. Dichloromethane and tetrahydrofuran were purified by passage through a bed of activated alumina. Ethanol (Absolute, 200 proof) and methanol (Certified ACS, ≥99.8%) were purchased from Fisher Scientific. 3-(4-Hydroxyphenyl)propan-1-ol, vanillin and all the other reagents were purchased from Sigma Aldrich. Purification of reaction products was carried out by flash chromatography using Agela Technologies flash silica (40-60 μm, 60 Å). Analytical thin layer chromatography was performed on Merck KGaA TLC silica gel 60 F_{254} glass plates (20 × 20 cm). Visualization was accomplished with UV light or ceric ammonium molybdate stain by heating. ^{1}H NMR and ^{13}C NMR were recorded on a Bruker Avance III 500 MHz. 1D NOE experiments were performed by an Agilent DD2 500 MHz. Mass spectra was obtained on a Waters Acquity UPLC. High resolution mass spectra were obtained on an Agilent 6210A LC-TOF (ESI mode). IR spectra were recorded on a Bruker Tensor 37 FT-IR (ATR). Optical rotations were measured on a Rudolph Research Analytical Autopol IV Automatic Polarimeter. Melting points were measured on a Thomas Hoover capillary melting point apparatus.

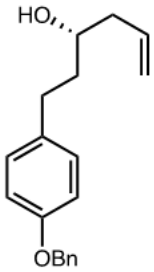


***3-(4-benzyloxyphenyl)propan-1-ol* 53**

To a 50mL flame-dried round-bottom flask was added 3-(4-hydroxyphenyl)-propan-1-ol **52** (716 mg, 4.70 mmol) and potassium carbonate (974 mg, 7.05 mmol). Absolute ethanol (10 mL, 200 Proof) was added to dissolve the aldehyde. Benzyl chloride (0.59 mL, 5.17 mmol) was added into the stirring suspension before the round-bottom was connected with a jacketed condenser. The reaction mixture was heated under nitrogen atmosphere and refluxed for 6 h. The white precipitate was filtered after the mixture was cooled down to room temperature. The filtrate was concentrated and purified by flash column chromatography (4:1 hexanes/EtOAc) to afford alcohol **53** as a white solid (mp 63-64 °C) in 95% yield (1.08 mg, 4.45 mmol).

Spectral data for **53**: ^{1}H NMR (500 MHz, $CDCl_3$) δ 1.26 (s, 1H), 1.87 (m, 2H), 2.66 (t, 2H, *J* = 7.5 Hz), 3.67 (t, 2H, *J* = 6.8 Hz), 5.05 (s, 2H), 6.91 (d, 2H, *J* = 7.8 Hz), 7.12 (d, 2H, *J* = 7.8 Hz), 7.32 (m, 1H), 7.38 (t, 2H, *J* = 7.8 Hz), 7.44 (d, 2H, *J* = 7.8 Hz); ^{13}C NMR (125 MHz, $CDCl_3$) δ 31.31, 34.56, 62.44,

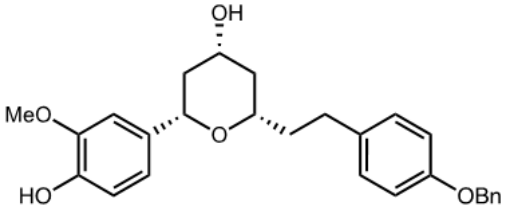
70.21, 114.94, 127.61, 128.04, 128.70, 129.47, 134.28, 137.33, 157.20; These spectral data match with those previously reported on this compound (Boll, P. M.; Hald, M.; Parmar, V. S.; Tyagi, O. D.; Bisht, K. S.; Sharma, N. K.; Hansen, S. *Phytochemistry*, **1992**, *31*, 1035).



***(S)-1-(4-benzyloxyphenyl)hex-5-en-3-ol* 54**

To an oven-dried sealed tube loaded with alcohol **53** (727 mg, 3.00 mmol), $[Ir(cod)Cl]_2$ (50.4 mg, 0.0750 mmol), (*R*)-BINAP (93.4 mg, 0.150 mmol), cesium carbonate (195 mg, 0.600 mmol) and 3-nitrobenzoic acid (50.1 mg, 0.300 mmol) in the glovebox was added THF (15 mL) followed by allyl acetate (3.2 mL, 30 mmol). The reaction mixture was allowed to stirred at 100 °C for 24 h and filtered to remove any insoluble material after it was cooled down. The filtrate was concentrated and the residue was purified by flash column chromatography (8:1 hexanes/EtOAc) afforded homoallylic alcohol **54** as an off-white solid (67-68 °C, 93% *ee*) in 65% yield (552 mg, 1.96 mmol). The enantiomeric ratio of **54** was determined by SFC as 93% *ee* [(Chiralpak ID column, CO_2:MeOH = 80:20, 3 mL/min, 210 nm), t_{minor} = 2.13 min, t_{major} = 2.53 min].

Spectral data for (*S*)-**7**: ^{1}H NMR (500 MHz, $CDCl_3$) δ 1.59 (s, 1H), 1.76 (m, 2H), 2.18 (m, 1H), 1.06 (m, 1H), 2.64 (dt, 1H, *J* = 14.2, 8.3 Hz), 2.75 (dt, 1H, *J* = 14.3, 7.3 Hz), 3.67 (m, 1H), 5.05 (s, 2H), 5.15 (dd, 2H, *J* = 14.1, 1.4 Hz), 5.82 (m, 1H), 6.91 (d, 2H, *J* = 8.7 Hz), 7.13 (d, 2H, *J* = 8.7 Hz), 7.33 (t, 1H, *J* = 7.4 Hz), 7.38 (t, 2H, *J* = 7.4 Hz), 7.44 (d, 2H, *J* = 7.4 Hz); ^{13}C NMR (125 MHz, $CDCl_3$) δ 31.28, 38.78, 42.22, 70.03, 70.20, 114.94, 118.46, 127.61, 128.03, 128.70, 129.48, 134.52, 134.78, 137.34, 157.17; LCMS (ESI-TOF) *m/z* 265.21 [(M–OH$^-$); calcd. for $C_{19}H_{21}O^+$: 265.1587]; IR (ATR) 3359br, 3027s, 2978s, 2907s, 2853s, 1612s, 1512vs, 1451s, 1384s, 1252vs, 1080s, 1042s cm^{-1}; $[\alpha]_D^{23}$ +85.3° (c 1.0, $CHCl_3$) on 93% *ee* of (*S*)-**7** (SFC).

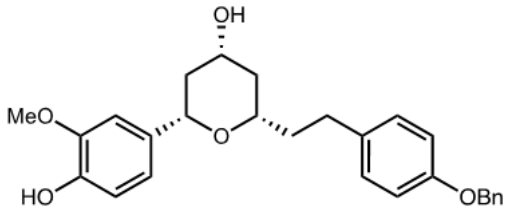


***(2S,4R,6S)-2-(4-(benzyloxy)phenethyl)-6-(4-hydroxy-3-methoxyphenyl)tetrahydro-2H-pyran-4-ol* SI-12**

To a flame-dried round-bottom flask was added a solution of homoallylic alcohol **54** (282 mg, 1.00 mmol), vanillin **55** (291 mg, 1.20 mmol) and $ReO_3(OSiPh_3)$ (25.5 mg, 0.05 mmol) in 5 mL CH_2Cl_2. The reaction was stirred at room temperature under nitrogen atmosphere for 48 h upon completion by TLC analysis. The reaction mixture was concentrated and the residue was purified by flash column chromatography (2:1 hexanes/EtOAc) afforded tetrahydropyranol **SI-12** as an off-white solid (135-136 °C, 52% *ee*) in 56% yield (295 mg, 0.562 mmol). The enantiomeric ratio of **SI-12** was determined by SFC as 52% *ee* [(Chiralpak ID column, CO_2:*i*-PrOH = 75:25, 3 mL/min, 210 nm), t_{major} = 8.27 min, t_{minor} = 9.72 min]. The reaction ran with homoallylic alcohol **54** (85 mg, 0.20 mmol, 93% *ee*), vanillin **55** (55 mg, 0.24 mmol) and $ReO_3(OSiPh_3)$ (7.6 mg, 0.010 mmol) in 1 mL CH_2Cl_2 at 5 °C for 72 h afforded tetrahydropyranol **SI-12** in 45% yield (39 mg, 0.091 mmol) and 72% *ee* (SFC).

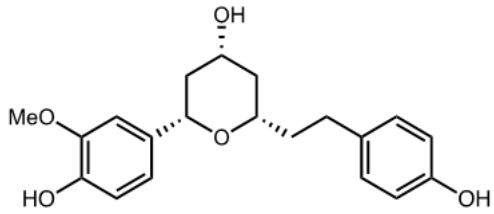
Spectral data for **SI-12**: ^{1}H NMR (500 MHz, $CDCl_3$) δ 1.24-1.35 (m, 1H), 1.44-1.56 (m, 2H), 1.74-1.83 (m, 1H), 1.92-2.05 (m, 2H), 2.16-2.22 (m, 1H), 2.63-2.79 (m, 2H), 3.40-3.49 (m, 1H), 3.87-3.99 (m, 1H), 3.91 (s, 3H), 4.27 (d, 1H, *J* = 11.0 Hz), 5.04 (s, 2H), 5.57 (s, 1H), 6.84-6.92 (m, 4H), 6.94 (s, 1H), 7.10 (d, 2H, *J* = 8.4 Hz), 7.32 (t, 1H, 7.3 Hz), 7.38 (t, 2H, *J* = 7.3 Hz), 7.43 (d, 2H, *J* = 7.3 Hz); ^{13}C NMR (125 MHz, $CDCl_3$) δ 30.92, 37.89, 41.07, 42.94, 56.06, 68.69, 70.21, 74.93, 77.27, 108.80, 114.23, 114.88, 119.12, 127.62, 128.04, 128.71, 129.53, 134.36, 134.53, 137.35, 145.16, 146.58,

157.14; LCMS (ESI-TOF) *m/z* 435.34 [(M+H$^+$); calcd. for $C_{27}H_{31}O_5^+$: 435.2171]; IR (ATR) 3434br, 3025s, 2948s, 2897s, 2835s, 1609s, 1511vs, 1451s, 1433s, 1383s, 1251vs, 1228s, 1158s, 1116s, 1068s, 1038s cm^{-1}; $[\alpha]_D^{22}$ –35.6° (c 0.54, MeOH) on 88% *ee* **SI-12** (SFC).



***(2S,4R,6S)-2-(4-(benzyloxy)phenethyl)-6-(4-hydroxy-3-methoxyphenyl)tetrahydro-2H-pyran-4-ol* SI-12**

A mixture of acetic acid (57 µL, 1.0 mmol), trimethylsilyl acetate (120 µL, 0.800 mmol) and boron trifluoride etherate (49 µL, 0.40 mmol) in cyclohexane (1.0 mL) was added to a solution of homoallylic alcohol **54** (56.5 mg, 0.200 mmol) and vanillin (36.5 mg, 0.240 mmol) in CH_2Cl_2 (1.0 mL) at 0 °C. The resulting solution was stirred under an argon atmosphere and followed by TLC analysis on completion. The reaction mixture was neutralized by $NaHCO_3$ sat. and the aqueous layer was extracted by CH_2Cl_2. The combined organic phase was washed with brine, dried with Na_2SO_4. After the solvent was remove under reduced pressure, the residue was dissolved in methanol (2 mL). Potassium carbonate (111 mg, 0.800 mmol) was added and the reaction mixture was stirred at room temperature for 5 h. After methanol was removed by reduced pressure, water was added and the aqueous was extracted with CH_2Cl_2. The combine organic phase was washed with brine, dried with Na_2SO_4 and concentrated by reduced pressure. The residue was purified by flash column chromatography (2:1 hexanes/EtOAc) afforded tetrahydropyranol **SI-12** as an off-white solid in 30% yield (26.3 mg, 0.0605 mmol). The enantiomeric ratio of **SI-12** was determined by SFC as 88% *ee* [(Chiralpak ID column, CO_2:*i*-PrOH = 75:25, 3 mL/min, 210 nm), t_{major} = 8.21 min, t_{minor} = 9.54 min].



***(2S,4R,6S)-2-(4-hydroxy-3-methoxyphenyl)-6-(4-hydroxyphenethyl)tetrahydro-2H-pyran-4-ol* 56**

To a flame-dried round-bottom flask was added *O*-benzyl engelheptanoxide C **SI-12** (52.1 mg, 0.120 mmol) and Pearlman's catalyst (33.7 mg) in methanol (1.2 mL). The reaction mixture was degassed by freeze-pump-thaw with three cycles, charged with hydrogen balloon and stirred at room temperature for 3 h. The reaction mixture was filtered and concentrated under reduced pressure. The residue was purified by flash column chromatography (2:1 hexanes/EtOAc) afforded engelheptanoxide C **56** as a semisolid in 98% yield (40.3 mg, 0.117 mmol).

Spectral data for **1c**: ^{1}H NMR (500 MHz, *d*^{6}-acetone) δ 1.22 (dd, 1H, *J* = 23.2, 11.6 Hz), 1.39 (dd, 1H, *J* = 23.2, 11.6 Hz), 1.68-1.78 (m, 1H), 1.80-1.89 (m, 1H), 1.93-2.00 (m, 1H), 2.07-2.14 (m, 1H), 2.58-2.73 (m, 2H), 3.40-3.47 (m, 1H), 3.76-3.92 (m, 2H), 3.85 (s, 3H), 4.28 (dd, 1H, *J* = 11.1, 1.5 Hz), 6.74 (d, 2H, *J* = 8.5 Hz), 6.79 (d, 1H, *J* = 8.0 Hz), 6.85 (dd, 1H, *J* = 8.4, 1.7 Hz), 7.01(dd, 1H, *J* = 8.4, 1.7 Hz), 7.03 (d, 2H, *J* = 8.5 Hz), 7.44 (s, 1H), 8.05 (s, 1H); ^{13}C NMR (125 MHz, *d*^{6}-acetone) δ 31.53, 39.11, 42.09, 44.37, 56.21, 68.42, 75.51, 78.04, 110.51, 115.34, 115.90, 119.38, 130.11, 133.81, 135.67, 146.50, 147.99, 156.21; HRMS (ESI-TOF) *m/z* 367.1522 [(M+Na$^+$); calcd. for $C_{20}H_{24}O_5Na^+$: 367.1521]; $[\alpha]_D^{22}$ –35.6° (c 0.54, MeOH) on 88% *ee* **1c** (SFC); Lit[4] $[\alpha]_D^{25}$ –7.44° (c 0.14, MeOH) on 100% *ee* material.

S17.3. Raw spectroscopic and chromatographic data.

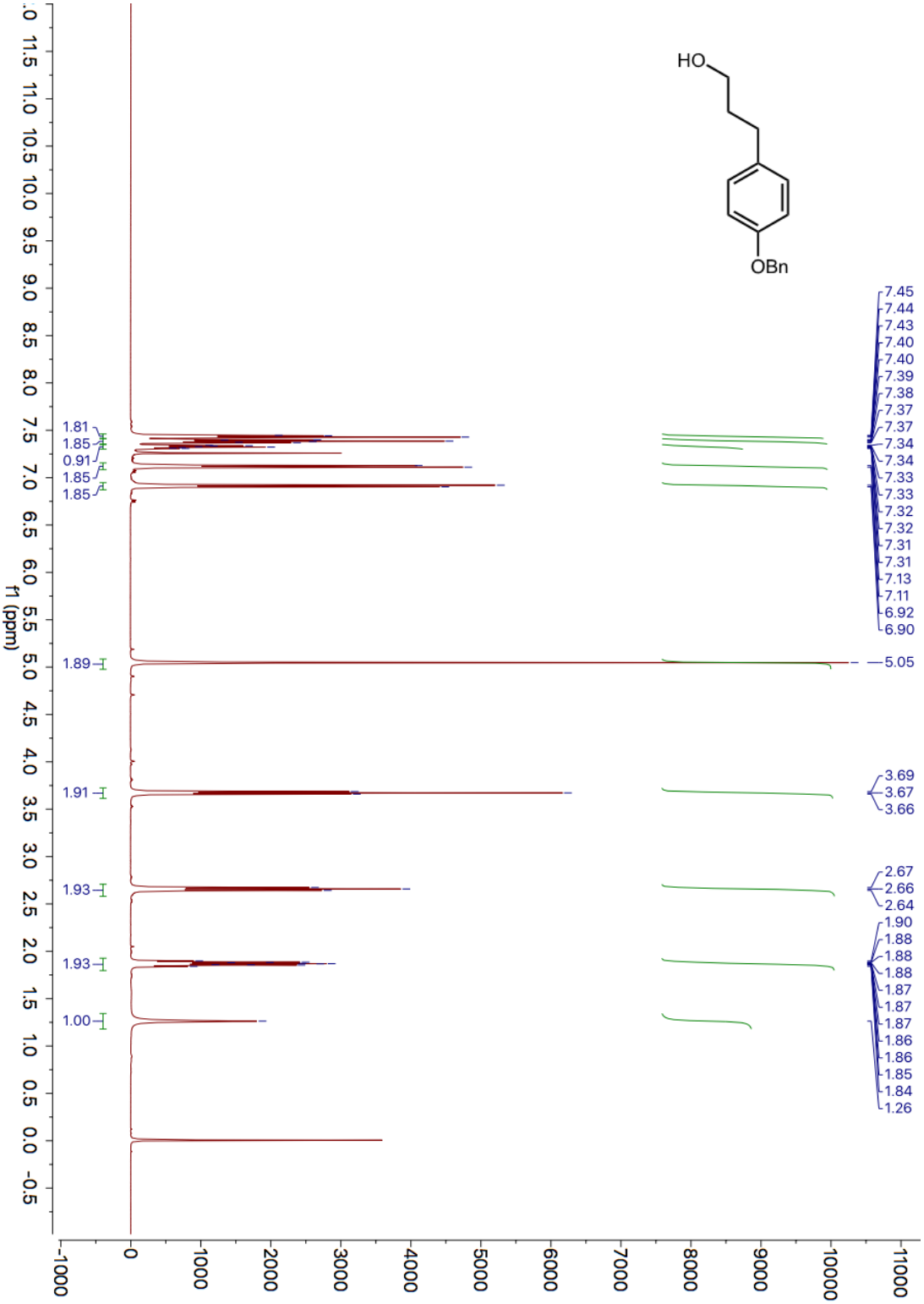


Figure S123. ^{1}H NMR spectrum of compound **53**.

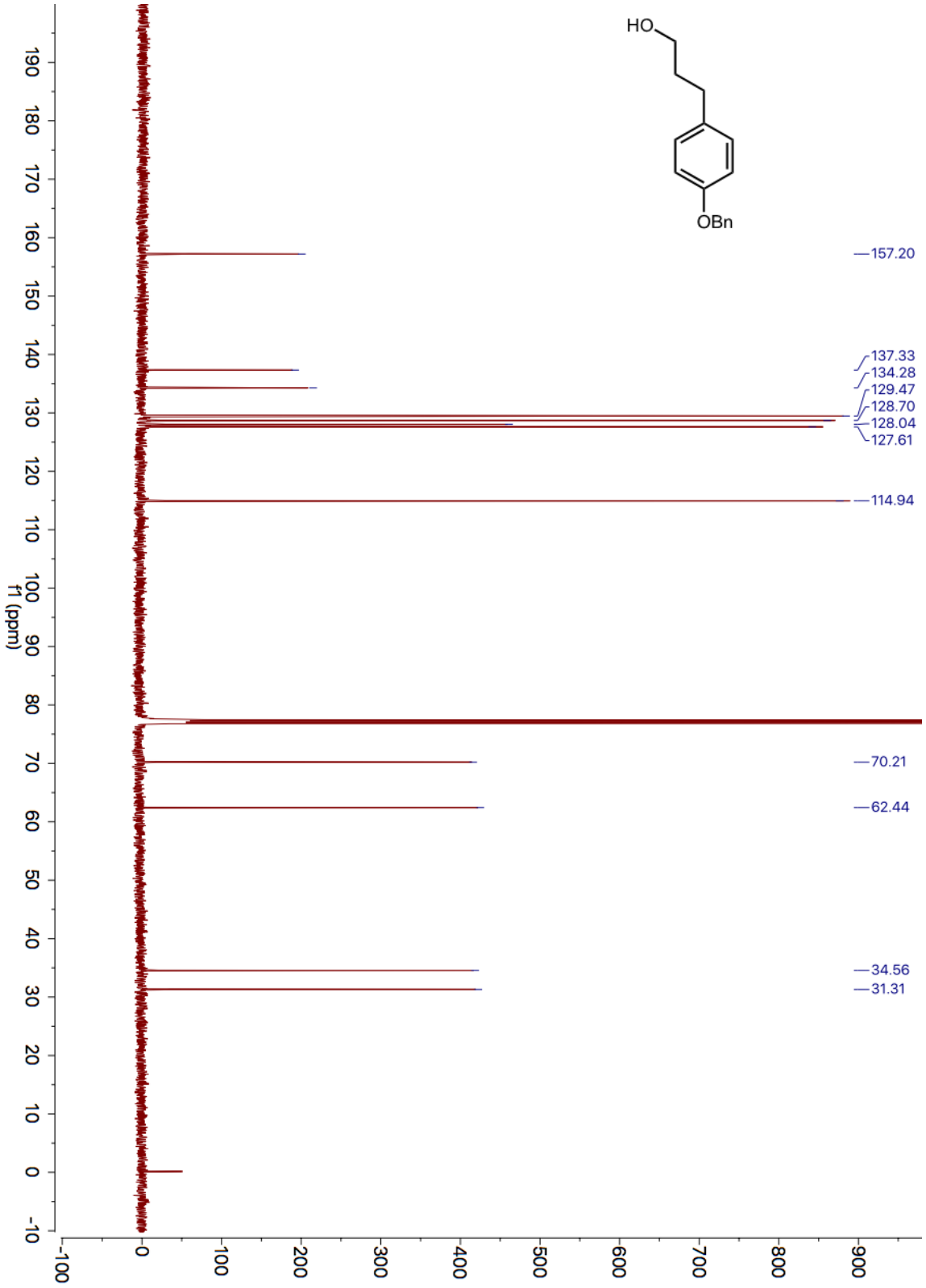


Figure S124. ^{13}C NMR spectrum of compound **53**.

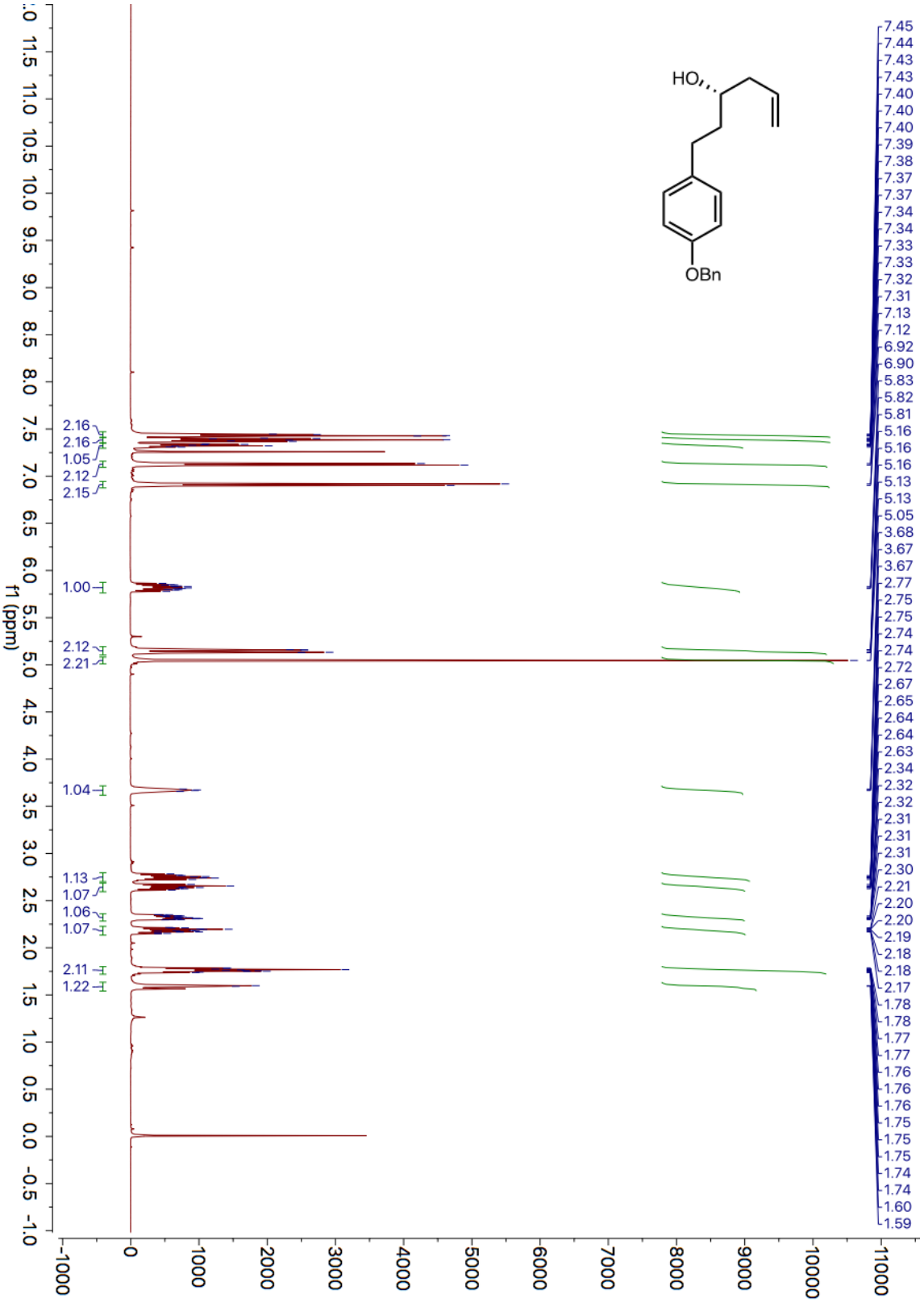


Figure S125. ^{1}H NMR spectrum of compound **54**.

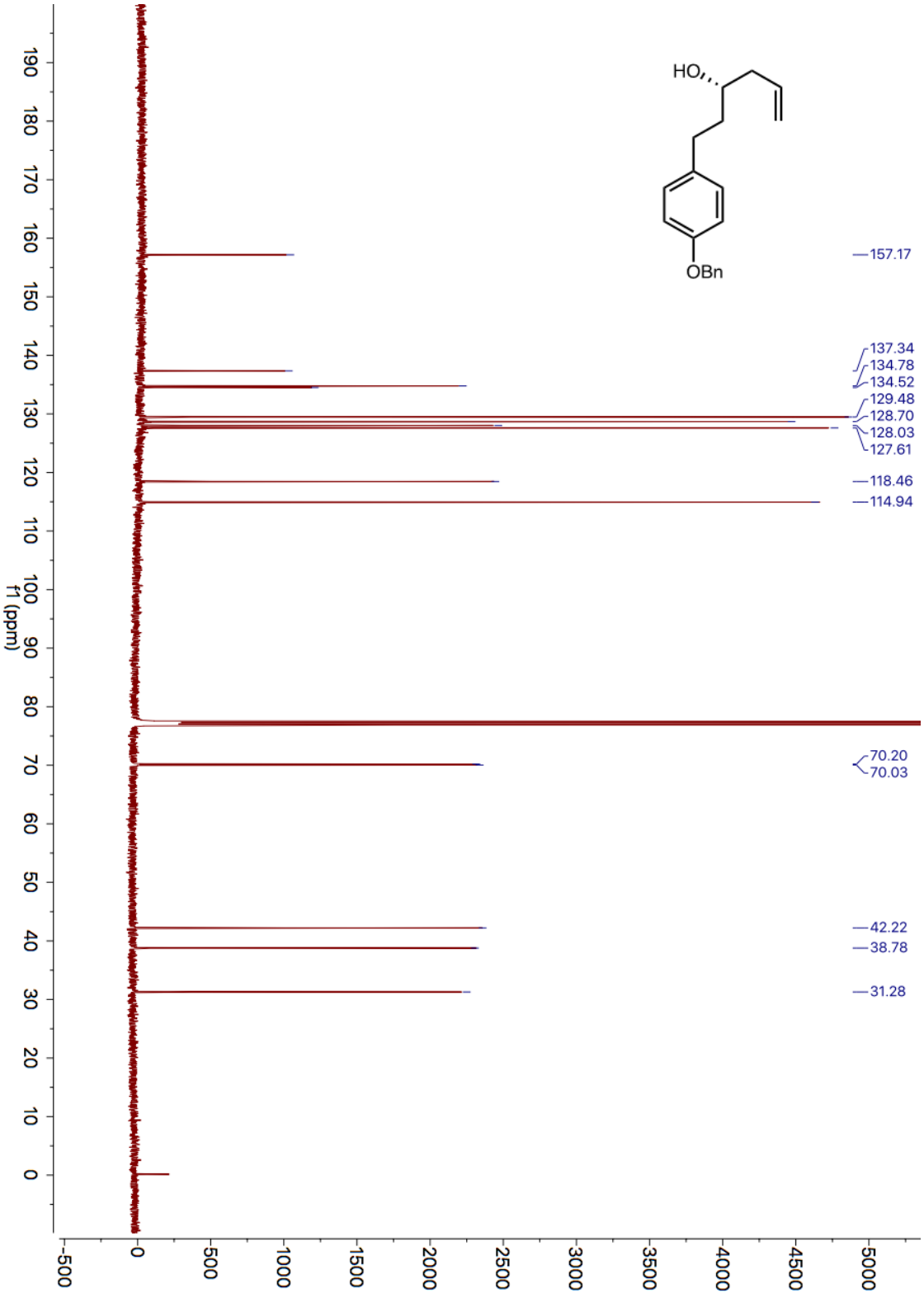


Figure S126. ^{13}C NMR spectrum of compound **54**.

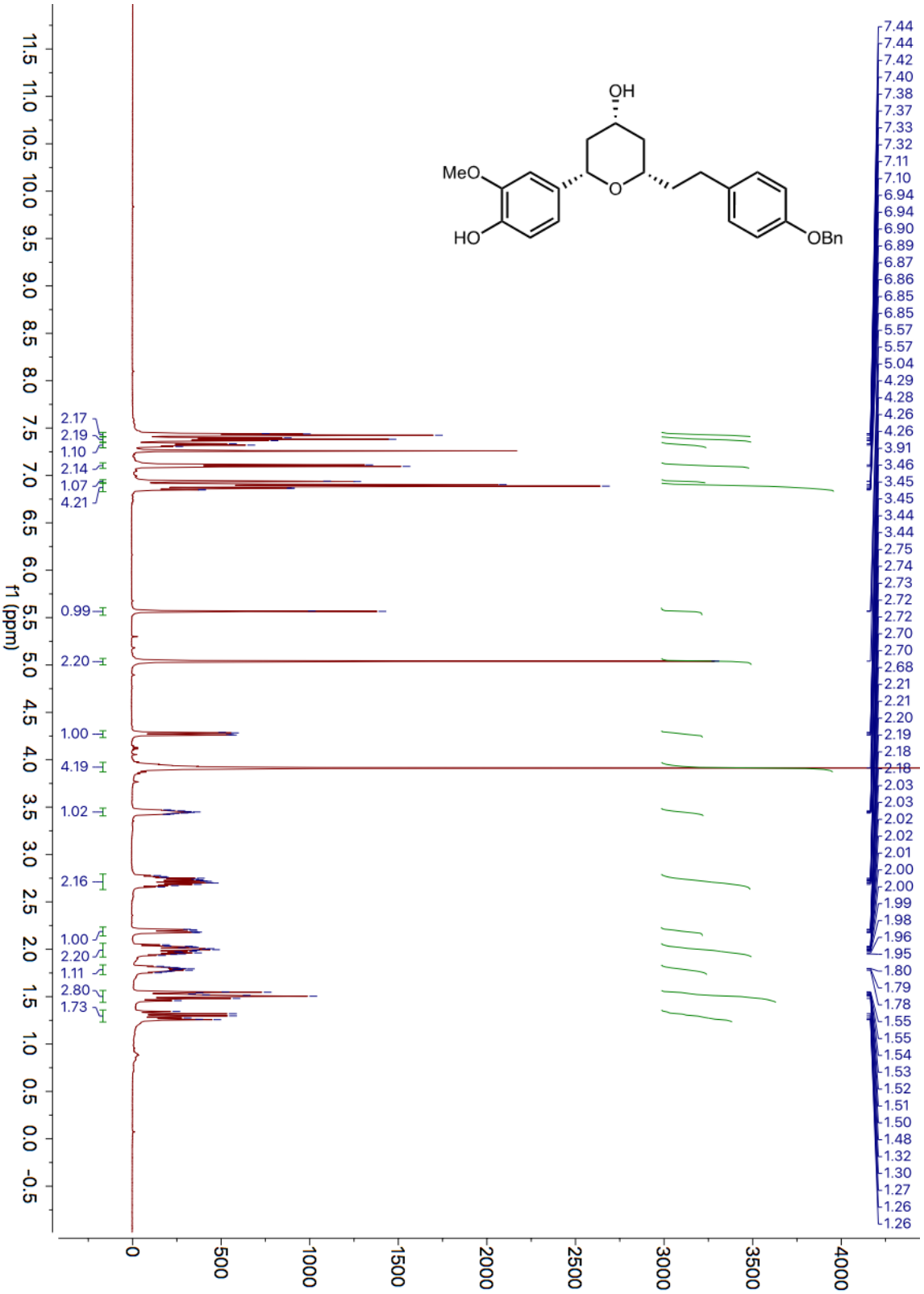


Figure S127. ^{1}H NMR spectrum of compound **SI-12**.

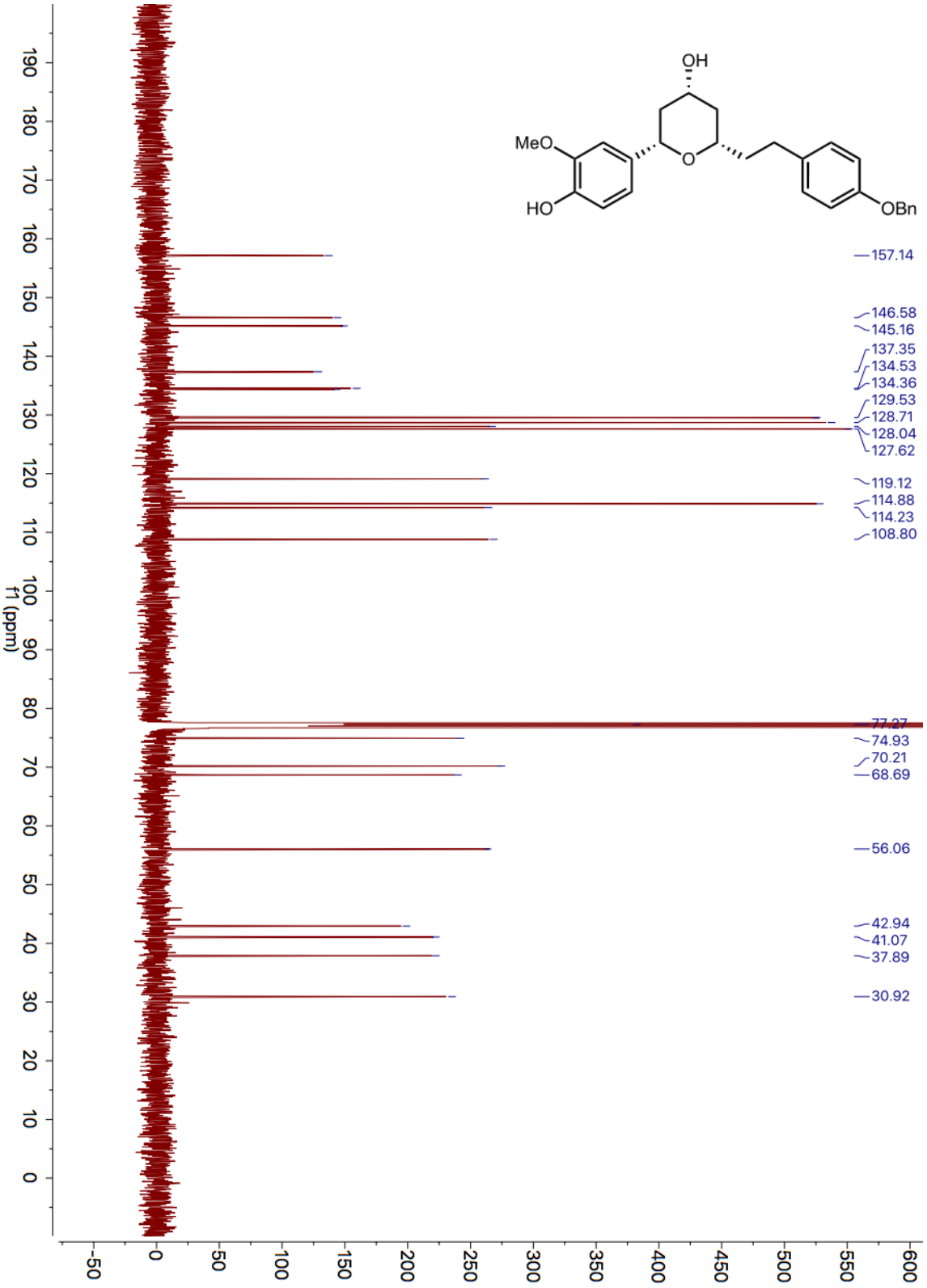


Figure S128. ^{13}C NMR spectrum of compound **SI-12**.

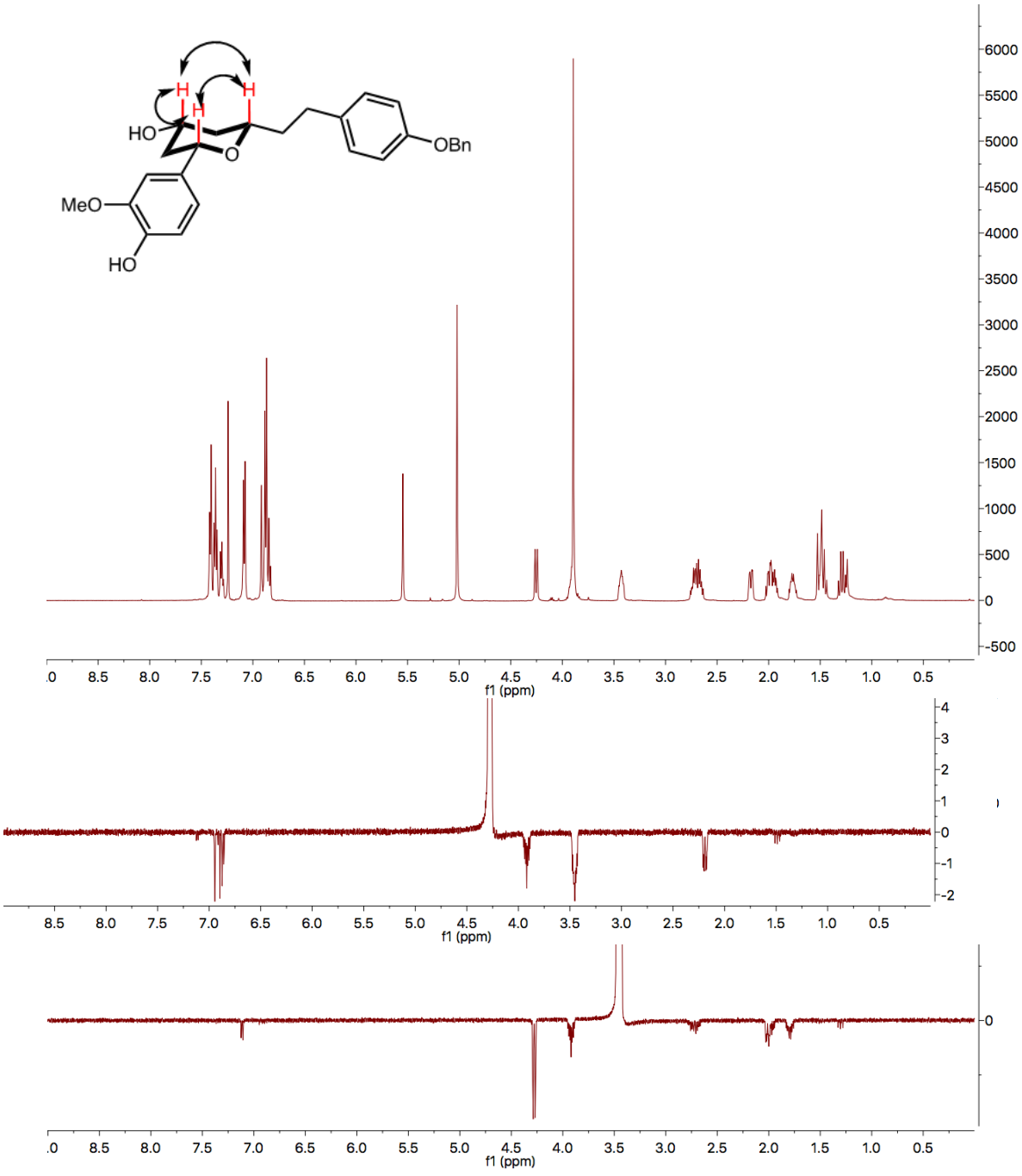


Figure S129. 1D NOE spectrum of **SI-12**.

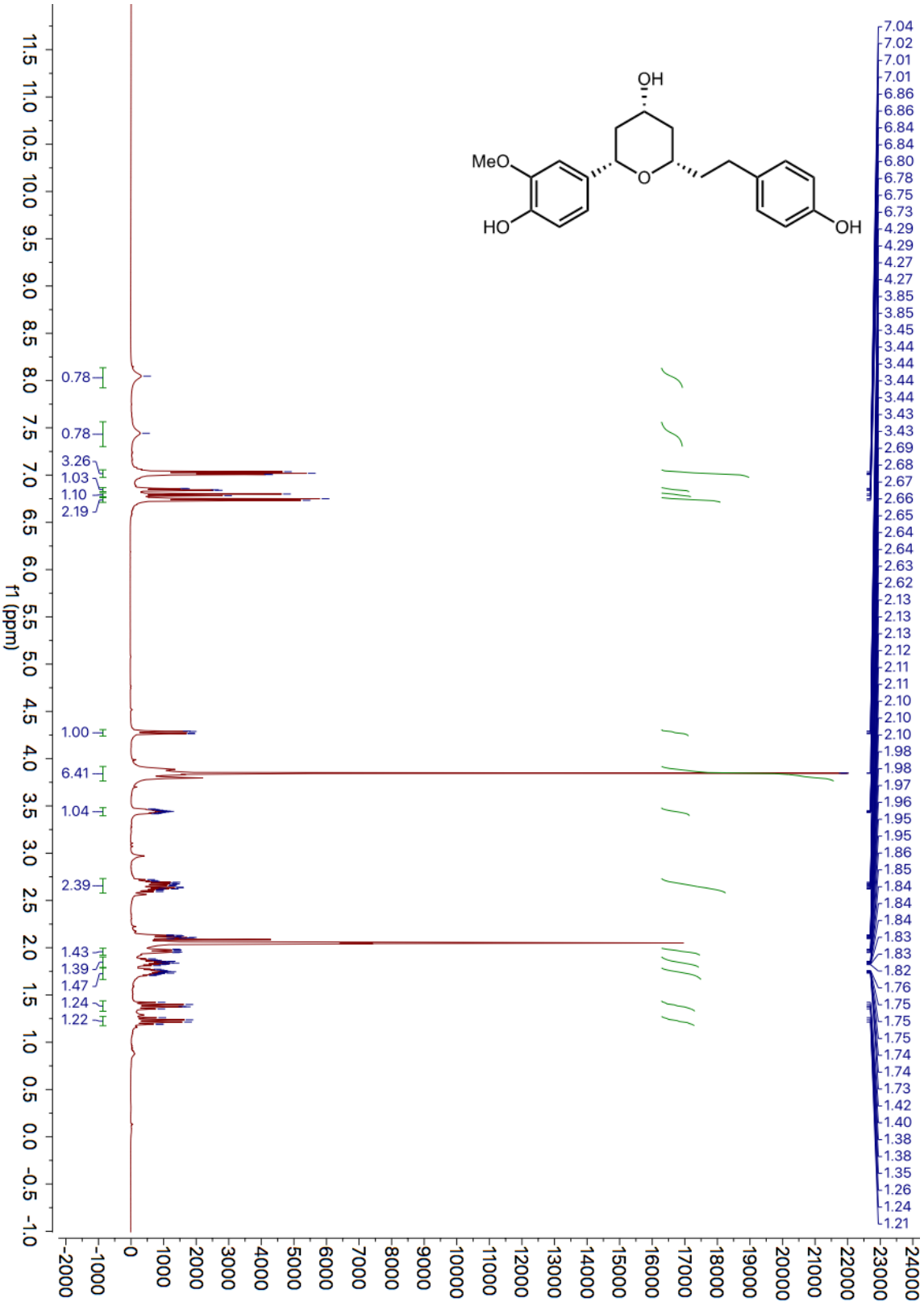


Figure S130. ^{1}H NMR spectrum of compound **56**.

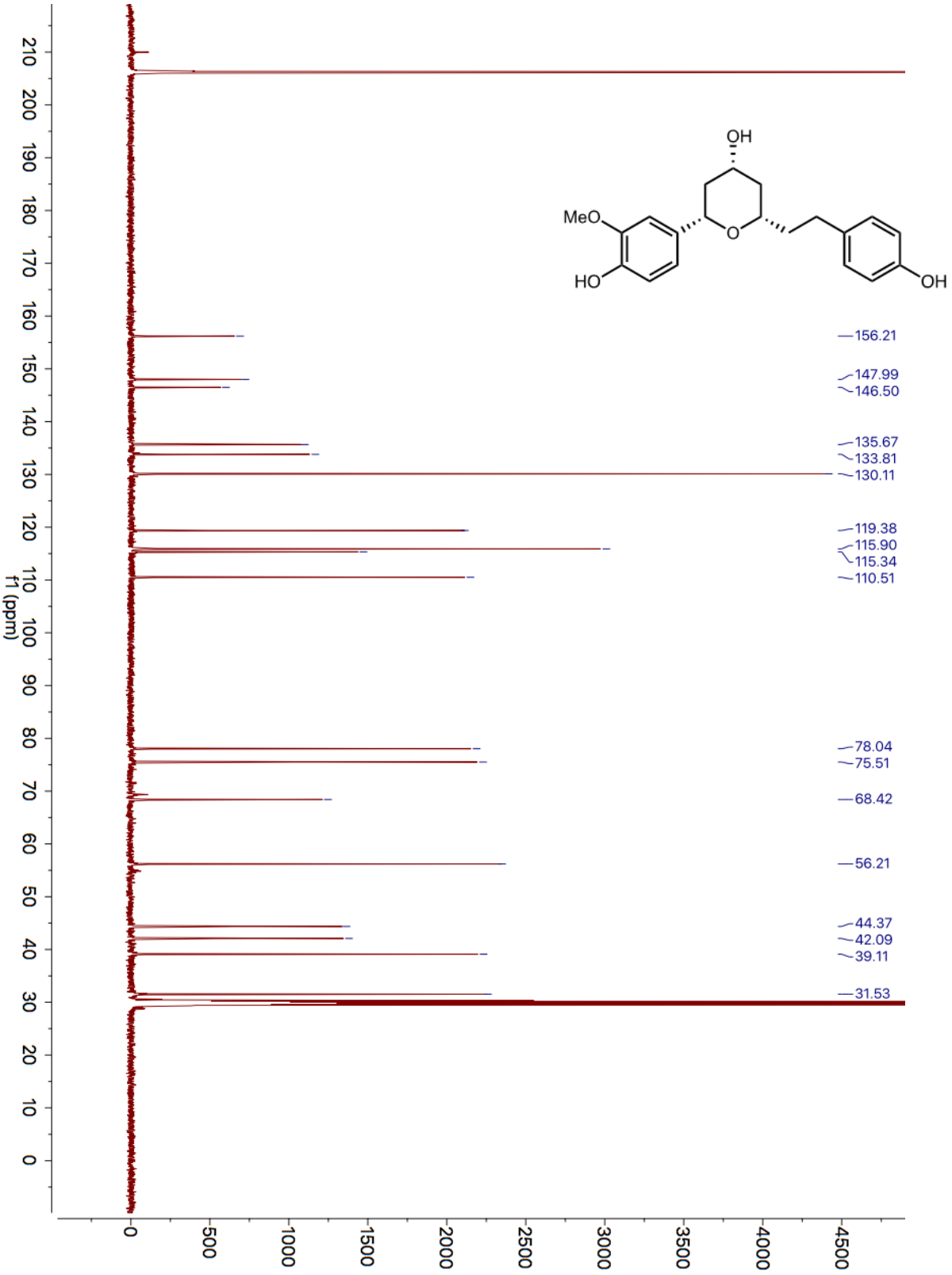


Figure S131. ^{13}C NMR spectrum of compound **56**.

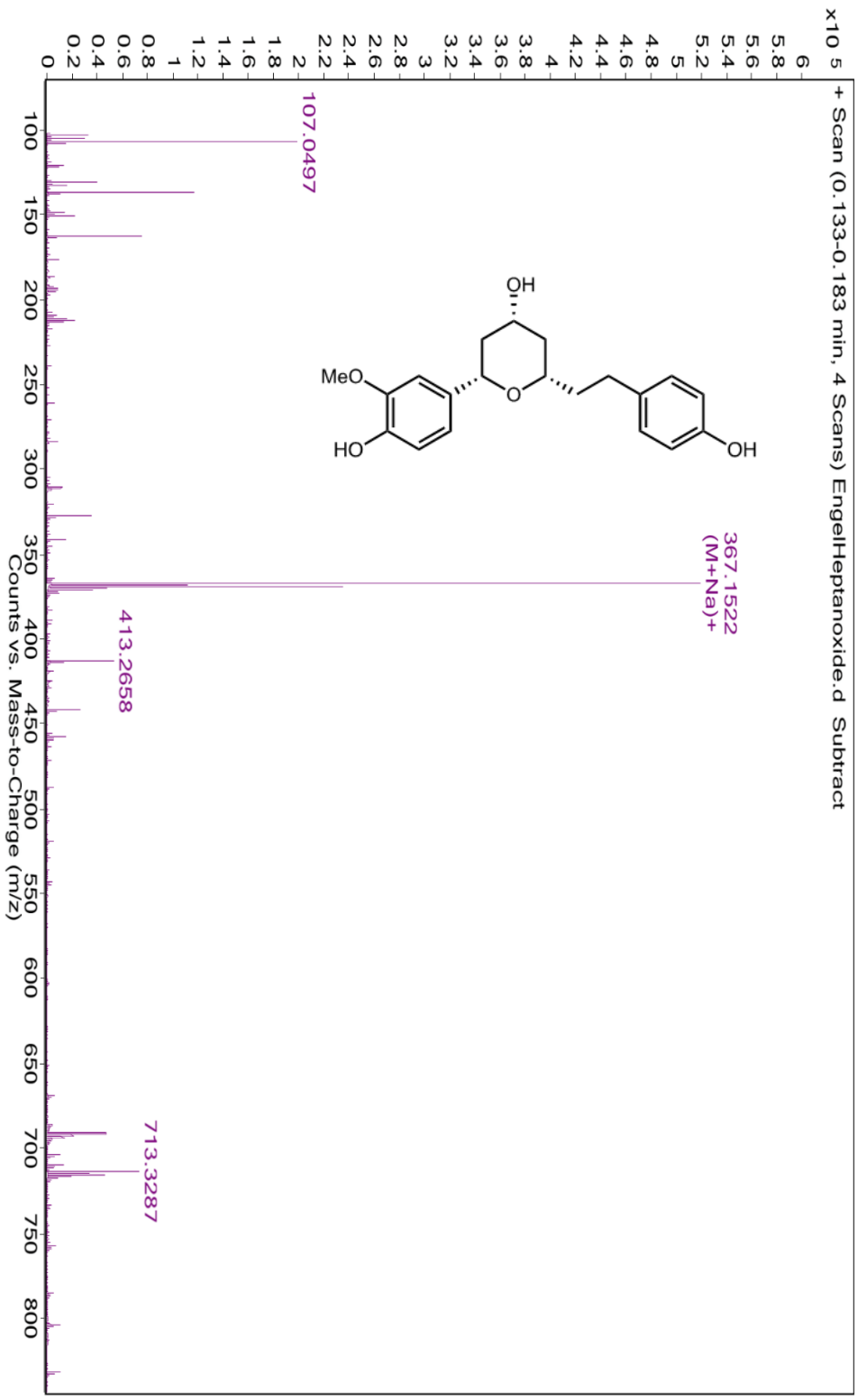


Figure S132. HRMS of compound **56**.

Section S18. Caption for Movie S1.

Movie S1. Retrosynthetic design with *Chematica*. **Part 1** focuses on a step-by-step design modality whereby the computer evaluates options at each step (via scoring functions) but it is the user who ultimately makes the choices how to navigate the synthetic "trees". The target – here, (*S*)-4-hydroxyduloxetine, same as described towards the end of the main text – is drawn at (0:03) and the "first generation" of possible precursors is returned at (0:09). Color coding of nodes: violet = unknown compounds, green = compounds already made and described in literature, red = commercially available chemicals, blue halos = protection is required to carry out the specific reaction. After displaying in a list view (0:10) and sorting with preference for cutting into equally-sized synthons (0:12), the user expands the second-generation options for the chiral alcohol (0:13). From these second-generation options, sorted according to how many stereogenic centers are created (0:20), the user chooses the ketone intermediate (0:21) for which Mannich reaction is then suggested as a preferred method of preparation (0:30).

Of course, such step-by-step searches can be very time consuming and largely rely on the user's expertise. Accordingly, they are more on the "educational" side of Chematica while the program's real power manifests itself in the fully automated modality illustrated in **Part 2.** After selecting/drawing the target (0:40), the user chooses scoring functions (here, from a predefined menu) and specifies the stop conditions (MWs, prices, popularities of the starting materials to be reached by the search) (0:41-0:44). After few minutes, complete pathways are returned. The two top-scoring paths – based on Buchwald-Hartwig and Mitsunobu chemistries – are displayed/scrutinized in detail (0:57-1:16); these two routes were also studied experimentally, as discussed in the main text and illustrated in the movie (1:19-1:23). In addition, there are many other viable pathways with lower scores (1:24-1:31). The user can display the prices and popularities of the individual molecules (numbers displayed over the nodes from 1:17 onwards). He/she also does not have to leave the Chematica environment to perform many other types of analyses, like the conformational analysis shown from 1:32-1:38.

Part 3, starting at 1:40, provides another example – not yet verified experimentally – of fully automated design of syntheses leading to imperanene, a natural product isolated from *Imperata Cylindrica* and used in traditional Chinese medicine as an anti-inflammatory and diuretic agent (*J. Nat. Prod.* **58**, 138-139, 1995). Typical syntheses of this target involve 8-13 steps (e.g., *Org. Lett*. **3**, 3021-3023, 2001). Chematica returns first pathways after ca. 70 iterations (ca. 2 min of real time and 1:51 of the movie). Details of the top-scoring pathway are displayed from 1:53 to 2:11. This five-step (including protection-deprotection; note a blue halo on the starting substrate) pathway is interesting since installation of the stereocenter is based on modern Krische methodology (*J. Am. Chem. Soc.* **136**, 8911-8914, 2014) and involves a somewhat counterintuitive removal of benzylic alcohol which leads to an intermediate participating in olefin metathesis to give imperanene in one step.

Section S19. Supplemental References

S1. Enders, D., and Schüßeler, T. (2002). Asymmetric synthesis of all stereoisomers of 7,11-dimethylheptadecane and 7-methylheptadecane, the female pheromone components of the spring hemlock looper and the pitch pine looper. Tetrahedron Lett. *43*, 3467–3470.

S2. Nicolaou, K.C., Sarabia, F., Ninkovic, S., and Yang, Z. (1997). Total synthesis of Epothilone A: The macrolactonization approach. Angew. Chem. Int. Ed. *36*, 525–527.

S3. Schkeryantz, J.M., and Danishefsky, S.J. (1995). Total synthesis of (+/-)-FR-900482. J. Am. Chem. Soc. *117*, 4722–4723.

S4. Huang, X., and Zhou, H. (2002). Novel tunable CuX_2-mediated cyclization reaction of cyclopropylideneacetic acids and esters for the facile synthesis of 4-Halomethyl-2(5H)-furanones and 4-Halo-5,6-dihydro-2H-pyran-2-ones. Org. Lett. *4*, 4419–4422.

S5. Angle, S.R., Fevig, J.M., Knight, S.D., Marquis, R.W., and Overman, L.E. (1993). Synthesis applications of cationic aza-Cope rearrangements. 24. The aza-Cope-Mannich approach to Strychnos alkaloids. Short stereocontrolled total syntheses of (+/-)-dehydrotubifoline and (+/-)-akuammicine. J. Am. Chem. Soc. *115*, 3966–3976.

S6. Shair, M.D., Yoon, T.Y., Mosny, K.K., Chou, T.C., and Danishefsky, S.J. (1996). The total synthesis of Dynemicin A leading to development of a fully contained bioreductively activated enediyne prodrug. J. Am. Chem. Soc. *118*, 9509–9525.

S7. Skoraczyński, G., Dittwald, P., Miasojedow, B., Szymkuć, S., Gajewska, E.P., Grzybowski, B.A., and Gambin, A. (2017). Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? Sci. Rep. *7*, #3582.

S8. Charest, M.G., Lerner, Ch.D., Brubaker, J.D., Siegel, D.R., and Myers, A.G. (2005). A convergent enantioselective route to structurally diverse 6-deoxytetracycline antibiotics. Science *308*, 395–398.

S9. Baran, P.S., and Richter, J.M. (2004). Direct coupling of indoles with carbonyl compounds: Short, enantioselective, gram-scale synthetic entry into the Hapalindole and Fischerindole alkaloid families. J. Am. Chem. Soc. *126*, 7450–7451.

S10. Wei, J.N., Duvenaud, D., and Aspuru-Guzik, A. (2016). Neural networks for the prediction of organic chemistry reactions. ACS Cent. Sci. *2*, 725–732.

S11. Segler, M.H.S., and Waller, M.P. (2017). Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chem. Eur. J. *23*, 5966–5971.

S12. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. *28*, 31–36.

S13. SMARTS theory manual, Daylight Chemical Information Systems Inc., Aliso Viejo, CA92656, USA. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, accessed 8rd November 2017.

S14. http://www.rdkit.org, accessed 8th November 2017.

S15. Jiang, M., Zhu, F., Xiang, H., Xu, X., Deng, L., and Yang, C. (2015). An efficient and practical approach to trifluoromethylthiolation of α-haloketones/α-haloarylmethanes. Org. Biomol. Chem. *13*, 6935–6939.

S16. Kong, D., Jiang, Z., Xin, S., Bai, Z., Yuan, Y., and Weng, Z. (2013). Room temperature nucleophilic trifluoromethylthiolation of benzyl bromides with (bpy)Cu(SCF3). Tetrahedron *69*, 6046–6050.

S17. Baldwin, J.E. (1976). Rules for ring closure. J. Chem. Soc. Chem. Commun. 734–736.

S18. Ishizaki, M., and Hoshino, O. (1994). Chiral pyridyl alcohol-promoted highly enantioselective and rapid addition of dialkylzinc to pyridinecarboxaldehydes. Chem. Lett. *23*, 1337–1340.

S19. Perron, Q., and Alexakis, A. (2007). Synthesis and application of a new pseudo C2-symmetric tertiary diamine for the enantioselective addition of MeLi to aromatic imines. Tetrahedron: Asymmetry *18*, 2503–2506.

S20. Ghosh, A.K., and Shevlin, M. (2004). The development of titanium enolate-based aldol reactions. In Modern Aldol Reactions, R. Mahrwald, ed. (Weinheim, Germany: Wiley-VCH Verlag GmbH), pp. 63–125.

S21. Shirai, F., and Nakai, T. (1988). A novel, double-asymmetric aldol approach to the synthesis of a 1β-methyl carbapenem antibiotic precursor. Tetrahedron Lett. *29*, 6461–6463.

S22. Shirai, F., Gu, J.-H., and Nakai, T. (1990). Diastereofacial selection in titanium tetrachloride-promoted aldol reactions with the silyl ketene acetal of methyl (R)-3-hydroxybutanoate. Chem. Lett. *19*, 1931–1934.
S23. Wuts, P.G.M. (2014). Reactivities, reagents, and reactivity charts. In Greene's Protective Groups in Organic Synthesis, P.G.M. Wuts, ed. (Hoboken, New Jersey: John Wiley & Sons, Inc.), pp. 1263–1309.
S24. Nicolaou, K.C., Simmons, N.L., Chen, J.S., Haste, N.M., and Nizet, V. (2011). Total synthesis and biological evaluation of marinopyrrole A and analogs. Tetrahedron Lett. *52*, 2041–2043.
S25. Hirao, H., and Ohwada, T. (2003). Theoretical study of reactivities in electrophilic aromatic substitution reactions: Reactive hybrid orbital analysis. J. Phys. Chem. A *107*, 2875–2881.
S26. Brown, J.J., and Cockroft, S.L. (2013). Aromatic reactivity revealed: beyond resonance theory and frontier orbitals. Chem. Sci. *4*, 1772-1780.
S27. Bader, R.F.W., and Chang, C. (1989). Properties of atoms in molecules: electrophilic aromatic substitution. J. Phys. Chem. *93*, 2946–2956.
S28. Galabov, B., Ilieva, S., Koleva, G., Allen, W.D., Schaefer III, H.F., and Schleyer, P. von R. (2013). Structure-reactivity relationships for aromatic molecules: electrostatic potentials at nuclei and electrophile affinity indices. Wiley Interdiscip. Rev. Comput. Mol. Sci. *3*, 37–55.
S29. Politzer, P., Abrahmsen, L., and Sjoberg, P. (1984). Effects of amino and nitro substituents upon the electrostatic potential of an aromatic ring. J. Am. Chem. Soc. *106*, 855–860.
S30. Koleva, G., Galabov, B., Wu, J.I., Schaefer III, H.F., and Schleyer, P. von R. (2009). Electrophile affinity: A reactivity measure for aromatic substitution. J. Am. Chem. Soc. *131*, 14722–14727.
S31. Hückel, E. (1931). Quantentheoretische beiträge zum benzolproblem. Z. Phys. *70*, 204–286.
S32. Van-Catledge, F.A. (1980). A Pariser-Parr-Pople-based set of Hückel molecular orbital parameters. J. Org. Chem. *45*, 4801–4802.
S33. Kruszyk, M., Jessing, M., Kristensen, J.L., and Jørgensen, M. (2016). Computational methods to predict the regioselectivity of electrophilic aromatic substitution reactions of heteroaromatic systems. J. Org. Chem. *81*, 5128–5134.
S34. Hansch, C., Leo, A., and Taft, R.W. (1991). A survey of Hammett substituent constants and resonance and field parameters. Chem. Rev. *91*, 165–195.
S35. Hammett, L.P. (1937). The Effect of structure upon the reactions of organic compounds. Benzene Derivatives. J. Am. Chem. Soc. *59*, 96–103.
S36. Sana, S., Tasneem, Ali, M.M., Rajanna, K.C., and Saiprakash, P.K. (2009). Efficient and facile method for the nitration of aromatic compounds by nitric acid in micellar media. Synth. Commun. *39*, 2949–2953.
S37. Njoroge, F.G., Vibulbhan, B., Pinto, P., Chan, T.-M., Osterman, R., Remiszewski, S., Del Rosario, J., Doll, R., Girijavallabhan, V., and Ganguly, A.K. (1998). Highly regioselective nitration reactions provide a versatile method of functionalizing benzocycloheptapyridine tricyclic ring systems: Application toward preparation of nanomolar inhibitors of farnesyl protein transferase. J. Org. Chem. *63*, 445–451.
S38. Ramana, M.M.V., Malik, S.S., and Parihar, J.A. (2004). Guanidinium nitrate: a novel reagent for aryl nitrations. Tetrahedron Lett. *45*, 8681–8683.
S39. Cruz, R.P.A., Ottoni, O., Abella, C.A.M, and Aquino, L.B. (2001). Regioselective acylations at the 2 and 6 position of N-acetylindole. Tetrahedron Lett. *42*, 1467–1469.
S40. Boruah, J.J., Das, S.P., Borah, R., Gogoi, S.R., and Islam, N.S. (2013). Polymer-anchored peroxo compounds of molybdenum and tungsten as efficient and versatile catalysts for mild oxidative bromination. Polyhedron *52*, 246–254.
S41. Baharfar, R., Alinezhad, H., Azimi, S., and Salehian, F. (2011). Regioselective and high-yielding bromination of phenols and anilins using N-bromosaccharin and Amberlyst-15. J. Chil. Chem. Soc. *56*, 863–865.
S42. Liang, D., Luo, H., Liu, Y.-F., Hao, Z.-Y., Wang, Y., Zhang, C.-L., Zhang, Q.-J., Chen, R.-Y., and Yu, D.-Q. (2013). Lysilactones A–C, three 6H-dibenzo[*b*,*d*]pyran-6-one glycosides from Lysimachia clethroides, total synthesis of Lysilactone A. Tetrahedron *69*, 2093–2097.
S43. Li, L., Qiu, D., Shi, J., and Li, Y. (2016). Vicinal diamination of arenes with domino aryne precursors. Org. Lett. *18*, 3726–3729.
S44. Paul, V., Sudalai, A., Daniel, T., and Srinivasan, K.V. (1994). Regioselective bromination of activated aromatic substrates with N-bromosuccinimide over HZSM-5. Tetrahedron Lett. *35*, 7055–7056.

S45. Kozic, J., Novák, Z., Římal, V., Profant, V., Kuneš, J., and Vinšová, J. (2016). Conformations, equilibrium thermodynamics and rotational barriers of secondary thiobenzanilides. Tetrahedron *72*, 2072–2083.

S46. Liu, J., Jiang, F., Jiang, X., Zhang, W., Liu, J., Liu, W., and Fu, L. (2012). Synthesis and antimicrobial evaluation of 3-methanone-6-substituted-benzofuran derivatives. Eur. J. Med. Chem. *54*, 879–886.

S47. Ducrot, P.-H. (1996). Efficient synthesis of Sordidin, a male pheromone compound emitted by cosmopolites Sordidus. Synth. Commun. *26*, 3923–3928.

S48. Gothard, C.M., Soh, S., Gothard, N.A., Kowalczyk, B., Wei, Y., Baytekin, B., and Grzybowski, B.A. (2012). Rewiring chemistry: Algorithmic discovery and experimental validation of one-pot reactions in the Network of Organic Chemistry. Angew. Chem. Int. Ed. *51*, 7922–7927.

S49. Hart, P., Nilsson, N., and Raphael, B. (1968). A Formal basis for the heuristic determination of minimum cost paths. IEEE Trans. Syst. Sci. Cybern. *4*, 100–107.

S50. Dijkstra, E.W. (1959). A note on two problems in connexion with graphs. Numer. Math. *1*, 269–271.

S51. Hopcroft, J., and Tarjan, R. (1973). Algorithm 447: efficient algorithms for graph manipulation. Commun. ACM *16*, 372–378.

S52. Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001) Binomial heaps. In Introduction to Algorithms, (MIT Press and McGraw-Hill, Cambridge, MA), pp. 455–475

S53. Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001) The Bellman-Ford algorithm. In Introduction to Algorithms, (MIT Press and McGraw-Hill, Cambridge, MA), pp. 588–592

S54. Cakmak, M., Mayer, P., and Trauner, D. (2011). An efficient synthesis of loline alkaloids. Nat. Chem. *3*, 543–545.

S55. Higashiyama, K., Nakahata, K., and Takahashi, H. (1994). Asymmetric synthesis of (+)-monomorine I by way of a diastereoselective reaction of 1,3-oxazolidine with a Grignard reagent. J. Chem. Soc. Perkin Trans. 1, 351-353.

S56. Ritter, F.J., Rotgans, I.E.M., Talman, E., Verwiel, P.E.J., and Stein, F. (1973). 5-methyl-3-butyl-octahydroindolizine, a novel type of pheromone attractive to Pharaoh's ants (Monomorium pharaonis (L.). Experientia *29*, 530–531.

S57. Risinger, A.L., Peng, J., Rohena, C.C., Aguilar, H.R., Frantz, D.E., and Mooberry, S.L. (2013). The Bat Flower: A source of microtubule-destabilizing and - stabilizing compounds with synergistic antiproliferative actions. J. Nat. Prod. *76*, 1923–1929.

S58. Kuranaga, T., Shirai, T., Baden, D.G., Wright, J.L.C., Satake, M., and Tachibana, K. (2009). Total synthesis and structural confirmation of Brevisamide, a new marine cyclic ether alkaloid from the dinoflagellate Karenia brevis. Org. Lett. *11*, 217–220.

S59. Li, Q., Mao, S., Cui, Y., and Jia, Y. (2012). Stereoselective synthesis of the C_5–C_{18} fragment of Halichomycin. J. Org. Chem. *77*, 4111–4116.

VIP Chemical Networks Very Important Paper

International Edition: DOI: 10.1002/anie.201712052
German Edition: DOI: 10.1002/ange.201712052

Discovery and Enumeration of Organic-Chemical and Biomimetic Reaction Cycles within the Network of Chemistry

*Michał D. Bajczyk[+], Piotr Dittwald[+], Agnieszka Wołos, Sara Szymkuć, and Bartosz A. Grzybowski**

Supplementary Information

CONTENTS:

Section S1. Additional algorithmic details and cycle statistics.

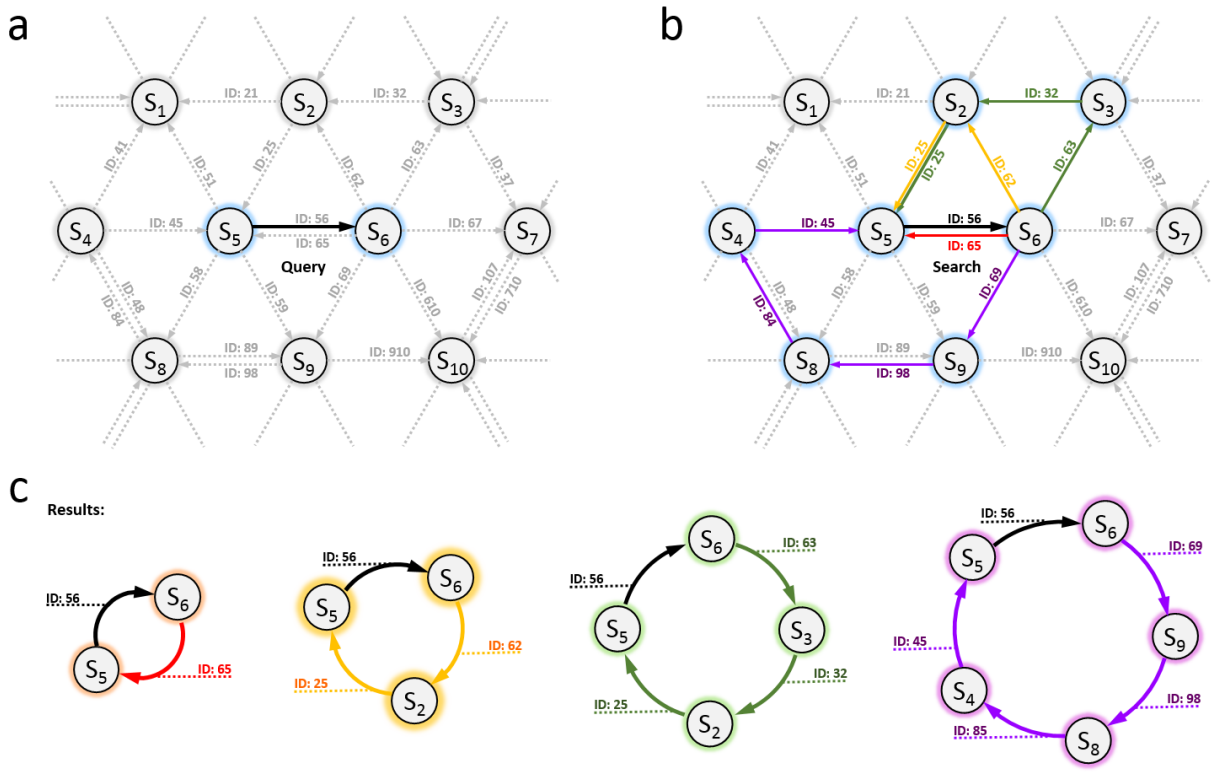


Figure S1. Scheme illustrating the algorithm used to identify reaction cycles within the NOC. **a)** Algorithm selects reaction $S_5 \rightarrow S_6$ (black arrow) linking substrate S_5 with product S_6. . **b)** Each "backward" path from S_6 to S_5 of desired length (here, $L - 1 = 1,2,3,4$) is identified using standard depth-limited searches. **c)** Together with initial reaction $S_5 \rightarrow S_6$ these paths create cycles of lengths $L = 2,3,4,5$. Algorithm is repeated for each possible pair S_i and S_j in the NOC.

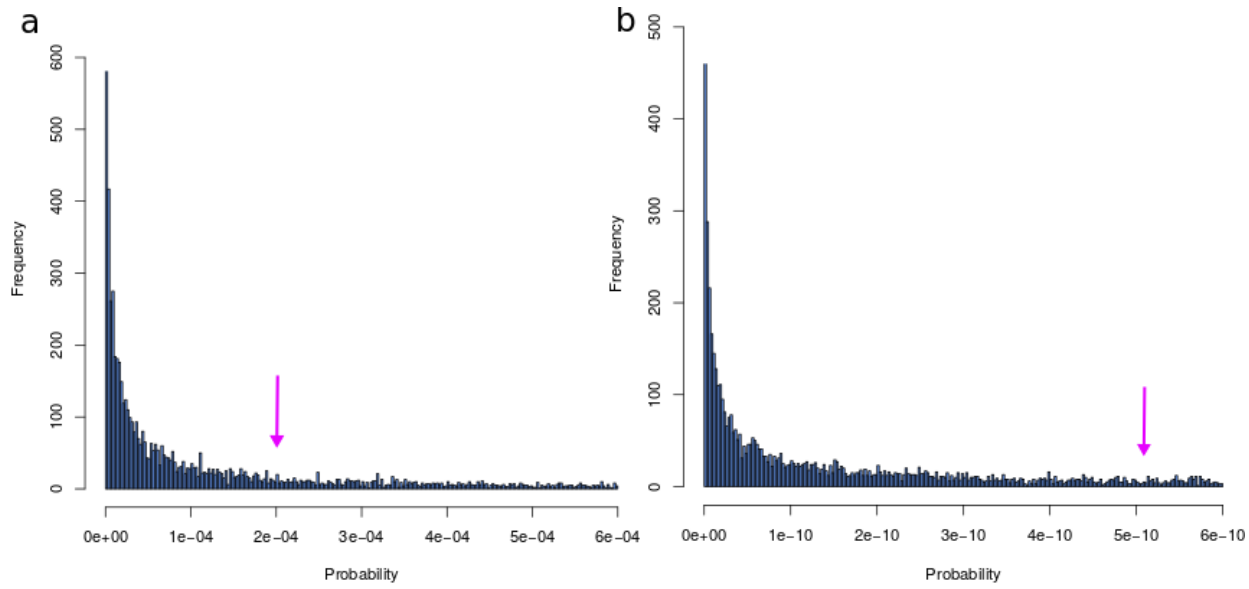


Figure S2. Distribution of probabilities of finding a desired cycle. The question asked was how probable would be a human user to discover a given (i.e., known to exists) cycle without *Cyclorg*, by navigating the NOC manually. Say, the cycle of interest is L = 3 steps long and involves sequence #1 → #2 → #3 → #1, where #1,#2,#3 are molecules at the cycle's nodes. As in the algorithm described in Figure S1, the user chooses the first reaction in the cycle (say, #1 → #2). From #2 which there are n_2 outgoing reactions. Only one of these reactions leads to the third molecule in the cycle, #3 – therefore, the probability of user choosing this reaction in $1/n_2$. Then, for molecule #3, there are n_3 outgoing reactions and the probability of choosing the one closing the cycle to #1 is $1/n_3$. The overall probability of finding the cycle is then the product of individual probabilities, $P = 1/(n_2 n_3)$. In general, the chance of finding a cycle of length L will be the product of $1/n_i$'s for each of the nodes involved (except for the first one, sine we chose our first reaction of interest). The plots give the distributions of such probabilities for (**a**) L =3 and (**b**) L = 5 cycles. The statistics are based on 10,000 randomly chosen main-substrate/main-product cycles. The distributions are very heavy tailed and in the graphs are truncated at $6 \cdot 10^{-4}$ for L = 3 and $6 \cdot 10^{-10}$ for L = 5. The median probabilities are indicated by pink arrows and are $\sim 2 \cdot 10^{-4}$ for L =3 and $\sim 5 \cdot 10^{-10}$ for L = 5. Of course, our analysis here assumes that the molecules in the first reaction we start from belong to a cycle – in reality, most molecules are not parts of any cycle and for such starting points the probabilities would be zero. On the other hand, for molecules belonging to many cycles, the probabilities we calculate do not consider a possibility of finding cycles other than the

specific one we are looking for. Still, the very low values of *P* tell us that manual searches are definitely not an efficient method of identifying cyclic reaction sequences.

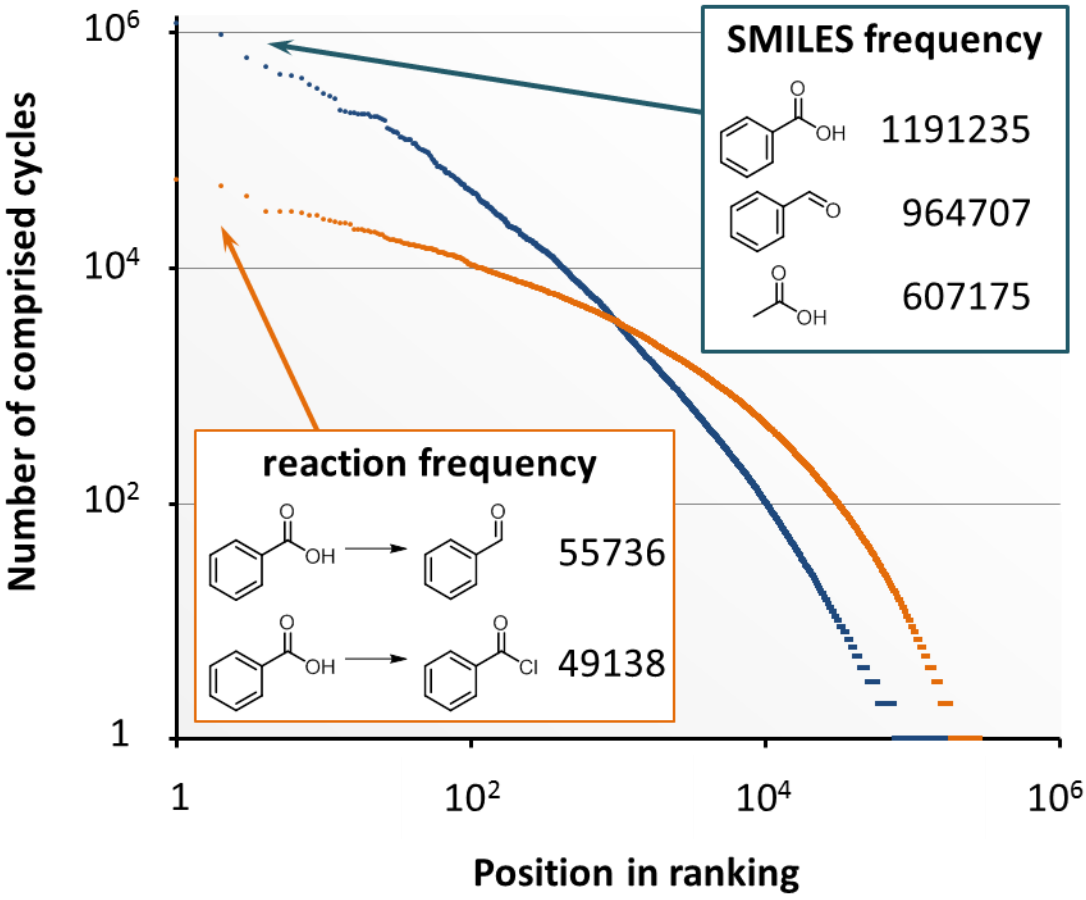


Figure S3. Frequency of occurrence of specific molecules and reactions in reaction cycles. The plot ranks 165,000 molecules (blue curve) and 278,000 reactions (orange curve) according to the numbers of main-substrate/main-product cycles in which they participate. The most popular molecules/reactions are simple ones. For example, the most popular molecule is benzoic acid (rank #1, participating in 1,191,235 cycles), followed by benzaldehyde (rank #2, participating in 964,707 cycles), and acetic acid (rank #3, participating in 607,175 cycles). Among the reactions, reduction of benzoic acid to benzaldehyde is ranked #1 and participates in 55,736 cycles followed by synthesis of benzoic acid chloride (rank #2, seen in 49,138 cycles). Note that the plot is doubly logarithmic and the distributions are heavily tailed (though not pure power-law) meaning that there are also many cycles in which unique molecules/reactions participate.

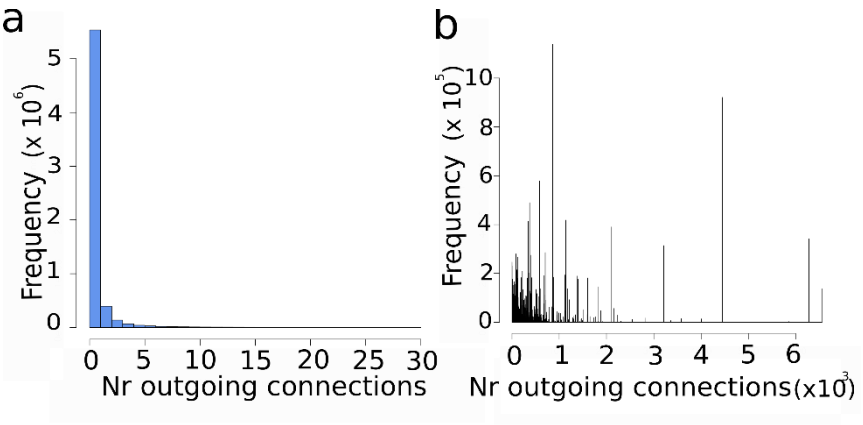


Figure S4. "Outgoing" connectivities of molecules in the entire NOC and in the cycles. a) The average number of outgoing connections of the molecules in the NOC (main product/main substrate considered) is low (median ≈ 1). The tail of the presented histogram is truncated for clarity. **b)** Distribution of the outgoing connectivity of molecules found in cycles of length 5 (frequencies are weighted by the times a given molecule is found in the cycles, cf. Figure S3). Note the horizontal scale is in thousands. The median expected connectivity is ca. 250. With this average number, the chance of closing a cycle of length 5 (assuming first reaction is chosen by the user and the remaining four are navigated "randomly", see Figure S1) can be estimated as $(250)^{-4} \sim 2.56 \cdot 10^{-10}$, which is the same order of magnitude as the value based on counting specific cycles in Figure S2b.

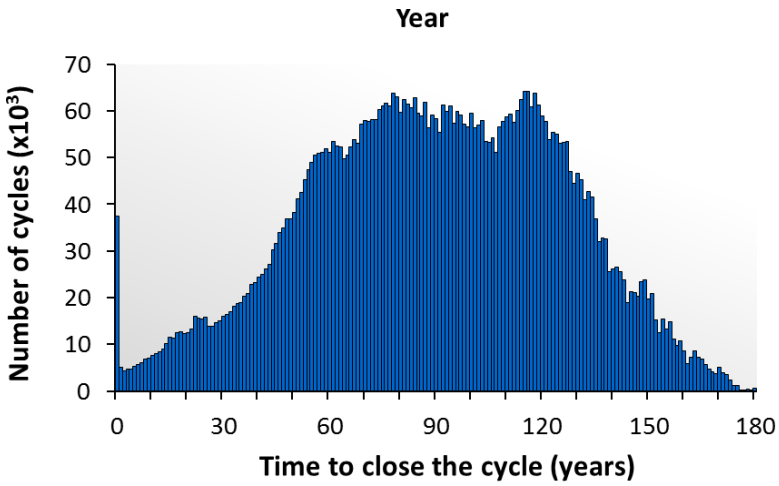


Figure S5. Distribution of times it takes to close a main-substrate/main-product cycle – that is, from the date the first reaction in the cycle was published to the time the last reaction, completing the cycle, was reported. Cycles take from one to 160 year to close, with average closure time of ~90 years.

Section S2. Using *Cyclorg* – a short tutorial.

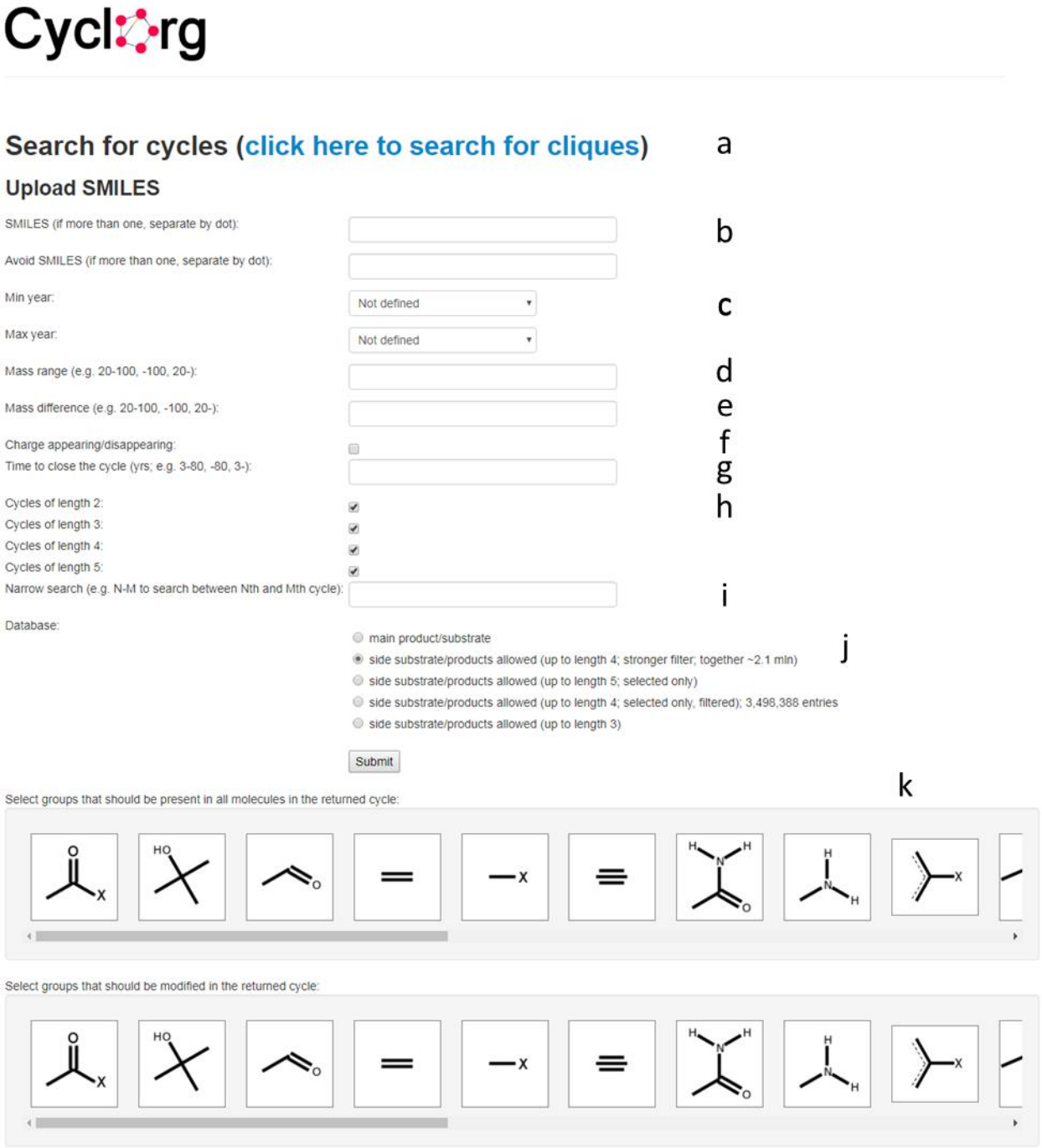


Figure S6. ***Cyclorg*'s main page**. Choices in the different input fields are as follows:
(a) = Specify whether you wish to search for cliques or cycles
(b) = Input the SMILES of one or more molecules that must be present in the cycle. Input the SMILES of one or more molecules that are to be avoided (i.e., cannot be present) in the cycle. If more than one molecule is entered, the SMILES need to be separated by dots. Input "[*]" means that any molecules are allowed in the cycle.
(c) = *Cyclorg* will search only for cycles involving reactions reported between "Min year" and "Max year".
(d) = Specify the lowest and the highest molecular weights of molecules in the cycle. "-100" means ,"100 or less"; "20-" means ,"20 or more".
(e) = Specify the minimum and maximum allowed difference between the masses of the lightest and the heaviest molecules in the cycle. For instance, specifying 10-200 means that all molecules in cycle's nodes will have MWs between 20 and 100. Notation "-100" means that all cycles with difference of masses between heaviest and lightest molecules of 100 or less will be shown. Notation "20-" means that only cycles with mass difference greater or equal than 20 will be shown.
(f) = Activate this option to limit searches to cycles in which charged species are created and then used (creation of charged species can be important in surface phenomena – e.g., in cycles powering rhythmic assembly disassembly of various species (see, for example, *Angew. Chem. Int. Ed.* **2010**, *49*, 8616-8619 or *Synlett* **2017**, *28*, 103-017).
(g) = Time that elapsed between publication of the earliest and the latest reactions in the cycle (see main text, Figure 1c). *Note*: This option is not available when searching for cliques.
(h) = Specify the length of cycles of interest (2-5) or sizes of cliques (3-8).
(i) = Search by numbers assigned to specific cycles in our cycle collection (starting from 0). *Cyclorg* displays first 1,000 cycles it finds during each search. Sometimes there are many more to display – in such cases, one can narrow the range of cycle numbers and perform the search for each such subset separately. In this way, *Cyclorg* will return 1000 cycles for each range queried. Note: this option is not available for clique searches.
(j) = Selection of cycle "databases". The user can chose either the cycles in which only the main/largest substrates and products of each reaction are retained or with this criterion relaxed (i.e., with cycles involving minority/small substrates/products of each reaction). The latter option will produce many useless cycles, but it will also allow finding cycles in which cycle's products are large and useful molecules. The numbers of cycles allowing for smaller reaction products are astronomical and the searches to identify them are ongoing (i.e., these databases are continuously being updated beyond current 18 million entries).
(k) = The panels list substructures the user would like to either (1) be present in every node of the cycle or (2) be modified at least once throughout the cycle.

Section S3. Examples of additional cycles.

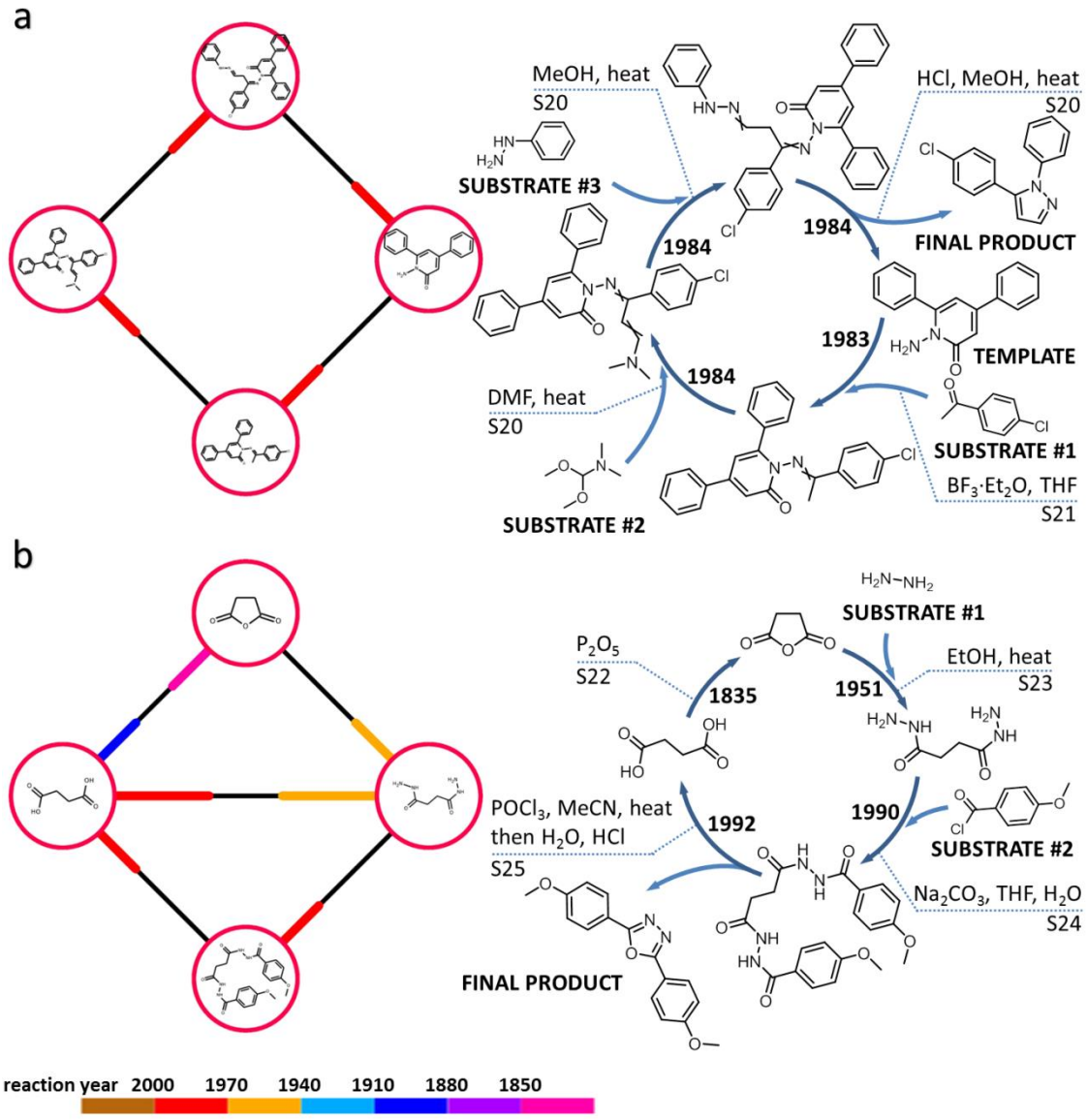


Figure S7. Closing reaction cycles over one year and over more than a century. Schemes on the left are raw outputs from *Cyclorg* (see **Movie S1**). Colored endings of the connections specify a 30-year period in which particular reactions were published (see color legend). If two ends are colored, it means that reactions in both directions are known. Note that in addition to full cycles, *Cyclorg* also displays "inner shortcuts" (e.g., inner arrows in the four-membered cycle in (b)). Schemes on the right elaborate on reaction details and provide literature references S# (see Section S5 below). **a)** In 1983, P. Molina's group published (*J. Heterocyclic Chem.* **1983**, *20*, 381-384) reaction between 1-amino-4,6-diphenyl-2-pyridone and methyl(p-chlorophenyl)ketone as the first step in the synthesis of pyrido-1,3,4-oxadiazine derivatives. Just one year later (*J. Heterocyclic Chem.* **1984**, *21*, 461-464), the same group used the product of this reaction as a starting material

for the syntheses of pyrazoles and isoxazoles. This cycle could potentially be operated continuously one-pot by unifying solvent to methanol and changing Lewis acid to a Bronsted acid. **b)** A cycle that took 157 years to complete. This cycle was opened in 1835 with the report (*Ann. Chim. Phys.* **1835**, *58*, 282-300) of synthesis of succinic anhydride. It was closed only in 1992 with a publication (*Liebigs Ann.* **1992**, *3*, 291-292) describing rearrangement–producing 2,5-diaryl-1,3,4-oxadiazoles and regenerating succinic acid. The cycle could be operated one pot (in acetonitrile and changing phosphorous pentoxide to DCC) but only once due to the water quench in the last step. The diaryl-oxadiazole scaffold produced by this cycle is of recent interest in molecular electronics for its electron-transporting and hole-blocking properties (*Org. Lett.* **2009**, *11*, 3072-3075; *Chem. Rev.* **2008**, *108*, 1245-1330).

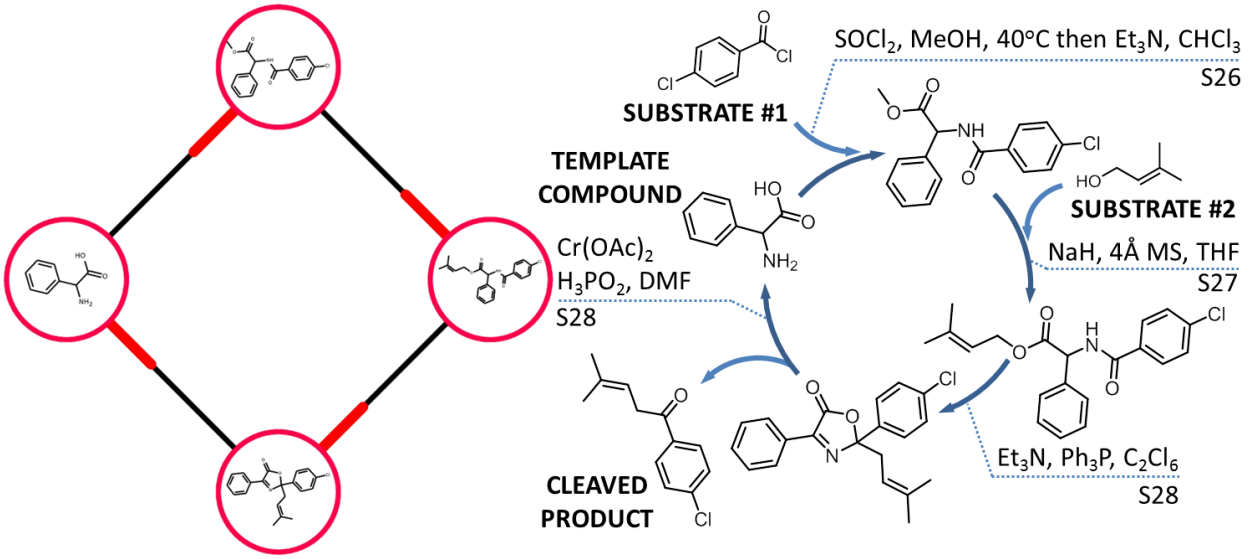


Figure S8. Example of buildup-rearrangement-cleavage tandem cycle. In many cycles, the reactions building up mass are followed by a rearrangement and then cleavage/release of cycles "product". In this example, 4-chloro-benzoyl chloride and 3-methyl-2-buten-1-ol substrates are successively added onto the phenylglycine "template" to give (4-chloro-benzoylamino)-phenyl-acetic acid 3-methyl-but-2-enyl ester that then rearranges (by a Claisen, then Cope types of rearrangement) into 2-(4-chlorophenyl)-2-(3,3-dimethylallyl)-4-phenyl-5(2H)-oxazolone, from which the 1-(p-chlorophenyl)-4-methyl-3-penten-1-one product is then released. This product is used in the synthesis of derivatives of cyano-featured dihydroisoxazoles known to exhibit antibacterial activity (*Org. Lett.* **2017**, *19*, 3255-3258) and of *gem*-bisprenyl-based building blocks (*Tetrahedron* **2013**, *69*, 7970-7974) of natural product analogs. The cycle can be extended to a broader scope of substrates than the one shown here; see *Tetrahedron* **1986**, *42*, 2063-2074.

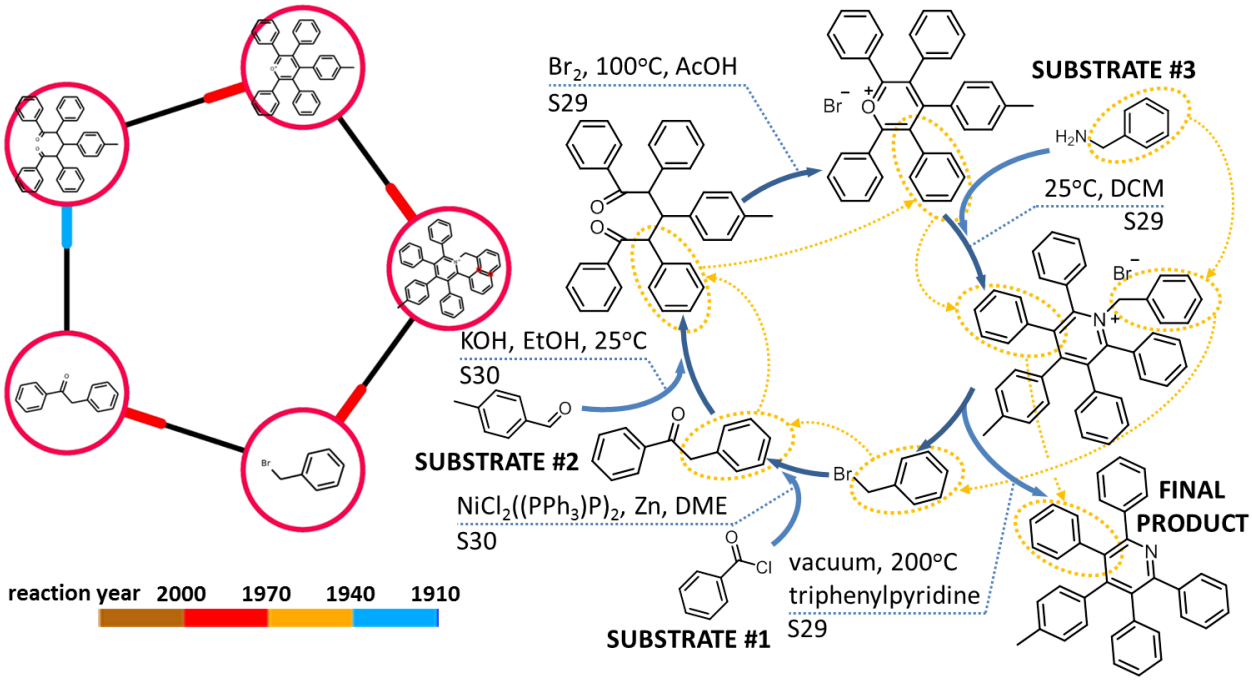


Figure S9. "Circulation" of atoms through a cycle identified by *Cyclorg* resembles the flow of matter through living systems. In this example, not a single atom stays in the cycle for longer than one full completion – all matter that enters leaves. The cycle starts with a Pd/Zn-catalyzed coupling of benzyl bromide with benzoyl chloride. Resulting ketone undergoes condensation with p-methylbenzaldehyde giving a pentaarylpentan-1,5-dione followed by a cyclization to a pyrylium bromide. Subsequently, a product from the previous step reacts with benzylamine yielding a 4-(p-tolyl)-1-benzyl-2,3,5,6-tetraphenyl pyridinium bromide which is subjected to a pyrolysis resulting in the formation of benzyl bromide (the substrate used in the cycle's first step) and 4-(p-tolyl)-2,3,5,6-tetraphenylpyridine. Blue arrows indicate reactions, while orange arrows and circles are used to highlight the movement of benzyl group – first introduced as reactant, then incorporated into main scaffold, and ultimately leaving in the cycle's product.

Section S4. Caption for Movie S1.

Movie 1. Cyclorg in action. The movie starts with the view of the graphical user interface, GUI. At **00:11**, the choice between searches for cliques vs. cycles is made. After choosing cycle searches, SMILES of a popular chiral auxiliary (see Figure 3a) is typed as query (**00:14-00:19**; 'O=C1N[C@@H](Cc2ccccc2)CO1') and minimal year of 1970 is chosen (00:21). Cycle lengths 2 and 3 are deselected (**00:26-00:28**) to allow only cycles comprising 4 or 5 steps, "Main product/substrate" database is chosen (**00:29**; this database is denoted as (2) in the main text) and the search commences at **00:30**. Five cycles are found and one is selected for closer inspection (**00:40-01:03**; the colored endings of lines on the graph show the directions of transformations; all participating reactions are shown below the cycle graph. Each row list a different reaction, "start" column displays substrates, "end," products, "yr" gives the year of earliest report An individual *Cyclorg* id (here 3799871) is copied for later use (**01:04-01:06**).

During the second search, all molecules are allowed by typing [*] into "SMILES" search query and the previously copied id is now pasted into 'Narrow Search' as the only search filter (**01:23-01:27**; *Note*: ranges of id's can also be input). Search commences at **01:32** with the same database as before to quickly retrieve the one desired cycle.

For the third search, all molecules are allowed in cycles except for those specified in 'Avoid SMILES' (**01:44-1:53;** 'CC(O)=O.O=Cc1ccccc1.OC(=O)c1ccccc1'; *Note*: multiple molecule SMILES are separated by dots). Molecules with MW > 600 are also barred (**01:58-02:00**) and the admissible difference in the MW's of the heaviest and the lightest molecule in the cycle is specified to be at least 200 (**02:01-02:02**; *Note*: this 200+ condition allows us to estimate the masses of products leaving the cycle). Only cycles that took 100 years of longer to close are selected (**02:04-02:05**). After selecting desired cycle lengths and databases, the carboxylic acid functional group is chosen from the menu of groups that need to be transformed/changed in the cycle (**02:13-02:18**). A legend is shown together with results (**02:36-02:40**) – colors of connection's endings in cycle graphs correspond to the dates respective reactions were reported (e.g., an orange ending means that reaction was first published between 1940 and 1969; for reaction published in 1970, the color would be red).

Section S5. Literature references to the reactions in all cycles described in the main text and in the SI.

[S1] R. Balasubramanian, M. V. George, Symmetry allowed π4s+π2s additions silacyclopentadienes. *Tetrahedron* **1973**, *29*, 2395-2404.

[S2] T. J. Barton, W. F. Goure, J. L. Witiak, W. D. Wulff, Observations and comments on the thermal behavior of 7-silanorbornadienes. *J. Organomet. Chem.* **1982**,. *225*, 87-106.

[S3] H. Ardill, R. Grigg, V. Sridharan, S. Surendrakumar, S. Thianpatanagu, S Kanajun, Iminium ion route to azomethine yields from primary and secondary amines. *J. Chem. Soc., Chem. Commun.* **1986**, 602-604.

[S4] R. Grigg, J. Idle, P. McMeekin, S. Surendrakumar, D. Vipond, X=Y-ZH systems as potential 1,3-dipoles. Part 12. Mechanism of formation of azomethine ylides via the decarboxylative route from α-amino acids. *J. Chem. Soc. Perkin Trans. 1* **1988**, 2703-2714.

[S5] M. Yoshimatsu, Y. Murase, A. Itoh, Y. Tanabe, O. Muraokay, Z-selective or stereospecific alkenylation reaction: a novel synthetic method for –fluoro-unsaturated esters. *Chem. Lett.* **2005**, *34*, 998-999.

[S6] a) H. Poleshner, M. Heydenreich, U. Schilde, Reactions of RSe–EMe3 (E = Si, Ge, Sn, Pb) with XeF2–RSe–F Equivalents in the Fluoroselenenylation of Acetylenes. *Eur. J. Inorg. Chem.* **2000**, *2000*, 1307-1313; b) J. P. Light II, M. Ridenour, L. Beard, J. W. Hershberger, Reactivity of allylic and vinylic silanes, germanes, stannanes and plumbanes toward SH2′ or SH2 substitution by carbon- or heteroatom-centered free radicals. *J. Organomet. Chem.* **1987**, *326*, 17-24.

[S7] Y. Nishiyama, H. Kawamatsu, S. Funato, K. Tokunaga, N. Sonoda, Phenyl tributylstannyl selenide as a promising reagent for introducion of the phenylseleno group. *J. Org. Chem.* **2003**, *68*, 3599–3602.

[S8] M. A. Arrica, T. Wirth, Fluorinations of α-seleno carboxylic acid derivatives with hypervalent (difluoroiodo)toluene. *Eur. J. Org. Chem.* **2005**, *2005*, 395–403.

[S9] B. C. Ranu, T. Mandal, S. Samanta, Indium(I) iodide-mediated cleavage of diphenyl diselenide. An efficient one-pot procedure for the synthesis of unsymmetrical diorganyl selenides. *Org. Lett.* **2003**, *5*, 1439-1441.

[S10] a) S. Dey, A. Sudalai, A concise enantioselective synthesis of (R)-selegiline, (S)-benzphetamine and formal synthesis of (R)-sitagliptin via electrophillic azidation of chiral imide enolates. *Tetrahedron: Asymmetry* **2015**, *26*, 67-72; b) D. A. Evans, T. C. Britton,

Electrophilic azide transfer to chiral enolates. A general approach to the asymmetric synthesis of alpha.-amino acids. *J. Am. Chem. Soc* **1987**, *109*, 6881–6883.

[S11] D. A. Evans, T. C. Britton, J. A. Ellman, R. L. Dorow, The asymmetric synthesis of α-amino acids. Electrophilic azidation of chiral imide enolates, a practical approach to the synthesis of (R)- and (S)-α-azido carboxylic acids. *J. Am. Chem. Soc.* **1990**, *112*, 4011–4030.

[S12] N. R. Treweeke, P. B. Hitchcock, D. A. Pardoe, S. Caddick, Controlling diastereoselectivity in the reactions of enantiomerically pure α-bromoacyl-imidazolidinones with nitrogen nucleophiles: substitution reactions with retention or inversion of configuration. *Chem. Commun.*, **2005**, 1868-1870.

[S13] L. N. Pridgen, J. Prol, B. Alexander, L. Gillyard, Single-pot reductive conversion of amino acids to their respective 2-oxazolidinones employing trichloromethyl chloroformate as the acylating agent: a multigram synthesis. *J. Org. Chem.* **1989**, *54*, 3231–3233.

[S14] a) T. Ikeda, T. Yasunaga, Kinetic behavior of L-arginine in interlamellar layes of montmorillonite in aqueous suspension. *J. Phys. Chem.* **1984**, *88*, 1253-1257; b) W. R. Boon, W. Robson, CCCXVIII The preparation of ornithine, ornithuric acid and α-benzoylornithine. *Biochem. J.* **1935**, *29*, 2684-2688.

[S15] a) A. Kjaær, P. O. Larsen, Amino acid studies. Part II. Structure and synthesis of albizziine (L-2-amino-3-ureidopropionic acid), an amino acid from higher plants. *Acta Chem. Scand.* **1959**, *13*, 1565-1572; b) A. F. Müller, F. Leuthardt, Nachweis der Citrullinbildung in

Mitochondriensuspensionen und Gewebsschnitten aus Leber durch Papierchromatographie. *Helv. Chim. Acta.* **1949**, *32*, 2289-2349.

[S16] K. Odo, Studies on derivatives of cyanamide. XLII. Synthesis of guanidine compounds and their derivatives. 6. Synthesis of arginine. *Nippon Kagaku Zasshi* **1953**, *74*, 774-775.

[S17] C. J. Marmion, T. Murphy, K. B. Nolan, Ruthenium(III) readily abstracts NO from L-arginine, the physiological precursor to NO, in the presence of H_2O_2. A remarkably simple model system for NO synthases. *Chem. Commun.* **2001**, 1870-1871.

[S18] K .J. Martinkus, Ch.-H. Tann, S. J. Gould, The biosynthesis of the streptolidine moiety in streptothricin F. *Tetrahedron* **1983**, *39*, 3493-3505.

[S19] a) L. H. Foley, G. Buechi, Biomimetic synthesis of dibromophakellin. *J. Am. Chem. Soc.* **1982**, *104*, 1776–1777; b) M. Wada, Citrullin, a new amino acid in the press juice of the watermelon, Citrullus vulgaris Schrad. *Biochemische Zeitschrift* **1930**, *224*, 420-429.

[S20] P. Molina, P. M. Fresneda, New synthesis of pyrazole and isoxazole derivatives. *J. Heterocyclic Chem.* **1984**, *21*, 461-464.

[S21] P. Molina, A. Ferao, P. M. Fresneda, A. Lorenzo, A. Tárraga, Synthesis of pyrido [2,1-f]-1,3,4-oxadiazine derivatives. *J. Heterocyclic Chem.* **1983**, *20*, 381-384.

[S22] F. Darcet, Essais sur l'acide succinique et sur quelques unes de ses combinaisons. *Ann. Chim. Phys.* **1835**, *58*, 282-300.

[S23] H. Feuer, B. Bachmann, E. H. White, The reactions of succinic anhydride with hydrazine hydrate. *J. Am. Chem. Soc.* **1951**, *73*, 4716–4719.

[S24] M. Al-Talib, H. Tashtoush, N. Odeh, Diacyl acid dihydrazides. *Magn. Reson. Chem.* **1990**, *28*, 1072-1075.

[S25] H. Tashtoush, M. Al-Talib, N. Odeh, Unusual rearrangement in the dehydration of aroylated butanodihydrazides. *Liebigs Ann.* **1992**, 291-292.

[S26] K.-D. Ginzel, P. Brungs, E. Steckhan, Indirect electrochemical α-methoxylation of N-acyl and N-carboalkoxy α-amino acid esters and application as cationic glycine equivalents. *Tetrahedron* **1989**, *45*, 1691-1701.

[S27] T. Miyashi, Y. Nishizawa, Y. Fujii, K. Yamakawa, M. Kamata, S. Akao, T. Mukai, The intramolecular nitrene type 1,1-cycloaddition reaction of allyl-substituted diazomethanes. *J. Am. Chem. Soc.* **1986**, *108*, 1617–1632.

[S28] U. Niewöhner, W. Steglich, Reduktive Spaltung von 3-Oxazolin-5-onen; Anwendung zur Synthese β, γ-ungesättigter Ketone aus *N*-Acyl-2-phenylglycin-allylestern. *Angew. Chem.* **1981**, *93*, 411-412.

[S29] A. R. Katritzky, F. Al-Omran, R. C. Patel, S. S. Thind, Improved methods of conversion of primary amines into bromides. J. Chem. Soc., Perkin Trans. 1 **1980**, 1890-1894.

[S30] T. Sato, K. Naruse, M. Enokiya, T. Fujisawa, Facile synthesis of benzyl ketones by reductive coupling of benzyl bromide and acyl chlorides in the presence of a palladium catalyst and zinc powder. *Chem. Lett.* **1981**, *10*, 1135-1138.

1. Synthetic Design with the *Chematica* Program – The Importance of Accurate Rules and of Higher-order Logic

Bartosz A. Grzybowski[ab], Sara Szymkuć[a], Karol Molga[a], Ewa P. Gajewska[a], and Agnieszka Wolos[a]

[a]Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, 02-224, Warsaw, Poland; E-mail: nanogrzybowski@gmail.com, [b]IBS Center for Soft and Living Matter and the Department of Chemistry, Ulsan National Institute of Science and Technology, UNIST - gil 50,Ulju-gun, 689-798, Ulsan, South Korea.

Keywords: Chematica · Computers · Retrosynthesis

SCIENTIFIC REPORTS

OPEN

Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?

Received: 16 December 2016
Accepted: 6 April 2017
Published online: 15 June 2017

G. Skoraczyński[1], P. Dittwald[2], B. Miasojedow[1], S. Szymkuć[2], E. P. Gajewska[2], B. A. Grzybowski[2,3] & A. Gambin[1]

Supplementary Information for Manuscript titled "*Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?"* by G. Skoraczyński[1†], P. Dittwald[2†], B. Miasojedow[1], S. Szymkuć[2], E.P. Gajewska[2], B.A. Grzybowski[2,3*], A. Gambin[1*]

[1)] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, 02-097 Warsaw, Poland

[2)] Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland

[3)] Center for Soft and Living Matter of Korea's Institute for Basic Science (IBS), Department of Chemistry, Ulsan National Institute of Science and Technology, Ulsan, South Korea

[†] The authors contributed equally

* e-mail:grzybor72@unist.ac.kr or aniag@mimuw.edu.pl

1. Methods.

This section provides details of the method used to compute the bounds on the accuracy of classifiers of reaction yields and durations.

Classification problem

One of most important tasks of machine learning methods is to predict the value of a certain characteristic, say y (whose evaluation is difficult and computationally expensive), based on vector of features x. For finite possible values of y one then has a classification problem and when y is a real number, such a problem is called regression. In the present work we focus on a binary classification problem, in which for a given chemical reaction we wish to predict whether its yield/duration are, respectively, high-low or long-short. Features used for classification include chemical descriptors, common substructures, information about solvent and temperature, and more.

The binary classification problem is widely described in the ML literature. There are many approaches to this problem including logistic regression, support vector machines (SVM)[1], random forests (RF)[2], *k*-Nearest Neighbors (kNN) and its modifications, etc. As discussed in the main text, RF gave the best performance. Here, our aim was to investigate whether the accuracy of these predictions can, in principle, be improved with some other (hypothetical) classifier architecture. As we show, it is not possible to achieve better accuracy unless some additional knowledge is provided. In order to prove this statement formally, we applied the method proposed recently by V. Berisha et al. in refs [3,4], which allows to estimate the probability of misclassification for the binary Bayes classifier.

Binary Bayes classifier and its accuracy

Let us consider the problem of classifying a feature vector $x \in R^p$, into one of classes $y \in \{0,1\}$. We denote conditional distributions by $f_0(x)$ and $f_1(x)$, respectively, and the prior probability of class 0 by p. The Bayes classifier $\delta(x)\colon R^p \to \{0,1\}$ assigns an observation x to a class with the highest posterior probability and maximizes probability of correct prediction. Although the Bayes classifier is usually unfeasible (since distributions f_0 and f_1 are unknown), its value lies in the fact that other ML techniques cannot achieve better accuracy than the Bayes classifier. Therefore it is reasonable to consider Bayes classifier error rate:

$$e^{Bayes} = P(\delta(x) \neq y)$$

as the measure of difficulty of a problem.

Efficient estimation of the Bayes error rate is complicated. Thus, instead of estimating e^{Bayes} directly, we introduce and then estimate sharp lower and upper bounds on e^{Bayes}. Bounds on e^{Bayes} are based on the following divergence measure $u(\cdot,\cdot)$:

$$u(f_0, f_1) = \int \frac{\left(p \cdot f_0(x) - (1-p) \cdot f_1(x)\right)^2}{p \cdot f_0(x) + (1-p) \cdot f_1(x)} dx$$

Having function u, Bayes error rate e^{Bayes} can be bounded according to Theorem 2 in ref [4]. This theorem states that:

$$\frac{1}{2} - \frac{1}{2}\sqrt{u(f_0, f_1)} \leq e^{Bayes} \leq \frac{1}{2} - \frac{1}{2}u(f_0, f_1) \, .$$

The function $u(f_0, f_1)$ can be estimated by the Friedman-Rafsky (FR) statistic[5]. This statistic entails building a minimum spanning tree (MST) on union of points from different classes and then calculating edges which are incident to vertices from both classes. The number of such edges constitutes a FR statistic. The spanning tree is a subgraph of a given graph, which is a tree (a connected graph with no cycles) incident to all vertices. Minimum spanning tree is a spanning tree which has minimal sum of weights on its edges.

Given the FR statistic, we can estimate function$u(f_0, f_1)$, and further bounds on the Bayes error rate e^{Bayes}. By theorem Theorem 1 from ref [4], we have

$$1 - 2\frac{FR(X_0, X_1)}{N_0 + N_1} \to u(f, g)$$

where $X_0 \in R^{N_0 \times dim}, X_1 \in R^{N_1 \times dim}$are samples from class 0 and 1, FR is Friedman-Rafsky statistic, and N_0, N_1 are numbers of points in class 0 and 1, respectively.

Bounds on classifier accuracy for yields and times of chemical reactions

Using methods described in the previous section, we estimate the Bayes error rate. To calculate the FR statistic for the set of descriptors, we split them into two subsets, for instance those associated with reactions with high and low yields (e.g., higher or lower than 0.65). As every descriptor is a multidimensional vector, the distances between them are calculated as Euclidean. After splitting points in multidimensional Euclidean space into two classes, we calculated Maximum Spanning Tree (MST) for the union of these two sets using Prim's algorithm[6]. With the MST at hand, we calculated the FR statistic, function u, and then Bayes error rates as described in the previous section. The procedure was repeated several times for different sample sizes with

randomly chosen points. For both reaction yield and duration time, the approximate Bayes error rates stabilize and appear to converge to the true Bayesian prediction error. The results obtained are summarized in Figure S1. The ca. 20% Bayes error rate estimate for our classification problem provides the formal proof that no other classifier can achieve better accuracy given the set of descriptors/fingerprints used to characterize molecules/reactions. For reaction duration dataset, the error's lower bound is smaller but still relatively high (ca. 18%). In Figure S2, analogous results based on the reaction fingerprints are presented. Note that in all cases estimates of upper and lower bounds on the Bayes error rates stabilize for large sample sizes. Thus, our estimates of e^{Bayes} are reliable. In addition, the PCA analysis also justifies the intrinsic complexity of the performed classification task, cf. Figure S3. The visualized data from different classes cannot be separated in the Euclidean space.

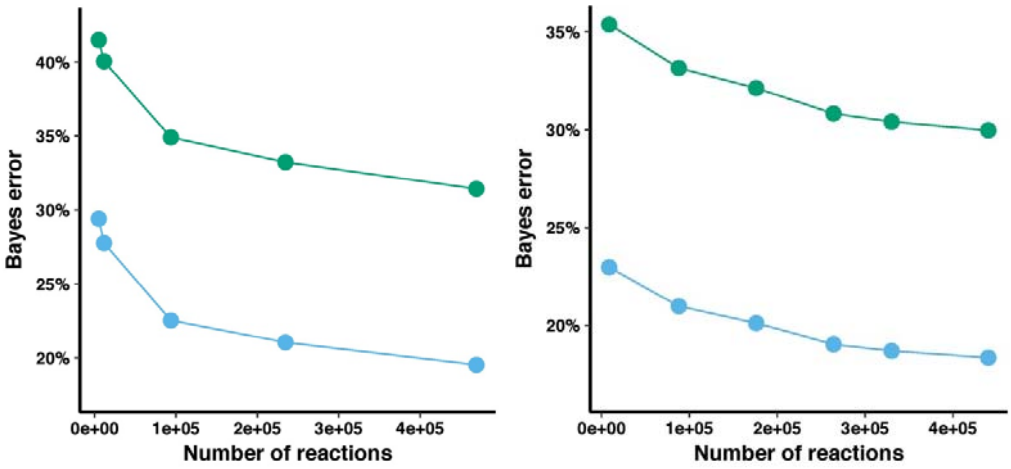


Figure S1. Upper and lower bounds on the Bayes error calculated based on molecular descriptors for different sizes of reaction sets. The left plot is for reaction yields, the right plot is for reaction times.

2. Additional results and analyses.

In addition to the Random Forest classifier, we also tested other machine learning methods. The classifier error for the Extreme Randomized Trees (ERT) was ca. 36% – that is, similar to RF but the classifier worked slower. For the Linear Support Vector Classification (parameter C = 1) the error was about 41%. As discussed in the main text, having constructed the classifiers we performed additional analyses based on the so-called Gini index[7], which indicated that classifiers' performance stabilizes when large sets of descriptors are used with the feature-importance score being stable over different algorithm runs. The results are summarized in Figure S4.

We also attacked the problem using Neural Networks[8,9]. First, we transformed the values of yields into the real line R by logit function ($log\left(\frac{yield}{1-yield}\right)$). Using feed-forward neural networks with single hidden layer and total 270 neurons in all layers, we fitted a linear model with transformed yields as a response variable and with fingerprints as explanatory variables. We used methods and algorithms described in section 8.10 of ref [10]. Finally, for new observation, we assigned a class to which the predicted value of yields belongs. The achieved accuracy was ca. 57% for yield prediction and ca. 74% for duration prediction, which is consistent with the performance of other classifiers.

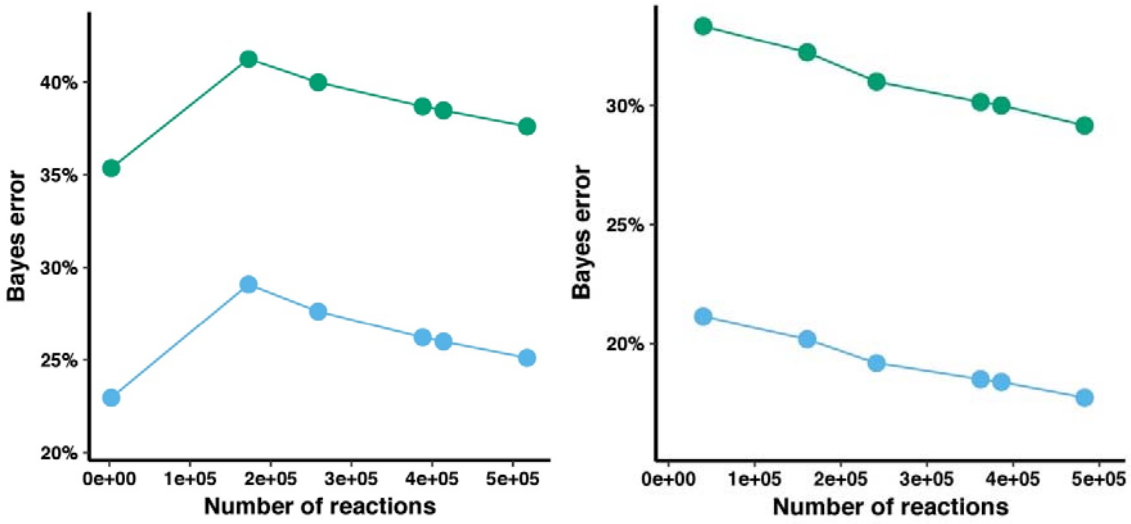


Figure S2. Upper and lower bounds on the Bayes error calculated based on reaction fingerprints for different sizes of reaction sets. The left plot is for reaction yields, the right plot is for reaction times. Note: The smallest error for the smallest number of reactions (in the left portion of the figure, for reaction yields) means that the number of data points was not sufficient to ensure good quality of the Bayes error estimation via the asymptotic theory of the Friedman-Rafsky statistics.

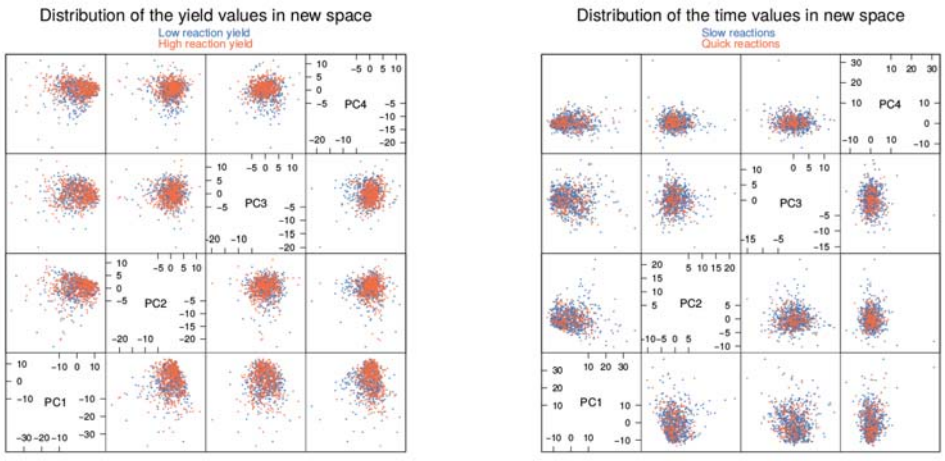


Figure S3. The Principal Component Analysis for reaction yield and duration datasets. Projections into 4 most significant components (explaining more than 50% of the variance) do not reveal any pattern.

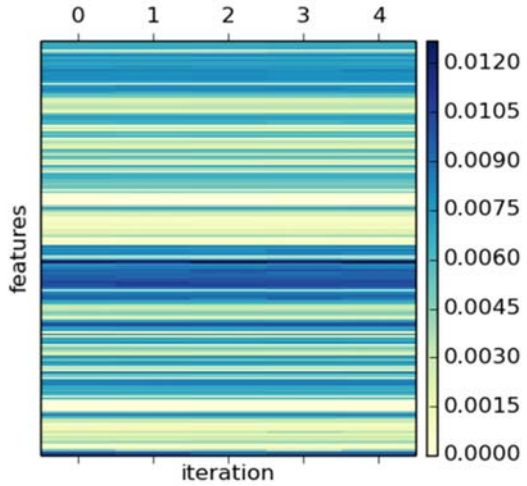


Figure S4. Gini index (GI) of chemical descriptors indicates the importance of a given feature for the classifier's decision. We observe, that GI does not change much between five independent runs of the Random Forest classifier.

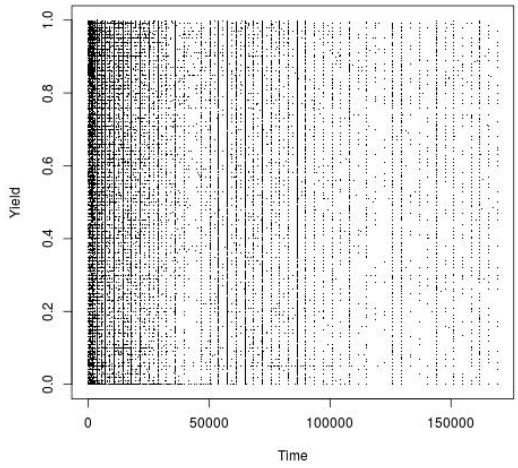


Figure S5. Scatterplot of reaction yields vs. times does not reveal any correlation. The calculated correlation coefficient was 0.06.

Supplementary references.

1. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273-297 (1995).
2. Breiman, L. Random forests. *Mach. Learn.* **45**, 5-32 (2001).
3. Berisha, V. & Hero, A.O. Empirical non-parametric estimation of the Fisher information. *IEEE Signal Process. Lett.* **22**, 988-992 (2015).
4. Berisha, V., Wisler, A., Hero, A.O. & Spanias, A. Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Trans. Signal Process.* **64**, 580-591 (2016).
5. Friedman, J.H. & Rafsky, L.C. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Stat.* **7**, 697-717 (1979).
6. Cormen, T.H., Leiserson C.E., Rivest, R.L. & Stein, C. *Introduction to algorithms* 2nd ed. (MIT Press and McGraw-Hill, 2001).

7. Menze, B.H. *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* **10**, #213 (2009).
8. Hastie T., Tibshirani, R. & Friedman, J. *The elements of statistical learning*, (Springer, 2009).
9. Haykin, S. *Neural networks: A comprehensive foundation* 2nd ed. (Prentice Hall, 1998).
10. Venables, W.N. & Ripley, B.D. *Modern applied statistics with S* 4th ed. (Springer, 2002).

RDKit descriptors

Table S1. Maximal set of RDKit descriptors (which, by being calculated for both substrates and products results in almost 400 descriptors) considered during classification tasks

Nr	Descriptor
1	MinAbsPartialCharge,
2	The number of radical electrons the molecule has (says nothing about spin state),
3	The average molecular weight of the molecule ignoring hydrogens,
4	MaxAbsEStateIndex,
5	MaxAbsPartialCharge,
6	MaxEStateIndex,
7	MinPartialCharge,
8	The exact molecular weight of the molecule,
9	The average molecular weight of the molecule,
10	The number of valence electrons the molecule has,
11	MinEStateIndex,
12	MinAbsEStateIndex,
13	MaxPartialCharge,
14	Calculate Balaban's J value for a molecule,
15	A topological index meant to quantify "complexity" of molecules.,
16	From equations (1),(9) and (10) of Rev. Comp. Chem. vol 2, 367-422, (1991),
17	Chi0n,
18	Chi0v,
19	From equations (1),(11) and (12) of Rev. Comp. Chem. vol 2, 367-422, (1991),
20	Chi1n,
21	Chi1v,
22	Chi2n,
23	Chi2v,
24	Chi3n,
25	Chi3v,

26	Chi4n,
27	Chi4v,
28	HallKierAlpha,
29	This returns the information content of the coefficients of the characteristic polynomial of the adjacency matrix of a hydrogen-suppressed graph of a molecule.,
30	Kappa1,
31	Kappa2,
32	Kappa3,
33	LabuteASA,
34	MOE Charge VSA Descriptor 1 (-inf < x < -0.30),
35	MOE Charge VSA Descriptor 10 (0.10 <= x < 0.15),
36	MOE Charge VSA Descriptor 11 (0.15 <= x < 0.20),
37	MOE Charge VSA Descriptor 12 (0.20 <= x < 0.25),
38	MOE Charge VSA Descriptor 13 (0.25 <= x < 0.30),
39	MOE Charge VSA Descriptor 14 (0.30 <= x < inf),
40	MOE Charge VSA Descriptor 2 (-0.30 <= x < -0.25),
41	MOE Charge VSA Descriptor 3 (-0.25 <= x < -0.20),
42	MOE Charge VSA Descriptor 4 (-0.20 <= x < -0.15),
43	MOE Charge VSA Descriptor 5 (-0.15 <= x < -0.10),
44	MOE Charge VSA Descriptor 6 (-0.10 <= x < -0.05),
45	MOE Charge VSA Descriptor 7 (-0.05 <= x < 0.00),
46	MOE Charge VSA Descriptor 8 (0.00 <= x < 0.05),
47	MOE Charge VSA Descriptor 9 (0.05 <= x < 0.10),
48	MOE MR VSA Descriptor 1 (-inf < x < 1.29),
49	MOE MR VSA Descriptor 10 (4.00 <= x < inf),
50	MOE MR VSA Descriptor 2 (1.29 <= x < 1.82),
51	MOE MR VSA Descriptor 3 (1.82 <= x < 2.24),
52	MOE MR VSA Descriptor 4 (2.24 <= x < 2.45),
53	MOE MR VSA Descriptor 5 (2.45 <= x < 2.75),
54	MOE MR VSA Descriptor 6 (2.75 <= x < 3.05),
55	MOE MR VSA Descriptor 7 (3.05 <= x < 3.63),
56	MOE MR VSA Descriptor 8 (3.63 <= x < 3.80),
57	MOE MR VSA Descriptor 9 (3.80 <= x < 4.00),
58	MOE logP VSA Descriptor 1 (-inf < x < -0.40),

59	MOE logP VSA Descriptor 10 (0.40 <= x < 0.50),
60	MOE logP VSA Descriptor 11 (0.50 <= x < 0.60),
61	MOE logP VSA Descriptor 12 (0.60 <= x < inf),
62	MOE logP VSA Descriptor 2 (-0.40 <= x < -0.20),
63	MOE logP VSA Descriptor 3 (-0.20 <= x < 0.00),
64	MOE logP VSA Descriptor 4 (0.00 <= x < 0.10),
65	MOE logP VSA Descriptor 5 (0.10 <= x < 0.15),
66	MOE logP VSA Descriptor 6 (0.15 <= x < 0.20),
67	MOE logP VSA Descriptor 7 (0.20 <= x < 0.25),
68	MOE logP VSA Descriptor 8 (0.25 <= x < 0.30),
69	MOE logP VSA Descriptor 9 (0.30 <= x < 0.40),
70	TPSA,
71	EState VSA Descriptor 1 (-inf < x < -0.39),
72	EState VSA Descriptor 10 (9.17 <= x < 15.00),
73	EState VSA Descriptor 11 (15.00 <= x < inf),
74	EState VSA Descriptor 2 (-0.39 <= x < 0.29),
75	EState VSA Descriptor 3 (0.29 <= x < 0.72),
76	EState VSA Descriptor 4 (0.72 <= x < 1.17),
77	EState VSA Descriptor 5 (1.17 <= x < 1.54),
78	EState VSA Descriptor 6 (1.54 <= x < 1.81),
79	EState VSA Descriptor 7 (1.81 <= x < 2.05),
80	EState VSA Descriptor 8 (2.05 <= x < 4.69),
81	EState VSA Descriptor 9 (4.69 <= x < 9.17),
82	VSA EState Descriptor 1 (-inf < x < 4.78),
83	VSA EState Descriptor 10 (11.00 <= x < inf),
84	VSA EState Descriptor 2 (4.78 <= x < 5.00),
85	VSA EState Descriptor 3 (5.00 <= x < 5.41),
86	VSA EState Descriptor 4 (5.41 <= x < 5.74),
87	VSA EState Descriptor 5 (5.74 <= x < 6.00),
88	VSA EState Descriptor 6 (6.00 <= x < 6.07),
89	VSA EState Descriptor 7 (6.07 <= x < 6.45),
90	VSA EState Descriptor 8 (6.45 <= x < 7.00),
91	VSA EState Descriptor 9 (7.00 <= x < 11.00),
92	CalcFractionCSP3((Mol)mol) -> float : returns the fraction of C atoms that are SP3 hybridized,

93	Number of heavy atoms a molecule.,
94	Number of NHs or OHs,
95	Number of Nitrogens and Oxygens,
96	CalcNumAliphaticCarbocycles((Mol)mol) -> int : returns the number of aliphatic (containing at least one non-aromatic bond) carbocycles for a molecule,
97	CalcNumAliphaticHeterocycles((Mol)mol) -> int : returns the number of aliphatic (containing at least one non-aromatic bond) heterocycles for a molecule,
98	CalcNumAliphaticRings((Mol)mol) -> int : returns the number of aliphatic (containing at least one non-aromatic bond) rings for a molecule,
99	CalcNumAromaticCarbocycles((Mol)mol) -> int : returns the number of aromatic carbocycles for a molecule,
100	CalcNumAromaticHeterocycles((Mol)mol) -> int : returns the number of aromatic heterocycles for a molecule,
101	CalcNumAromaticRings((Mol)mol) -> int : returns the number of aromatic rings for a molecule,
102	Number of Hydrogen Bond Acceptors,
103	Number of Hydrogen Bond Donors,
104	Number of Heteroatoms,
105	Number of Rotatable Bonds,
106	CalcNumSaturatedCarbocycles((Mol)mol) -> int : returns the number of saturated carbocycles for a molecule,
107	CalcNumSaturatedHeterocycles((Mol)mol) -> int : returns the number of saturated heterocycles for a molecule,
108	CalcNumSaturatedRings((Mol)mol) -> int : returns the number of saturated rings for a molecule,
109	RingCount,
110	Wildman-Crippen LogP value,
111	Wildman-Crippen MR value,
112	Number of aliphatic carboxylic acids,
113	Number of aliphatic hydroxyl groups,
114	Number of aliphatic hydroxyl groups excluding tert-OH,
115	Number of N functional groups attached to aromatics,
116	Number of Aromatic carboxylic acide,
117	Number of aromatic nitrogens,

118	Number of aromatic amines,
119	Number of aromatic hydroxyl groups,
120	Number of carboxylic acids,
121	Number of carboxylic acids,
122	Number of carbonyl O,
123	Number of carbonyl O, excluding COOH,
124	Number of thiocarbonyl,
125	Number of C(OH)CCN-Ctert-alkyl or C(OH)CCNcyclic,
126	Number of Imines,
127	Number of Tertiary amines,
128	Number of Secondary amines,
129	Number of Primary amines,
130	Number of hydroxylamine groups,
131	Number of XCCNR groups,
132	Number of tert-alicyclic amines (no heteroatoms, not quinine-like bridged N),
133	Number of H-pyrrole nitrogens,
134	Number of thiol groups,
135	Number of aldehydes,
136	Number of alkyl carbamates (subject to hydrolysis),
137	Number of alkyl halides,
138	Number of allylic oxidation sites excluding steroid dienone,
139	Number of amides,
140	Number of amidine groups,
141	Number of anilines,
142	Number of aryl methyl sites for hydroxylation,
143	Number of azide groups,
144	Number of azo groups,
145	Number of barbiturate groups,
146	Number of benzene rings,
147	Number of benzodiazepines with no additional fused rings,
148	Bicyclic,
149	Number of diazo groups,
150	Number of dihydropyridines,
151	Number of epoxide rings,

152	Number of esters,
153	Number of ether oxygens (including phenoxy),
154	Number of furan rings,
155	Number of guanidine groups,
156	Number of halogens,
157	Number of hydrazine groups,
158	Number of hydrazone groups,
159	Number of imidazole rings,
160	Number of imide groups,
161	Number of isocyanates,
162	Number of isothiocyanates,
163	Number of ketones,
164	Number of ketones excluding diaryl, a,b-unsat. dienones, heteroatom on Calpha,
165	Number of beta lactams,
166	Number of cyclic esters (lactones),
167	Number of methoxy groups -OCH3,
168	Number of morpholine rings,
169	Number of nitriles,
170	Number of nitro groups,
171	Number of nitro benzene ring substituents,
172	Number of non-ortho nitro benzene ring substituents,
173	Number of nitroso groups, excluding NO2,
174	Number of oxazole rings,
175	Number of oxime groups,
176	Number of para-hydroxylation sites,
177	Number of phenols,
178	Number of phenolic OH excluding ortho intramolecular Hbond substituents,
179	Number of phosphoric acid groups,
180	Number of phosphoric ester groups,
181	Number of piperdine rings,
182	Number of piperzine rings,
183	Number of primary amides,
184	Number of primary sulfonamides,

185	Number of pyridine rings,
186	Number of quarternary nitrogens,
187	Number of thioether,
188	Number of sulfonamides,
189	Number of sulfone groups,
190	Number of terminal acetylenes,
191	Number of tetrazole rings,
192	Number of thiazole rings,
193	Number of thiocyanates,
194	Number of thiophene rings,
195	Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes),
196	Number of urea groups

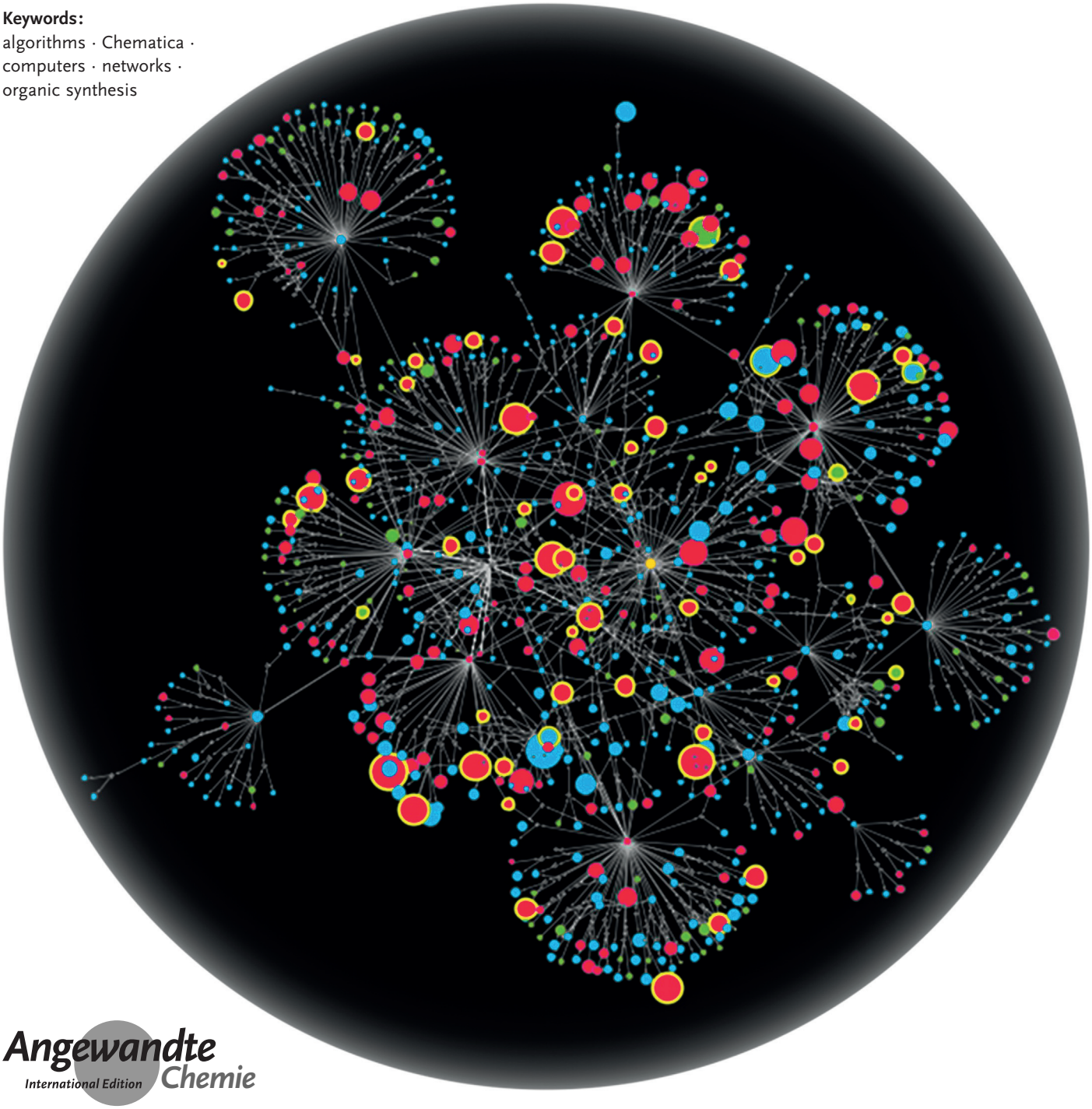
Computer-Aided Synthetic Planning

International Edition: DOI: 10.1002/anie.201506101
German Edition: DOI: 10.1002/ange.201506101

Computer-Assisted Synthetic Planning: The End of the Beginning

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

Keywords:
algorithms · Chematica · computers · networks · organic synthesis



Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

Computer-Assisted Synthetic Planning: The End of the Beginning

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

anie_201506101_sm_miscellaneous_information.pdf
anie_201506101_sm_Movie1_NOC_Travel.mp4
anie_201506101_sm_Movie2_Taxol.mp4
anie_201506101_sm_Movie3_Aripiprazole.mp4
anie_201506101_sm_Movie4_Syntaurus.mp4

Supplementary Information

CONTENTS:

Section S1. Setting up searches in Chematica.

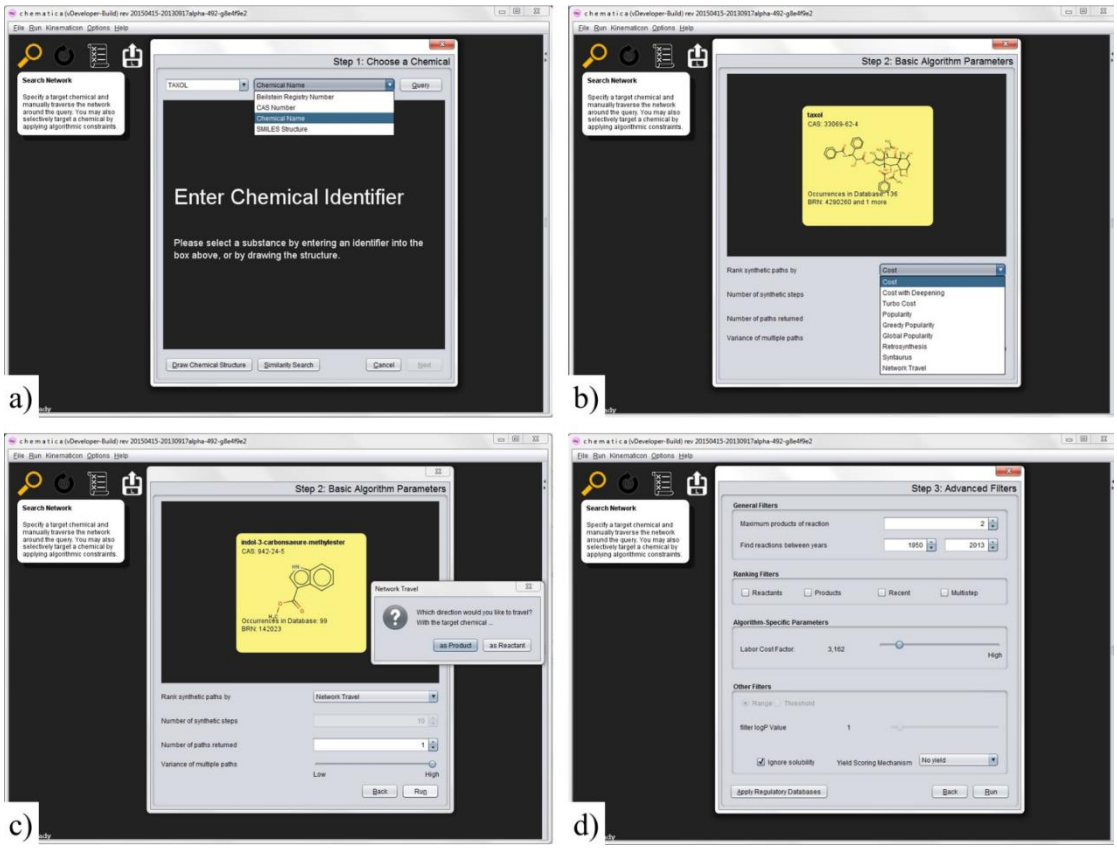


Figure S1. Setting up network searches in Chematica. **a)** In the starting window, the molecule of interest is specified by its identifier (common name, SMILES string, CAS number, or Beilstein identifier) or can be drawn in JAVA-based structure editor. Next, the user specifies the search algorithm – here for the target being **b)** Taxol the user might search for the minimal cost synthesis (see Figure 8 in the main text); for **c)** methyl indole-3-carboxylate, the user might wish to specify the Network Travel algorithm, as in Figure 6 in the main text. **d)** For searches like cost-minimization, the user can specify various additional parameters/constraints including time constraints, cost of labor vs. cost of substrates (slider in the middle, here set for labor being ca. three times more expensive than chemicals), solubility of the participating substances ("filter logP Value"), toxicity data ("Apply Regulatory Databases") and more. For more details, see main text.

Section S2. Manual network traversal in Chematica.

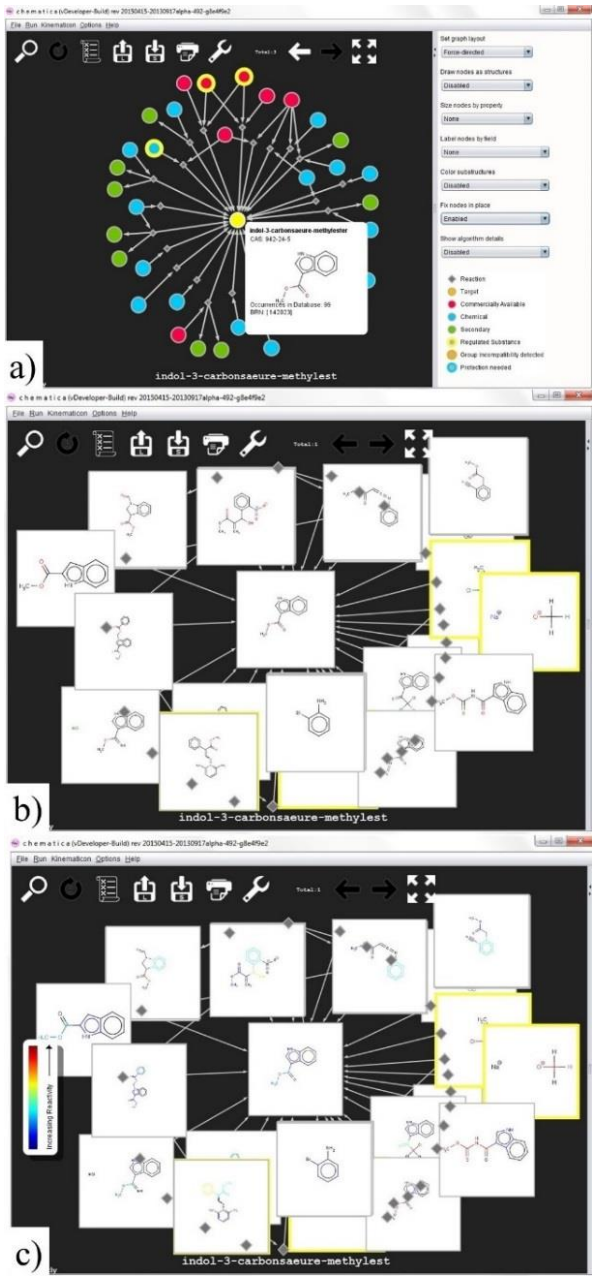


Figure S2. a) Reactions that lead to methyl indole-3-carboxylate. **b)** The same network but with nodes displayed as molecular structures. **c)** Again, the same network but with the functional groups colored according to their reactivity (for background literature, see ref [S1]). Less reactive groups (e.g., phenyls) are colored blue.

Section S3. Quantifying molecules' popularity in Chematica.

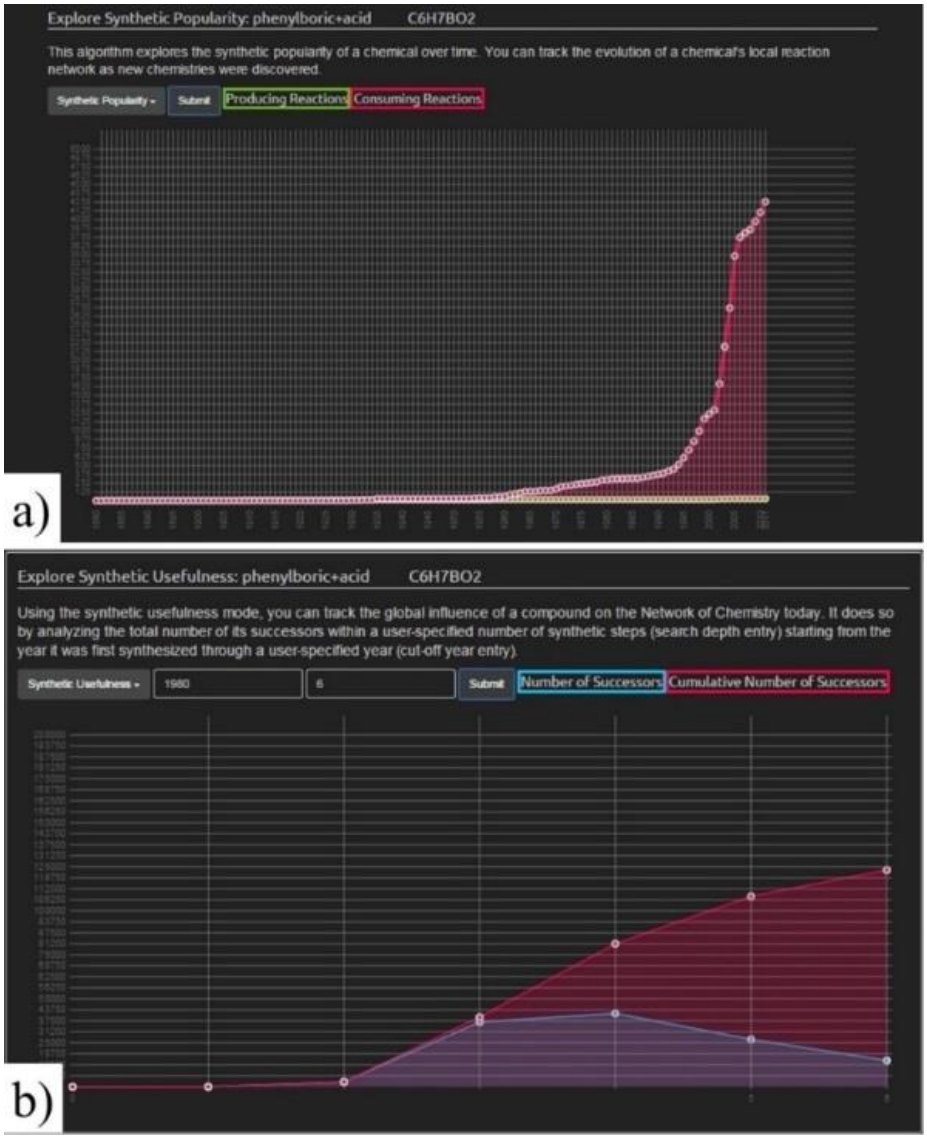


Figure S3. a) The plot illustrating time changes in the synthetic popularity of phenylboronic acid. Green line quantifies the number of reactions producing phenylboronic acid from 1880 to 2015 – as seen, there is no appreciable increase in the number of new ways of making this simple compound. However, the number of reactions in which phenylboronic acid was used as a substrate (red markers) increased dramatically after year 2000 – the reader will no doubt

correlate this increase with the development and growing popularity of efficient methods for palladium-catalyzed formation of aryl-aryl bonds by Akira Suzuki (Nobel Prize in 2010). **b)** Synthetic usefulness is another network measure telling us how many other molecules (red markers) can be made from a given molecule in *n* number of steps. The plot is for phenylboronic acid from which as many as ~120 000 other molecules can be made within 6 steps. The violet markers tell us how many molecules are made between m-th and m+1 steps. This type of a plot tells us about the proximity of the nearest "hub" molecule (here, at 3-4 steps away from phenylboronic acid).

Section S4. Constraints on Chematica's NOC searches.

In addition to the cost vs. labor parameter discussed in Section 1.3 of the main text, Chematica's SOCS scheme supports five types of constraints that come across as most useful in synthetic planning:

(i) Maximum number of reaction products (see Figure S4a, sub-menu marked as "#1) can filter out reactions with by-products. In some cases it can filter out reactions in which both stereoisomers were isolated – by doing so, this filter prefers enantioselective reactions (though it must be noted that many papers do not report the minority stereoisomer even if isolated);

(ii) The time span of the reactions to be considered (in Figure S4a marked as "#2") is useful in considering seasoned vs. modern syntheses);

(iii) Solubility (marked as "#5") limits the searches to molecules having only a certain logP value (octanol/water partition coefficient). One option for the user is to specify a threshold logP such that only substances below this value are considered in searches. This is useful in finding pathways for which the reactions are likely to proceed in polar solvents, especially water (low logP values) which is desirable for "green" syntheses. Another option is to specify a range of logP's such that all molecules in the pathway fit within this range – all reactions comprising pathways identified in this way are expected to proceed in solvents of similar polarity. For both cases, the values of logP for all molecules are calculated using a modified version of a highly predictive atom contribution model developed by Viswanadhan et al [S2].

(iv) Application of regulatory databases (in Figure S4a marked #6) opens a sub-window (#7) in which the user can choose toxic or regulated substances from four different lists (48 of the most dangerous precursors to chemical weapons from Australia Group list; 285 chemicals from the US Department of Homeland Security, DHS, regulated chemicals list; 985 substances

from the US Environmental Protection Agency, EPA, list; and 929 chemicals from the EPA List of Lists). The selected lists or their user-specified subsets are then excluded from the searches;

(v) Avoidance of any given substance can also be specified (Figure S5). This type of constraint is useful if optimized pathways are always "funneled" through a particular substance (or substances) and one wishes to force the searches into other synthetic possibilities.

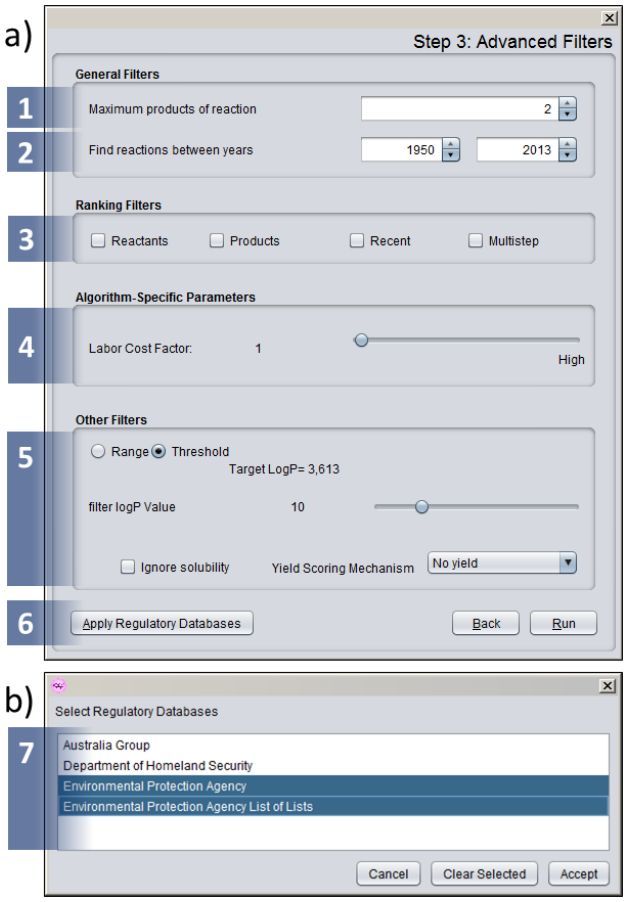


Figure S4. Chematica's sub-menu window in which various search constraints can be specified: #1 specifies maximum allowable number of products of each reaction; #2 defines the range of years in which the reactions of interest were published; #3 is a family of advanced filters favoring (but not completely prohibiting) certain reactions (e.g., those that have minimal

possible numbers of reactants); #4 slider defining how much the user values the cost of labor vs. the cost of starting materials; #5 panel allowing the user to specify solubilities of molecules to be considered during reaction planning; #6 a button and #7 the list of regulatory databases to be used to avoid toxic/regulated substances in the syntheses designed.

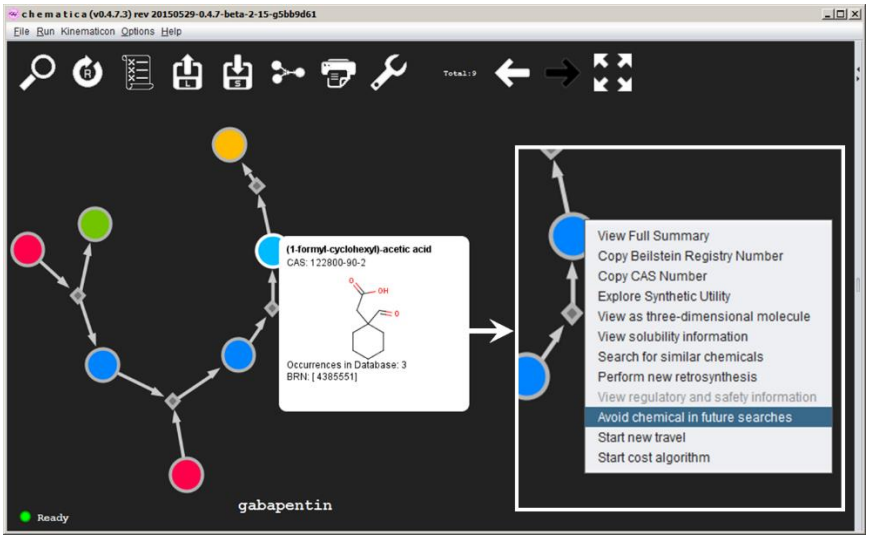


Figure S5. A screenshot from Chematica illustrates how any particular substance can be excluded from/avoided in subsequent searches.

Section S5. Multi-target optimization. In Sections1.4 and 1.5 of the main text we described optimization of syntheses leading to one specific target. In chemical industry, however, one might wish to simultaneously optimize pathways leading to multiple targets such that their syntheses share many common substrates/intermediates. The problem with this type of searching is that the numbers of possibilities of each individual pathway (approximately) multiply, and the total numbers of possible syntheses can well exceed 10^{100}. For such astronomical numbers, deterministic algorithms exploring all options (cf. Sections 1.3.3 and 1.4) are simply inadequate and one has to resort to probabilistic searches such as Monte Carlo, MC, used widely in statistical physics and molecular modeling to probe large numbers of possible states to find global minima (for theory of MC methods, see [S3]). Below, we address this problem based on our recent study described in detail in ref [25a].

Briefly, after the targets are specified, the search is initialized with some randomly generated, "guess" synthesis plan. This plan is then altered using two types of Monte Carlo moves: (*i*) reaction insertion/removal or (*ii*) substrate insertion/removal. Those moves that decrease the total synthesis cost are accepted unconditionally; those that increase the cost are accepted according to the so-called Metropolis criterion – that is, with the probability proportional to $\exp(-\beta C_{tot})$, where β is an adjustable parameter analogous to inverse temperature in a physical system, $\beta \propto 1/T$. In this way, the searches are not trapped in local cost minima and each viable synthesis plan j is visited with probability $p_j \propto \exp(-\beta C_{tot}(j))$. To guide the search even more efficiently towards the global cost minimum, the so-called Simulated Annealing[S3c] MC is performed in which the value of parameter β is initially low (such that many local minima can be effectively

explored) but is then gradually increased such that the search explores only low-cost plans with any significant probability, ultimately evolving towards the globally optimal cost.

Figure S6 illustrates the performance of the algorithm in optimizing the portfolio of 51 products of a small synthetic company (ProChimia Surfaces, www.prochimia.com) specializing in the synthesis of thiols, disulfides and silanes for self-assembled monolayers. ProChimia is a particularly suitable candidate since it is owned by one of the authors (B.A.G.), and we had full access to the synthetic procedures it used before our optimization. As a benchmark for further comparisons, we first optimized the syntheses of all 51 molecules individually to obtain the average synthesis cost per gram, $C_{tot}^{0} \sim \$40$. We then applied the global optimization procedure in which the synthesis cost gradually decreased (Figure S6a) until reaching an optimal *collective* synthesis plan (Figure S6c), for which the average cost per gram was $C_{tot} \sim \$21.5$ (i.e., 45% less compared to the individually optimized syntheses, Figure S6b). A general feature of this type of collective optimization is that the savings they offer increase with the number of targets as there are more opportunities to exploit common reactions and intermediates. In ProChimia's case, many syntheses go through undecylenic bromide (Br-$(CH_2)_7$-CH=CH_2) because alkenes are useful handles to convert to other functional groups such as thiols and silanes. On the other end of the chain, the haloalkyl functionality serves as an important precursor to azides, amines, amides, sulfonates, etc. It should be noted, however, that for targets very distant from one another on the Network (i.e., usually structurally very disparate molecules), the chances of syntheses reaching common substrates become low – although such cases are industrially rare and even for relatively diverse target sets the savings are still on the order of 10%. While collective optimizations are not a matter of seconds to minutes (and are run separately from

Chematica), they are also not prohibitive in the sense that one can typically optimize a company's synthetic portfolio within one-two days on a multi-core computer.

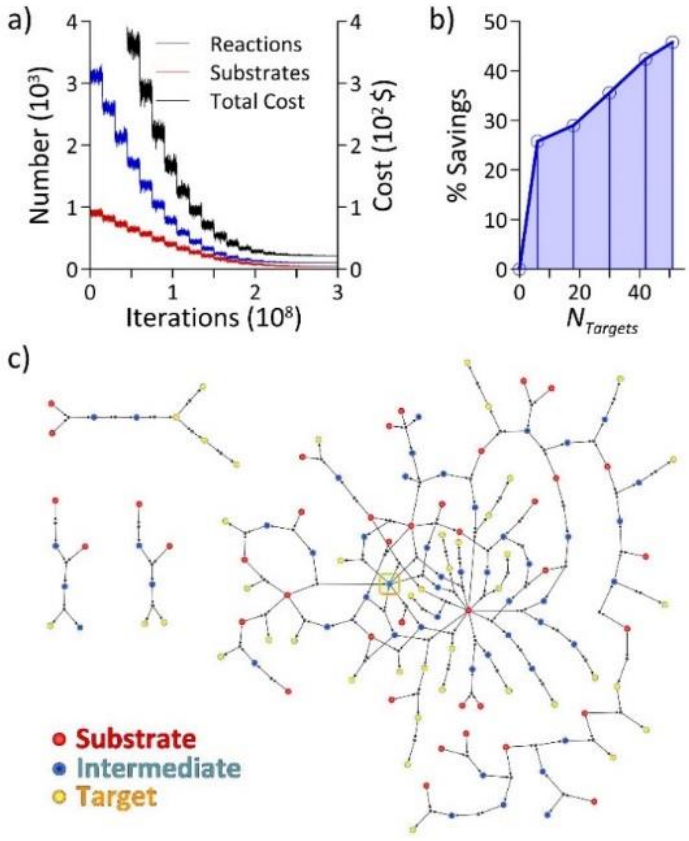


Figure S6. Global synthesis optimization by Monte Carlo algorithms. **a)** As the synthesis plan is gradually optimized (by Monte Carlo Simulated Annealing, see text), the numbers of reactions and substrates involved in the synthesis decrease steadily (but not necessarily monotonically) with the number of the optimization Monte Carlo moves attempted; similarly, the total cost of synthesis decreases to a (near) optimal value. **b)** Percent savings as a function of the number of target compounds. Here, sets of 6, 18, 30, 42, and 51 compounds are chosen from the commercial portfolio of ProChimia's compounds; see SI to ref [25a] for the list of compounds. (c) Network schematic for the optimal synthesis plan for 51 Prochimia products for a reaction cost

of C^{o}_{rxn}= 10. Note that for most products, the algorithm finds a common synthetic "tree" in which key intermediates are shared in the synthesis of different products. The node enclosed by an orange box is undecylenic bromide – one of the "hub" intermediates discussed in the main text. Figure is reproduced by permission from [25a].

Section S6. Network rewiring: One-pot reactions. Another important problem that can be addressed by searching known chemistries is the possibility of combining individual reactions in the network into one-pot sequences, effectively "rewiring" the network and creating synthetic "shortcuts". One-pot reactions[S4] are central to modern synthesis (in both academic and industrial settings) as they save resources and time by avoiding isolation, purification, characterization, and production of chemical waste after each synthetic step. Sometimes, such reactions are identified by chance or, more often, by careful inspection of individual steps that are to be "wired together" – this latter process, however, can be quite complicated given that it is necessary to consider all potential cross-reactivities of all molecules participating in the reaction sequence as well as the compatibility of solvents and reaction conditions between individual steps.

In work described in detail in ref [25b], we taught the computer several types of rules about cross-reactivity and reaction compatibility. Say, we wish to establish whether two individual reactions A → B and B → C can be combined into a one-pot sequence leading directly from A to C without isolation of any intermediates (Figure S7). After ascertaining that such a direct connection has not yet been reported in the literature/NOC, we apply several screening "filters":

Filer **#1** examines the compatibility of functional groups on all molecules participating in a putative sequence. To do so, a house-written program unambiguously partitions each of the

molecules into functional groups taken from a list of 322 common chemical functionalities (Fig. S8a, for details see ref [25b]). The constituent groups are then compared against a 322x322 "master" matrix where all possible group combinations are classified as mutually unreactive (i.e., compatible; grey entries in Figure S8b) or reactive (incompatible, red entries). Filter **#2** verifies whether reaction conditions required in each step permit undesired reactions of functional groups in other steps. These rules are summarized in the form of a table comprising 97 typical reaction types/conditions vs. 322 functional groups (see SI to ref. [25b]). Filter **#3** checks for the compatibility of solvents using well-known solvent miscibility tables. Filter **#4** checks for anhydrous vs. aqueous conditions (e.g., in Gattermann reactions which install aldehyde groups in aryl systems under aqueous conditions, subsequent one-pot steps cannot involve water-sensitive reactants or reagents such as Grignard compounds, alkali metal hydrides, organolithium reagents, etc.), filter **#5** checks for oxidizing vs. reducing conditions, filter **#6** determines acid-base compatibility, filter **#7** checks for the incompatibilities in terms of hydride/proton sources, and filter **#8** checks for the compatibility of chemical groups on the reagents used (akin to filter **#1** for substrates/products). Overall, the rules stored in the filter tables comprise over 15,000 chemical criteria to evaluate candidate one-pot sequences.

In reference [25b], we applied this algorithm to several small networks of reactions and then verified its predictions by performing the one-pot syntheses identified. Perhaps the most striking example is the rewiring of a synthetic network of inhibitors of phosphoinositide 3-kinase delta (PI3Kδ), a key enzyme in the signaling pathway involved in airway inflammation[S5]. Figure S9 shows the network of syntheses of several PI3Kδ inhibitors, inhibitor precursors, or closely related compounds with four two-step, eight three-step, and one four-step one-pot sequences predicted and then validated experimentally as indicated by the experimental yields next to the

colored arrows. The most telling testimony of the method's effectiveness is the prediction – and then execution – of a four-step one-pot sequence combining cyclization, chlorination, alkynylation, and arylation. This sequence, indicated by a violet arrow in Figure S9 was carried out with an overall 49% yield. We note that the algorithm-identified sequences in Figure S9 provide an attractive approach for large-scale preparations, since they allow for flexible and regioselective introduction of substituents using acyclic precursor **5** and a substituted aniline **6c**.

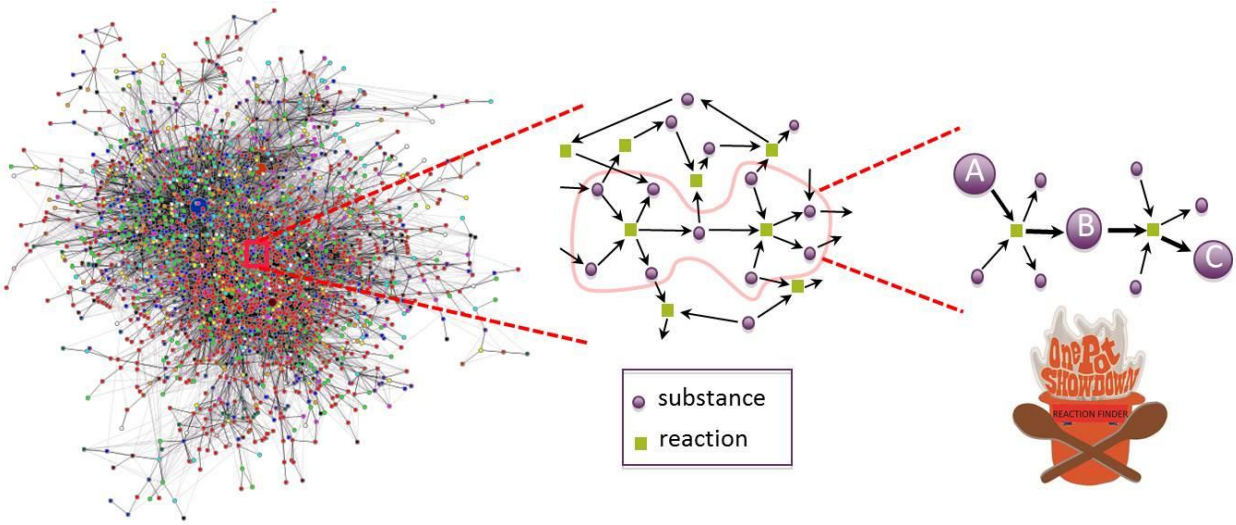


Figure S7. Networks and one-pot reactions. The left-side picture illustrates a relatively-large subset of the NOC with nodes colored according to the date these substances were reported. The fact that there are many colors illustrates that chemistry has been created by many independent "agents"/chemists. Can these diverse reactions be combined into sequences? To potentially do so, we focus on certain smaller sub-networks (middle) for which we would like to find synthetic shortcuts in the form of reaction sequences that can be executed in one pot, without isolation of intermediates. The simplest such a sequence is illustrated in the rightmost picture. Here, "rewiring" A → B and B → C reactions into a one-pot sequence leading directly from A to C

requires that A does not react with C, that substrates of the first reaction do not react with substrates of the second reaction, etc. See text for the list of other conditions.

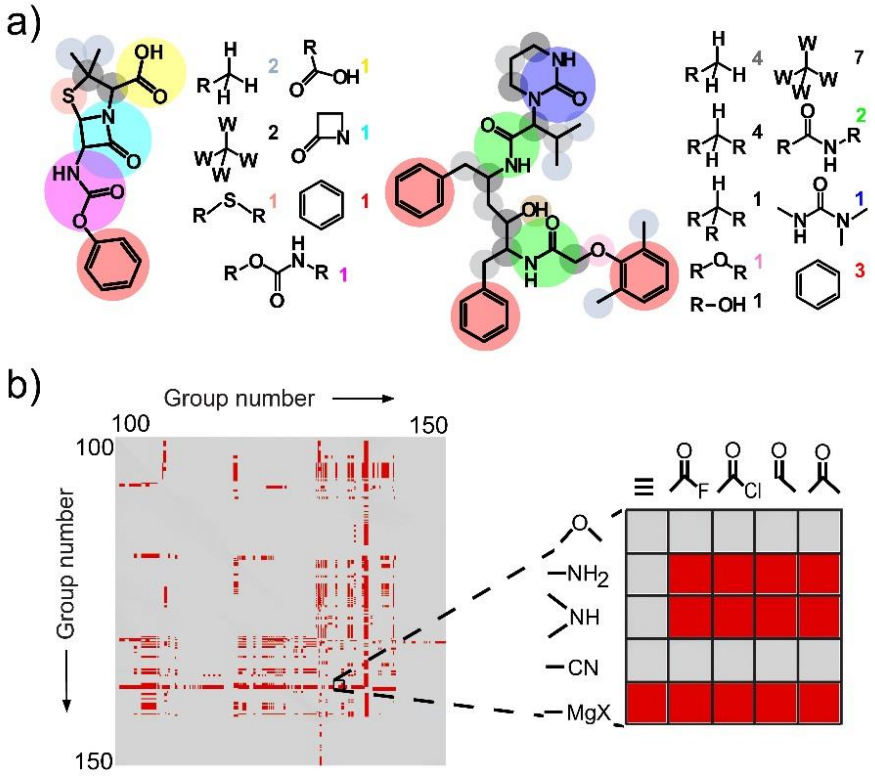


Figure S8. Illustration of automated molecule-to-group partitioning and functional group compatibility check. **a)** Examples of algorithmic partitioning of molecules into specific functional groups. The full list of possible 322 groups is included in the SI to ref [25b]. **b)** A large fragment (*left*) and further magnification (right) of the group compatibility 322 x 322 "master" matrix used to determine the compatibility or incompatibility of groups involved in a putative one-pot sequence under typical reaction conditions. The zoomed fragment contains some familiar group combinations and illustrates their well-known reactivity trends (e.g., ethers are

poor nucleophiles and generally unreactive, primary and secondary amines, on the other hand, are reactive towards all kinds of electrophiles, etc.). Figure reproduced by permission from [25b].

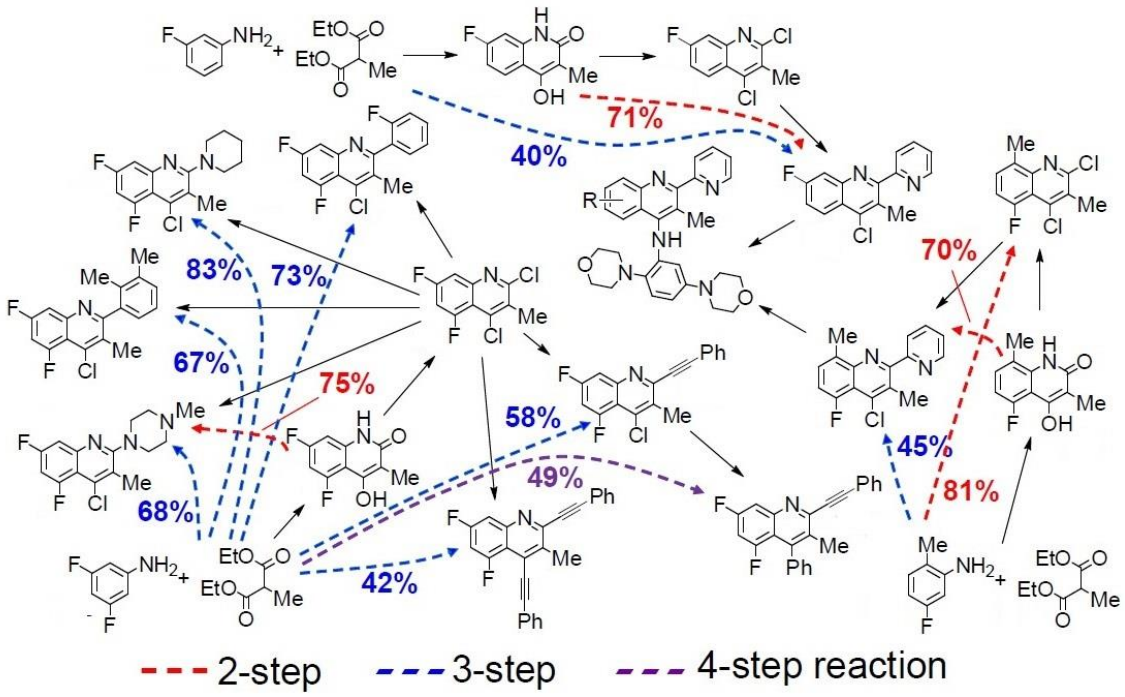


Figure S9. Rewiring networks of individual reactions involving PI3Kδ inhibitors and closely related compounds. Literature-reported, individual reactions correspond to black arrows; 2-step sequences are represented by red arrows, 3-step sequences by blue arrows, and a 4-step sequence is denoted by a purple arrow. Numbers next to the arrows correspond to the yields achieved when the predicted reactions were verified experimentally. Figure adapted by permission from [25b].

Section S7. "Black swans" of chemistry – examples of specialized but important reactions.

One example of a reaction that is relatively rarely used but still practically important is a cycloaddition shown in Figure S10a and having only 128 literature examples in Reaxys. Still, it enables facile preparation of *syn*-1,3-aminoalcohols which are useful building blocks and common motifs present in a number of natural products including cephalotaxine family. Likewise, without a "specialized" (only 44 literature examples in Reaxys) dehydrative aromatization of 1,4-cyclohexanediol derivatives, the size-selective synthesis of cycloparaphenylenes would be quite tedious if not impossible dues to strain factors (Figure S10b).

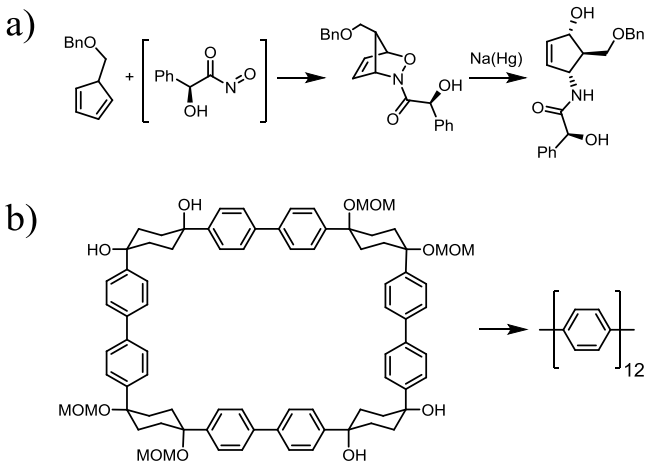
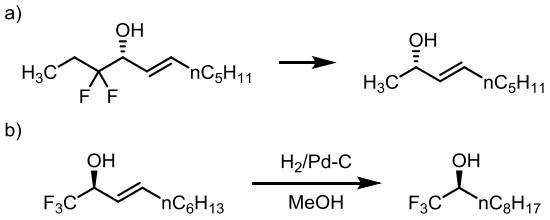
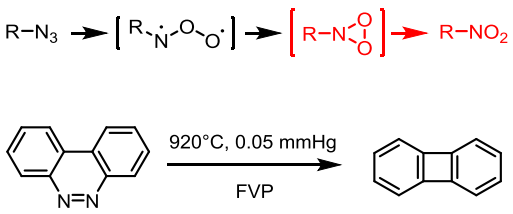
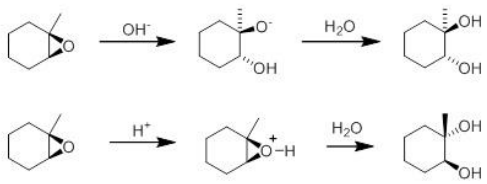


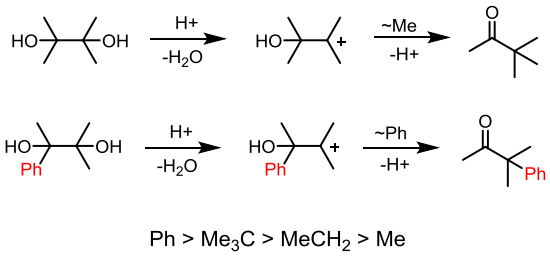
Figure S10. The importance of specialized reactions in the synthesis of nontrivial molecular cores. **a)** Cycloaddition of chiral *N*-acylnitroso compound (generated in situ) and cyclopentadiene followed by reduction with sodium amalgam. Although this reaction is not widely used, it allows for efficient preparation of *syn*-1,3-aminoalcohols which are useful in the synthesis of natural products. **b)** Synthesis of cycloparaphenylenes. Although this class of

macromolecules attracts considerable research interest (mostly in the context of molecular electronics, separation of fullerenes, and templates for nanotubes formation) there are only a few methods for their preparation. The last step of an efficient and size-selective synthesis involves dehydrative aromatization of a 1,4-cyclohexanediol units performed by treating an appropriate cyclic substrate with *p*-toluenesulfonic acid in *m*-xylene under microwave irradiation and at 150 °C. Without this specialized transformation, preparation of cycloparaphenylenes would be quite impossible due to geometry factors and generation of large ring strains.

Section S8. Partial list of problems encountered during automated extraction of reaction "cores" from repositories of literature-reported reactions.

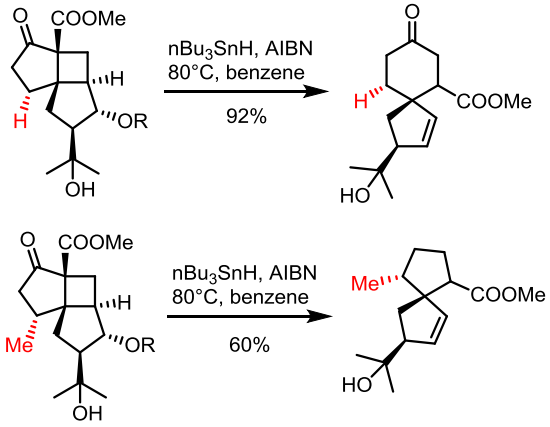
	Example	Comment
Human entry errors		The example shows (a) a reaction as entered into one of the major commercial databases and (b) the actual reaction from the source paper. The reaction "core" machine-extracted from the database would clearly be nonsensical.
Reaction conditions "useless" for general synthetic planning		In the top example, the reaction core (red; again, example taken from a major commercial DB) would suggest that dioxaziridines generated from azides can be precursors to nitro compounds. In reality, this specific photochemical reaction proceeds only at extremely low temperatures (77 K). The second example[S6a,b] might suggest that cinnolines easily generated from 2,2'-dinitrobiphenyls can be intermediates in preparation of biphenylenes. However, this specific reaction occurs only at very high temperatures thus limiting it use in more complicated molecules.
Differences in reaction conditions and substrate-dependence		Although epoxide opening ("reaction core") may lead to the same product under different conditions, there are examples in which the structure of the substrate itself dictates different outcomes under acidic and basic conditions. Machine extraction would treat such outcomes as distinct reactions whereas, in reality, they are the same reaction type, just "context-sensitive".

Effects of proximal substituents

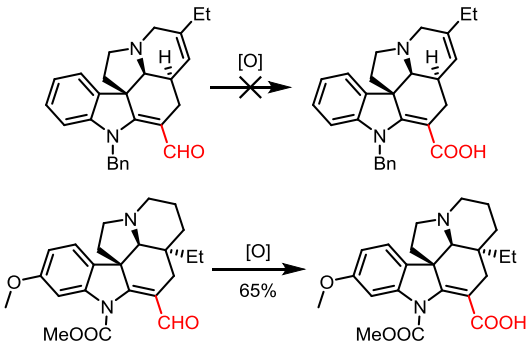


This is an extension of the previous category whereby the same reaction type (here, acid catalyzed pinacol rearrangement) yields different products irrespective of conditions but depending on the substituents present. If the reaction core is extracted too narrowly, these differences will not be taken into account; if the core is extracted too broadly, the machine will "learn" a multitude of specific variants rather than one reaction. Only an expert chemist can teach the computer a general rule in which order of migration depends on the migratory aptitude of the substituents.

Steric and electronic effects of distant substituents

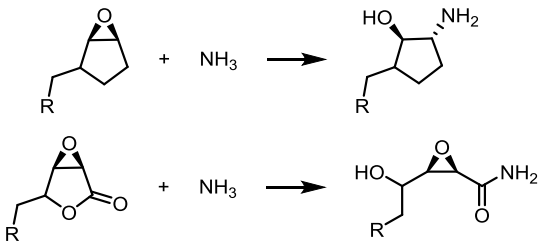


Although the reaction "core" seems to span only five bonds, the reaction is singular to the entire cyclobutylmethyl motif with even minor alterations (here, presence of methyl group) resulting in dramatic changes in steric hindrance and the overall reaction outcome.



In the second example, minor alterations in N-protection and the concomitant electronic effects either allow or prohibit oxidation of aldehyde to carboxylic acid. For more examples where such small differences have dramatic influence on reactivity see [S6c].

Presence of cross-reactive groups



With the core extracted (here, epoxide opening[S6d] and double-bond reduction[S6e]), reaction might or might not proceed in a specific molecule because of the presence of other cross-reactive groups (also see main text, Section 2.2.3). Although opening of the epoxide by ammonia could proceed smoothly in the first substrate, presence of

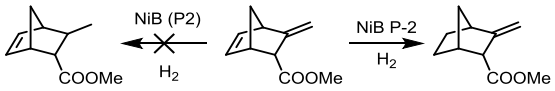
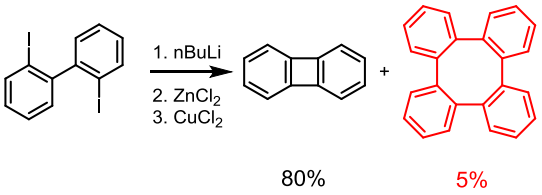
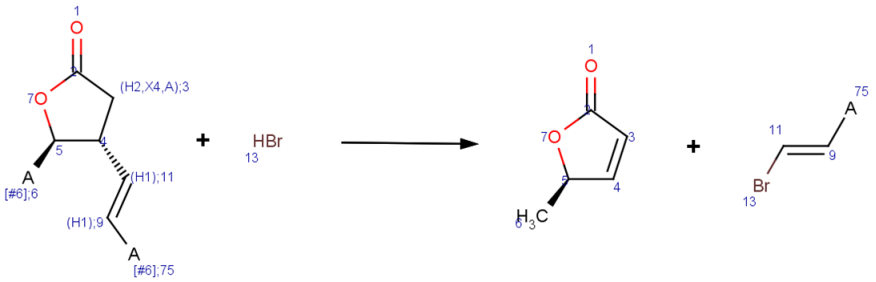
		lactone in the second substrate leads to the formation of an amide instead of an anti-aminoalcohol. In the second example, selective reduction of exocyclic alkene is impossible if the substrate molecule contains a strained double bond which reacts more readily under given conditions.
Relative abundances of products		In this example, a machine would extract two possible products although only one – here, obtained in 80% yield – is synthetically relevant. Correcting for such cases is possible but only by telling the machine to systematically score the extracted cores according to yields. If the machine is allowed to extracts the reaction leading to the minority product, such a transform will only lead to false positives during reaction planning[S6f].
Stereo and regiochemistry	Multiple examples (see main text)	This is one of the key problems of machine extraction: The reaction cores might have the stereo-/regiochemistry specified but these configurations are generally not conserved when applied to specific molecules (see Section 2.2.4).

Table S1. Partial list of problems encountered during automated extraction of reaction "cores" from repositories of literature-reported reactions.

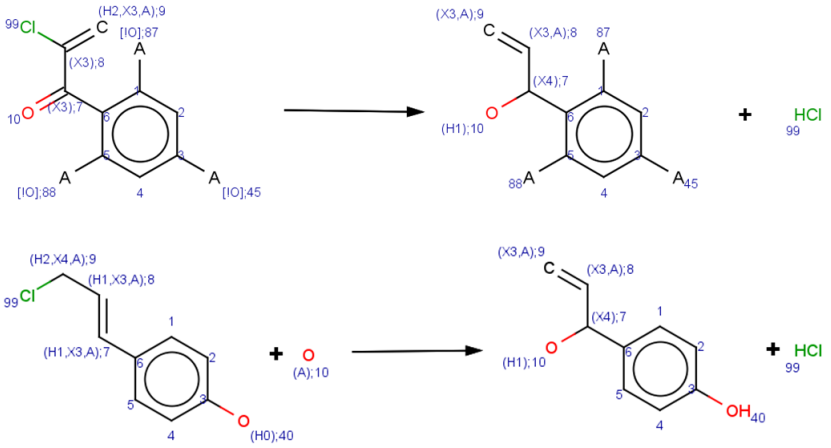
Section S9. Additional examples of transformations coded to account for steric or electronic effects.



rxn_id: 8827,
name: "Asymmetric 1,4-Addition Of Organocuprates ",
reaction_SMARTS:"[O:1]=[C:2]1[C:3][C@H:4](/[CH:11]=[CH:9]/[#6:75])[C@@H:5]([#6:6])[O:7]1.[Br:13]>>[O:1]=[C:2]1[C:3]=[C:4][C@@H:5]([#6:6])[O:7]1.[*:75]/[C:9]=[C:11]/[Br:13]"
products: ["[O]=[C]1[C][C@H](/[CH]=[CH]/[#6])[C@@H]([#6])[O]1", "[Br]"]
groups to protect: ["[#6][CH2][OH]", "[#6][SX2H]", "[OH][CX4]C[CX4][OH]", "[OH][CX4][CX4][OH]", "[#6][C]([#6])([#6])[OH]", "[CX4,c][NH2]", "[#6]C([#6])=O", "[#6][CH]=O", "[#6]C([OH])=O", "[#6][CH]([#6])[OH]", "[CX4,c][NH][CX4,c]", "[c][OH]", "[OH][c][c][OH]"]
protection_conditions_code: ["A01", "OC3"]
incompatible_groups:["[#6][N+]#[C-]", "c[N+]#[N]", "[#6][SX3](=O)[OH]", "[NX3]=[NX3]", "[#6][SX2][SX2][#6]", "[#6][CH2][Cl,Br,I]", "[#6]C(=O)[Cl,Br,I]", "[#6]O[OH]", "[CX4][Cl,Br,I]", "[#6]C(=[SX1])[#6]", "[#6][CH]=[SX1]", "[#6]OO[#6]", "[#6]N=C=[O,S]", "[CX4]1[O,N][CX4]1", "[#6][CH]([#6])[Cl,Br,I]", "[#6]C(=[O])OC(=[O])[#6]", "[#6]C([#6])([#6])[Cl,Br,I]", "[CX3]=[NX2][*!O]", "[#6]=[N+]=[N-]", "[#6][SX2][#6]"]
typical reaction conditions: "1.nBuLi.2.CuI.3.Electrophile",
general references: "10.1002/chem.200500513 AND 10.1016/j.tet.2011.09.022 AND 10.1021/jo00036a041",

Figure S11. Steric effects. Although, at first glance, 1,4-addition of organocuprates might seem a very general reaction, steric effects need to be taken into account in its more sophisticated variants. In the example shown, addition of vinyl cuprate (generated from vinyl bromide via lithiation-transmetallation sequence) occurs *anti*- to the already present stereocenter. This highly diastereoselective transformation is important for the synthesis of polycyclic natural products (e.g., Trauner's *Guanacastepene*, Jamison's *Terpestacin*, Mander's *Sordaricin*, Feringa's *Prostaglandin* PGE_1, Crimmins' *Silphinene*, or Paquette's *Capnellene*), and allows for the introduction of one (as shown) or even two or three stereocenters via trapping of the generated enolate (without using chiral catalysts or auxiliaries). Coding transformations of this type is quite an art in itself as it requires specifying proper stereochemistry of the *entire* reaction motif as well as all permissible atoms types and substituents (e.g., here, atoms #6 and #75 can be any carbon, atom #3 is C-sp^3 with two H's, etc.). We also note that the particular entry shown is just one of

about twenty entries that together cover this class of reactions with proper stereochemistry in all possible variants.



rxn_id: 7452,
name: " Oxidation-halogenation of allylic alcohols",
reaction_SMARTS:"[c:1]([*!O:87])1[c:2][c:3](-[*!O:45])[c:4][c:5]([*!O:88])[c:6]1/[CX3:7](=[O:10])[CX3:8]([Cl:99])=[CX3H2:9]>>[c:1]([*:87])1[c:2][c:3](-[*:45])[c:4][c:5]([*:88])[c:6]1[CX4:7]([OH:10])[CX3:8]=[CX3:9].[Cl:99]"
products:["[c]([*!O])1[c][c](-[*!O])[c][c]([*!O])[c]1/[CX3](=[O])[CX3]([Cl])=[CX3H2]"]]
groups to protect: ["[#6][CH2][OH]", "[#6][SX2H]", "[OH][CX4][CX4][OH]", "[CX4,c][NH][CX4,c]", "[CX4,c][NH2]", "[OH][CX4]C[CX4][OH]", "[#6][CH]([#6])[OH]", "[c][OH]", "[#6][C]([#6])([#6])[OH]"]
protection_conditions_code: ["OX6", "NNB4"]
incompatible_groups:["[#6][N+]#[C-]", "c[N+]#[N]", "[#6][SX3](=O)[OH]", "[NX3]=[NX3]", "[CX3]=[NX2][*!O]", "[#6]C(=O)[Cl,Br,I]", "[#6]O[OH]", "[#6]C(=[SX1])[#6]", "[#6][CH]=[SX1]", "[#6]OO[#6]", "[#6]N=C=[O,S]", "[CX4]1[O,N][CX4]1", "[#6]C(=[O])OC(=[O])[#6]", "[CX3]=[NX2][O]", "[#6]=[N+]=[N-]"]
typical reaction conditions: " oxalyl.chloride.DMSO.Et3N.DCM.-78C",
general references: "10.1021/jo0711992",

Figure S12. Electronic effects. Although remote substituents typically have minor effects on reaction outcomes, there are cases in which they have to be taken into account. In the example shown, a Swern-type reaction of allylic alcohol may lead to two different products depending on the presence or the absence of phenolic oxygen(s) influencing the conjugated *pi*-system. As in Figure S11, each atom and substituent type must be carefully determined. For example, in Syntaurus' record #7452 shown here, preparation of vinyl chloride is possible only if there are no phenols at *o*- or *p*- positions. Preparation of allyl chloride is possible only if there are *o*- or *p*- phenolic oxygens (Record #7450, only graphical output shown; *o*- variant not shown).

Section S10. Comparison of matrix vs. SMILES notation of molecules.

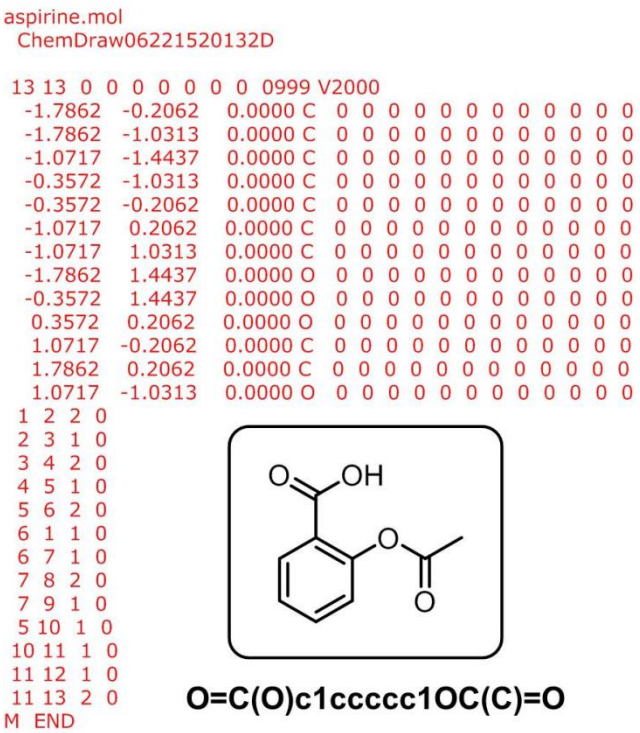


Figure S13. Comparison of a ".mol" file (*red*) and SMILEs notation (*bold black*) for the same molecule (aspirine).

Section S11. Example of QM calculations in Syntaurus. While there are many types of computational methods available for such calculations, it must be remembered that during retrosynthetic planning the computer typically considers very large numbers of possibilities (thousands to millions) and only methods operating on the timescales of ms per molecule are practical. We found that simple Hückel-type approach with parameters for heteroatoms (ref [46] in the main text) gives the best tradeoff between accuracy and speed – the choice of this method is ultimately justified by its satisfactory performance for different types of aromatic systems.

A typical result of Hückel-type calculations in Syntaurus is illustrated in Figure S12a, in which the synthetic target (central yellow node) is 4-nitrophenyl 4-methylbenzoate phenyl ester. Among many synthetic possibilities Syntaurus suggests, one is nitration of phenyl 4-methylbenzoate (yellow node circled with a white halo). The feasibility of such an electrophilic aromatic substitution is judged by the program based on the delocalization energies it calculates. Here, positions labelled in the figure as C2, C3, C14, C15 (deactivated by the ester substituent) and C10, C12 (in *meta* position to phenolic oxygen) have high delocalization energies and the sites available for nitration are the low-energy C9, C11, C13 positions – consequently, Syntaurus decides that the nitration reaction it considered at position C11 is allowable. We note, however, that if the synthetic target were, e.g., 3-nitrophenyl 4-methylbenzoate rather than 4-nitrophenyl 4-methylbenzoate and the nitration would have to occur at C10, the program would disallow this reaction (and it would not even be displayed among synthetic options). Naturally, such calculations are not limited to substituted benzenes and Syntaurus uses them for arbitrary aromatic and conjugated systems. Figure S12b has an example in which the program considers electrophilic nitration of methyl 3-methyl-1*H*-indole-6-carboxylate as a potential route to a methyl 3-methyl-2-nitro-1*H*-indole-6-carboxylate target. Because the position labelled as C10

has a very low delocalization energy, the nitration reaction at this position is indeed judged as feasible.

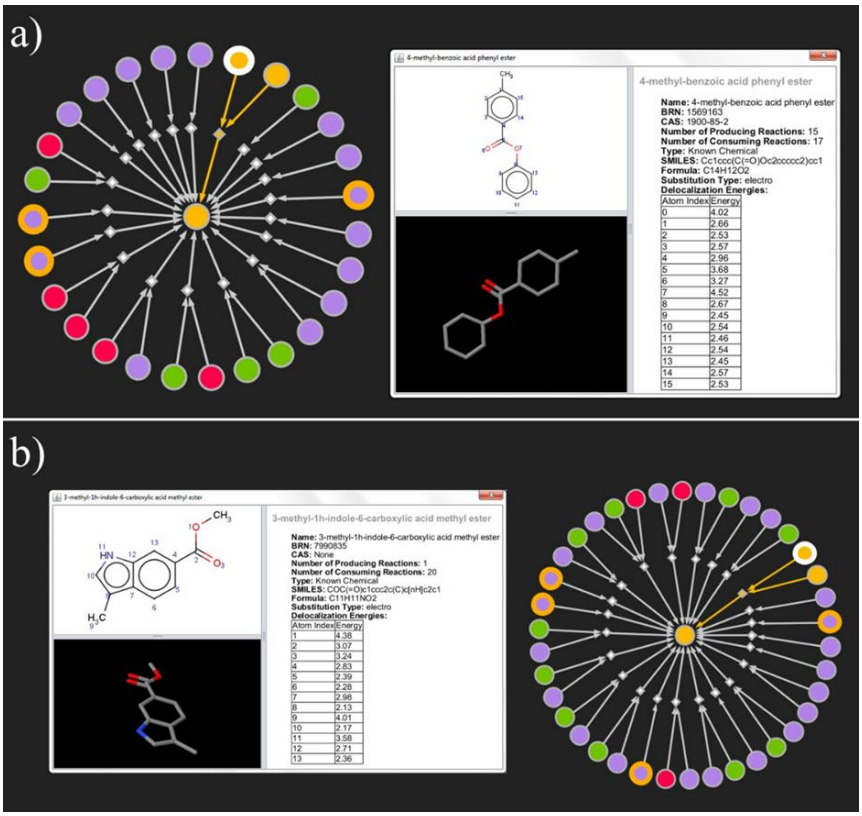


Figure S14. Supporting synthetic planning with basic Quantum Mechanics. The image in **a)** shows the screenshot of Chematica's main window displaying retrosynthetic options leading in one step to the 4-nitrophenyl 4-methylbenzoate phenyl ester target. Violet nodes denote unknown substances, green nodes denote known substances, red nodes stand for commercially available substrates, yellow nodes are currently selected molecules, and the orange halos indicate incompatibility conflicts. In the current example, one of the synthetic possibilities is electrophilic aromatic substitution (nitration) of phenyl 4-methylbenzoate (yellow node circled with a white halo) for which Chematica calculates electron delocalization energies to decide the feasibility of the proposed reaction. These energies are tabulated in the right portion of the „Molecule

Summary" subwindow shown. **b)** Analogous calculation of delocalization energies (but for a more complex aromatic system) allows Chematica to determine the outcome of a proposed nitration reaction at position labelled in the sub-window as C10.

Section S12. Nonsensical motifs.

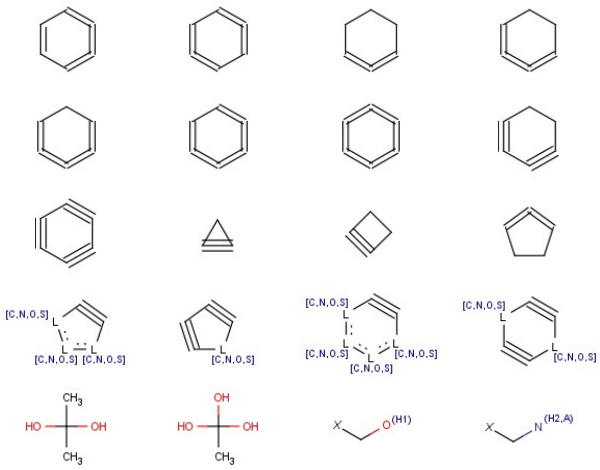
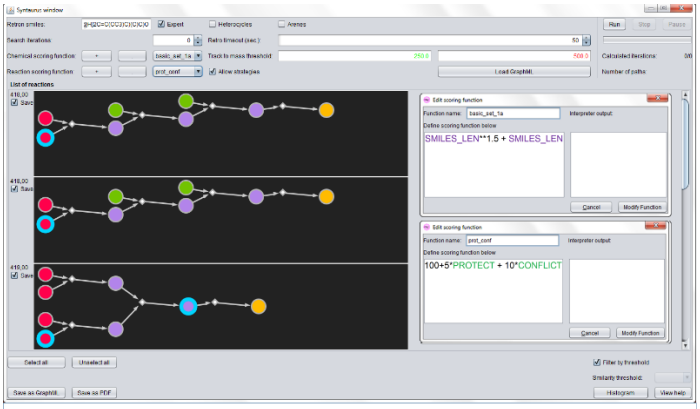


Table S2. Small selection of "impossible motifs" in Syntaurus. These motifs are forbidden due to known chemical instability problems (e.g., in geminal diols, triols, haloalcohols and amines; X denotes list of all halogens) or due to steric factors (e.g., small-ring cyclic allenes or alkynes).

Section S13. Variables available in Syntaurus to define scoring functions.



Variable	Description
PROTECT	+1 penalty for each protection needed (in molecules indicated by blue halos in the node representation of pathways)
CONFLICT	+1 penalty for each group incompatibility detected (in molecules indicated by orange halos in the node representation)
YIELD	Estimation of reaction yield based on a thermodynamic model (cf. [56])
EXCLUDE	Takes a list of reaction ID's and then (i) if defined with a „+" sign, excludes these reactions from synthetic searches or (ii) if defined with a „-" sign, promotes these reactions in searches
MASS	Mass of each substrate
SMILES_LEN	Length of a molecule's SMILES (in characters); similar to MASS but accounts for the overall complexity of the molecule implicit in the parentheses, @ and @@, and other special characters in the SMILES. (e.g. compare CCC(O)C vs. CC[C@@H](O)C)
STEREO	Number of stereocenters in each substrate
KNOWN	+1 if molecule is known in the NOC, 0 otherwise
BUY	+1 if molecule is commercially available, 0 otherwise
WEIRD	+1 penalty if a molecule is unknown and (i) has molecular mass less than 100 and (ii) the ratio of selected (i.e. S, P, Se, O, N, Si, Sn, B, As) heteroatoms to carbons this molecule contains is greater than 1.5

Figure S15. Setting-up scoring functions and variables available in Syntaurus. Top portion shows the Syntaurus' window in which the scoring functions for the searches are input. In the particular example, CSF = $SMILES\_LEN^{1.5}$ + SMILES_LEN and RSF = 10 + 5·PROTECT + 10·CONFLICT. The pathways shown are actual top scoring syntheses of tetrahydrocannabinol. The bottom table describes the variables that can be used to define RSF (green variables) and CSF (violet).

Section S14. Step-by-step design of epicolactone's synthesis (cf. Figure 15 in the main text) guided by Syntaurus. Figure S16a below illustrates the structure of epicolactone as input into Syntaurus' step-by-step retrosynthetic module with all the filters turned off (Figure S16b). Figure S17 displays all the synthetic possibilities then generated within one synthetic step. Red nodes denote commercially available chemicals, green nodes are known molecules, violet nodes stand for unknown molecules, blue halos indicate the need for protection chemistry, and orange halos indicate that serious cross-reactivity conflicts have been detected for a particular reaction.

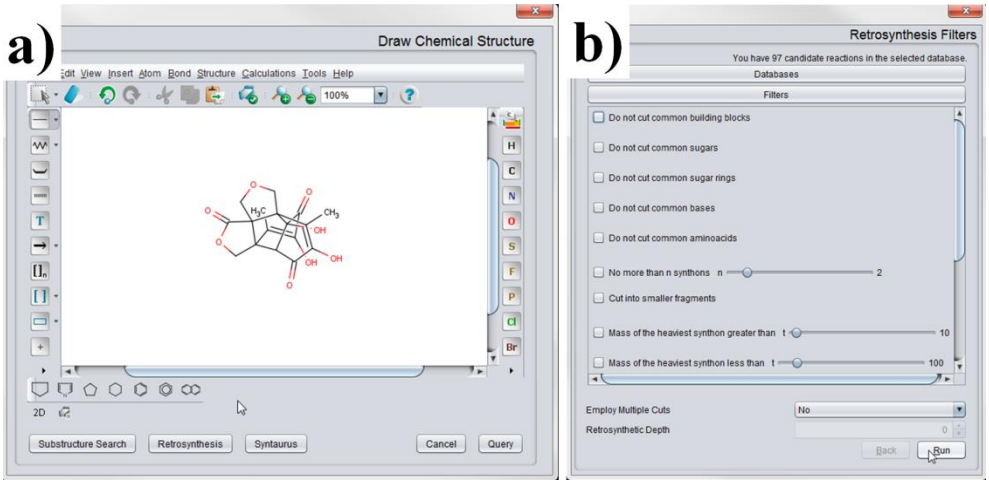


Figure S16.

Next, the results were displayed in the list format. Since the target is a "caged," polycyclic molecule, the results were scored/ranked according to the **RINGS** variable giving favorable score for each ring created (Figure S18).

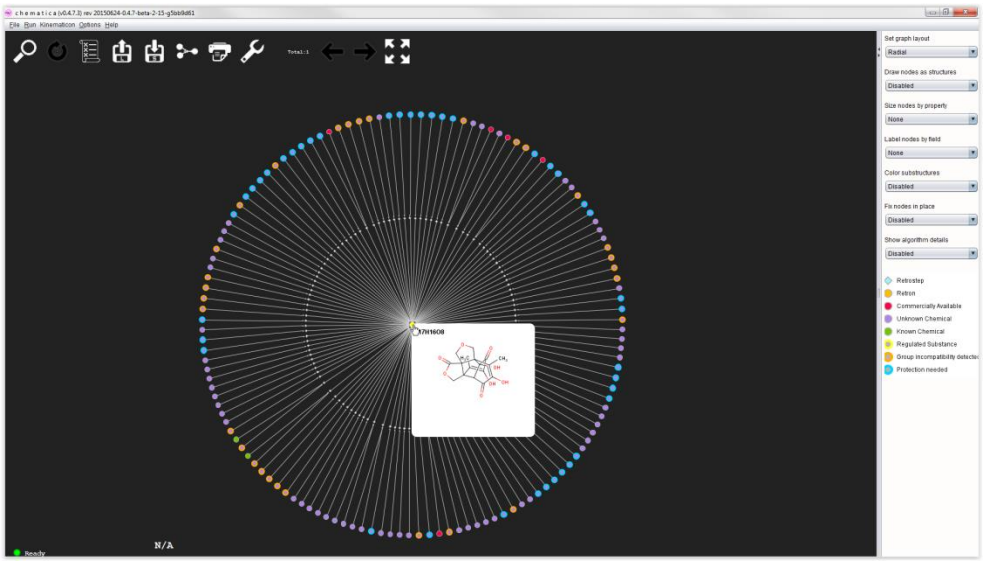


Figure S17.

Of the two top candidates, one entailed synthesis of a lactone moiety via reaction of alcohol with hydroxamate (or another activated carboxylic acid derivative). Although this transformation increases the number of rings, no significant simplification of the overall structure is achieved. (Figure S18). On the other hand, the second-ranked option ("vinylogous aldol") is clearly simplifying the target structure – consequently, it was chosen for further inspection.

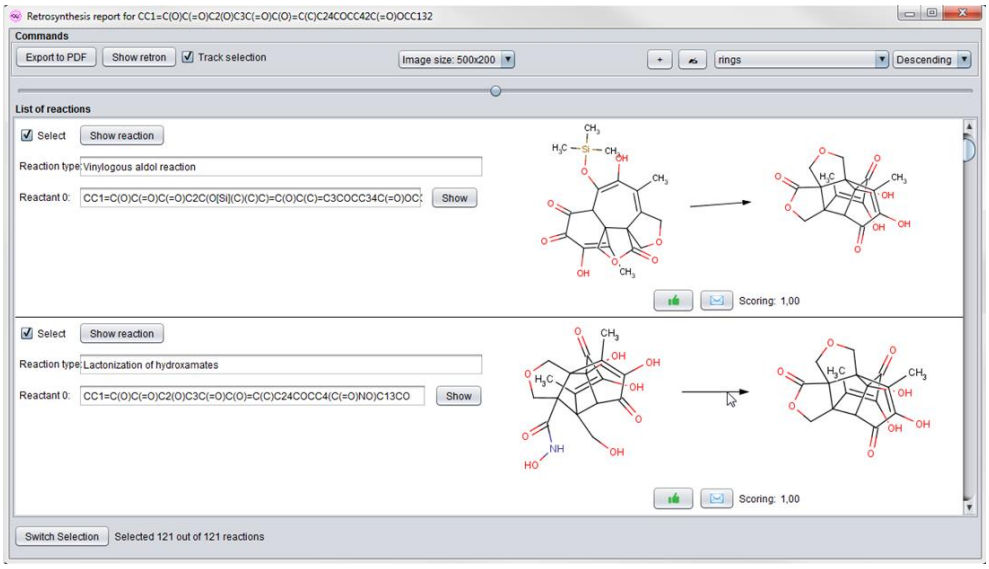


Figure S18.

Subsequently, formation of the silyl enol ether was easily accomplished by the reaction of enolate with TMSCl. The commercial availability of TMSCl was immediately obvious (and confirmed by its node being colored red and also by its top ranking score using the **BUY** function promoting commercially available chemicals, Figure S19).

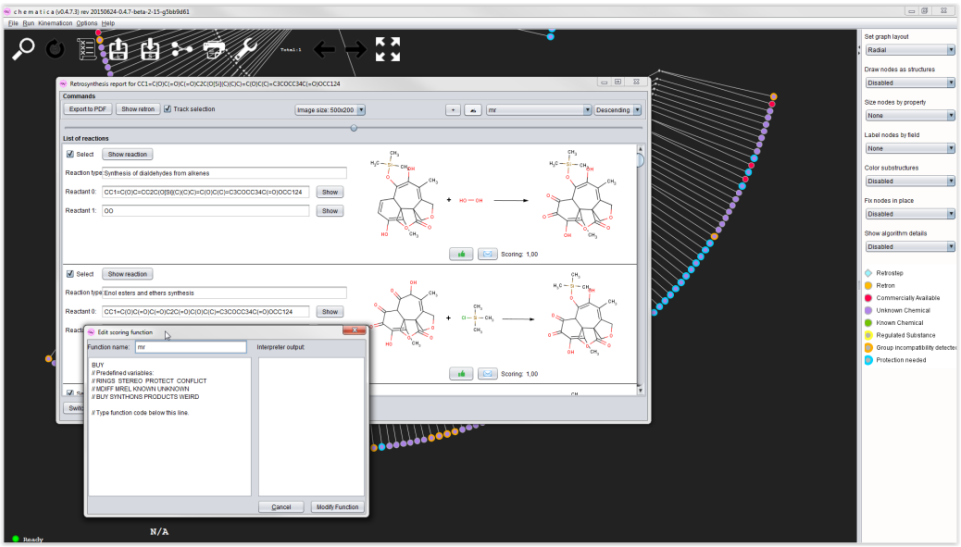


Figure S19.

The next step was a bottle-neck in the search and took most time. Specifically, as no obvious skeletal cuts were identified, we considered possible tautomerism between ketone and enol forms of the carbonyl moieties present (Figure S20).

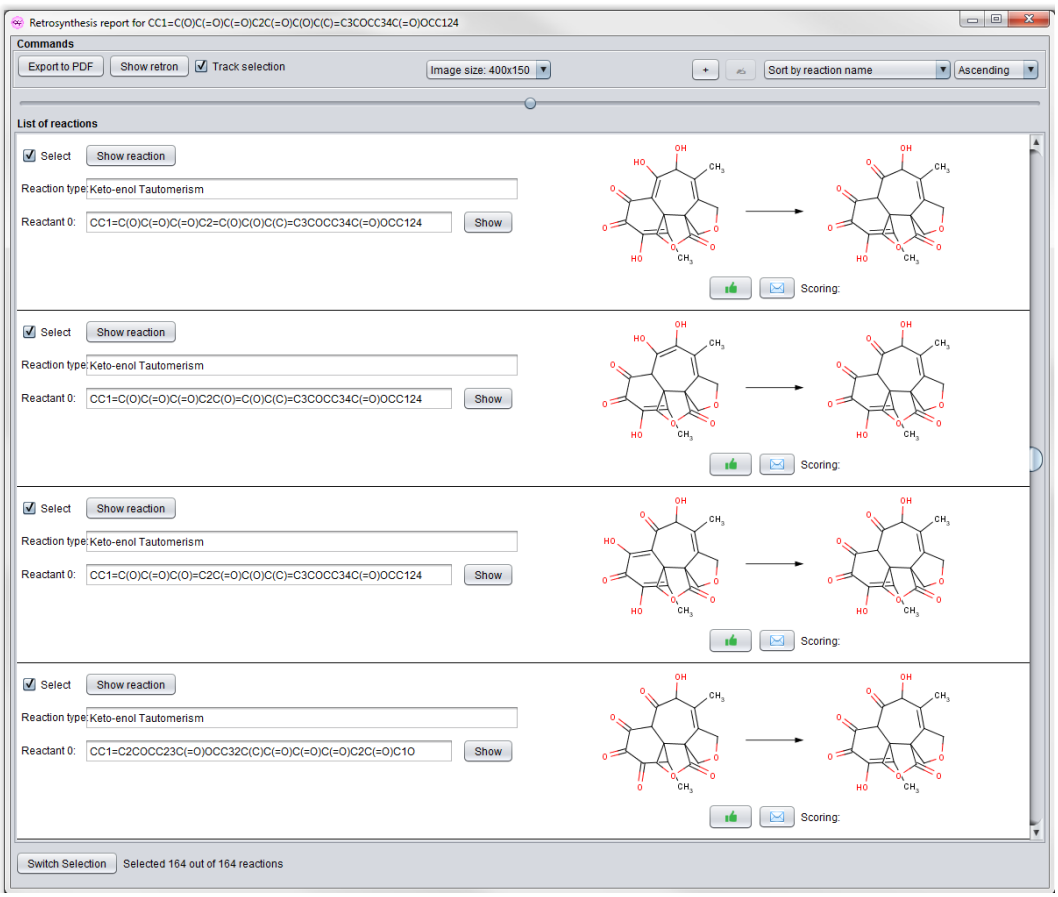


Figure S20.

After inspecting reactions leading to various enolates (total time ca. 30 min), we focused on the "dihydroxyalkene" path (Figure S21) since it led to two promising retro-Claisen condensations (Figure S22) simplifying the seven membered carbocycle to more synthetically amenable *6-5-6* or *5-6-6* ring systems (option denoted by the red arrow in Figure S22).

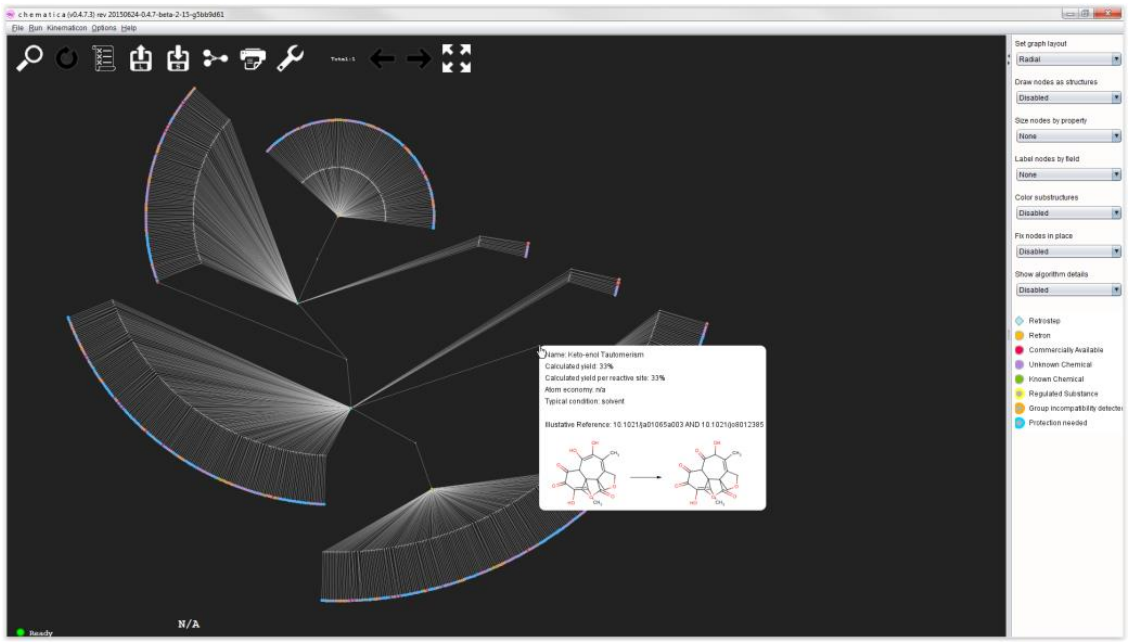


Figure S21.

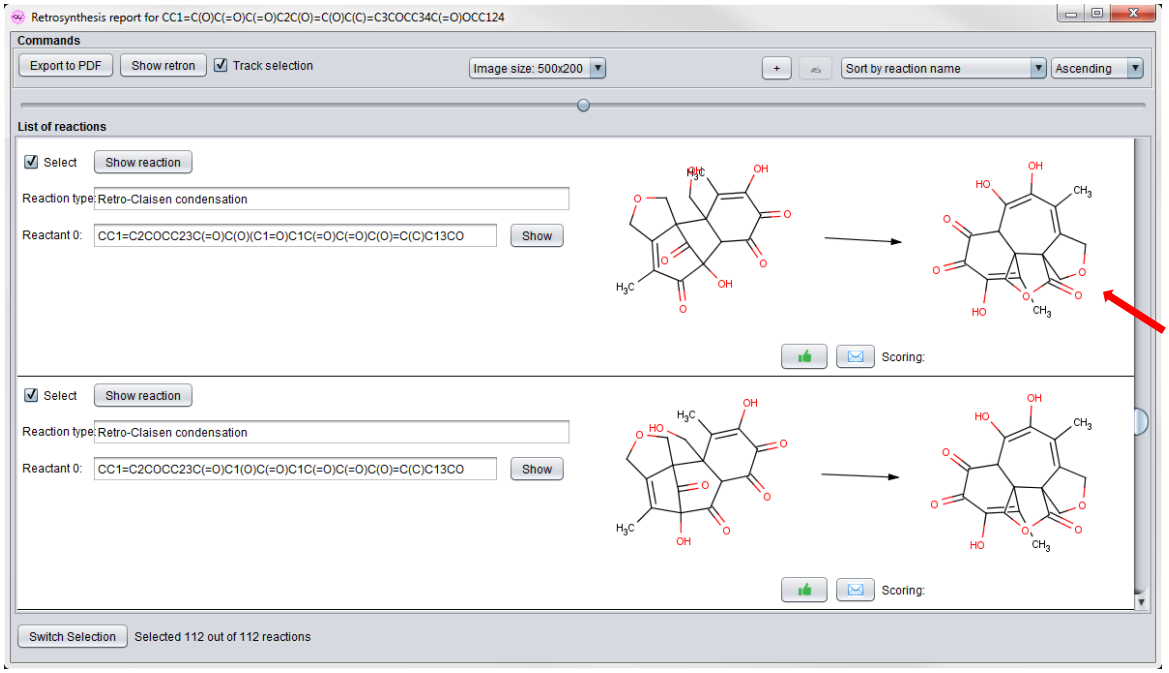


Figure S22.

Focusing on the first of the retro-Claisens in Figure S22, we used either RINGS (as before) or MREL functions (the latter promoting cuts into substrates of similar sizes; very often useful for smaller molecules) to identify as the top choice an elegant oxidative phenol coupling[S7] of two relatively simple substrates (Figure S23), each makeable in one step from known and/or commercially available compounds. In this way, we identified the first plausible pathway.

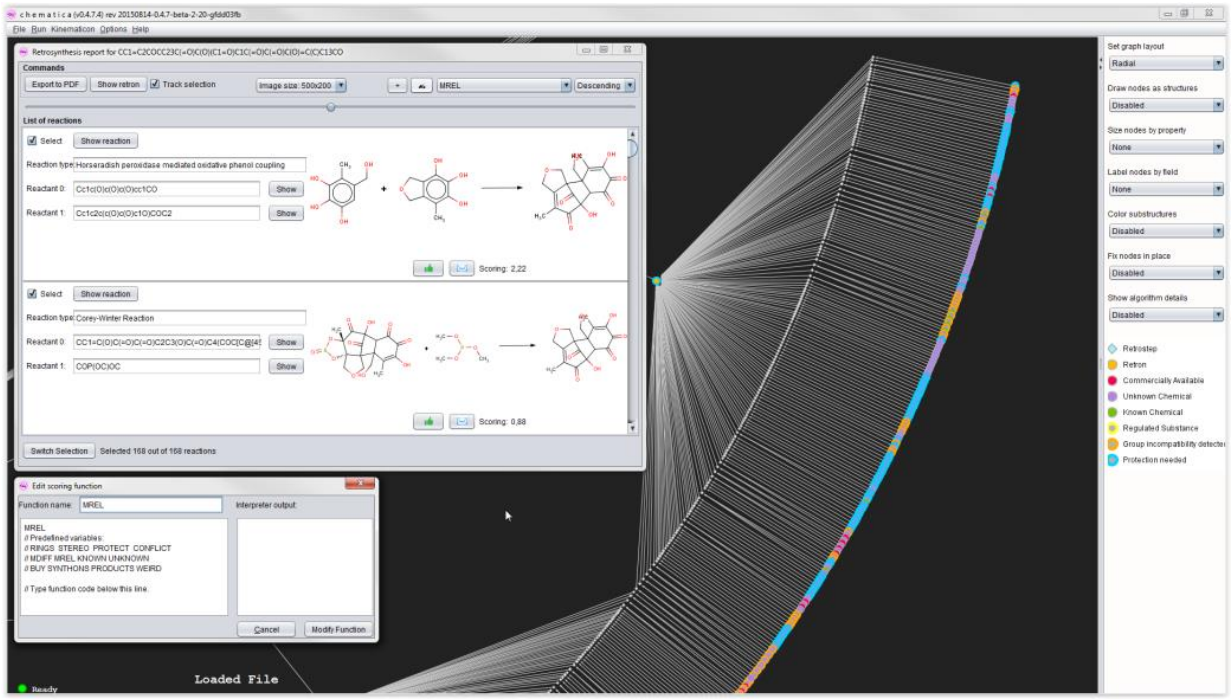


Figure S23.

However, on closer inspection we found the step shown in Figure S23 to be somewhat problematic – in particular, we anticipated that statistical mixture of the coupling products may be obtained due to similar substitution patterns on both coupling partners (similar types of substituents with no clear electronic differentiation influencing benzene rings). We therefore backtracked to the *6-7-(5)-5* intermediate (Figure S24a) and – with the possibility of a later oxidative coupling in mind – looked for any options for which differentiation using traceless

groups would be possible. We rapidly identified the possibility of introducing a carboxylic acid as traceless differentiating group (Figure S24b, red arrow).

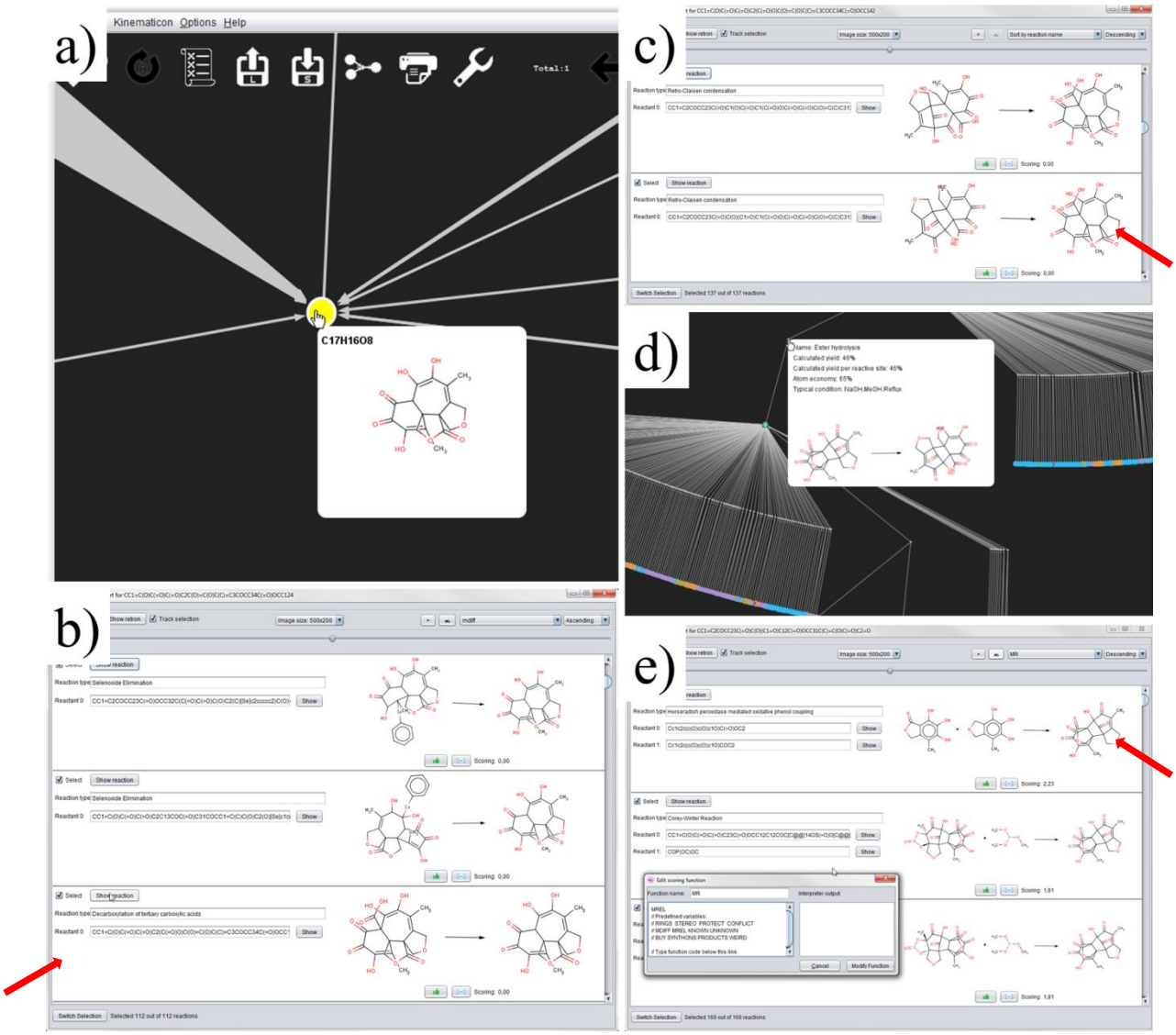


Figure S24.

The rest of the analysis was straightforward and involved, as before, retro-Claisen condensation (Figure 24c), opening of a lactone (Figure 24d), and oxidative coupling (Figure 24e) – again,

guided by the RINGS and/or MREL functions. Finally, the substrates indicated by the red arrow in Figure 24e could both be made from the same, known precursor – one via etherification and one via oxidative lactonization of 1,4 diols (Figure S25). Also as before, these choices were favored by the RINGS function.

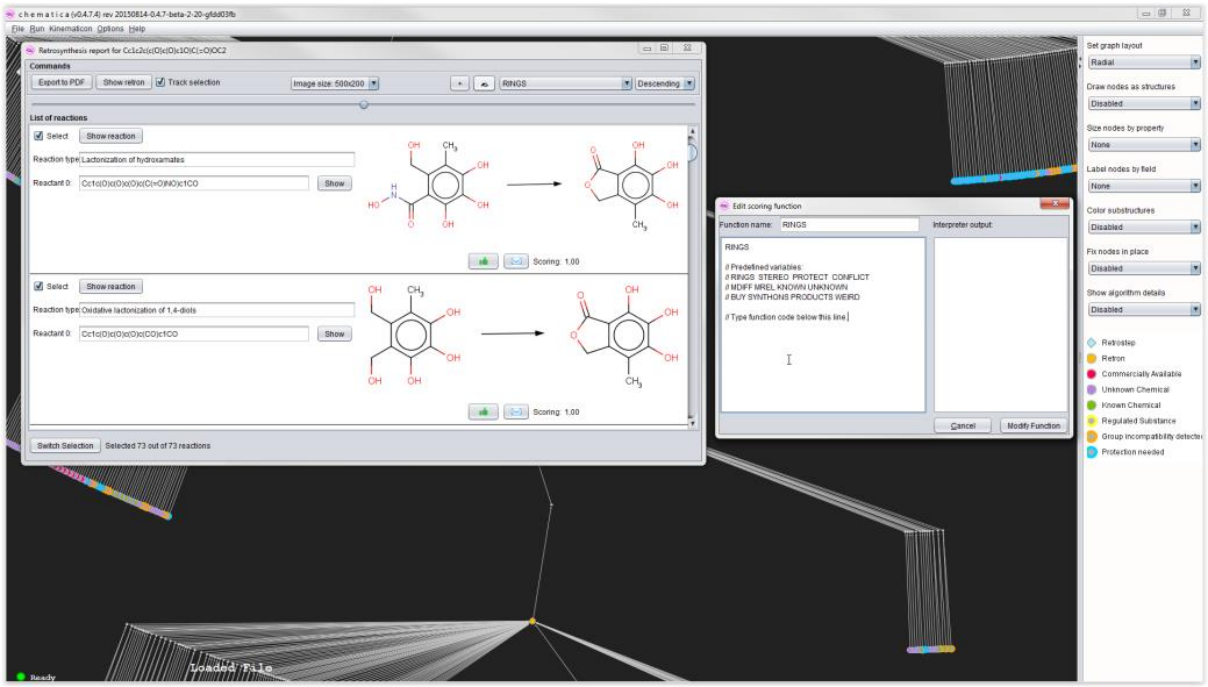


Figure S25.

In summary, we navigated the tree of syntheses shown in Figure S26. These searches were facilitated by the use of, mostly, a simple RINGS function suitable for polycyclic compounds. In the entire process, we inspected few tens of synthetic possibilities with most favorable scores (instead of millions if the searches were completely random/unguided), and completed the synthesis within few-hours time. We ventured into one branch that gave a synthesis that was ultimately problematic but learned from it the oxidative phenol coupling which was useful in the synthetic route ultimately chosen. We wish to stress that while the process required a

“ chemically savvy user,” it was “blind” as we did not know the “correct” pathway beforehand – it was only after we completed the synthetic design that we shared our synthetic solution with Prof. Trauner who then disclosed to us his own approach to the problem. Finally, the synthesis described could also be found by Syntaurus’ fully automated search though it took 1220 iterations and ca. 12 hrs (with chemicals’ scoring functions promoting smaller, buyable, or known substrates, CSF = SMILES_LEN*(1 - BUY)*(1 - KNOWN) and reaction scoring functions heavily penalizing any reactions with cross-reactivity conflicts, RSF = 10.0 + 10000*CONFLICT).

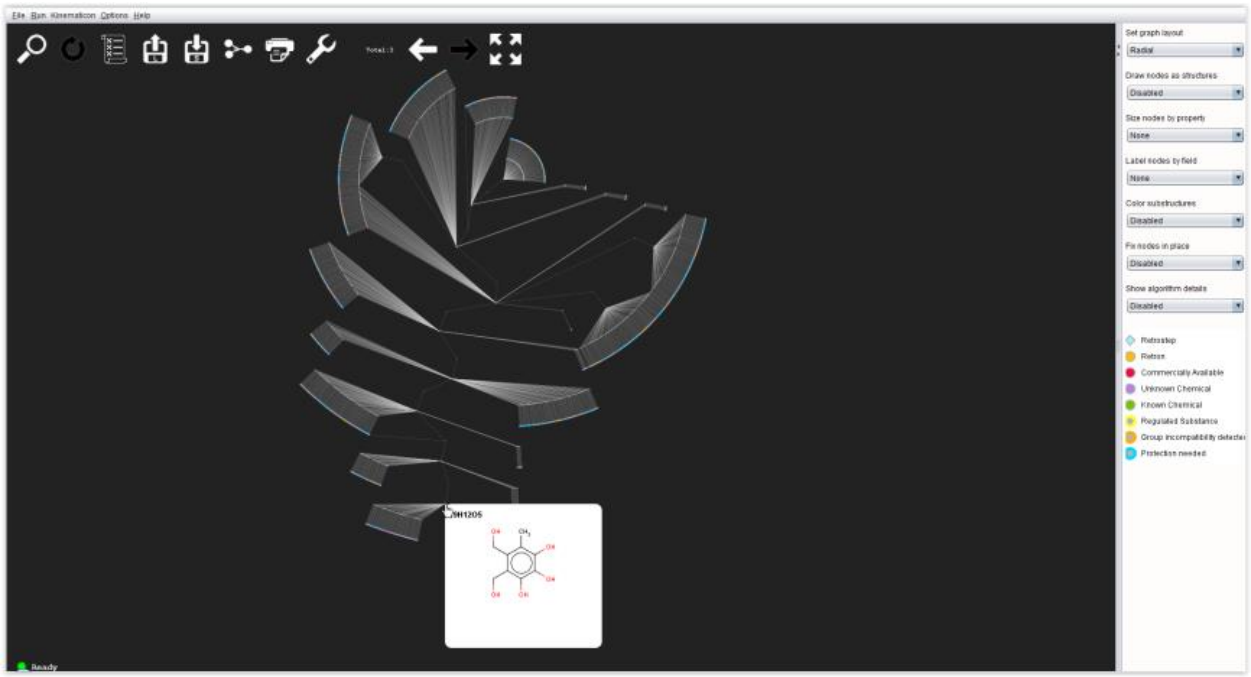


Figure S26.

Section S15. Syntaurus "rediscovers" published pathways.

Figure S27a shows Syntaurus-designed synthesis of a butyrylcholinesterase inhibitor based on the evodiamine scaffold[S8a]. In the first step, commercially available 2-amino-5-hydroxybenzoic acid is *N*-alkylated with methyl iodide[S8b]. Because free hydroxyl and carboxylic acid groups are present, the program warns the user (blue halo on the node) that these groups should be protected (under the given reaction conditions, Syntaurus suggests, respectively, methyl/benzyl/*t*-butyl ester and methyl/methoxymethyl/benzyl ether as most suitable protecting groups). The second building block (cyclic imine) is readily prepared via formylation of commercially available tryptamine under treatment with dimethylformamide [S8c] followed by Bischler-Napieralski reaction[S8d]. Direct imine acylation[S8e] between thus prepared 3,4-dihydro-β-carboline and *o*-anthranilic acid derivative gives the key intermediate, which, following deprotection, yield the target molecule after treatment with commercially available 3-methoxyphenylisocyanate [S8a].

Figure S27b has a simple synthesis leading to CJ-15801, an inhibitor of multiple-drug-resistant *Staphylococcus aureus* strains. Interestingly, the solution Syntaurus suggests matches exactly the pathway published in 2004 by Porco *et. al.* [S9]. In the first step, commercially available (*R*)-pantolactone is opened via reaction with ammonia. The second building block is prepared by hydroiodination of propiolic acid leading exclusively to (*E*)-regioisomer. Synthesis of the target molecule is accomplished via copper mediated coupling of amide with β-iodoacrylate carried out after deprotection of sensitive carboxyl and 1,3-diol moieties.

In the next example, in Figure S27c, Syntaurus rediscovered the published[S10] synthesis of an experimental KOR (kappa-opioid receptor) agonist (see also real-time Movie S4). The program correctly determines the optimal pathway comprising the key, intramolecular Diels-Alder

reaction in step c following the formation of an amide from dicarboxylic acid in step b (the published path started from a related anhydride). The necessary starting material is prepared via amination of an activated alcohol (step a).

Figure S27d shows a route leading to (-) -curvularin, a fungal macrocyclic lactone isolated from *Penicillium* species[S11a], and previously prepared using Friedel-Crafts acylation and alkene metathesis as the macrolactonisation step[S11b]. Syntaurus' strategy resembles recent synthesis[S11c] involving annulation of 1,3-ketoester moiety of another natural product from the *Diplodialide* class with an aryne precursor. Specifically, preparation of the benzyne building block is accomplished starting from 1,3,5-trihydroxybenzene undergoing bromination (step d) followed by subsequent silylation and triflation (steps e, f)[S11d]. Presence of other hydroxyl groups requires their protection and the program suggest methyl ether, methoxymethyl ether, or benzyl ether as the top-three protection group candidates. Synthesis of β-ketoester bearing diplodialide scaffold participating in the final step is carried out in a three-step linear synthesis starting from known methyl 3-oxopent-4-enoate and (*S*)-hept-6-en-2-ol undergoing cross metathesis[S11e] (step a). Further reduction of the thus obtained enone[S11f] (step b) leads to the formation of a saturated cyclisation precursor which is lactonised[S11g] (step c) to give the desired 10-membered ring. Finally, insertion of the aryne precursor into 1,3-ketoester initiated by fluoride anion[S11c,h] yields (step g) the desired target after deprotection of phenolic oxygens.

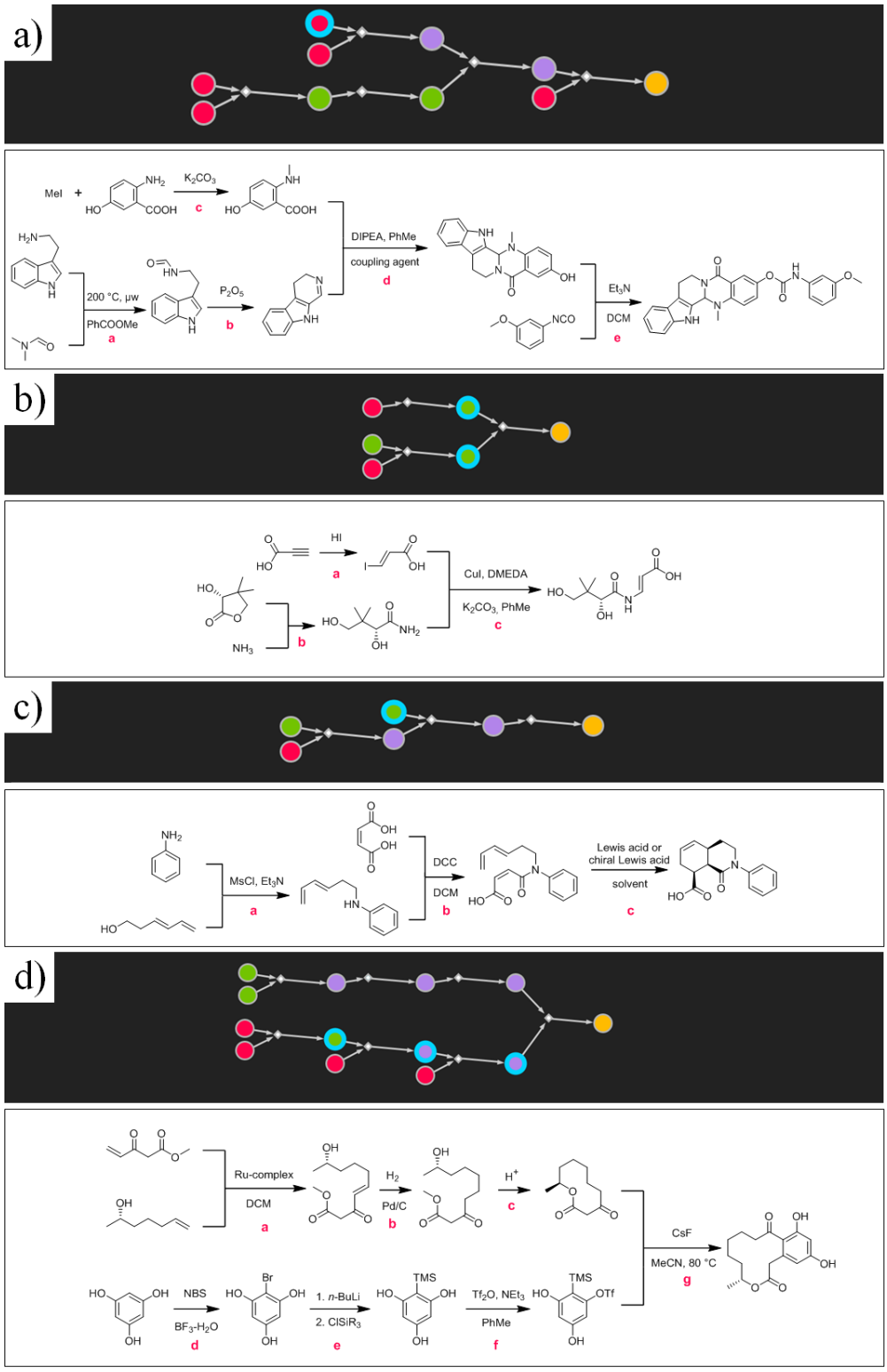


Figure S27. Syntaurus (blindly) rediscovers published syntheses. **a)** Synthesis of evodiamine based butyrylcholinesterase inhibitor[S8a] found with CSF = SMILES_LEN$^{3/2}$ and RSF = 10 +

40·PROTECT + 100·CONFLICT. **b)** Synthesis of CJ-15801[S9] identified with CSF = $SMILES\_LEN^2$ and RSF = 50 + 20·$CONFLICT^2$ + 10·PROTECT. **c)** Synthesis of an experimental KOR (κ-opioid receptor[S10a] identified using CSF = $SMILES\_LEN^{3/2}$ + 50·RINGS and RSF = 100 + 5·PROTECT + 10·CONFLICT. See also Movie S4. **d)** Synthesis of (-)-curvularin[S11a] found using CSF = (SMILES_LEN·RINGS + 10·RINGS·MASS)·(1-BUY)·(1-KNOWN) and RSF = 100 + 10000·CONFLICT. Color coding of nodes: red = commercially available; green = known in the NOC; violet = unknown, yellow = target; blue halos = protection required. In all cases, reaction conditions suggested by the program (in reality, displayed by clicking on the reaction nodes) are listed above the reaction arrows.

Section S16. Additional comments on the Diels-Alder reaction in juvabione synthesis (Figure 20c, step d).

The Diels-Alder step merits some further consideration. On one hand, a similar approach based on Diels-Alder reaction followed by Claisen-Ireland rearrangement and addition of a Grignard reagent to acyl chloride was used in 2000 by Neier *et. al*. for the preparation of juvabione[S12a]. In 1990 Fujii *et. al.* also prepared juvabione using Diels-Alder reaction between the proposed ester and methyl-vinyl ketone[S12b]. We note that although the Diels-Alder step involves an electron poor diene and a neutral dienophile, there were some reports where using either electron poor[S12b] or electron rich[S12c] dienophiles led to the desired products (see also arguments in [S12d]).

Section S17. Quantifying the effectiveness of scoring functions.

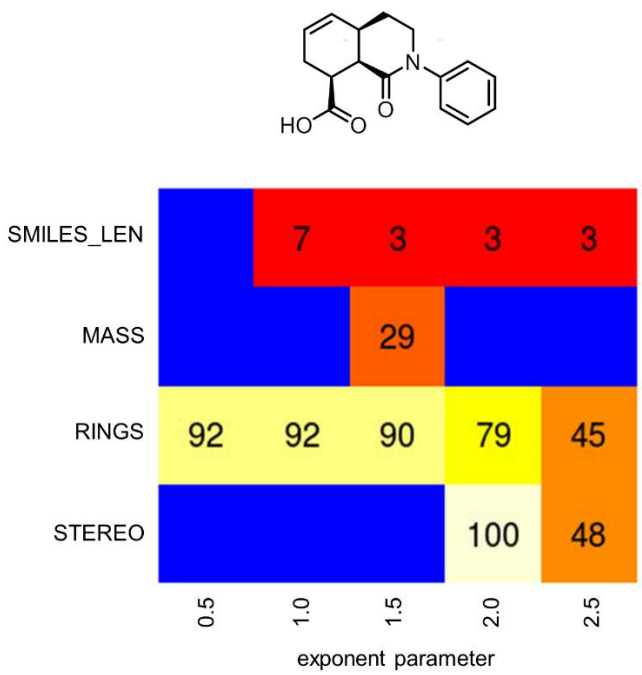


Figure S28. To determine the most "effective" scoring functions (i.e., identifying viable synthetic pathways in minimal numbers of iterations) the influence of different scoring variables available in Syntaurus was examined. In the example shown, RSF was set to a constant value (10) while various chemicals' scoring functions were of the form CSF =VAR^n, where VAR could be STEREO, RINGS, MASS or SMILES_LEN, and the exponent n was varied between 0.5 and 2.5. The searches were considered as "successful" when the pathway denoted in Figure S27c for κ-opioid receptor agonist was identified. The numbers in each (VAR,n) entry correspond to the numbers of algorithm iterations that were needed to find the synthesis. The most effective scoring functions, $SMILES\_LEN^{1.5}$, $SMILES\_LEN^{2}$, and $SMILES\_LEN^{2.5}$ found the synthesis already after three iterations. Blue entries denote that the synthesis was not found even after 100 iterations.

Section S18. Generation of multiple synthetic solutions.

The example in this section illustrates generation of multiple qualitatively different syntheses leading to a tubulin assembly inhibitor, OXi8006. Although preparation of this compound has been published[S13a], its synthesis starting from isovanillin required tedious preparation of starting materials while a shorter pathway[S13b] relied on a Pd-mediated carbonylative cyclisation-arylation sequence using highly toxic and operationally problematic CO. In about 3 minutes, Syntaurus generated tens of alternative, highly convergent plans, of which some are shown in Figure S29 and are discussed in detail in the figure caption.

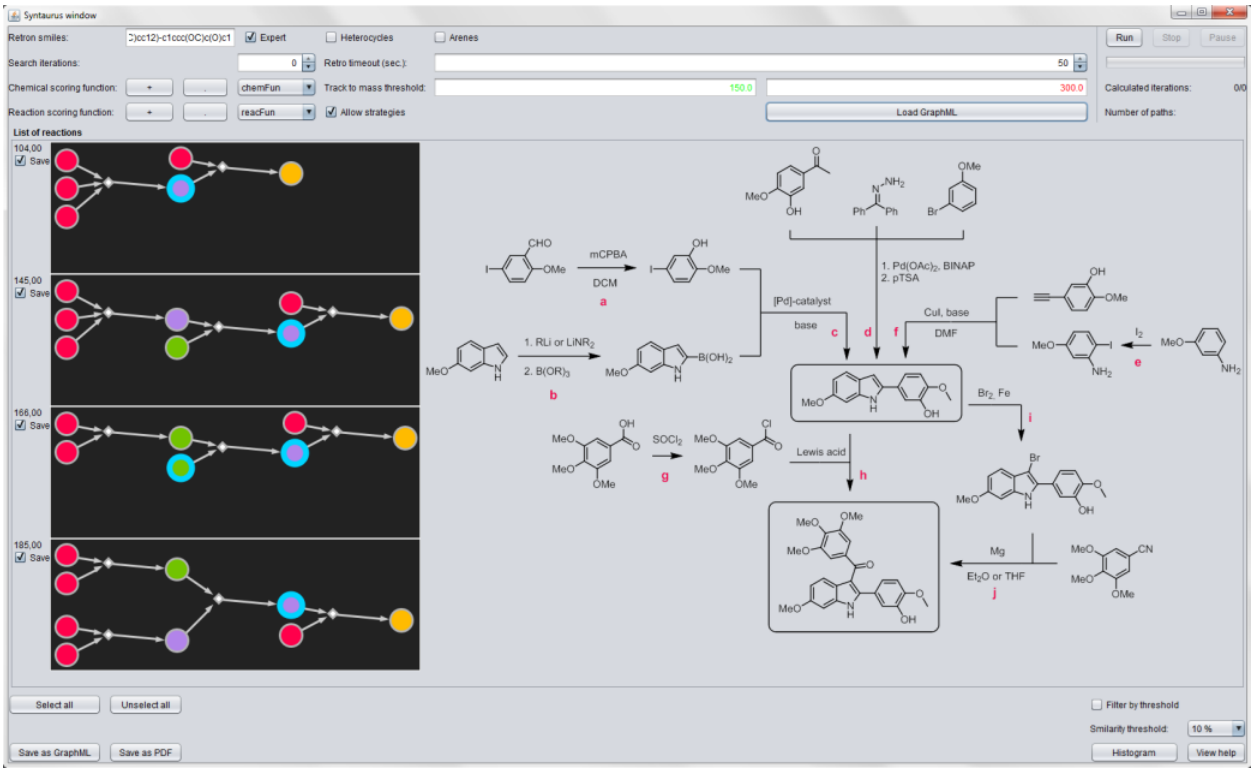


Figure S29. Multiple synthetic solutions leading to the same target. Syntaurus' window shows several top-scoring syntheses of OXi8006[S14a,b] (ca. 50 synthetic plans designed within 3 min.

using CSF = SMILES_LEN$^{3/2}$ + 20·RINGS and RSF = 40 + 10·PROTECT + 50·CONFLICT). In an elegant step labelled "d," Syntaurus proposed synthesis of the key 2-arylindole intermediate employing three-component coupling of appropriate acetophenone, aryl bromide and benzophenone hydrazone as nitrogen source (step d)[S14c]. Another promising possibility starts from 6-methoxyindole undergoing facile metalation at C-2 under Katritzky's carbon dioxide protection/activation methodology[S14d]. Subsequent quench with trialkoxy borate followed by hydrolysis gives boronic acid (step b)[S14e,f] necessary for Suzuki coupling which can be accomplished under a wide range of conditions[S14g] (choice of palladium catalyst, base, solvent, additives, and boronic acid derivatives including esters, MIDA-esters, etc.). Proposed synthesis of the aryl iodide partner involving conversion of benzaldehyde to phenol[S14h] allows for starting from a commercially available substrate (though a bit complicated one and makeable in the Chematica's network module from simpler materials). The third proposed pathway makes the key intermediate via the Castro protocol in step f [S14i,j] starting from (4-methoxy-3-hydroxyphenyl)acetylene and 3-methoxyaniline undergoing regioselective iodination[S14k] in step e. Conversion of the 2-arylindole intermediate into target molecule can be accomplished via Friedel-Crafts acylation[S13a] (step h) with acyl chloride prepared in step g from commercially available trimethylgallic acid. Another possibility is C-3 bromination[S14l] (step i) followed by transformation to organolithium or Grignard reagent[S14m] and reaction with appropriate benzonitrile[S14n]. Color coding of nodes: red = commercially available; green = known in the NOC; violet = unknown, yellow = target; blue halos = protection required.

Supplementary References:

[S1] B. Kowalczyk, K. J. M. Bishop, S. K. Smoukov, B. A. Grzybowski, *J. Phys. Org. Chem.* **2009**, *22*, 897–902.

[S2] V. Viswanadhan, A.K. Ghose, G.R. Reyankar, R.K. Robins, *J. Chem. Inf. Comp. Sci.* **1989**, *29*, 163-172.

[S3] a) K. Binder, D.W. Heermann, *Monte Carlo Simulation in Statistical Physics: An Introduction*, Springer, Heidelberg, **2010**; b) J. M. Hammersley, D. C. Handscomb, *Monte Carlo Methods*, Methuen & Co., London, **1975**; c) L. Ingber, *Math. Comp. Model.* **1993**, *18*, 29-57.

[S4] a) S. Kobayashi, *Chem. Soc. Rev.* **1999**, *28*, 1-15; b) J.-C. Wasilke, S. J. Obrey, R. T. Baker, C. G. Bazan, *Chem. Rev.* **2005**, *105*, 1001-1020; c) A. Domling, *Chem. Rev*. **2006**, *106*, 17-89; d) D. Enders, C. Wang, J. W. Bats, *Angew. Chem. Int. Ed.* **2008**, *47*, 7539-7542; e) M. Galibert, O. Renaudet, P. Dumy, D. Boturyn, *Angew. Chem. Int. Ed.* **2011**, *50*, 1901-1904; f) A. Grirrane, A. Corma, H. Garcia, *Science* **2008**, *322*, 1661-1664; g) N. Z. Burns, P. S. Baran, R. W. Hoffmann, *Angew. Chem. Int. Ed.* **2009**, *48*, 2854-2867; h) P. A. Wender, V. A. Verma, T. J. Paxton, T. H. Pillow, *Acc. Chem. Res.* **2008**, *41*, 40-49; i) T. Newhouse, P. S. Baran, R. W. Hoffmann, *Chem. Soc. Rev.* **2009**, *38*, 3010-3021.

[S5] a) Amgen Inc.. Patent: US2010/331293 A1, **2010** b) Amgen Inc.; M. Brown, Y. Chen, T. D. Cushing, F. Gonzalez Lopez De Turiso, X. He, T.J. Kohn, J.W. Lohman, V. Pattaropong, J. Seganish, Y. Shin, J.L. Simard, Patent: WO2010/151737 A2, **2010** b) S. J. Park, K. H. Min, C. L. Lee, *Respirology* **2008**, *13*, 764–771.

[S6] a) T. Harder, P. Wessig, J. Bendig, R. Stösser, *J. Am. Chem. Soc.* **1999**, *121*, 6580–6588; b) J. A. H. MacBride, S. Kanoktanaporn, *J. Chem. Res., Miniprint*, **1980**, *6*, 2901-2910; c) M. A. Sierra, M. C. de la Torre, *Angew. Chem. Int. Ed*. **2000**, *39*, 1538-1559; d) N. S. Mani, C. A.

Townsend, *J. Org. Chem.* **1997**, *62*, 636–640; e) M. Bertrand, H. Monti, K. Chang Huong, *Tetrahedron Lett.* **1979**, *20*, 15–18; f) M. Iyoda, S. M. H. Kabir, A. Vorasingha, Y. Kuwatani, M. Yoshida, *Tetrahedron Lett.* **1998**, *39*, 5393–5396.

[S7] S. Sang, J. D. Lambert, S. Tian, J. Hong, Z. Hou, J. H. Ryu, R. E. Stark, R. T. Rosen, M. T. Huang, C. S. Yang, et al., *Bioorganic Med. Chem.* **2004**, *12*, 459–467.

[S8] a) U. Huan, B. Klin, F. H. Darras, J. Heilmann, M. Decker, *Eur. J. Med. Chem.* **2014**, *81*, 15–21; b) Q.-G. Ji, D. Yang, Q. Deng, Z.-Q. Ge, L.-J. Yuan, *Med. Chem. Res.* **2014**, *23*, 2169-2177; c) T. Lebleu, H. Kotsuki, J. Maddaluno, J. Legros, *Tetrahedron Lett.* **2014**, *55*, 362–364; d) M. Bertrand, G. Poissonnet, M. H. Théret-Bettiol, C. Gaspard, G. H. Werner, B. Pfeiffer, P. Renard, S. Léonce, R. H. Dodd, *Bioorg. Med. Chem.* **2001**, *9*, 2155–2164; e) W. P. Unsworth, C. Kitsiou, R. J. K. Taylor, *Org. Lett.* **2013**, *15*, 258–261.

[S9] C. Han, R. Shen, S. Su, J. A. Porco, *Org. Lett.* **2004**, *6*, 27–30.

[S10] S. R. Slauson, R. Pemberton, P. Ghosh, D. J. Tantillo, J. Aubé, *J. Org. Chem*. **2015**, *80*, 5260-5271.

[S11] a) Y. Yao, M. Hausding, G. Erkel, T. Anke, U. Förstermann, H. Kleinert, *Mol. Pharmacol.* **2003**, *63*, 383–391; b) S. Elzner, D. Schmidt, D. Schollmeyer, G. Erkel, T. Anke, H. Kleinert, U. Förstermann, H. Kunz, *ChemMedChem* **2008**, *3*, 924–939 and references cited therein; c) P. M. Tadross, S. C. Virgil, B. M. Stoltz, *Org. Lett.* **2010**, *12*, 1612–1614; d) H. Yoshida, T. Morishita, J. Ohshita, *Chem. Lett.* **2010**, *39*, 508–509; e) P. Dewi-Wülfing, J. Gebauer, S. Blechert, *Synlett* **2006**, 487–489; f) G. Mehta, K. Pallavi, J. D. Umarye, *Chem. Commun.* **2005**, 4456–4458; g) T. Ishida, K. Wada, *J. Chem. Soc. Perkin Trans.* **1979**, 323–327; h) U. K. Tambar, B. M. Stoltz, *J. Am. Chem. Soc.* **2005**, *127*, 5340–5341.

[S12] a) N. Soldermann, J. Velker, O. Vallat, H. Stoeckli-Evans, R. Neier, *Helv. Chim. Acta* **2000**, *83*, 2266–2276; b) M. Fujii, T. Aida, M. Yoshihara, A. Ohno, *Bull. Chem. Soc. Jpn.* **1990**, *63*, 1255–1257; c) W. Poly, D. Schomburg, H. M. R. Hoffmann, *J. Org. Chem.* **1988**, *53*, 3701–3710; d) Analysis of the frontier orbital energies of the proposed substrates places EWG-diene/alkyl-substitued dienophile (ΔE = 8.8 eV) closer to EWG-diene/enamine (ΔE = 8.7 eV; neglecting influence from methyl groups) than to EWG/methyl-vinyl ketone pair (ΔE = 9.3 eV) suggesting that the proposed reaction might be feasible. See K. N. Houk, *J. Am. Chem. Soc.* **1973**, *95*, 4092–4094.

[S13] S. R. Slauson, R. Pemberton, P. Ghosh, D. J. Tantillo, J. Aubé, *J. Org. Chem*. **2015**, *80*, 5260-5271.

[S14] a) M. B. Hadimani, M. T. MacDonough, A. Ghatak, T. E. Strecker, R. Lopez, M. Sriram, B. L. Nguyen, J. J. Hall, R. J. Kessler, A. R. Shirali, L. Liu, C. M. Garner, G. R. Pettit, E. Hamel, D. J. Chaplin, R. P. Mason, M. L. Trawick, K. G. Pinney, *J. Nat. Prod.* **2013**, *76*, 1668–1678; b) B. L. Flynn, E. Hamel, M. K. Jung, *J. Med. Chem.* **2002**, *45*, 2670–2673; c) S. Wagaw, B. H. Yang, S. L. Buchwald, *J. Am. Chem. Soc.* **1998**, *120*, 6621–6622; d) A. R. Katritzky, K. Akutagawa, *Tetrahedron Lett.* **1985**, *26*, 5935–5938; e) E. Vazquez, I. W. Davies, J. F. Payack, *J. Org. Chem.* **2002**, *67*, 7551–7552; f) R. Dandu, M. Tao, K. A. Josef, E. R. Bacon, R. L. Hudkins, *J. Heterocycl. Chem.* **2007**, *44*, 437–440; g) N. S. Kumar, E. M. Dullaghan, B. B. Finlay, H. Gong, N. E. Reiner, J. J. P. Selvam, L. M. Thorson, S. Campbell, N. Vitko, A. R. Richardson, R. Zoraghi, R. N. Young, *Bioorg. Med. Chem.* **2014**, *22*, 1708–1725; h) S. Centonze-Audureau, F. H. Porée, J. F. Betzer, J. D. Brion, A. Pancrazi, J. Ardisson, *Synlett* **2005**, 6, 981–985; i) R. D. Stephens, C. E. Castro, *J. Org. Chem.* **1963**, *28*, 3313–3315; j) X. Yu, E.-J. Park, T. P. Kondratyuk, J. M. Pezzuto, D. Sun, *Org. Biomol. Chem.* **2012**, *10*, 8835-8847; k) H. Shen, K. P.

C. Vollhardt, *Synlett* **2012,** 23, 208-214; l) V. Bocchi, G. Palla, *Synthesis* **1982**, *1982*, 1096–1097; m) M. Amat, S. Hadida, S. Sathyanarayana, J. Bosch, *J. Org. Chem.* **1994**, *59*, 10–11; n) K. Harikrishna, A. Rakshit, I. S. Aidhen, *Eur. J. Org. Chem.* **2013**, 4918–4932.

MOVIE DESCRIPTIONS

Movie 1. Basic „travel" over the NOC (cf. Figure 6 in the main text). This real-time movie illustrates the simplest exploration of the Network of Organic Chemistry – namely, visualization of all reactions in the synthetic vicinity of a desired molecule (here, 3-benzoylindole derivative). In the movie, only few nodes are expanded but the network thus generated is already quite complex. This is precisely the point of this illustration: To show how complex the network of synthetic possibilities can become within just few synthetic steps. To search through such networks, one needs algorithms described in the main text.

Movie 2. Most cost-effective synthesis of Taxol found in the NOC (cf. Figure 8 in the main text). The search for the said synthesis starts with specifying the target (here, Taxol), the desired optimization criterion (here, „Cost"), and the length of allowable syntheses (here, up to 50 steps). The movie is in real time and the entire search takes ca. 7 seconds within which the program considers hundreds of millions of possible synthetic plans (exact numbers displayed from 00:37 to 00:48 sec). Different display modalities are used (00:48) including 2D molecular drawings and 3D models of individual molecules (01:00-01:20). Another display modality („unconstrained nodes") is introduced at 01:25 after which reaction arrows are colored according to the year in which the syntheses were published (till 02:04). Next, the nodes are sized according to molecular mass (sizing according to „synthetic popularity" is also available; cf. main text). At 02:29, an option to display all molecules as 2D cartoons is activated.

Movie 3. Retrosynthesis of Aripiprazole (cf. Figure 14 in the main text). Synthesis of aripiprazole is designed using step-by-step searches guided by scoring functions. The search is set up and at 00:06 sec the program returns 72 possible syntheses for the first retrosynthetic step. Various filters can be used to display only some of these options; here, we deselect most of them with the exception of „Cut Into Smaller Fragments" (00:10 sec). The options meeting this condition are displayed at 00:19 with nodes colored red for commercially available chemicals, green for molecules known in the NOC, and violet for unknown molecules. Reactions displayed in the form of a list (00:25) are sorted according to a simple scoring function (favoring commercially available and known substrates with similar size; input of the function is shown from 00:29 to 00:47). The two display modes „talk to one another" (i.e., highlighting in the list also highlights the specific reaction in the network; 00:52). Already in this first search step, we identify an option starting from two known substrates (00:52) – the synthesis of the known substance could be pursued from the NOC's module (cf. Movie 2) but here we continue retrosynthesis, until finding two red nodes. The green node is thus further expanded into 51 new options (00:55) of which some are filtered out using the criterion „Cut into Smaller Fragments". We already find one buyable (red) node and one known substrate but decide to expand further the synthesis of the known compound. Repeating the procedure (analogously to the preceding steps) guides us toward simple substrates that engage in the formation of an oxime followed by the Beckmann rearrangement. Synthesis of second building block is elaborated in similar manner starting from the expansion of *N*-arylpiperazine at 2:23 and leading to commercially available piperazine and dichlorobromobenzene (2:36). Although the latter is commercially available, we decided to prepare it from cheaper dichloroaniline via Sandmeyer reaction. The entire pathway is thus completed in six steps.

Movie 4. Syntaurus' fully automated design of a synthesis leading to kappa-opioid receptor (cf. Figure S27c). After choosing the target (here, copy-pasting its SMILES, 00:08), the Syntaurus' main window is opened (00:13) where the user specifies the Chemicals Scoring Function (00:15-00:29; CSF = $SMILES\_LEN^{1.5}$ + 50*RINGS) and then the Reaction Scoring Function (00:30-00:46; RSF = 100 + 5*PROTECT + 10*CONFLICT). The number of iterations is set to „0" (00:56) – a special character telling Syntaurus to search until stopped by the user. Functions are set, the searches are set up without strategies, the molecular weights of end-point substrates are limited to, at most, 150 (both for known/green and buyable/red chemicals), and the search commences at 01:00. The number of iterations performed and the number of pathways found are displayed in the upper-right corner. First 6 viable syntheses are found already after 4 iterations (01:19) but we let the algorithm to search more, until after 8 iterations some 26 viable pathways are found (01:38). Everything you see is in real time! These 26 synthetic plans are displayed at 01:49 and for one of them all steps are displayed until 02:09 (the synthesis is the same as in Figure S27c). Scrolling down shows more syntheses that are progressively longer and have worse scores (e.g., notice nodes with orange halos for which serious cross-reactivity conflicts are identified). At 02:24 the user sets a threshold for the similarity of pathways such that only significantly different ones are displayed. Finally, at 02:35, a synthesizability histogram is displayed (cf. Section 2.7 in the main text) which shows how many paths having given scores were identified. The rest of the movie is just rescaling the axes of the histogram plot.

Angewandte International Edition Chemie

Reaction Yields

International Edition: DOI: 10.1002/anie.201503890
German Edition: DOI: 10.1002/ange.201503890

A Priori Estimation of Organic Reaction Yields**

*Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, Sara Szymkuć, Piotr Dittwald, Karol Molga, and Bartosz A. Grzybowski**

Abstract: *A thermodynamically guided calculation of free energies of substrate and product molecules allows for the estimation of the yields of organic reactions. The non-ideality of the system and the solvent effects are taken into account through the activity coefficients calculated at the molecular level by perturbed-chain statistical associating fluid theory (PC-SAFT). The model is iteratively trained using a diverse set of reactions with yields that have been reported previously. This trained model can then estimate a priori the yields of reactions not included in the training set with an accuracy of ca. ±15%. This ability has the potential to translate into significant economic savings through the selection and then execution of only those reactions that can proceed in good yields.*

Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

A Priori Estimation of Organic Reaction Yields**

*Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, Sara Szymkuć, Piotr Dittwald, Karol Molga, and Bartosz A. Grzybowski**

anie_201503890_sm_miscellaneous_information.pdf

Supporting Information

S.1. Additional considerations

S.1.1. Statistics of reactions under kinetic vs. thermodynamic control.

In our model, thermodynamic considerations are relevant at the stage of calculating initial/guess free energies which are then subject to multidimensional optimization. The thermodynamic portion of the analysis is important insofar as it produces initial/guess group contributions which, as explained in the main text, ensure convergence of the training data and predictability for the test-set data. This importance of the thermodynamic basis of the model can reflect the fact that majority of chemical reactions are under thermodynamic control. Even though this statement might sound "intuitive," the statistics of organic reactions under thermodynamic vs. kinetic control have never been, to the best of our knowledge, analyzed or even approximated in any systematic fashion. Hence, we investigated this issue further.

We begin by noting that even though it might be tempting to perform such analyses based on reaction temperatures or times (using "common wisdom" that heating a reaction mixture for a long time must surely overcome all kinetic barriers), such arguments are very naïve and misleading – indeed, there are several classes of reactions which, even for long reaction times and relatively high temperatures can lead to kinetic products (a prominent example being Diels-Alder leading to endo products).

Therefore any realistic analysis requires examination of reactions class by class – and in specific borderline cases, examination of individual reactions. We performed such searches of chemical literature using the Reaxys database. The results of these analyses substantiate the preponderance of thermodynamic vs. kinetic control in organic chemistry, with the counts of reactions that can be unambiguously assigned as kinetically controlled in the thousands vs. tens of millions of total reactions reported.

First, we ran Reaxys searches on reaction classes for which there is consensus as to them being under thermodynamic control. Some typical counts are:

- reduction of aromatic nitro group to amine group 81 000
- Suzuki coupling between Cl/Br/I and boronic acid/ester 26 670
- synthesis of boronic acids from bromoarenes 13 310
- condensation of carboxylic acids with amines 63 710
- hydrolysis of esters to carboxylic acids 264 800
- reduction of ketones to secondary alcohols 15 750
- aromatic electrophilic substitution such as:
    - acetylation 2690
    - bromination 13960

Please note that there are many more populous classes (e.g., multiple aromatic substitutions with exception of kinetically controlled C-1 sulphonation of naphthalene) – the problem with such searches is, unfortunately, that Reaxys search engine simply times out for very high numbers of hits.

Next, we inspected classes of reactions which are known to proceed under either thermodynamic or kinetic control. For these reactions, we inspected Reaxys entries manually and in cases of any ambiguities, consulted the original literature sources:

(i) Wittig-type Reaction

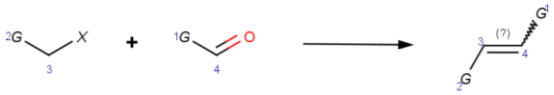


Figure S.1.

The general Reaxys query illustrated in Figure S.1 above gave **1249** reactions which were all manually inspected. During this inspection, only the results with specified regiochemistry of the product and with conditions matching the Wittig-type reaction were taken into account. Screenshots in Figures S.2 and S.3 have the examples of the thermodynamic (E) and kinetic (Z) products. The analysis revealed the preponderance of the thermodynamic over kinetic products (**575 vs 75** hits).

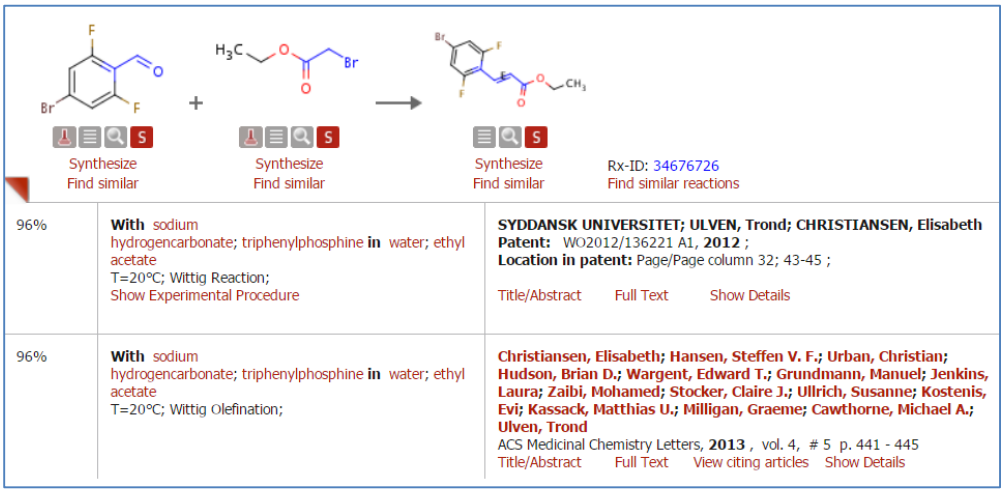


Figure S.2. An example of a Reaxys-retrieved Wittig reaction leading to a thermodynamic product.

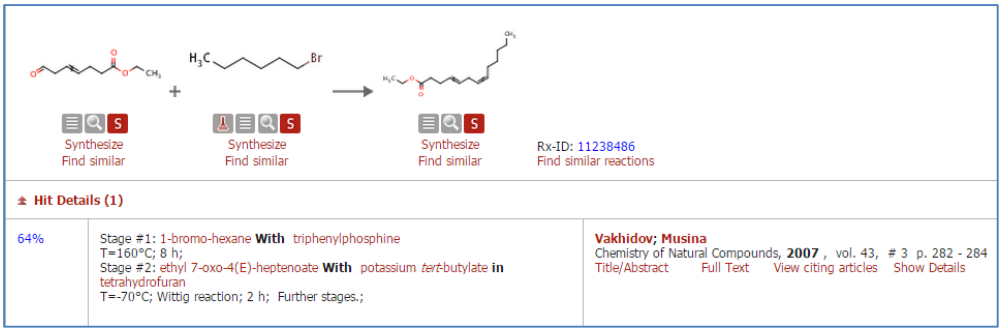


Figure S.3. An example of a Reaxys-retrieved Wittig reaction leading to a kinetic product.

(ii) Synthesis of silyl enol ethers from ketones

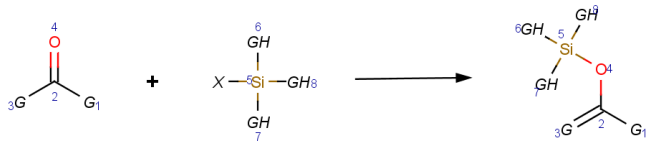


Figure S.4.

The general Reaxys query illustrated in Figure S.4 above gave **3930** reactions which were all manually inspected – examples that did not represent the searched reaction or in which the starting material was not a ketone (but a carbonyl compound of a different type, like an ester or thioester) were filtered out. After filtering, products were classified as thermodynamic when the enol's double bond was formed on the more substituted side of a molecule, when enols formed from symmetrical substrates, and for enols with only one possible deprotonation site. Products were classified as kinetic when the enol double bond was formed on the less substituted side of a molecule. In addition, reactions where

formation of a silyl enol ether could be accompanied by migration of an existing double bond (α,β-unsaturated ketone) were also investigated and classified.

Screenshots in Figures S.5 and S.6 have the examples of the kinetic and thermodynamic products, respectively. The analysis of the entire set of reactions revealed the preponderance of the thermodynamic over kinetic products (**1436 vs 415** hits).

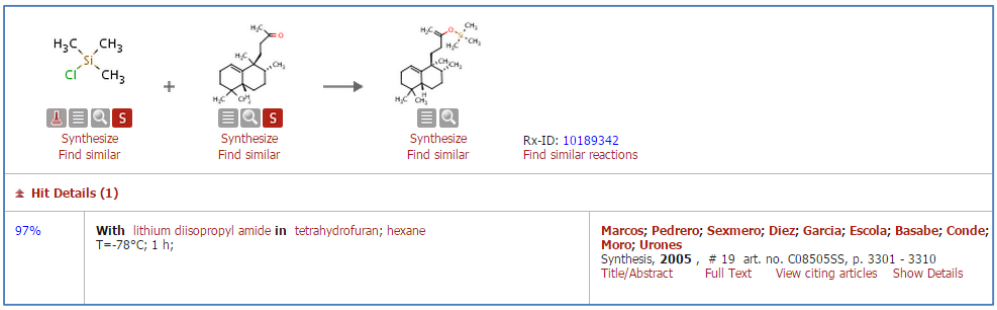


Figure S.5. An example of a Reaxys-retrieved synthesis of silyl enol ethers from ketones leading to a kinetic product.

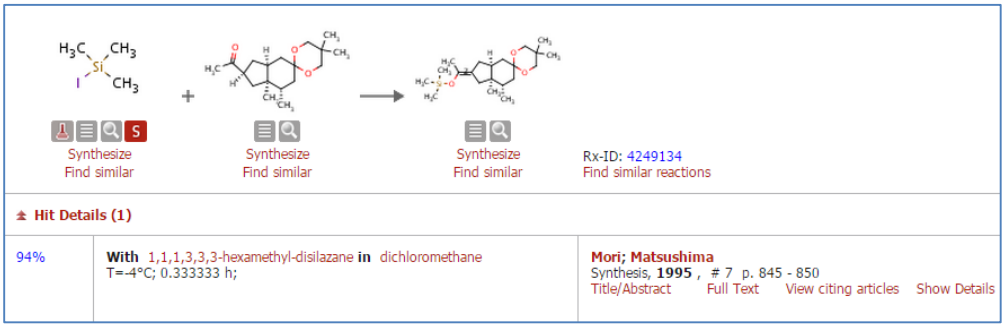


Figure S.6. An example of a Reaxys-retrieved synthesis of silyl enol ethers from ketones leading to a thermodynamic product.

3. Diels-Alder reaction.

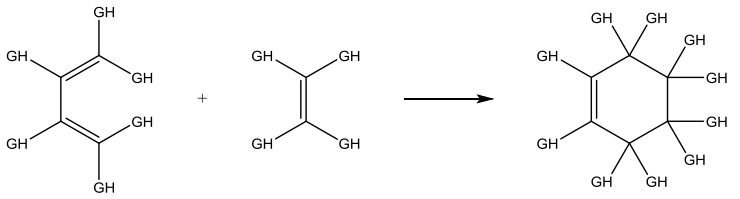


Figure S.7. Reaxys query for all reactions relevant to the Diels-Alder product pattern. GH denotes any atom including hydrogen.

For the general query denoting the Diels-Alder reaction (cf. Figure S.7 above), Reaxys database returned 12 184 results. Since as of March 2015 Reaxys stores ca. 36 million reactions, this number of hits corresponds to 0.34‰ of all transformations. Filtering by "Reaction Type" (RXD.TYP) field and limiting to reactions to those labelled as "Diels-Alder reaction", " Diels Alder cycloaddition", "cycloaddition," or "asymmetric Diels-Alder reaction" narrowed the results to 1436 reactions. Manual examination of first 500 entries ranked by Reaxys-Rank algorithm (reactions marked as the "best described") allowed us to classify each reaction into one of three groups: reactions with unspecified stereochemistry (for which assignment of thermodynamic vs. kinetic products was impossible; see examples in Figure S.8), reactions forming kinetic (endo) products (Figure S.9a), and reactions leading to thermodynamic (exo) products (Figure S.9b). The analysis revealed that 207 hits (41%) belonged to the first class, 206 hits (41%) corresponded to kinetic (endo) products, and the thermodynamically favored exo product was found in **87** examples (18%), from which 36 (7.2%) led to symmetrically substituted cyclohexene.

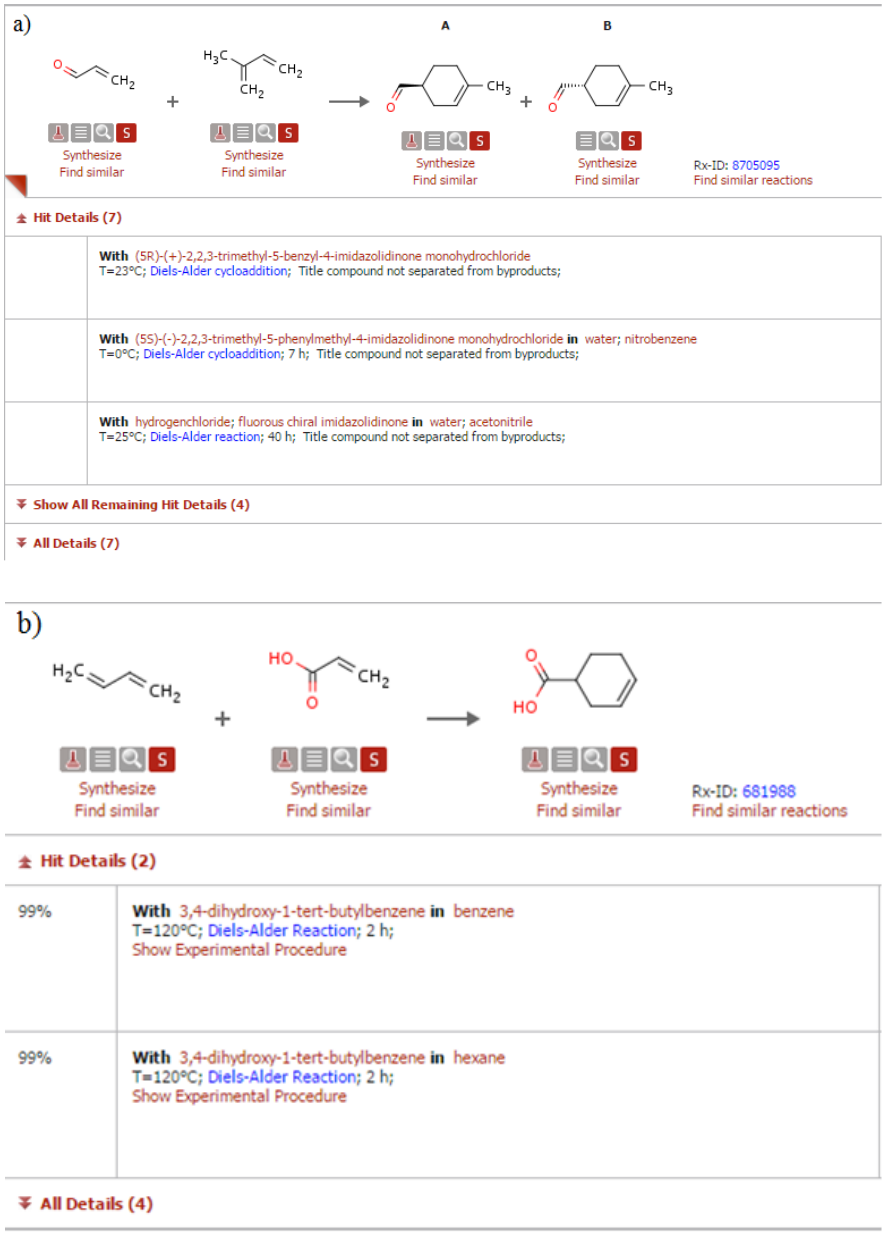


Figure S.8. Screenshots from Reaxys searches of the Diels-Alder reactions show examples of reactions which could not be unambiguously assigned as thermodynamic vs. kinetic products due to several factors: a) both products are listed but without any yields; b) simplified representation of products without stereochemical information reflects the lack of such information in source publications or formation of both products in different ratios under different conditions examined.

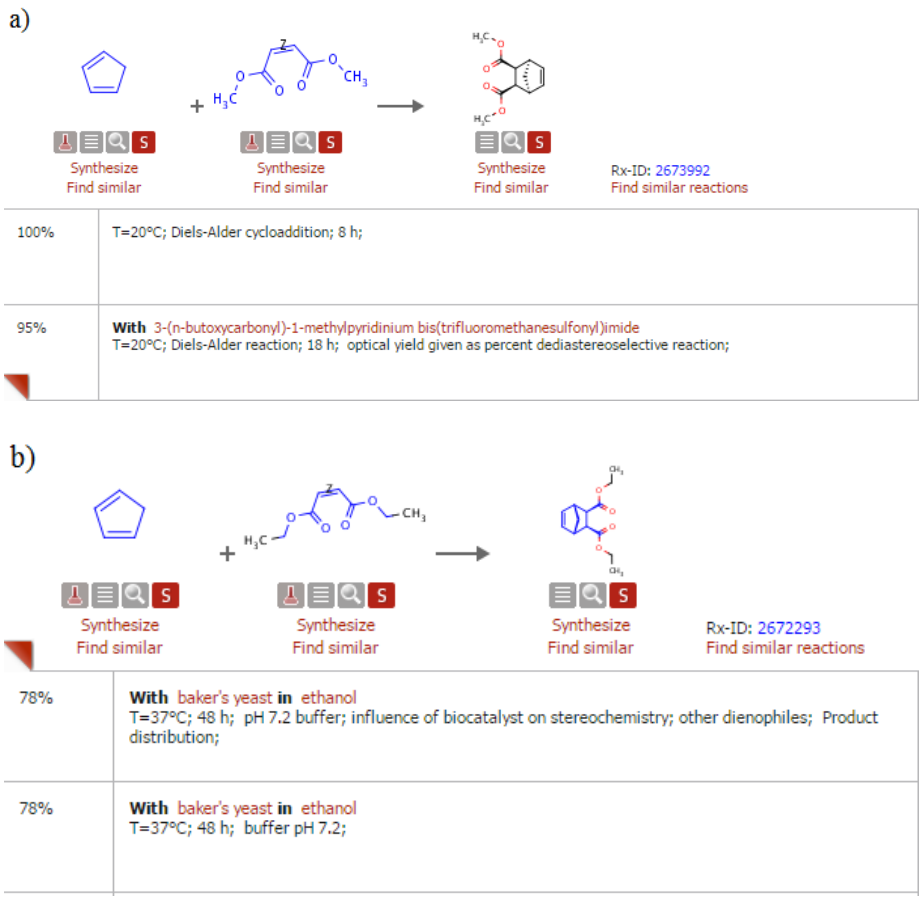


Figure S.9. Examples of Reaxys-retrieved Diels-Alder reactions leading (**a**) kinetic endo or (**b**) thermodynamic exo products.

S.1.2. Additional figure illustrating the diversity of the training set.

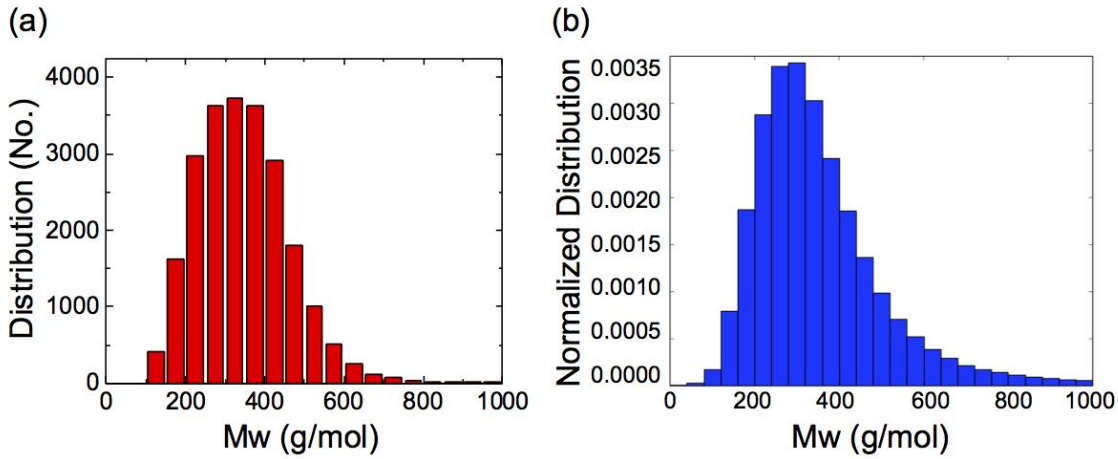


Figure S.10. Molecular weight distribution of the heaviest substrate (reactant or product) in reactions comprising (a) our training set of 23,000 reactions and (b) the entire database of ca. 9 M[1] data point from which our training set was randomly selected

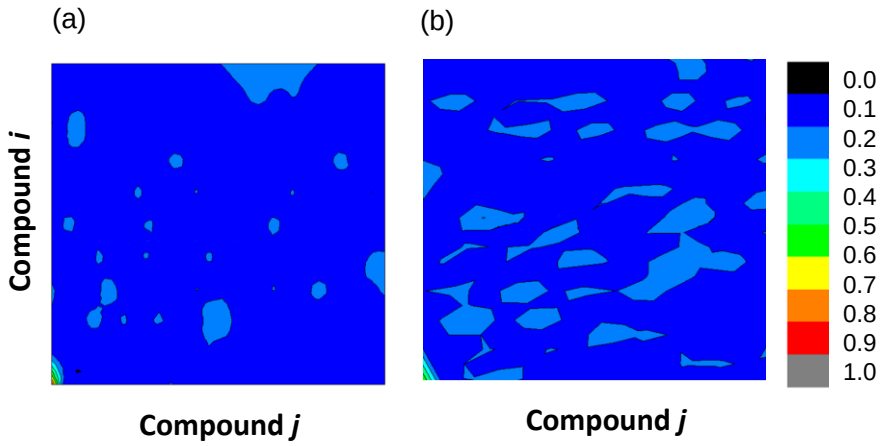


Figure S.11 A map of pairwise Tanimoto coefficients calculated for the reaction (a) substrates and (b) products for ca. 23K reactions existed in the training set indicates that the training set is structurally diverse (i.e., blue color corresponds to low similarity)

It is also interesting to investigate whether yield predictions are statistically better/worse for certain types of structurally similar molecules/reactions. To answer this question, we performed clustering analysis illustrated here for a random subset of 500 reactions in a test set (results are similar for larger sample sizes but visually clear presentation of the clustering trees becomes problematic; raw data is available from the authors upon request). For these reactions, we selected SMILES for the heaviest reactants and products, and calculated for both of them the so-called extended fingerprints (implementation from rcdk package, an interface to CDK libraries in R statistical language; see R. Guha, R., Chemical Informatics Functionality in R, *J. Stat. Soft*. **6**, 18, 2007, also www.r-project.com) which are well-known structural descriptors. We then quantified pairwise similarity between all reactants (or products) by calculating the Tanimoto coefficients. These analyses allowed us to construct the pairwise molecule-to-molecule distance matrix on which we performed hierarchical clustering using the so-called Ward's algorithm (this algorithm was chosen as it results in clusters of comparable sizes). The results of the clustering were summarized in the form of the so-called hierarchical clustering trees (a.k.a., dendrograms). For example, Figure S.12a shows a dendrogram for heaviest reactants. The tree pruned at a given level produces subtrees reflecting the clustering structure whereby structurally similar molecules in different leafs are also color-coded by the absolute error in yield prediction, $err = |\xi_{pred} - \xi_{exp}|$ (red, orange, yellow and green correspond to *err* within intervals (0;0.05], (0.05;0.1], (0.1;0.2], and (0.2;0.5], respectively). No systematic clusters of colors are detected. To emphasize this lack of any systematic clustering, we make a comparison with dendrogram in Figure S.12b where the leafs were intentionally colored to resemble the clustering structure. The lack of clustering

is further emphasized in the heatmaps in Figures S.12c and S12.d which plot the matrices of clustered Tanimoto distances between heaviest reactants (Figure S.12c) and products (Figure S.12d). The clustering dendrograms are shown on the axes and the yellow-red scale corresponds to descending similarity. As seen, other than the diagonal corresponding to the distances between identical molecules, there are no systematic clusters.

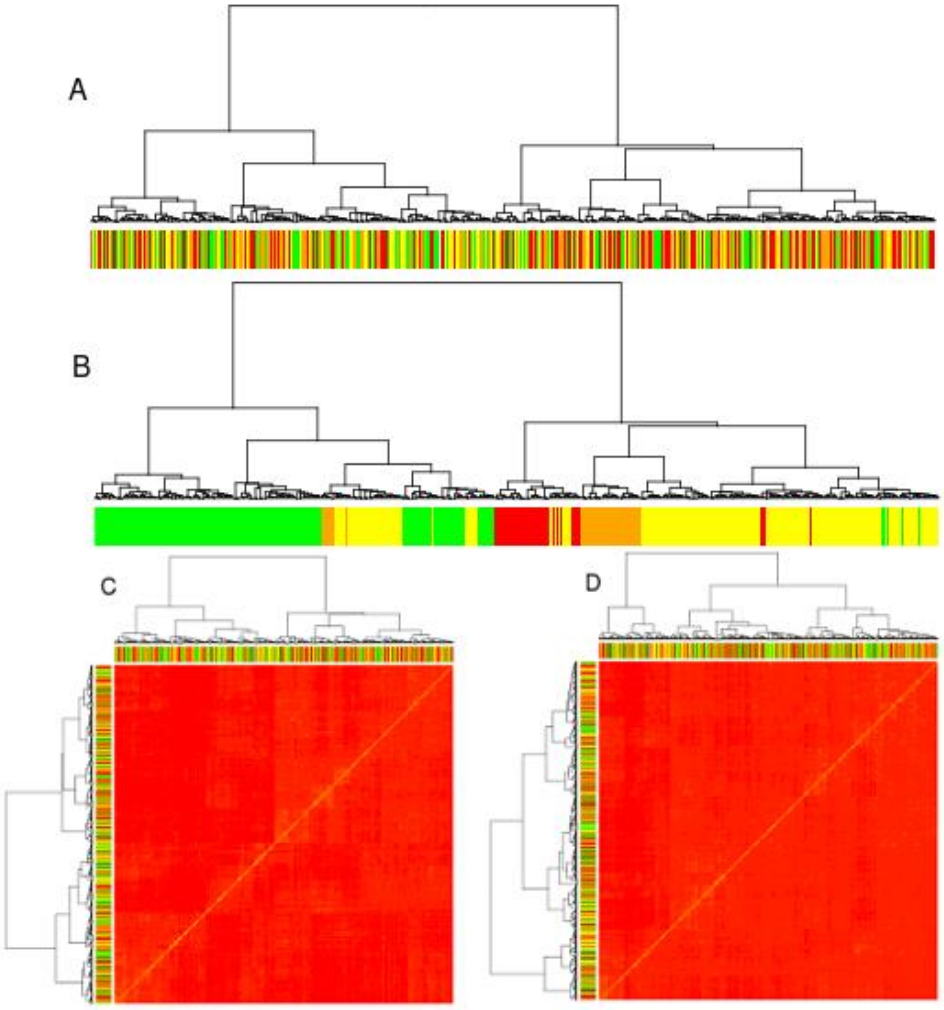


Figure S.12. Clustering analysis reveals no structural similarities between molecules having similar levels of yield-prediction error.

S.2. Thermodynamically guided routine to optimize the group free energies of formation.

In this section we narrates in more detail the algorithm used to calculate the groups' free energies of formation at given experimental temperature. The first three steps were as in the main text:

(1) From the reported literature reactions we chose and manually curated a training set of ca. 23.000 reactions whose reaction conditions, full stoichiometries, and experimental yields were all reported in literature. The set was chosen at random with the proviso that both small (Mw as low as 150 g/mol) as well as large (Mw as high as 1000 g/mol) were amply represented (cf. Figure S1 above).

(2) All participating molecules were decomposed into 296 functional groups listed in Table S.2. The decomposition procedure was hierarchical in the sense that functional groups were matched against the molecule of interest in the descending order of their complexity (i.e., more complex groups were matched first).

(3) These functional groups (j) were assigned initial values Gibbs free energies of formation at 298 K, $g_j^{form,298K}$, getting aid from Hukkerikar et al.[2] reported values.

The fourth step was as follows:

(4) Thereafter, applying the additive formulation, free energies of formation, $G_i^{form,298K}$, of all substrates and products, i, at 298 K were calculated using the functional groups given values. $\Delta G_{Calc}^{rxn,298K}$ and $-RT\ln K_{Calc}^{rxn,298K}$ were thus calculated employing the correct stoichiometry coefficients, ν_i:

$$-RT \ln K_{Calc}^{rxn,298\,K} = \Delta G_{Calc}^{rxn,298\,K} = \sum_i \nu_i \, . G_i^{form,298\,K} \quad \text{(S1)}$$

These reaction free energies were then translated to the actual temperatures at which the reaction was reported to obtain $\Delta G_{Calc}^{rxn,T}$ and $-RT \ln K_{Calc}^{rxn,T}$ trough heat capacities and enthalpies reported for liquids as explained and tabulated in Section S.4 and equations below.

$$\ln \frac{K_{Calc}^{rxn,T}}{K_{Calc}^{rxn,298K}} = \frac{\left(\Delta G_{Calc}^{rxn,T} - \Delta G_{Calc}^{rxn,298\,K}\right)}{-RT} = \int_{298}^{T} \frac{\Delta H_{Calc}^{rxn}}{RT^2} dT \quad \text{(S2)}$$

$$\Delta H_{Calc}^{rxn} = \sum_i \nu_i \, H_i^{form} + \int_{298\,K}^{T} Cp_i \, dT \quad \text{(S3)}$$

$$Cp^{\ell}(T) = Cp_0^{\ell}(T) + \sum_i \Gamma_i \, Cp_i^{\ell}(T) \quad \text{(S4)}$$

$$Cp_{0,i}^{\ell}(T) = a_{0,i} + b_{0,i}\left(\frac{T}{100}\right) + c_{0,i}\left(\frac{T}{100}\right)^2 \quad \text{(S5)}$$

where, $a_{0=105.94\times10^{-3}\,KJ/mol}$, $b_{0=-5.4\times10^{-3}\,KJ/(mol.K)}$, and $c_{0=-7.24\times10^{-3}\,KJ/(mol.K^2)}$ are constant for all compounds and the values of a_i, b_i, and c_i coefficients are listed in Table S.4 for all functional groups. The enthalpy of formation H_i^{form} is calculated by the structure-based approach presented by Hukkerikar et al.[2].

S.3. PC-SAFT molecular theory

The fugacity coefficient of the species were derived using the Perturbed-Chain Statistical Associated Fluid Theory (PC-SAFT) which was originally developed by Kleiner[3], Vrabec[4], and Gross[5]. Here we briefly describe the basis of this theory.

Let us begin by writing out the governing describing the reduced residual Helmholtz free energy, A^{res}/Nk_BT as,

$$\frac{A^{res}}{Nk_BT} = \frac{A^{hc}}{Nk_BT} + \frac{A^{disp}}{Nk_BT} + \frac{A^{assoc}}{Nk_BT} \quad \text{(S6)}$$

where N and k_B are the total number of molecules and Boltzmann constant, respectively, and with $\frac{A^{hc}}{Nk_BT}$, $\frac{A^{disp}}{Nk_BT}$, $\frac{A^{assoc}}{Nk_BT}$ denoting the hard-chain fluid term, dispersion attraction contribution, and short-range association due to hydrogen bonding. The hard-chain fluid reference contribution has been derived by Shukla and Chapman[6], Chapman et al.[7], and Mansoori et al.[8] as

$$\frac{A^{hc}}{Nk_BT} = \bar{m}\frac{A^{hs}}{Nk_BT} - \sum_i x_i\,(m_i - 1)\ln g_{jj}^{hs}\left(\sigma_{jj}\right) \quad \text{(S7)}$$

where the residual Helmholtz free energy for hard-sphere fluid is given on a per-segment basis

$$\frac{A^{hs}}{Nk_BT} = \frac{1}{\xi_0}\left[\frac{3\xi_1\xi_2}{1-\xi_3} + \frac{\xi_2^3}{\xi_3(1-\xi_3)^2} + \left(\frac{\xi_2^3}{\xi_3^2} - \xi_0\right)\ln(1-\xi_3)\right] \quad \text{(S8)}$$

where $\bar{m}$ is the mean segment number in the mixture

$$\bar{m} = \sum_i x_i m_i \quad \text{(S9)}$$

In Eq S12, x_i is the mole fraction. The radial distribution function (RDF) of the hard-sphere fluid is defined as

$$g_{ij}^{hs} = \frac{1}{1-\xi_3} + \left(\frac{d_i d_j}{d_i + d_j}\right)\frac{3\xi_2}{(1-\xi_3)^2} + \left(\frac{d_i d_j}{d_i + d_j}\right)^2\frac{2\xi_2^2}{(1-\xi_3)^3} \quad \text{(S10)}$$

With the number density (ρ) dependent parameter ξ_n expressed as

$$\xi_n = \frac{\pi}{6}\rho \sum_i x_i m_i d_i^n \quad n \in \{0,1,2,3\} \tag{S11}$$

$$\eta = \xi_3 = \frac{\pi}{6}\sum_i x_i m_i d_i^3 \tag{S12}$$

The density at a given system pressure, P^{sys} is determined iteratively by adjusting the reduced density (packing fraction), η until the calculated pressure is equal to the system pressure, P^{sys}. The number density of the molecules ρ has a direct relationship with packing fraction η as the following

$$\rho = \eta\left(\frac{\frac{6}{\pi}}{\sum_i x_i m_i d_i^3}\right) \tag{S13}$$

For a converged value of η , the molar density $\tilde{r}$ in units of mol/m^3 is obtained from

$$\tilde{\rho} = \frac{\rho}{N_{AV}}\left(10^{10}\frac{\text{Å}}{m}\right)^3 \tag{S14}$$

A temperature-dependent segment diameter d_i given by

$$d_i = \sigma_i[1 - 0.12\exp(-3\frac{\varepsilon_i}{k_B T})] \tag{S15}$$

The dispersion attraction term due to the chain-like shape of the molecule was derived by Gross and Sadowski[9] as the first and the second order thermodynamic perturbation theory (TPT) terms

$$\frac{A^{disp}}{Nk_B T} = -2\pi\rho I_1(\bar{m},\eta)\overline{m^2\varepsilon\sigma^3} - \pi\rho\bar{m}C_1 I_2(\bar{m},\eta)\overline{m^2\varepsilon^2\sigma^3} \tag{S16}$$

The compressibility C_1 is given by

$$C_1 = \left(1 + Z^{hc} + \rho\frac{\partial Z^{hc}}{\partial\rho}\right)^{-1} = \left(1 + \bar{m}\frac{8\eta-2\eta^2}{(1-\eta)^4} + (1-\bar{m})\frac{20\eta-27\eta^2+12\eta^3-2\eta^4}{[(1-\eta)(2-\eta)]^2}\right)^{-1} \tag{S17}$$

$$I_1(\bar{m},\eta) = \sum_{i=0}^{6} a_j(\bar{m})\,\eta^i \tag{S18}$$

$$I_1(\bar{m},\eta) = \sum_{i=0}^{6} b_j(\bar{m})\,\eta^i \tag{S19}$$

and

$$a_i(\bar{m}) = a_{0i} + \frac{\bar{m}-1}{\bar{m}} a_{1i} + \frac{\bar{m}-1}{\bar{m}} \frac{\bar{m}-2}{\bar{m}} a_{2i} \quad \text{(S20)}$$

$$b_i(\bar{m}) = b_{0i} + \frac{\bar{m}-1}{\bar{m}} b_{1i} + \frac{\bar{m}-1}{\bar{m}} \frac{\bar{m}-2}{\bar{m}} b_{2i} \quad \text{(S21)}$$

The model constants $a_{0i}, a_{1i}, \ldots$ are given in Table S.1.

Table S.1. Universal Model Parameters for Equations S20 and S21.

I	a_{0i}	a_{1i}	a_{2i}	b_{0i}	b_{1i}	b_{2i}
0	0.910563	-0.3084	-0.09061	0.724095	-0.57555	0.097688
1	0.636128	0.186053	0.452784	2.238279	0.69951	-0.25576
2	2.686135	-2.503	0.59627	-4.00258	3.892567	-9.15586
3	-26.5474	21.41979	-1.72418	-21.0036	-17.2155	20.64208
4	97.75921	-65.2559	-4.13021	26.85564	192.6723	-38.8044
5	-159.592	83.31868	13.77663	206.5513	-161.826	93.62677
6	91.29777	-33.7469	-8.67285	-355.602	-165.208	-29.6669

The hydrogen bonding contribution due to the association and solvation between molecules is accounted for by the Wertheim theory of hydrogen bonding[10] for association and solvation complexation networks. The hydrogen bonding contribution to the chemical potential is given by

$$\frac{\mu_i^{assoc}}{Nk_BT} = \sum_{X_{A_i}} \ln X_{A_i} - \frac{\rho}{2} \frac{\partial \ln \Delta}{\partial n_i} \sum_{j=1}^{C} n_j \sum_{B_j} \left(1 - X_{B_j}\right) \quad \text{(S22)}$$

and hence,

$$\frac{A^{assoc}}{Nk_BT} = \sum_{i=1}^{C} n_i \sum_{A_i} \left(\ln X_{A_i} - \frac{1-X_{A_i}}{2}\right) \quad \text{(S23)}$$

where X_{A_i} is the fraction of A-type sites in species *i* that are not bonded to other molecular sites (B-types). The fractions of sites are obtained from

$$X_{A_i} = \frac{1}{1+\rho \sum_{j=1}^{C} x_j \sum_{B_j} X_{B_j} \Delta^{A_i B_j}} \quad \text{(S24)}$$

where the temperature, density, and composition dependent association strength $\Delta^{A_i B_j}$ accounts for the equilibrium constant independent on the extent of bonding

For further details about PC-SAFT molecular theory, we refer readers to the original works published by Kleiner[3], Vrabec[4], and Gross[5].

S.4. Estimation of PC-SAFT molecular theory parameters

The non-ideality of the reaction system are accounted for through activity coefficients calculated by the PC-SAFT molecular theory. The primary need to solve for any properties for a particular molecule is to obtain the parameters by those the physical properties of that molecule can be described, i.e. shape factor or number of segments per chain assuming that molecules are chains composed of spherical segments (m), temperature-independent segment diameter (σ), and depth potential (ε/k). For associating molecules three additional parameters are needed; number of associating cites (Nd), the associating energy depth (ε^{HB}/k), and effective association volume (k^{AD}). For molecules whose parameters are not available, sets of thermo-physical properties of the material such as vapor pressure, liquid and vapor densities, second virial coefficient, internal energy, or coexistence PVT data are required to optimize the equation of state parameters. Optimizing the Equation of State parameters on thermo-physical data is a fairly standard method and is suitable for small molecules such as solvents. However, this procedure depends on and is restricted by subjected to the availability of experimental data, which is usually troublesome for large molecules, an alternative approach similar to the structured-base methodology suggested by us[11] is useful for the compounds with no reported experimental properties.

Herein, we adapted this methodology with minor differences as it is explained below to estimate the PC-SAFT molecular theory parameters for thousands of compounds and reactions described in chemical-organic literature for which yields and full stoichiometries are known. For a molecule with known molecular structure, one can

estimate the liquid molar volume and heat of vaporization at room temperature as well as normal boiling point using the approach provided by Hukkerikar et al.[2]

$$V_L^{298}\left(\frac{cc}{mol}\right) = 1000\{0.0123 + \sum_i \Gamma_i(\Delta V_L^{298})_i\} \quad \text{(S25)}$$

$$H_{vap}^{298}\left(\frac{KJ}{mol}\right) = 10.4327 + \sum_i \Gamma_i\left(\Delta H_{vap}^{298}\right)_i \quad \text{(S26)}$$

$$T_b(K) = 244.79\ln(\sum_i \Gamma_i(\Delta T_b)_i) \quad \text{(S27)}$$

To solve for the three parameters: m , σ, and ε/k , as well as liquid density at 1 bar, four constraints can be defined as explained in detail by Emami et el[11], elsewhere. The first constraint is to match the estimated liquid compressibility factors (Z_L^{298}) with the following equation

$$Z_L^{298} = \frac{0.1\,V_L^{298}}{298R} \quad \text{(S28)}$$

Second, the internal energy departure function should match with the internal energy computed from liquid molar volume and heat of vaporization according to the following relationships:

$$\delta = \left(\frac{H_{vap}^{298} - 298R}{V_L^{298}}\right)^{1/2} \quad \text{(S29)}$$

$$\frac{\delta^2 V_L^{298}}{298R} = -\frac{U}{298R} \quad \text{(S30)}$$

The internal energy (U) is derived from the Helmholtz energy calculated by PC-SAFT. The last two equations are isofugacity criterion for the vapor and liquid phase and vapor pressure equality to 1 bar at normal boiling point.

$$\ln\varphi^L = \ln\varphi^V \quad \text{(S31)}$$

$$P_{Calc}^{sat}(@\,nBT) = 1bar \quad \text{(S32)}$$

where,

$$\ln\varphi = \beta\mu^{res}(T,V) - \ln Z \quad \text{(S33)}$$

$$\frac{\mu^{res}(T,V)}{k_B T} = \tilde{A}^{res} + (Z-1) \quad \text{(S34)}$$

and

$$\tilde{A}^{res} = \frac{A^{res}}{N k_B T} \quad \text{(S35)}$$

For associating molecules there is need for three additional parameters, i.e., Number of association Nd, association energy (ε^{HB}/k), and effective association volume (k^{AD}). Number of association (e.g., the number of hydrogen bonds formed by donor and acceptor sites) can be determined form the structure of a molecule. We estimate the dimensionless hydrogen bonding volume correlated as

$$k^{AD} = 0.035 \frac{\pi}{6} \quad \text{(S36)}$$

The hydrogen bond energy depths for different associating cites are represented by Pimentel and McClellan[12] as listed in Table S.3.

S.5. Functional groups

Table S.2. List of the functional groups into which the molecules are decomposed.

	Name	Structure (SMILES)
1	aC-CHn-X(n[1,2])-X:Halogen	c[CX4;H1,H2][F,Br,I,Cl]
2	C-F3	[CX4]([F])([F])([F])
3	CH(cyc)-CL	[CX4;R;H1][Cl]
4	CH(cyc)-F	[CX4;R;H1][F]
5	CHm=CHn-F(m,n[0,2])	[C;!R]=[C;!R]F
6	CHm=CHn-Br(m,n[0,2])	[C;!R]=[C;!R]Br
7	(CHn=C)cyc-Cl(n[0..2])	[CX3;R]=[CX3;H0;R][Cl]
8	CHm=CHn-Cl(m,n[0,2])	[C;!R]=[C;!R][Cl]
9	aC-Cl	[c][Cl]
10	aC-F	c[F]
11	aC-I	c[I]
12	aC-Br	c[Br]
13	RCF2	[CX4]([F])([F])
14	RCF	[CX4]([F])
15	HF	[FH]
16	CCL3	[CX4;H0]([Cl])([Cl])[Cl]
17	CCL2	[CX4;H0]([Cl])([Cl])
18	CL<C=C>	[Cl;$([C;!R]=[C;!R])]
19	CCL	[CX4;H0][Cl]
20	CHCL	[CX4;H1][Cl]
21	CH2CL	[CX4;H2][Cl]
22	HCL	[HCl]
23	CL2	[Cl][Cl]
24	HI	[HI]
25	I2	[I][I]
26	HBr	[HBr]
27	Br2	[Br][Br]
28	F except as above	F
29	Cl except as above	Cl
30	I except as above	I
31	Br except as above	Br
32	SiO	[SiH0;!R][O;!R]
33	Si	[Si;!R]
34	aC-SO3(sulfonate)-aC	c[SX4;!R](=O)(=O)[OX2]c

35	aC-SO2-aC	c[SX4;H0](=O)(=O)c
36	aC-S-aC(different-rings)	[c][S;H0;!R][c]
37	aC-CHn-SH(n[1,2])	cC[SX2;H1]
38	aC-CHn-S-(n[1,2])	cC[SX2;H0]
39	CHm(cyc)-S-CHn(cyc)(m,n[0,1])	[CX4;R][SX2;R][CX4;R]
40	CH(cyc)-S	[C;R;H1][SX2;H0;!R]
41	SO3(sulfonate)	[SX4](=[O])(=[O])([O])
42	aC-SO2	c[SX4](=[O])(=[O])
43	aC-SO	c[SX3](=[O])
44	aC-SH	c[SX2;H1]
45	aC-S-	c[SX2;H0]
46	RSO2	[S;X4;H0;R](=[O;!R])(=[O;!R])
47	RS	[S;X2;R]
48	SO2	[SX4;!R](=[O;!R])(=[O;!R])
49	SO	[SX3;!R](=[O;!R])
50	SO(ring)	[SX3;R](=[O;!R])
51	CH3S	[CX4;H3][SX2;H0]
52	CH2S	[CX4;H2][SX2;H0]
53	CH2SH	[CX4;H2][SX2;H1]
54	>C=S	[CX3]=[S;!R]
55	S	[#16]
56	P=O	[P]=[O]
57	aC-NHCONH-aC(diff-rings)	[cH0][N;H1;!R][C;!R](=O)[N;H1;!R][cH0]
58	aC-CO-Ncyc(different-rings)	[cH0][C;!R](=O)[N;R]
59	aC-NH-aC(different-rings)	c[NH;!R]c
60	aC-N-CHcyc(different-rings)	c[N;!R][CH;R
61	N-multiring	[N$(*([#6])([#6])([#6]))]
62	aC-NHn(cyc)(fused-rings)(n[0,1])	[c;R2,R3][n;H0,H1;R2,R3]
63	PYRIIDINE.FUSED[2]	c1ccc2ncccc2c1
64	PYRIIDINE.FUSED[2-iso]	c1ccc2cnccc2c1
65	NH-(CHn)3-COOH(n[0,2])	[NX3;H1][C;!R][C;!R][C;!R][C;!R](=O)[OH]
66	NH2-(CHn)3-OH(n[0,2])	[NH2;!R][C;!R][C;!R][C;!R][OH]
67	NHk-(CHn)3-NH2(k[0,1];n[0,2])	[NX3;!R][C;!R][C;!R][C;!R][NX3;H2]
68	aC-CHn-NHm(n[1,2],m[0,2])	c[CX4;H1,H2][NH0]
69	aC-CHn-NHm(n[1,2],m[0,2])	c[CX4;H1,H2][NH1]
70	aC-CHn-NHm(n[1,2],m[0,2])	c[CX4;H1,H2][NH2]
71	aCaNaC	cnc
72	aC-CHn-CN(n[1,2])	cC#N
73	aC-CHn-CONH2(n[1,2])	cCC(=O)[NH2]

74	(CH2NHCH2)(cyc)	[CH2;R][NH;R][CH2;R]
75	CH(cyc)-NH2	[C;R;H1][NH2;!R]
76	CH(cyc)-NH-CHn(n[0,3])	[C;R;H1][NX3;H1;!R][C;!R]
77	CH(cyc)-CN	[C;R;H1][C;!R]#[N;!R]
78	>N(cyc)-CH2	[N;R;H0][CX4;H2;!R]
79	NC-CHn-COO(n[1,2])	[N;!R]#[C;!R][CX4;H1,H2;!R][C;!R](=[O;!R])[O;!R]
80	CHm(NH)CHn(NH2)(m,n[1,2])	[CX4;H1,H2;!R]([NX3;H1])[CX4;H1,H2;!R]([NX3;H2])
81	CHn(OH)CHm(NH)(m[0,1],n,p[0,2])	[CX4;!R]([OH])C[NX3;H0]
82	CHn(OH)CHm(NH)(m[0,1],n[0,2])	[CX4;!R]([OH])C[NX3;H1]
83	CHn(OH)CHm(NH2)(m[0,1],n[0,2])	[CX4;!R]([OH])C[NX3;H2]
84	CHm(N)-COOH(m,n[0,2])	[CX4;H0,H1,H2;!R]([NX3;H0])[C;!R](=O)[OH]
85	CHm(NH)-COOH(m,n[0,2])	[CX4;H0,H1,H2;!R]([NX3;H1])[C;!R](=O)[OH]
86	CHm=CHn-CN(m,n[0,2])	[C;!R]=[C;!R][C;!R]#N
87	aC-NHCONH	c[NX3;H1;!R][C;!R](=O)[NX3;H1;!R]
88	aC-NHCO	c[NX3;H1;!R][CX3;H0;!R](=O)
89	aC-CONH2	c[CX3;H0;!R](=O)[NX3;H2;!R]
90	aC-CONH	c[CX3;H0;!R](=O)[NX3;H1;!R]
91	ACNH2	[c][NX3;H2;!R]
92	CON(CH2)2	[CX3;!R](=O)[NX3;!R][CX4;H2;!R][CX4;H2;!R]
93	CONHCH2	[CX3;!R](=O)[NX3;H1;!R][CX4;H2;!R]
94	CONHCH3	[CX3;!R](=O)[NX3;H1;!R][CX4;H3;!R]
95	HCONH	[CX3;!R;H1](=O)[NX3;H1;!R]
96	CONH2	[CX3;!R](=O)[NX3;H2;!R]
97	CNO2	[O-][N+](=O)C
98	aN-in-aromatic-ring	[n]
99	ACNO2	[c][NX3;!R](=O)[O;!R]
100	NHCO-except-as-above	[NX3;H0;!R][C;!R](=O)
101	CH2NCO	[CX4;H2;!R][N;!R]=[C;!R]=O
102	CNH2	[CX4;H0;!R][NX3;H2;!R]
103	CNOH	[CX3;H0;!R]=[N;!R][OX2;H1;!R]
104	CH2NH2	[CX4;H2;!R][NX3;H2;!R]
105	CHNH2	[CX4;H1;!R][NX3;H2;!R]
106	CH3NH	[CX4;H3;!R][NX3;H1;!R]
107	CH2NH	[CX4;H2;!R][NX3;H1;!R]
108	CHNH	[CX4;H1;!R][NX3;H1]
109	CH3N	[CX4;H3][NX3;H0]
110	CH2N	[CX4;H2][NX3;H0]
111	CH-N	[CX4;H1]-[N]

112	CH2CN	[CX4;H2][C]#[N]
113	CHCN	[CX4;H1][C]#N
114	CCN	[CX4;H0][C]#N
115	HCN	[CX2;H]#[N]
116	CH=N	[CX3;H1;!R]=[N;!R]
117	C=N	[CX3;H0;!R]=[N;!R]
118	ONO	[O;!R][N;!R]=[O;!R]
119	CN-except-as-above	[C;!R]#N
120	RCH=N	[C;X3;H1;R]=[N;X2;H0;R]
121	RC=N	[C;X3;H0;R]=[N;X2;H0;R]
122	(R)C=N	[C;X3;H0;R]=[N;X2;H0;!R]
123	RNH	[NX3;H1;R]
124	RN	[NX3;H0;R]
125	>NH	[#7;X3;H1
126	NH2-except-as-above	[NX3;H2]
127	N=N	[#7]=[#7]
128	NH3	[NX3;H3]
129	aC-CHm-CO-aC(different-rings)(m[0,2])	[cH0][C;!R][C;!R](=O)[cH0]
130	aC-CO-aC(different-rings)	[cH0][C;!R](=O)[cH0]
131	aC-CO-CHn(cyc)(different-rings)(n[0,1])	[cH0][C;!R](=O)[CX4;H1;R]
132	aC-CHn-O-CHm-aC(different-rings)(n,m[0,1])	[cH0]-[C;!R]-O-[C;!R]-[cH0]
133	aC-O-CHn-aC(different-rings)(n[0,2])	[cH0]-O[C;!R][cH0]
134	aC-O-aC(different-rings)	[cH0]-O-[cH0]
135	aC-CO(cyc)(fused-rings)	[c$(*(c)(c)([C;R]))][C$(*(c)([C,O,S,N;R])(=O))](=O)
136	aC-O(cyc)(fused-rings)	[c$(*(c)(c)([O,o;R]))][#8;R;H0]
137	OH-(CHn)3-OH(n[0,2])	[OH][C;!R][C;!R][C;!R][OH]
138	CHp-O-(CHn)3-OH(n,p[0,2])	[C;!R]O[C;!R][C;!R][C;!R][OH]
139	CHp-O-(CHn)4-OH(n,p[0,2])	[C;!R]O[C;!R][C;!R][C;!R][C;!R][OH]
140	aC-CHn-CHO(n[1,2])	c[CX4;!R][CX3;H1;!R](=O)
141	aC-CHn-OOC(n[1,2])	c[CX4;!R]O[CX3;H0;!R](=O)
142	aC-CHn-COOH (n[1,2])	c[CX4;!R][C;!R](=O)[OH]
143	aC-CHn-COO(n[1,2])	c[CX4;!R][C;!R](=O)[OX2;H0]
144	aC-CHn-CO-(n[1,2])	c[CX4;!R][CX3;H0;!R](=O)
145	aC-CHn-OH(n[1,2])	c[CX4;!R][OX2;H1]
146	aC-CHn-O-(n[1,2])	c[CX4;!R][OX2;H0]
147	AC-O-CHm(m[0,3])	c[OX2;!R][CX4;!R]
148	(COOCH2)(cyc)	[C;R](=O)[#8][CH2;R]
149	(CH2OCH2)(cyc)	[CH2;R][O;R][CH2;R]
150	(CH2OCH)(cyc)	[CH2;R][O;R][CH1;R]

151	(CHn=C)(cyc)-CO(n[0,2])	[C;R][CX3;H0;R][CX3;H0;!R](=[O;!R])
152	CH(cyc)-COO	[C;R;H1][C;!R](=[O;!R])[OX2;H0;!R]
153	CH(cyc)-COOH	[C;R;H1][C;!R](=[O;!R])[OX2;H1;!R]
154	CH(cyc)-OOC	[C;R;H1][OX2;H0;!R][CX3;H0;!R](=[O;!R])
155	CH(cyc)-OH	[CX4;R;H1][OH;!R]
156	CH(cyc)-CO	[C;R;H1][CX3;H0;!R](=O)
157	CH(cyc)-CHO	[C;R;H1][CX3;H1;!R](=O)
158	CH(cyc)-O	[C;R;H1][OX2;H0;!R]
159	C(cyc)-OH	[C;R;H0][OH;!R]
160	OH-CHn-COO(n[0,2])	[OH][CX4;H0,H1,H2;!R][CX3;!R](=O)[OX2;H0]
161	HO-CHn-COOH(n[1,2])	[OH][CX4;H0,H1,H2;!R][C;!R](=O)[OH]
162	CHm(OH)CHn(OH)(m,n[0,2])	[CX4;H0,H1,H2;!R]([OH])[CX4;H0,H1,H2;!R]([OH])
163	CH3COOCHm(m[0,1])	[CX4;H3;!R][CX3;H0;!R](=O)[OX2;H0][CX4;H0,H1;!R]
164	CHmCOOH(m[0,1])	[CX4;H0,H1;!R][CX3;H0;!R](=O)[OX2;H1]
165	CHmCHO(m[0,1])	[CX4;H1,H0;!R][CX3;H1;!R]=O
166	CH3COCH2	[CX4;H3;!R][CX3;H0;!R](=O)[CX4;H2;!R]
167	CH3COCHm(m[0,1])	[CX4;H3;!R][CX3;H0;!R](=O)[CX4;H0,H1;!R]
168	(CHdCHOCHdCH)(cyc)	[C;R][CH1;R][O;R][C;R][CH1;R]
169	CHm-O-CHn=CHp(m,n,p[0,3])	[CX4;H0,H1,H2;!R;O1]O[C;!R]=[C;!R]
170	CHn=CHm-COO-CHp(m,n,p[0,3])	[C;!R]=[C;!R][C;!R](=O)O[C;!R]
171	(CHn=CHm)cyc-COOH	[C;R]=[C;R][C;!R](=O)[O;!R]
172	CHm=CHn-COOH(m,n,p[0,3])	[C;!R]=[C;!R][C;!R](=O)[OH]
173	CHm=CHn-CHO(m,n,[0,2])	[C;!R]=[C;!R][CX3;H1;!R](=O)
174	CHOH	[CX4;H1;!R][OX2;H1]
175	aC-COOH	[cX3;H0][CX3;H0;!R](=O)[OX2;H1;!R]
176	aC-COO	[cX3;H0][CX3;H0;!R](=O)[OX2;H0;!R]
177	aC-CO	[cX3;H0][CX3;H0;!R](=O)
178	aC-CHO	[cX3;H0][CX3;H1;!R](=O)
179	aC-OCO	[cX3;H0][OX2;H0;!R][C;!R](=O)
180	aC-O	[cX3;H0][OX2;H0;!R]
181	ACOH	[c][OH]
182	aC=O	[c](=O)
183	RCO	[#6;X3;R;H0]=O
184	CH2OCHO	[CX4;H2;!R][O;!R][CX4;H1;!R]-[O;!R]
185	CH3COOCH2	[CX4;H3;!R][CX3;H0;!R](=O)[OX2;!R][CX4;H2;!R]
186	CH2COOCH3	[CX4;H2;!R][CX3;H0;!R](=O)[OX2;!R][CX4;H3;!R]

187	CH2COOCH2	[CX4;H2;!R][CX3;H0;!R](=O)[OX2;!R][CX4;H2;!R]
188	CHCOOCH2	[CX4;H1;!R][CX3;H0;!R](=O)[OX2;!R][CX4;H2;!R]
189	CCOOCH2	[CX4;H0;!R][CX3;H0;!R](=O)[OX2;!R][CX4;H2;!R]
190	OCH2CH2O	[OX2;H0;!R]-[CX4;H2;!R][CX4;H2;!R]-[OX2;H0;!R]
191	OCH2CH2OH	[OX2;H0;!R][CX4;H2;!R][CX4;H2;!R][OX2;H1;!R]
192	CO3(carbonate)	[CX3;H0;!R](=[O])([OX2;H0;!R])([OX2;H0;!R])
193	CH3COO	[CX4;H3;!R][C](=[O])[OX2;H0;!R]
194	CH2COO	[CX4;H2;!R][C](=O)[OX2;H0;!R]
195	CHCOO	[CX4;H1;!R][C](=O)[OX2;H0;!R]
196	CCOO	[CX4;H0;!R][C](=O)[OX2;H0;!R]
197	CH3CO	[CX4;H3;!R][CX3;H0]=[O]
198	CH2CO	[CX4;H2;!R][CX3;H0]=[O]
199	CHOCH	[CX4;H1]-[O]-[CX4;H1]
200	CHCO	[CX4;H1][C](=O)
201	COOH	[CX3;H0](=O)[OH;!R]
202	COO	[CX3;H0](=[O])[OX2;H0;!R]
203	CCO	[CX4;H0;!R][C](=O)
204	C2H2O	O1[CX4;H0][CX4;H2]1
205	C2HO	O1[CX4;H0][CX4;H1]1
206	CH3O	[CX4;H3]-[OX2;H0]
207	CH2O	[CX4;H2]-[OX2;H0]
208	CHO	[CX3;H1]=[O]
209	C-O	[CX4;H0]-[OX2;H0]
210	CH-O	[CH]-[OX2;H0]
211	OH	[OX2;H1]
212	O2	[OX1]=[OX1]
213	CO2	[O]=[C]=[O]
214	=O	[#8X1]
215	O	[#8]
216	aC-(CHn=CHm)(cyc)(fused-rings)(n,m[0,1])	c[CX3;H0,H1;R]=[CX3;H0,H1;R]
217	aC-(CHm=CHn)-aC(different-rings)(m,n[0,2])	c[C;!R]=[C;!R]c
218	aC-(CHn)2-aC(different-rings)(n[0,2])	[c;H0][C;!R][C;!R][c;H0]
219	aC-CHm-aC(different-rings)	[c;H0][C;!R][c;H0]
220	aC-aC(different-rings)	c-c
221	aC-(CHn)2-CHcyc(different-rings)(n[0,2])	[c;H0][C;!R][C;!R][CH1;R]
222	aC-(CHn)-CHcyc(different-rings)(n[0,2])	[c;H0][C;!R][CH1;R]
223	aC-CHn,cyc(different-rings)(n[0,1])	[c;R1][CX4;H0,H1,H2;R1]

224	aC-CHn,cyc(fused-rings)(n[0,1])	[c;R2][CX4;H0,H1,H2;R1]
225	AROMFUSED[2]s1	[cH1]1[cH1][cH1][cH0]2[cH1][cH1][cH1][cH0][cH0]2[cH1]1
226	AROMFUSED[2]s2	[cH1]1[cH1][cH1][cH0]2[cH1][cH1][cH0][cH1][cH0]2[cH1]1
227	aC-CH(CH3)2	c[CH;!R]([CH3;!R])([CH3;!R])
228	aCCH(CH2)2	c[CH;!R]([CH2;!R])([CH2;!R])
229	aCCH(CH3)(CH2)	c[CH;!R]([CH3;!R])([CH2;!R])
230	aC-C(CH3)3	c[CX4;H0;!R]([CH3])([CH3])([CH3])
231	AROMRINGs1s2s3s4s5	[c;H0]1[c;H0][c;H0][c;H0][c;H0][cH]1
232	AROMRINGs1s2s3s4	[c;H0]1[c;H0][c;H0][c;H0][cH][cH]1
233	AROMRINGs1s2s3s5	[c;H0]1[c;H0][c;H0][cH][c;H0][cH]1
234	AROMRINGs1s2s4s5	[c;H0]1[c;H0][cH][c;H0][c;H0][cH]1
235	AROMRINGs1s2s3	[c;H0]1[c;H0][c;H0][cH][cH][cH]1
236	AROMRINGs1s2s4	[c;H0]1[c;H0][cH][c;H0][cH][cH]1
237	AROMRINGs1s3s5	[c;H0]1[cH][c;H0][cH][c;H0][cH]1
238	AROMRINGs1s2	[c;H0]1[c;H0][cH][cH][cH][cH]1
239	AROMRINGs1s3	[c;H0]1[cH][c;H0][cH][cH][cH]1
240	AROMRINGs1s4	[c;H0]1[cH][cH][c;H0][cH][cH]1
241	aC-CH=CH2	c[CX3;H1]=[CX3;H2]
242	aC-CH=CH	c[CX3;H1]=[CX3;H1]
243	aC-C=CH2	c[CX3;H0]=[CX3;H1]
244	aC-C#CH	c[CX2;H0]#[CX2;H1]
245	aC-C#C	c[CX2;H0]#[CX2;H0]
246	ACCH3	[cX3;H0][CX4;H3]
247	ACCH2	[cX3;H0][CX4;H2]
248	ACCH	[cX3;H0][CX4;H1]
249	aC-C	[cX3;H0][CX4;H0;!R]
250	aC-fuseArmtcRing	[c$(*(c)(c)(c));R2]
251	aC-fuseNonArmtcSubrng	[c$(*(c)(c)(C));R2]
252	ACH	[cX3;H1]
253	AC	[cX3;H0]
254	CHcyc-Chcyc(different-rings)	[CX4$(*([C;R])([C;R])([C;R]));R1;H1][CX4$(*([C;R])([C;R])([C;R]));R1;H1]
255	CH multiring	[CX4;H1;R2,R3]
256	C multiring	[CX4;H0;R2,R3,R4]
257	CH(cyc)-C=CHn(n[1,2])	[C;R;H1][CX3;H0;!R]=[CX3;H1,H2]
258	3-membered-ring	C1CC1
259	4-membered-ring	C1CCC1
260	5-membered-ring	C1CCCC1
261	6-membered-ring	C1CCCCC1

262	RC-CH2CH3	[CX4;R][CH2;!R][CH3;!R]
263	RC-CH2CH2CH3	[CX4;R][CH2;!R][CH2;!R][CH3;!R]
264	CH(cyc)-CH3	[C;R;H1][CX4;H3;!R]
265	CH(cyc)-CH2	[C;R;H1][CX4;H2;!R]
266	CH(cyc)-CH	[C;R;H1][CX4;H1;!R]
267	CH(cyc)-C	[C;R;H1][CX4;H0;!R]
268	RC-CH3	[CX4;R;H0][CH3;!R]
269	C(cyc)-CH2	[C;R;H0][CX4;H2;!R]
270	RCH=CH	[#6;X3;H1;R]=[#6;X3;H1;R]
271	(R)C=CH	[#6;X3;H0;R
272	RCH=C	[#6;X3;H1;R]=[#6;X3;H0;R]
273	RC=C	[#6;X3;H0;R]=[#6;X3;H0;R]
274	(R)C=C	[#6;X3;H0;R]=[CX3;H0;!R]
275	RCH2=C	[#6;X3;H2]=[#6;X3;H0;R]
276	RCH2	[#6;X4;H2;R]
277	RCH	[#6;X4;H1;R]
278	RC	[#6;X4;H0;R]
279	CH3-CHm=CHn(m,n[0,2])	[CX4;H3;!R][CX3;!R]=[CX3;!R]
280	CH2-CHm=CHn(m,n[0,2])	[CX4;H2;!R][CX3;!R]=[CX3;!R]
281	CHp-CHm=CHn(m,n[0,2];p[0,1])	[CX4;H0,H1;!R][CX3;!R]=[CX3;!R]
282	<CH3>2CH	[CX4;H1;!R]([CX4;H3;!R])([CX4;H3;!R])
283	<CH3>3C	[CX4;H0;!R]([CX4;H3;!R])([CX4;H3;!R])([CX4;H3;!R])
284	CH2=CH	[CX3;H2;!R]=[CX3;H1;!R]
285	CH=CH	[CX3;H1;!R]=[CX3;H1;!R]
286	CH2=C	[CX3;H2;!R]=[CX3;H0;!R]
287	CH=C	[CX3;H1;!R]=[CX3;H0;!R]
288	CH3	[CX4;H3;!R]
289	CH2	[CX4;H2;!R]
290	CH	[CX4;H1;!R]
291	C	[CX4;H0;!R]
292	CH#C	[CX2;H1]#[CX2;H0;!R]
293	C#C	[CX2;H0;!R]#[CX2;H0;!R]
294	H2	[H][H]
295	Ethylene	[CX3;H2]=[CX3;H2]
296	Ethyne	[CX2;H]#[CX2;H]

S.6. Group contribution parameters

The formulas below were used to determine Gibbs free energy of formation, heat of vaporization, liquid molar volume, and boiling temperature.[2] The parameters in Table S.3 for the liquid molar volume, heat of vaporization, boiling temperature, and hydrogen bond data are taken from the work published by Hukkerikar et al.[2] and Nguyen et al.[13], respectively. For larger functional groups absent in the above works, we extended the parameters based on similarity of the structure or by the additivity rule from smaller groups.

$$\frac{\Delta G^f}{RT} = 8.5016 + \sum_i \Gamma_i \Delta G_i^f \quad \text{(S37)}$$

$$H_{vap}^{298}\left(\frac{KJ}{mol}\right) = 10.4327 + \sum_i \Gamma_i (\Delta H_{vap}^{298})_i \quad \text{(S38)}$$

$$V_L^{298}\left(\frac{cc}{mol}\right) = 1000(0.0123 + \sum_i \Gamma_i (\Delta V_L^{298})_i) \quad \text{(S39)}$$

$$T_b(K) = 244.79 \ln(\sum_i \Gamma_i (\Delta T_b)_i) \quad \text{(S40)}$$

Table S.3. Parameters for the free energy of formation, $\frac{\Delta G_{i,298}}{RT}$, liquid molar volume, $\Delta V_{298}^{L}(cc/\text{mol})$, heat of vaporization, $\Delta H_{298}^{vap}(\frac{KJ}{mol})$, number of acceptors, nA, number of donors, nD, energy of hydrogen bond, $\varepsilon^{HB}/k(K)$.

	Name	$\mathrm{D}G_{i,298}$ kJ/mol	$\mathrm{D}V_{298}^{L}$ cc/mol	$\mathrm{D}H_{298}^{vap}$ kJ/mol	$\mathrm{D}T_b^{1bar}$ K	nA	nD	e^{HB}/k K
1	aC-CHn-X(n[1,2])-X:Halogen	-45.359	0.0001	12.377	0.225	0	0	0.0
2	C-F3	-2.804	0.0412	1.868	1.112	0	0	0.0
3	CH(cyc)-CL	-7.326	0.0240	10.913	-0.050	0	0	0.0
4	CH(cyc)-F	-74.677	0.0182	1.546	-0.094	0	0	0.0
5	CHm=CHn-F(m,n[0,2])	-81.679	0.0069	-5.756	0.044	0	0	0.0
6	CHm=CHn-Br(m,n[0,2])	25.739	0.0002	7.401	-0.181	0	0	0.0
7	(CHn=C)cyc-Cl(n[0..2])	-1.458	-0.0079	0.384	-0.099	0	0	0.0
8	CHm=CHn-	-0.126	0.0027	3.391	-0.024	0	0	0.0

	Cl(m,n[0,2])							
9	aC-Cl	1.252	0.0258	10.184	1.665	0	0	0.0
10	aC-F	0.907	0.0167	4.310	0.683	0	0	0.0
11	aC-I	1.689	0.0361	4.310	2.944	0	0	0.0
12	aC-Br	3.303	0.0300	13.613	2.239	0	0	0.0
13	RCF2	-3.010	0.0294	4.603	0.512	0	0	0.0
14	RCF	8.815	0.0121	-0.829	0.426	0	0	0.0
15	HF	-4.159	0.0209	7.6965	292.67	1	1	3422.2
16	CCL3	-60.459	0.0621	17.951	3.106	0	0	0.0
17	CCL2	-63.766	0.0448	18.972	2.131	0	0	0.0
18	CL<C=C>	-6.080	0.0181	6.018	1.356	0	0	0.0
19	CCL	3.839	0.0188	8.145	0.822	0	0	0.0
20	CHCL	9.691	0.0267	10.872	1.483	0	0	0.0
21	CH2CL	4.467	0.0317	12.461	2.254	0	0	0.0
22	HCL	-0.471	0.030563	9.2885	188.15	1	1	245.5
23	CL2	-77.899	0.045506	20.4097	239.12	0	0	0.0
24	HI	-0.421	0.045718	20.0188	237.55	0	0	0
25	I2	115.442	0.063835	41.8782	457.56	0	0	0
26	HBr	1.804	0.037073	17.9048	206.45	0	0	0
27	Br2	11.444	0.05148	29.7888	331.9	0	0	0
28	F except as above	7.630	0.0123	-3.349	0.718	0	0	0.0
29	Cl except as above	5.715	0.0181	6.018	1.356	0	4	6744.1
30	I except as above	4.641	0.0246	14.364	2.652	0	4	1631.2
31	Br except as above	6.044	0.0213	9.808	2.028	0	4	4516.0
32	SiO	-25.187	0.0283	12.462	-0.342	0	0	0.0
33	Si	-13.115	0.0221	10.229	-0.447	0	0	0.0
34	aC-SO3(sulfonate)-aC	30.620	0.0334	27.960	6.677	0	0	0.0
35	aC-SO2-aC	19.253	0.0694	27.960	6.677	0	0	0.0
36	aC-S-aC(different-rings)	-0.041	0.0477	17.320	0.280	0	0	0.0
37	aC-CHn-SH(n[1,2])	16.484	0.0032	22.239	0.159	1	1	1912.4
38	aC-CHn-S-(n[1,2])	-2.349	0.0394	24.324	-0.140	0	0	0.0
39	CHm(cyc)-S-CHn(cyc)(m,n[0,1])	0.628	0.0167	22.954	2.714	0	0	0.0
40	CH(cyc)-S	2.233	0.0031	18.059	-0.107	0	0	0.0
41	SO3(sulfonate)	11.175	0.0552	15.242	5.014	0	0	0.0
42	aC-SO2	-0.816	0.0214	21.601	4.368	0	0	0.0
43	aC-SO	-3.148	0.0214	21.601	4.368	0	0	0.0
44	aC-SH	5.317	0.0269	16.633	2.523	1	1	1912.4
45	aC-S-	-0.765	0.0317	15.242	1.783	0	0	0.0
46	RSO2	-3.917	0.0164	75.280	3.656	0	0	0.0

47	RS	-4.517	0.0101	13.164	1.886	0	0	0.0
48	SO2	-0.650	0.0143	75.280	3.656	0	0	0.0
49	SO	13.580	0.0143	15.242	4.762	0	0	0.0
50	SO(ring)	-9.805	0.0237	15.242	5.064	0	0	0.0
51	CH3S	-4.472	0.0356	14.289	2.503	0	0	0.0
52	CH2S	-4.132	0.0294	15.753	1.955	0	0	0.0
53	CH2SH	-0.117	0.0341	15.109	2.588	1	1	1912.4
54	>C=S	-3.090	0.0197	93.482	1.428	0	0	0.0
55	S	-1.460	0.0157	13.164	1.195	0	0	0.0
56	P=O	3.992	0.0089	6.970	0.427	0	0	0.0
57	aC-NHCONH-aC(diff-rings)	17.367	0.0515	60.857	-0.013	2	2	1258.2
58	aC-CO-Ncyc(different-rings)	1.640	0.0216	34.954	3.028	0	0	0.0
59	aC-NH-aC(different-rings)	-1.059	0.0039	31.118	0.200	1	1	629.1
60	aC-N-CHcyc(different-rings)	0.412	0.0147	27.387	2.346	0	0	0.0
61	N-multiring	0.977	-0.0001	-0.004	-0.406	0	0	0.0
62	aC-NHn(cyc)(fused-rings)(n[0,1])	0.467	-0.0013	22.492	-0.155	0	0	0.0
63	PYRIIDINE.FUSED[2]	14.451	-0.0034	2.206	-0.151	0	1	5032.7
64	PYRIIDINE.FUSED[2-iso]	-6.961	-0.0034	-3.964	-0.206	0	1	5032.7
65	NH-(CHn)3-COOH(n[0,2])	48.120	0.0554	35.150	6.883	2	3	4152.1
66	NH2-(CHn)3-OH(n[0,2])	17.908	-0.0030	43.985	5.156	2	2	4328.2
67	NHk-(CHn)3-NH2(k[0,1];n[0,2])	19.532	0.0532	36.328	1.982	2	2	1761.5
68	aC-CHn-NHm(n[1,2],m[0,2])	0.205	0.0013	2.650	-0.042	0	0	0.0
69	aC-CHn-NHm(n[1,2],m[0,2])	1.052	0.0013	2.650	-0.042	1	1	629.1
70	aC-CHn-NHm(n[1,2],m[0,2])	5.607	0.0013	2.650	-0.042	1	1	1761.5
71	aCaNaC	2.075	0.0194	0.019	5.291	0	0	0.0
72	aC-CHn-CN(n[1,2])	3.843	0.0013	32.281	0.137	0	1	1509.8
73	aC-CHn-CONH2(n[1,2])	25.140	0.0384	57.556	-0.163	1	1	1761.5
74	(CH2NHCH2)(cyc)	3.456	0.0027	23.513	3.131	1	0	629.1
75	CH(cyc)-NH2	4.563	0.0036	-8.479	-0.356	1	1	1761.5
76	CH(cyc)-NH-CHn(n[0,3])	1.093	0.0027	24.910	-0.241	1	1	629.1
77	CH(cyc)-CN	19.026	0.0061	-3.945	0.328	0	1	1509.8
78	>N(cyc)-CH2	-5.180	0.0042	20.935	-0.174	1	1	1107.2

79	NC-CHn-COO(n[1,2])	38.489	-0.0012	42.765	0.336	0	2	3652.3
80	CHm(NH)CHn(NH2)(m,n[1,2])	21.830	0.0057	2.423	0.282	2	2	2390.6
81	CHn(OH)CHm(NH)(m[0,1],n,p[0,2])	34.932	0.0009	53.882	4.471	1	2	2566.7
82	CHn(OH)CHm(NH)(m[0,1],n[0,2])	5.911	0.0009	53.882	4.471	2	2	3195.8
83	CHn(OH)CHm(NH2)(m[0,1],n[0,2])	2.240	0.0009	53.882	4.471	2	2	4328.2
84	CHm(N)-COOH(m,n[0,2])	52.971	0.0155	38.338	-0.061	1	2	3523.0
85	CHm(NH)-COOH(m,n[0,2])	6.230	0.0155	38.338	-0.061	1	2	4152.1
86	CHm=CHn-CN(m,n[0,2])	-0.786	0.0012	3.162	-0.107	0	1	1509.8
87	aC-NHCONH	23.524	0.0317	125.036	0.750	2	2	1952.4
88	aC-NHCO	-3.581	0.0326	131.395	5.318	1	1	629.1
89	aC-CONH2	7.470	0.0349	131.395	6.314	1	1	2113.7
90	aC-CONH	-3.266	0.0326	131.395	5.433	1	1	2113.7
91	ACNH2	7.217	0.0260	21.681	2.864	1	1	1530.0
92	CON(CH2)2	0.658	0.0533	38.818	3.383	0	0	0.0
93	CONHCH2	-4.468	0.0533	51.416	5.024	0	1	2113.7
94	CONHCH3	15.766	0.0380	47.990	4.220	0	1	2113.7
95	HCONH	-26.000	0.0194	44.146	5.125	1	1	2113.7
96	CONH2	16.104	0.0147	46.396	5.288	1	1	2113.7
97	CNO2	16.481	0.0504	-39.116	2.131	1	2	1107.2
98	aN-in-aromatic-ring	0.033	0.0052	11.803	1.051	0	0	0.0
99	ACNO2	2.427	0.0308	24.083	3.471	0	0	0.0
100	NHCO-except-as-above	-1.489	0.0317	45.564	4.486	1	2	1761.5
101	CH2NCO	8.175	0.0402	29.546	2.885	0	0	0.0
102	CNH2	26.438	0.0144	12.867	0.886	1	1	1761.5
103	CNOH	0.836	0.0227	125.036	3.111	1	1	2566.7
104	CH2NH2	7.291	0.0262	14.100	2.264	1	1	1018.0
105	CHNH2	6.485	0.0214	14.571	1.437	1	1	709.0
106	CH3NH	6.193	0.0279	13.407	1.986	1	1	629.1
107	CH2NH	4.389	0.0246	12.355	1.269	1	1	653.0
108	CHNH	-5.163	0.0182	12.409	0.594	1	1	629.1
109	CH3N	1.452	0.0265	13.292	0.999	0	0	0.0
110	CH2N	3.913	0.0190	8.400	0.332	0	0	0.0
111	CH-N	4.451	0.0128	12.462	0.021	0	0	0.0
112	CH2CN	5.889	0.0314	21.902	3.561	0	1	629.1
113	CHCN	0.719	0.0254	23.692	2.707	0	1	629.1
114	CCN	13.751	0.0211	20.577	1.723	0	1	1509.8

115	HCN	13.891	0.0254	26.956	298.850	1	1	1660.8
116	CH=N	-1.839	0.0182	12.409	1.110	0	1	1610.5
117	C=N	-7.494	0.0265	13.292	0.784	0	0	0.0
118	ONO	-6.364	0.0128	26.206	1.733	0	0	0.0
119	CN-except-as-above	6.026	0.0158	21.121	2.738	0	1	1509.8
120	RCH=N	100.068	0.0220	12.161	5.540	0	1	1610.5
121	RC=N	0.099	0.0220	12.161	4.940	0	0	0.0
122	(R)C=N	7.352	0.0369	19.569	1.260	0	0	0.0
123	RNH	3.333	0.0035	16.133	1.701	0	1	629.1
124	RN	-5.204	0.0017	16.133	1.100	0	0	0.0
125	>NH	-0.354	0.0061	18.400	1.987	0	1	1610.5
126	NH2-except-as-above	2.904	0.0096	18.400	1.987	1	1	1761.5
127	N=N	-9.414	-0.0986	15.745	-1.559	0	0	0.0
128	NH3	11.671	0.0250	19.871	239.720	1	1	3321.6
129	aC-CHm-CO-aC(different-rings)(m[0,2])	31.853	0.0436	29.981	0.052	0	0	0.0
130	aC-CO-aC(different-rings)	0.590	-0.0073	25.179	0.052	0	0	0.0
131	aC-CO-CHn(cyc)(different-rings)(n[0,1])	24.531	0.0258	23.716	0.335	0	0	0.0
132	aC-CHn-O-CHm-aC(different-rings)(n,m[0,1])	56.245	0.0010	21.520	-0.048	0	1	2143.3
133	aC-O-CHn-aC(different-rings)(n[0,2])	7.626	0.0104	-0.800	47.972	0	1	2143.3
134	aC-O-aC(different-rings)	11.609	0.0017	11.917	-0.229	0	1	2143.3
135	aC-CO(cyc)(fused-rings)	1.396	0.0043	18.821	-0.182	0	0	0.0
136	aC-O(cyc)(fused-rings)	-0.430	0.0042	12.929	-0.287	0	0	0.0
137	OH-(CHn)3-OH(n[0,2])	0.274	0.0250	0.769	5.073	2	2	5133.4
138	CHp-O-(CHn)3-OH(n,p[0,2])	4.436	0.0478	32.773	4.542	1	2	4710.0
139	CHp-O-(CHn)4-OH(n,p[0,2])	53.122	0.0478	32.773	4.542	1	2	4710.0
140	aC-CHn-CHO(n[1,2])	43.217	-0.0009	138.941	0.291	1	1	2315.0
141	aC-CHn-OOC(n[1,2])	-1.153	0.0064	28.003	-0.033	0	0	0.0
142	aC-CHn-COOH(n[1,2])	2.588	0.0271	-404.162	0.205	1	2	3523.0
143	aC-CHn-COO(n[1,2])	-9.168	0.0064	28.003	0.132	0	1	2143.3

144	aC-CHn-CO-(n[1,2])	1.926	0.0365	23.622	0.165	0	0	0.0
145	aC-CHn-OH(n[1,2])	4.361	0.0010	10.360	0.071	1	1	2566.7
146	aC-CHn-O-(n[1,2])	1.595	-0.0013	10.360	0.091	0	1	2143.3
147	AC-O-CHm(m[0,3])	-1.397	-0.0013	10.360	0.091	0	1	2143.3
148	(COOCH2)(cyc)	5.943	0.0191	3.690	0.827	0	1	2143.3
149	(CH2OCH2)(cyc)	0.313	0.0394	13.951	2.314	0	1	2143.3
150	(CH2OCH)(cyc)	2.779	0.0394	15.156	2.013	0	1	2143.3
151	(CHn=C)(cyc)-CO(n[0,2])	62.917	0.0366	14.076	0.420	0	0	0.0
152	CH(cyc)-COO	1.925	0.0031	0.000	0.113	0	1	2143.3
153	CH(cyc)-COOH	6.086	0.0050	20.031	0.181	1	2	3523.0
154	CH(cyc)-OOC	3.655	-0.0368	0.000	-0.348	0	0	0.0
155	CH(cyc)-OH	0.951	-0.0007	3.927	-0.338	1	1	2566.7
156	CH(cyc)-CO	-1.488	-0.0067	3.257	0.018	0	0	0.0
157	CH(cyc)-CHO	-4.780	0.0048	77.428	-0.092	1	1	2315.0
158	CH(cyc)-O	-3.426	0.0062	4.095	-0.040	0	0	0.0
159	C(cyc)-OH	0.331	-0.0719	28.737	-0.367	1	1	2566.7
160	OH-CHn-COO(n[0,2])	16.820	0.0012	46.568	-0.217	1	2	4710.0
161	HO-CHn-COOH(n[1,2])	28.777	0.0171	43.908	-0.159	2	3	6089.7
162	CHm(OH)CHn(OH)(m,n[0,2])	-0.564	0.0020	-5.631	0.065	2	2	5133.4
163	CH3COOCHm(m[0,1])	64.794	0.0023	-2.086	-0.121	0	1	2143.3
164	CHmCOOH(m[0,1])	8.371	0.0050	10.126	-0.033	1	2	3523.0
165	CHmCHO(m[0,1])	10.617	0.0021	16.052	-0.057	1	1	2315.0
166	CH3COCH2	15.179	0.0003	2.068	-0.038	0	0	0.0
167	CH3COCHm(m[0,1])	10.396	0.0003	3.857	-0.212	0	0	0.0
168	(CHdCHOCHdCH)(cyc)	16.513	0.0714	3.857	-0.212	0	1	2143.3
169	CHm-O-CHn=CHp(m,n,p[0,3])	2.375	0.0005	6.399	0.057	0	1	2143.3
170	CHn=CHm-COO-CHp(m,n,p[0,3])	8.899	0.0019	16.050	0.086	0	1	2143.3
171	(CHn=CHm)cyc-COOH	-10.880	8.4548	-266.369	0.250	1	2	3523.0
172	CHm=CHn-COOH(m,n,p[0,3])	8.897	0.0046	12.729	0.136	1	2	3523.0
173	CHm=CHn-CHO(m,n,[0,2])	13.262	-0.0018	22.401	0.219	1	1	2315.0
174	CHOH	1.328	0.0001	1.068	-0.185	1	1	2011.0
175	aC-COOH	3.818	0.0478	21.495	4.594	1	2	3523.0
176	aC-COO	0.585	0.0312	23.019	2.119	0	1	2143.3

177	aC-CO	-0.343	0.0208	22.849	2.252	0	1	2143.3
178	aC-CHO	5.742	0.0187	44.751	2.671	1	1	2315.0
179	aC-OCO	3.363	0.0219	23.019	2.049	0	0	0.0
180	aC-O	-10.523	0.0084	14.027	1.148	0	1	2143.3
181	ACOH	0.041	0.0170	35.655	2.546	1	1	2566.7
182	aC=O	-3.505	0.0175	5.559	1.970	0	0	0.0
183	RCO	3.945	0.0120	18.229	2.202	0	1	2143.3
184	CH2OCHO	19.632	0.0395	20.054	3.325	1	2	#REF!
185	CH3COOCH2	5.852	0.0589	24.966	3.198	0	1	2143.3
186	CH2COOCH3	2.195	0.0589	22.626	3.040	0	1	2143.3
187	CH2COOCH2	5.255	0.0589	29.833	3.093	0	1	2143.3
188	CHCOOCH2	2.389	0.0589	26.701	2.025	0	1	2143.3
189	CCOOCH2	20.202	0.0589	22.987	1.157	0	1	2143.3
190	OCH2CH2O	1.202	0.0456	17.641	1.950	2	2	4630.0
191	OCH2CH2OH	12.482	0.0417	31.206	3.818	1	2	2392.0
192	CO3(carbonate)	19.491	0.0231	20.330	2.208	0	0	0.0
193	CH3COO	36.855	0.0423	20.165	2.620	0	1	2143.3
194	CH2COO	-8.930	0.0364	21.012	2.118	0	1	2143.3
195	CHCOO	4.943	0.0284	21.899	1.447	0	1	2143.3
196	CCOO	35.496	0.0224	18.185	0.579	0	1	2143.3
197	CH3CO	4.798	0.0347	12.850	2.691	0	1	2143.3
198	CH2CO	-1.521	0.0283	16.738	1.967	0	1	2143.3
199	CHOCH	0.342	0.0272	10.711	0.900	1	1	2315.0
200	CHCO	-3.379	0.0199	16.121	1.193	0	1	2143.3
201	COOH	-0.170	0.0207	15.136	3.974	1	2	3523.0
202	COO	0.463	0.0196	16.843	1.580	0	1	2143.3
203	CCO	1.566	0.0281	14.591	0.593	0	1	2143.3
204	C2H2O	10.224	0.0290	15.576	1.785	1	1	2143.3
205	C2HO	27.600	0.0290	15.576	1.785	1	1	2315.0
206	CH3O	1.813	0.0281	7.455	1.584	0	1	2143.3
207	CH2O	-2.018	0.0228	8.821	0.975	0	1	2143.3
208	CHO	3.949	0.0167	12.798	2.102	1	1	2315.0
209	C-O	-1.137	0.0049	9.328	-0.339	0	1	2143.3
210	CH-O	0.268	0.0205	10.296	0.327	0	1	2143.3
211	OH	4.822	0.0042	23.971	2.248	1	1	2566.7
212	O2	14.520	0.028023	0.000	90.188	0	0	0.0
213	CO2	8.956	0.035443	5.3113	2.2476	0	0	0.0
214	=O	-5.572	0.0104	-0.800	1.138	0	1	2143.3
215	O	4.137	0.0104	-0.800	1.138	0	1	2143.3
216	aC-	-0.609	-0.0017	13.799	-0.365	0	0	0.0

	(CHn=CHm)(cyc)(fused-rings)(n,m[0,1])							
217	aC-(CHm=CHn)-aC(different-rings)(m,n[0,2])	33.804	-0.0043	50.436	0.281	0	0	0.0
218	aC-(CHn)2-aC(different-rings)(n[0,2])	45.766	-0.0037	2.819	0.541	0	0	0.0
219	aC-CHm-aC(different-rings)	0.177	0.0001	3.084	0.058	0	0	0.0
220	aC-aC(different-rings)	2.150	0.0070	2.506	-0.007	0	0	0.0
221	aC-(CHn)2-CHcyc(different-rings)(n[0,2])	49.149	0.0462	20.856	-0.213	0	0	0.0
222	aC-(CHn)-CHcyc(different-rings)(n[0,2])	-11.260	0.0462	20.856	-0.213	0	0	0.0
223	aC-CHn,cyc(different-rings)(n[0,1])	-0.468	0.0017	0.000	-0.272	0	0	0.0
224	aC-CHn,cyc(fused-rings)(n[0,1])	0.102	-0.0025	-1.237	-0.356	0	0	0.0
225	AROMFUSED[2]s1	55.145	-0.0069	-1.910	-0.161	0	2	2000.0
226	AROMFUSED[2]s2	144.653	0.0063	-1.910	-0.183	0	2	2000.0
227	aC-CH(CH3)2	-4.948	0.0014	1.972	0.130	0	0	0.0
228	aCCH(CH2)2	58.544	0.0014	1.972	0.130	0	0	0.0
229	aCCH(CH3)(CH2)	61.310	0.0014	1.972	0.130	0	0	0.0
230	aC-C(CH3)3	12.079	0.0019	-1.830	0.169	0	0	0.0
231	AROMRINGs1s2s3s4s5	166.448	-0.0148	0.190	0.061	0	1	1000.0
232	AROMRINGs1s2s3s4	103.233	-0.0118	6.707	0.141	0	1	1000.0
233	AROMRINGs1s2s3s5	146.349	-0.0085	6.707	0.077	0	1	1000.0
234	AROMRINGs1s2s4s5	153.340	-0.0051	6.707	0.132	0	1	1000.0
235	AROMRINGs1s2s3	-16.567	-0.0035	4.132	0.058	0	1	1000.0
236	AROMRINGs1s2s4	-21.176	-0.0048	4.722	0.063	0	1	1000.0
237	AROMRINGs1s3s5	141.652	-0.0036	6.707	0.018	0	1	1000.0
238	AROMRINGs1s2	-3.095	-0.0069	6.707	0.018	0	1	1000.0
239	AROMRINGs1s3	-13.818	-0.0049	6.707	0.018	0	1	1000.0
240	AROMRINGs1s4	-3.026	-0.0015	6.707	0.018	0	1	1000.0
241	aC-CH=CH2	10.636	0.0409	5.888	1.886	0	0	0.0
242	aC-CH=CH	1.411	0.0287	3.952	1.878	0	0	0.0
243	aC-C=CH2	-0.091	0.0297	3.382	1.417	0	0	0.0
244	aC-C#CH	-16.355	0.0347	12.734	1.808	0	0	0.0

245	aC-C#C	-2.808	0.0413	17.730	2.051	0	0	0.0
246	ACCH3	4.457	0.0302	6.994	1.262	0	0	0.0
247	ACCH2	-1.528	0.0229	8.995	0.853	0	0	0.0
248	ACCH	-1.114	0.0162	9.008	0.027	0	0	0.0
249	aC-C	18.515	0.0095	13.769	-0.605	0	0	0.0
250	aC-fuseArmtcRing	0.479	0.0046	5.797	1.253	0	0	0.0
251	aC-fuseNonArmtcSubrng	-7.215	0.0035	5.525	1.161	0	0	0.0
252	ACH	0.058	0.0120	4.098	0.733	0	0	166.7
253	AC	-0.057	0.0071	6.359	0.832	0	0	0.0
254	CHcyc-Chcyc(different-rings)	37.501	0.0118	9.790	-0.104	0	0	0.0
255	CH multiring	0.317	-0.0002	-1.641	0.125	0	0	0.0
256	C multiring	-6.260	-0.1393	-0.150	0.231	0	0	0.0
257	CH(cyc)-C=CHn(n[1,2])	91.083	-0.0028	-0.739	0.022	0	0	0.0
258	3-membered-ring	8.459	0.0262	13.480	1.542	0	0	0.0
259	4-membered-ring	3.507	0.0382	17.170	2.257	0	0	0.0
260	5-membered-ring	-7.019	0.0502	20.860	2.972	0	0	0.0
261	6-membered-ring	6.756	0.0622	24.550	3.686	0	0	0.0
262	RC-CH2CH3	-0.004	0.1725	1.327	0.792	0	0	0.0
263	RC-CH2CH2CH3	62.682	0.1891	6.128	1.370	0	0	0.0
264	CH(cyc)-CH3	5.524	0.0029	0.247	-0.168	0	0	0.0
265	CH(cyc)-CH2	-7.607	0.0019	-0.287	-0.130	0	0	0.0
266	CH(cyc)-CH	24.297	-0.0062	1.058	0.032	0	0	0.0
267	CH(cyc)-C	39.577	0.0056	1.971	-0.064	0	0	0.0
268	RC-CH3	2.421	-0.0682	0.794	0.002	0	0	0.0
269	C(cyc)-CH2	10.420	-0.0722	1.545	0.174	0	0	0.0
270	RCH=CH	5.000	0.0246	7.441	1.377	0	0	0.0
271	(R)C=CH	-0.655	0.0269	10.761	0.912	0	0	0.0
272	RCH=C	1.453	0.0020	5.948	0.925	0	0	0.0
273	RC=C	-10.911	-0.0361	5.948	0.662	0	0	0.0
274	(R)C=C	28.961	-0.0006	8.355	0.427	0	0	0.0
275	RCH2=C	9.927	0.0313	3.976	1.253	0	0	0.0
276	RCH2	0.334	0.0160	3.690	0.715	0	0	0.0
277	RCH	-2.726	0.0059	4.895	0.414	0	0	0.0
278	RC	2.903	0.1393	3.436	-0.381	0	0	0.0
279	CH3-CHm=CHn(m,n[0,2])	5.185	0.0027	5.714	-0.006	0	0	0.0
280	CH2-	-1.681	0.0012	5.364	-0.106	0	0	0.0

	CHm=CHn(m,n[0,2])							
281	CHp-CHm=CHn(m,n[0,2];p[0,1])	39.057	0.0020	5.225	-0.020	0	0	0.0
282	<CH3>2CH	1.496	0.0021	-0.232	0.056	0	0	0.0
283	<CH3>3C	1.446	0.0045	1.480	0.046	0	0	0.0
284	CH2=CH	5.825	0.0333	-0.471	1.495	0	0	0.0
285	CH=CH	-0.126	0.0244	-2.407	1.200	0	0	0.0
286	CH2=C	26.502	0.0230	-2.976	1.031	0	0	0.0
287	CH=C	31.329	0.0108	-5.634	0.765	0	0	0.0
288	CH3	0.227	0.0238	1.614	0.922	0	0	0.0
289	CH2	-0.115	0.0166	4.801	0.578	0	0	0.0
290	CH	-1.501	0.0084	5.755	-0.119	0	0	0.0
291	C	-1.434	-0.0015	4.918	-0.650	0	0	0.0
292	CH#C	-0.841	0.0159	6.375	1.588	0	0	0.0
293	C#C	0.378	0.0159	11.371	1.272	0	0	0.0
294	H2	10.835	28.568	0	20.39	0	0	0.0
295	Ethylene	-1.303	49.321	6.8431	169.41	0	0	0.0
296	Ethyne	47.606	42.209	6.8431	20.39	0	0	0.0

Formulas for the group contributions to the enthalpy of formation and heat capacity and implemented based on Equations S41-S43.[14] The parameters are taken from the work published by Kolska et al.[14]. For larger functional groups absent in the above work, we extended the parameters based on similarity of the structure or by the additivity rule from smaller groups.

$$H^f\left(\frac{KJ}{mol}\right) = 83.9657 + \sum_i \Gamma_i (\Delta H^f)_i \quad \text{(S41)}$$

$$C_p^\ell(T) = C_{p0}^\ell(T) + \sum_i \Gamma_i \, C_{p,i}^\ell(T) \quad \text{(S42)}$$

$$C_{p0,i}^\ell(T) = a_{0,i} + b_{0,i}\left(\frac{T}{100}\right) + c_{0,i}\left(\frac{T}{100}\right)^2 \quad \text{(S43)}$$

Table S.4. Heat Capacity Group Contribution Parameters

	Name	$D H^f_{298}$ kJ/mol	*a* J/(mol.K)	*b* J/(mol.K^2)	*c* J/(mol.K^3)
1	aC-CHn-X(n[1,2])-X:Halogen	-5.084	57.09	-45.85	9.65
2	C-F3	-708.662	-29.59	55.77	-6.9
3	CH(cyc)-CL	-161.494	-78.46	40.5	-4.38
4	CH(cyc)-F	-362.113	-78.46	40.5	-4.38
5	CHm=CHn-F(m,n[0,2])	42.559	43.05	-39.85	7.54
6	CHm=CHn-Br(m,n[0,2])	12.098	43.05	-39.85	7.54
7	(CHn=C)cyc-Cl(n[0..2])	50.050	43.05	-39.85	7.54
8	CHm=CHn-Cl(m,n[0,2])	-0.999	-32.11	-6.64	5.2
9	aC-Cl	-27.439	18.32	11.79	-1.69
10	aC-F	-173.328	8.2	14.34	-2.09
11	aC-I	84.300	15.62	25.52	-5.08
12	aC-Br	28.183	6.08	22.28	-3.6
13	RCF2	-500.727	12.93	19.38	-2.34
14	RCF	-264.327	35.87	-45.37	12.02
15	HF	-273.3	-43.41708	29.09534	-9.42E-01
16	CCL3	-142.227	57.68	16.51	-1.3
17	CCL2	-120.295	43.52	24.65	-4.95
18	CL<C=C>	-88.598	39.99	-5.52	-0.14
19	CCL	-12.451	93.83	-23.16	3.38
20	CHCL	-59.584	93.83	-23.16	3.38

21	CH2CL	-110.048	10	21.75	-2.18
22	HCL	-92.31	-58.64237	60.40135	-7.24
23	CL2	0.000	-41.99575	56.02672	-8.86094
24	HI	0	-27.37453	43.25987	-7.25991
25	I2	0	-25.271	51.4	-7.24
26	HBr	0	-48.26905	52.44007	-7.25277
27	Br2	0	-68.37574	84.25338	-13.94047
28	F except as above	-289.217	-11.58	12.82	-1.08
29	Cl except as above	-88.598	39.99	-5.52	-0.14
30	I except as above	19.997	-95.22	85.97	-14.69
31	Br except as above	-41.477	-21.39	32.75	-5.29
32	SiO	-208.391	-108.784	75.69	-11.37
33	Si	-208.391	-78.744	51.4	-7.24
34	aC-SO3(sulfonate)-aC	-327.517	-137.3166	131.3234	-19.7866
35	aC-SO2-aC	-327.517	366.5917	-152.4483	20.9417
36	aC-S-aC(different-rings)	30.131	-509.98	283.05	-41.17
37	aC-CHn-SH(n[1,2])	9.593	506.25	-237.78	34.29
38	aC-CHn-S-(n[1,2])	24.627	527.88	-249.95	35.08
39	CHm(cyc)-S-CHn(cyc)(m,n[0,1])	#REF!	47.26	-37.5	6.61
40	CH(cyc)-S	-77.103	-45.91	27.08	-4.31
41	SO3(sulfonate)	-426.487	-138.98	129.66	-21.45
42	aC-SO2	-345.781	365.76	-153.28	20.11
43	aC-SO	-16.863	395.8	-177.57	24.24
44	aC-SH	30.900	40.68	15.13	-2.91
45	aC-S-	12.962	99.88	-18.58	1.85
46	RSO2	-318.096	-108.94	105.37	-17.32
47	RS	-4.207	-2.8	24.33	-3.96
48	SO2	-318.096	-108.94	105.37	-17.32
49	SO	-66.348	-78.9	81.08	-13.19
50	SO(ring)	-301.685	-78.9	81.08	-13.19
51	CH3S	-40.401	45.3	7.02	-0.08
52	CH2S	12.962	53.18	8.7	-2.35
53	CH2SH	-46.550	31.55	20.87	-3.14
54	>C=S	99.363	54.06	22.96	-6.19
55	S	-4.207	-48.86	56.79	-9.06
56	P=O	60.854	-78.9	81.08	-13.19
57	aC-NHCONH-aC(diff-rings)	-157.616	69.92	33.05	-6.13
58	aC-CO-Ncyc(different-rings)	-71.778	-105.24	154.97	-27.01

59	aC-NH-aC(different-rings)	-20.041	565.25	-274.99	39.49
60	aC-N-CHcyc(different-rings)	63.718	145.84	-46.2	5.92
61	N-multiring	12.973	51.85	-3.48	0.44
62	aC-NHn(cyc)(fused-rings)(n[0,1])	2.234	90.55	-16.34	2.06
63	PYRIIDINE.FUSED[2]	12.210	-25.43333333	13.00666667	-1.4
64	PYRIIDINE.FUSED[2-iso]	12.076	-25.43333333	13.00666667	-1.4
65	NH-(CHn)3-COOH(n[0,2])	-511.878	60.79	47.75	-5.16
66	NH2-(CHn)3-OH(n[0,2])	-0.649	61.29	9.08	4.07
67	NHk-(CHn)3-NH2(k[0,1];n[0,2])	215.717	126.97	35.28	-10.64
68	aC-CHn-NHm(n[1,2],m[0,2])	-13.470	126.97	35.28	-10.64
69	aC-CHn-NHm(n[1,2],m[0,2])	-13.470	270.04	-18.77	-4.32
70	aC-CHn-NHm(n[1,2],m[0,2])	-13.470	96.42	54.1	-9.9
71	aCaNaC	0.000	61.37	-47.23	9.64
72	aC-CHn-CN(n[1,2])	-24.328	522.57	-247.6	35.66
73	aC-CHn-CONH2(n[1,2])	-207.863	132.13	1.44	-0.61
74	(CH2NHCH2)(cyc)	-33.756	271.87	-166.98	24.72
75	CH(cyc)-NH2	25.227	-198.45	149.07	-26.01
76	CH(cyc)-NH-CHn(n[0,3])	48.296	-147.48	129.95	-22.3
77	CH(cyc)-CN	55.151	113.61	-57.97	10.89
78	>N(cyc)-CH2	-2.647	-192.82	175.48	-27.81
79	NC-CHn-COO(n[1,2])	-199.797	49.51	28.15	1.64
80	CHm(NH)CHn(NH2)(m,n[1,2])	0.321	443.66	-91.64	1.26
81	CHn(OH)CHm(NH)(m[0,1],n,p[0,2])	0.629	-194.13	102.9	-13.55
82	CHn(OH)CHm(NH)(m[0,1],n[0,2])	0.629	-194.13	102.9	-13.55
83	CHn(OH)CHm(NH2)(m[0,1],n[0,2])	0.629	-194.13	102.9	-13.55
84	CHm(N)-COOH(m,n[0,2])	9.813	60.79	47.75	-5.16
85	CHm(NH)-COOH(m,n[0,2])	9.813	234.41	-25.12	0.42
86	CHm=CHn-CN(m,n[0,2])	-46.987	9.56	-3.63	2.19
87	aC-NHCONH	-97.770	79.98	21.43	-2.91
88	aC-NHCO	-135.800	28.13	24.91	-3.35
89	aC-CONH2	-176.920	155.62	-20.78	1.65
90	aC-CONH	-135.800	155.62	-20.78	1.65
91	ACNH2	3.666	133.67	-25.76	2.28
92	CON(CH2)2	-224.927	126.05	22.45	-7.48

93	CONHCH2	-179.267	252.93	-34.81	-1.57
94	CONHCH3	-239.927	45.69	55.78	-10.41
95	HCONH	-147.255	82.95	7.53	-1.76
96	CONH2	-236.697	82.95	7.53	-1.76
97	CNO2	-4.246	-99.91	-99.91	-99.91
98	aN-in-aromatic-ring	54.183	10.79	5.74	-0.98
99	ACNO2	-31.208	380.42	-171.77	23.67
100	NHCO-except-as-above	-185.285	82.95	7.53	-1.76
101	CH2NCO	-29.200	109.86	19.24	-7.89
102	CNH2	52.450	66.17	11.97	-2.34
103	CNOH	-147.255	37.015	5.965	32.955
104	CH2NH2	-49.197	48.21	27.05	-4.95
105	CHNH2	2.674	34.31	39.4	-7.34
106	CH3NH	-16.011	14.59	44.77	-8.21
107	CH2NH	28.577	221.83	-45.82	0.63
108	CHNH	70.206	78.76	8.23	-5.69
109	CH3N	55.844	-46.47	75.91	-13.72
110	CH2N	103.574	78.76	8.23	-5.69
111	CH-N	-208.391	78.76	8.23	-5.69
112	CH2CN	53.668	36.93	7.45	1.85
113	CHCN	112.337	42.4	9.25	0.04
114	CCN	127.276	47.87	11.05	-1.77
115	HCN	108.190	35.9025	17.4925	11.0825
116	CH=N	70.206	23.935	23.935	23.935
117	C=N	55.844	23.935	23.935	23.935
118	ONO	-133.896	-99.91	-99.91	-99.91
119	CN-except-as-above	128.047	9.56	-3.63	2.19
120	RCH=N	42.085	23.935	23.935	23.935
121	RC=N	27.942	23.935	23.935	23.935
122	(R)C=N	203.408	23.935	23.935	23.935
123	RNH	24.786	-209.01	172.27	-28.22
124	RN	87.128	-209.01	172.27	-28.22
125	>NH	5.158	51.85	-3.48	0.44
126	NH2-except-as-above	5.158	51.85	-3.48	0.44
127	N=N	181.949	-99.91	-99.91	-99.91
128	NH3	-16.400	861.8139	-621.79385	116.52189
129	aC-CHm-CO-aC(different-rings)(m[0,2])	-130.072	152.95	-23.39	2.36
130	aC-CO-aC(different-rings)	18.212	578.47	-275.95	38.64
131	aC-CO-CHn(cyc)(different-rings)(n[0,1])	-231.802	97.91	-5.87	0.67

132	aC-CHn-O-CHm-aC(different-rings)(n,m[0,1])	57.891	-30.04	24.29	-4.13
133	aC-O-CHn-aC(different-rings)(n[0,2])	59.096	-30.04	24.29	-4.13
134	aC-O-aC(different-rings)	-8.883	-476	250.62	-34.67
135	aC-CO(cyc)(fused-rings)	46.370	103.77	-17.3	1.21
136	aC-O(cyc)(fused-rings)	18.859	103.77	-17.3	1.21
137	OH-(CHn)3-OH(n[0,2])	1.258	-776.13	474.81	-74.6
138	CHp-O-(CHn)3-OH(n,p[0,2])	-254.898	83.63	-10.86	7.77
139	CHp-O-(CHn)4-OH(n,p[0,2])	-254.898	83.63	-10.86	7.77
140	aC-CHn-CHO(n[1,2])	-136.941	80.28	4.92	-1.05
141	aC-CHn-OOC(n[1,2])	0.023	61.76	14.61	0.94
142	aC-CHn-COOH (n[1,2])	169.113	61.76	14.61	0.94
143	aC-CHn-COO(n[1,2])	0.023	61.76	14.61	0.94
144	aC-CHn-CO-(n[1,2])	-179.557	450.77	-226.43	32.98
145	aC-CHn-OH(n[1,2])	-1.313	-233.61	171.75	-28.9
146	aC-CHn-O-(n[1,2])	4.386	-60.08	48.57	-8.26
147	AC-O-CHm(m[0,3])	4.386	56.52	26.51	-5.35
148	(COOCH2)(cyc)	-409.360	-174.57	123.71	-22.84
149	(CH2OCH2)(cyc)	#REF!	-126.93	100.97	-20.41
150	(CH2OCH)(cyc)	#REF!	49.63	-16.59	4.07
151	(CHn=C)(cyc)-CO(n[0,2])	-292.430	80.92	-16.07	3.06
152	CH(cyc)-COO	-380.089	116.63	-33.64	8.49
153	CH(cyc)-COOH	-505.892	116.63	-33.64	8.49
154	CH(cyc)-OOC	53.849	116.63	-33.64	8.49
155	CH(cyc)-OH	62.903	-173.4	96.39	-11.09
156	CH(cyc)-CO	-281.287	135.15	-43.33	6.5
157	CH(cyc)-CHO	-193.142	135.15	-43.33	6.5
158	CH(cyc)-O	-72.674	74.01	-30.05	4.57
159	C(cyc)-OH	-64.767	31.28	-11.21	6.36
160	OH-CHn-COO(n[0,2])	-483.359	195.44	29.41	-9.18
161	HO-CHn-COOH(n[1,2])	-667.466	195.44	29.41	-9.18
162	CHm(OH)CHn(OH)(m,n[0,2])	-9.440	-169.45	150.35	-31.31
163	CH3COOCHm(m[0,1])	-17.089	-338.6	142.49	-16.75
164	CHmCOOH(m[0,1])	0.974	-127.32	62.69	-7.4
165	CHmCHO(m[0,1])	-2.747	-364.47	252.44	-42.14
166	CH3COCH2	-5.694	-71.15	54.13	-9.95
167	CH3COCHm(m[0,1])	-0.756	-84.87	63.96	-12.01
168	(CHdCHOCHdCH)(cyc)	-0.756	-151.24	110.58	-20.98

169	CHm-O-CHn=CHp(m,n,p[0,3])	-17.230	45.65	-52.96	11.86
170	CHn=CHm-COO-CHp(m,n,p[0,3])	-40.610	63.7	-55.53	10.43
171	(CHn=CHm)cyc-COOH	176.587	31.34	21.22	2.07
172	CHm=CHn-COOH(m,n,p[0,3])	2.020	850.61	-562.94	92.16
173	CHm=CHn-CHO(m,n,[0,2])	-207.076	163.18	-105.6	17.22
174	CHOH	-12.929	-393.77	284.02	-47.66
175	aC-COOH	-379.234	-738.97	370.11	-39.06
176	aC-COO	-282.861	87.72	-6.65	0.79
177	aC-CO	-85.931	103.77	-17.3	1.21
178	aC-CHO	-119.335	26.21	18.4	-2.5
179	aC-OCO	-282.861	471.3	-215.04	30.15
180	aC-O	-65.836	40.33	23.3	-5.76
181	ACOH	-179.270	178.51	-34.86	1.98
182	aC=O	49.707	31.1	11.01	-2.2
183	RCO	-208.391	33.03	8.24	-1.69
184	CH2OCHO	-336.815	-40.23	40.1	-5.38
185	CH3COOCH2	-442.215	-21.27	34.02	-9.4
186	CH2COOCH3	-442.717	21.12	-35.22	10.86
187	CH2COOCH2	-506.407	16.41	-37.05	11.15
188	CHCOOCH2	-323.884	7.84	20.47	-2.85
189	CCOOCH2	-323.884	-23.93	32.22	-4.45
190	OCH2CH2O	-295.457	23.69	-11.32	0.99
191	OCH2CH2OH	-385.535	-5.75	59.46	-4.83
192	CO3(carbonate)	-513.687	130.69	-29.16	4.45
193	CH3COO	-421.564	58.31	-4.75	5.65
194	CH2COO	-358.678	91.18	23.69	-9.1
195	CHCOO	-303.234	7.84	20.47	-2.85
196	CCOO	-303.234	7.84	20.47	-2.85
197	CH3CO	-213.078	127.88	-47.29	10.19
198	CH2CO	-154.663	65.01	15.15	-3.53
199	CHOCH	75.527	71.9	-13.95	3.29
200	CHCO	-91.194	67.08	10.77	-3.23
201	COOH	-432.997	12.58	20.7	-0.21
202	COO	-307.193	-3.4	43.61	-7.28
203	CCO	-91.194	-23.93	32.22	-4.45
204	C2H2O	-167.929	49.63	-16.59	4.07
205	C2HO	-167.929	49.63	-16.59	4.07
206	CH3O	-195.368	7.24	25.2	-2.53

207	CH2O	-147.729	42.67	9.68	-1.36
208	CHO	-165.775	31.1	11.01	-2.2
209	C-O	-29.577	31.1	11.01	-2.2
210	CH-O	-98.994	54.36	3.9	-1.66
211	OH	-213.819	13.08	-17.97	9.02
212	O2	0.000	0.50658	-86.36543	78.64939
213	CO2	-393.51	1174.58453	-969.31549	210.12214
214	=O	0.222	-30.04	24.29	-4.13
215	O	0.222	-30.04	24.29	-4.13
216	aC-(CHn=CHm)(cyc)(fused-rings)(n,m[0,1])	8.105	-361	234.7	-37.84
217	aC-(CHm=CHn)-aC(different-rings)(m,n[0,2])	-5.325	-134.98	110.23	-17.46
218	aC-(CHn)2-aC(different-rings)(n[0,2])	2.202	-162.91	94.57	-13.04
219	aC-CHm-aC(different-rings)	3.475	-162.91	94.57	-13.04
220	aC-aC(different-rings)	4.477	-949.7	542.18	-78.94
221	aC-(CHn)2-CHcyc(different-rings)(n[0,2])	-64.712	99	-33.17	6.41
222	aC-(CHn)-CHcyc(different-rings)(n[0,2])	-64.712	99	-33.17	6.41
223	aC-CHn,cyc(different-rings)(n[0,1])	72.466	-352.63	223.53	-34.99
224	aC-CHn,cyc(fused-rings)(n[0,1])	12.613	-352.63	223.53	-34.99
225	AROMFUSED[2]s1	5.225	-392.02	161.3	-16.51
226	AROMFUSED[2]s2	12.377	-426.19	177.11	-18.53
227	aC-CH(CH3)2	5.437	5.51	-12.78	3.67
228	aCCH(CH2)2	5.437	5.51	5.51	5.51
229	aCCH(CH3)(CH2)	5.437	5.51	5.51	5.51
230	aC-C(CH3)3	5.589	43.76	-23.35	4.39
231	AROMRINGs1s2s3s4s5	1.600	73.7	-55.26	10.65
232	AROMRINGs1s2s3s4	20.771	-70.85	47.97	-7.66
233	AROMRINGs1s2s3s5	15.526	-32.41	20.35	-2.69
234	AROMRINGs1s2s4s5	12.146	-76.59	53.81	-8.73
235	AROMRINGs1s2s3	6.140	22.01	-14.75	2.98
236	AROMRINGs1s2s4	1.127	-15.05	7.75	-0.47
237	AROMRINGs1s3s5	-5.880	-60.51	29.24	-3.03
238	AROMRINGs1s2	-5.880	-5.82	1.12	0.26
239	AROMRINGs1s3	-5.880	-49.93	25.88	-2.94
240	AROMRINGs1s4	-5.880	-32.01	14.84	-1.3

241	aC-CH=CH2	64.545	-0.29	38.36	-5.08
242	aC-CH=CH	123.256	-134.98	110.23	-17.46
243	aC-C=CH2	126.246	-62.49	82	-13.56
244	aC-C#CH	246.700	58.32	-2.92	1.88
245	aC-C#C	279.780	58.32	-2.92	1.88
246	ACCH3	-33.270	20.23	9.63	-0.43
247	ACCH2	31.782	49.18	-6.09	1.15
248	ACCH	82.221	96.82	-30	3.37
249	aC-C	142.857	52.7	-15.1	1.17
250	aC-fuseArmtcRing	20.230	384.23	-151.04	15.62
251	aC-fuseNonArmtcSubrng	13.163	539.17	-291.04	41.64
252	ACH	-0.673	-1.28	8.17	-0.43
253	AC	49.485	474.7	-258.65	37.43
254	CHcyc-Chcyc(different-rings)	-145.791	-16.53	32.62	-7.48
255	CH multiring	60.144	-82.22	50.32	-8.17
256	C multiring	96.182	-82.22	50.32	-8.17
257	CH(cyc)-C=CHn(n[1,2])	65.752	-33.33	7.74	0.06
258	3-membered-ring	-175.062	202.24	-97.25	16.86
259	4-membered-ring	-204.333	196.38	-85.82	16.32
260	5-membered-ring	-233.604	190.52	-74.39	15.78
261	6-membered-ring	-262.875	184.66	-62.96	15.24
262	RC-CH2CH3	-5.386	23.64	27.67	-3.4
263	RC-CH2CH2CH3	-26.037	39.83	30.88	-2.99
264	CH(cyc)-CH3	86.878	-60.66	37.5	-5.87
265	CH(cyc)-CH2	78.653	-50.85	31.05	-4.97
266	CH(cyc)-CH	80.554	-151.74	87.41	-13.13
267	CH(cyc)-C	0.014	-35.95	20.41	-3.2
268	RC-CH3	-23.344	24.35	-20.88	4.24
269	C(cyc)-CH2	-26.067	24.35	24.35	24.35
270	RCH=CH	37.072	18.76	0.52	2.28
271	(R)C=CH	154.010	18.76	0.52	2.28
272	RCH=C	41.683	49.82	-27.08	5.26
273	RC=C	41.683	49.82	-27.08	5.26
274	(R)C=C	212.539	49.82	-27.08	5.26
275	RCH2=C	41.683	43.14	-15.85	5.44
276	RCH2	-29.271	-5.86	11.43	-0.54
277	RCH	-72.896	104.05	-54.34	8.7
278	RC	116.357	18.2	6.76	-2.66
279	CH3-CHm=CHn(m,n[0,2])	-4.507	22.93	-33.28	8.05

280	CH2-CHm=CHn(m,n[0,2])	-3.809	22.37	-33.23	7.89
281	CHp-CHm=CHn(m,n[0,2];p[0,1])	-1.856	-0.34	-15.08	5.06
282	<CH3>2CH	-2.834	2.3	-6.45	1.55
283	<CH3>3C	-4.168	-4.29	8.08	-2.02
284	CH2=CH	23.783	-14.86	47.46	-8.04
285	CH=CH	82.420	18.12	40.76	-9.36
286	CH2=C	78.179	-13.26	64.86	-13.5
287	CH=C	138.648	-0.58	80.98	-21.01
288	CH3	-84.039	-10.75	17.7	-1.15
289	CH2	-20.651	16.19	3.21	0.41
290	CH	37.653	50.97	-19.12	3.71
291	C	96.182	53.24	-26.31	4.51
292	CH#C	187.142	25.94	-0.53	3.95
293	C#C	230.295	4.16	28.24	-4.93
294	H2	0	84.79645	-1828.17454	4909.00569
295	Ethylene	52.51	22.06167	-10.82192	19.85289
296	Ethyne	210.68	68.8302	-45.886	17.83998

S.7. Additional considerations for large-scale reactions

In the main text, our analysis is on the data typically reported for the laboratory scale reactions. Although theoretically the yield is not expected to vary with the scale at which the reaction is carried out, in practice considerations such as reactor design might have an effect. In this section, we consider an illustrative example for a stirred tank reactor in which the geometrical changes in the tank and agitator and heat dissipation effects are predicted to lead to scale-dependent yields. We show that such predictions are in line with the available – but limited – experimental data available in literature and relevant to chemical industry.

Let us begin by writing out the mass and energy balances for a simple stirred-tank reactor.[15]

$$\frac{dN_i}{dt} = (\dot{N}_i)_{in} - \left(\dot{N}_i\right)_{out} + \sum_{j=1}^{M} \nu_{ij}\dot{X}_j \quad \text{(S44)}$$

$$\frac{dU}{dt} = \sum_{i=1}^{C}(\dot{N}_i\overline{H_i})_{in} - \sum_{i=1}^{C}\left(\dot{N}_i\overline{H_i}\right)_{out} + \dot{Q} \quad \text{(S45)}$$

where $\dot{N}_i$ is the molar rate of species *i*, ν_{ij} is the stoichiometric coefficient for the species *i* in reaction *j*, $\dot{X}_j$ is the molar rate of reaction j, *U* is the internal energy, $\overline{H}_i$ is the partial molar enthalpy for species *i*, M is the number of independent chemical reactions, *C* is the number of components, and $\dot{Q}$ is the heat flow rate. After some rearrangement, the equations for the general, non- equilibrium conditions can be written as follows

$$V\frac{dC_i}{dt} = q\{(C_i)_{in} - C_i\} + V\sum_{j=1}^{M} \upsilon_{ij} r_j \quad \text{(S46)}$$

$$V\frac{d}{dt}(\sum_i C_i\overline{H}_i) = q(\sum_i (C_i\overline{H}_i)_{in} - \sum_i C_i\overline{H}_i) + \dot{Q} \quad \text{(S47)}$$

where *V* is the volume, *q* is the volumetric flow rate, *r* is the rate of the reaction, and *C* is the concentration. At steady-state, the mass balance becomes

$$C_i = (C_i)_{in} + \frac{V}{q}\sum_{j=1}^{M} \nu_{ij} r_j \quad \text{(S48)}$$

Hence, the simplified form of the energy balance for steady-state condition becomes

$$\dot{Q} = V \sum_{j=1}^{M} \Delta_{rxn,j} H(T_{in})\, r_j + q \sum_i C_i (H_i(T) - H_i(T_{in})) \quad \text{(S49)}$$

For a batch reactor[15] with no inlet and/or outlet mass flows, the design equations become

$$\frac{dC_i}{dt} = \sum_{j=1}^{M} \nu_{ij} r_j \quad \text{(S50)}$$

$$C_P \frac{dT}{dt} = -V \sum_{j=1}^{M} \Delta_{rxn,j} H(T) r_j + \dot{Q} \quad \text{(S51)}$$

where $\mathbf{C}_p$ is the effective heat capacity of the fluid in the reactor.

For tubular reactors[15], the mass and energy balances can be written as

$$\frac{DC_i}{Dt} = \sum_{j=1}^{M} \nu_{ij} r_j \quad \text{(S52)}$$

$$\frac{D(\sum C_i \bar{U}_i)}{Dt} = \frac{1}{A}\dot{\aleph} \quad \text{(S53)}$$

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \nu \frac{\partial}{\partial z} \quad \text{(S54)}$$

where $\dot{\aleph}$ is the heat flow rate per unit length of the reactor.

The overall steady-state mass and energy balances for all of the aforementioned reactors are

$$(N_i)_{out} = (N_i)_{in} + \sum_{j=1}^{M} \nu_{ij} X_j \quad \text{(S55)}$$

$$Q + W = \sum_{j=1}^{C} (N_i \bar{H}_i)_{out} - \sum_{j=1}^{C} (N_i \bar{H}_i)_{out} \quad \text{(S56)}$$

To calculate the reaction temperature, the following equation should be solved

$$\sum_{j=1}^{C} (N_i)_{out} = \int_{T_{in}}^{T_{ad}} C_{P,i} dT + \sum_{j=1}^{M} [\Delta_{rxn,j} H(T_{in})] \quad \text{(S57)}$$

Now, assume that we aim to scale up a reaction design from a pilot size to industrial size. In such situation the yield of the reaction will vary based on several factors. Mixing time of the reactants (t_{mix}) versus reaction half time (the time necessary for the reaction to achieve 50% yield

denoted by $t_{1/2}$), the size and power of the impeller (for CST and batch reactors), the shape of the scaled-up reactor compared to the pilot one, heat dissipation leading to temperature changes, and many other factors might all affect the reaction yield Lets assume we desire to scale-up a continuous stirred flow reactor (CSTR). We can define a scaling factor, S, as[16]:

$$S = \frac{V_{Large\ Scale}}{V_{Pilot\ Scale}} \quad \text{(S58)}$$

Mixing is one primary factor to consider for designing a large-scale CSTR reactor. If, for instance, we seek to scale up each dimension of the reaction tank by the factor (including the diameter of the agitator) of 10 ($S=V_{\text{scale up}}/V_{\text{pilot}} = 10^3$), and keep the rotational velocity of the impeller the same, we will need to increase the impeller power by a factor of 10^5 to keep the same mixing as in the pilot size reaction since the impeller before and after scale up should have the same "power number" as defined below for the agitators operating at high Reynolds numbers[16]:

$$po = \left(\frac{P_I}{\rho N_I^3 D_I^5}\right)_{Pilot} = \left(\frac{P_I}{\rho N_I^3 D_I^5}\right)_{scale\ up} \quad \text{(S59)}$$

where ρ is the density of the solution to be agitated, and P, N and D are the power, rotational velocity, and the diameter of the impeller.

If, on the other hand, we keep the power per volume of the tank constant, i.e., increase the power 10^3 times as the volume increases by the same factor, the speed of the impeller has to decrease by the factor of $S^{-2/9}$ ($\sim 1000^{-2/9} \sim 0.2$). This means that keeping the power per volume constant, the impeller will have to rotate five times slower to achieve the same mixing as for the small-scale reactor (equivalent to five times larger mixing time) – otherwise the yield of reaction could decrease significantly (e.g., from 100% to 20%).

In another scenario, let's assume the reaction is not isothermal as most of the reactions are. For such reactions, we would need a heat jacket with inlet and outlet stream to control the temperature. The heat generation in the reactor can be calculated as:

$$\dot{q}_{generated} = -V\Delta H\,\Re \quad \text{(S60)}$$

where V is the volume of the reactor, ΔH is the enthalpy of the reaction, and $\Re$ is the reaction rate. A heat jacket should be applied to supply the heat required for the reaction,

$$Q = hA\Delta T \quad \text{(S61)}$$

where h is the heat transfer coefficient, A is the area of the jacket through which the heat transfer occur, and ΔT is the difference between the inlet and outlet temperatures of heating/cooling fluid that run into the heat jacket. The dimensionless number for heat transfer is the so-called the Nusselt number:

$$Nu = \frac{hD_I}{k} \quad \text{(S62)}$$

where h is the convective hear transfer coefficient, D_I is the impeller diameter, and k is the conductive heat transfer coefficient.

Since the Nusselt number should remain constant upon scale-up, one can calculate the scale factor for h in the case of linear scale-up (similar geometrical scale-up in all dimension) as:

$$\frac{h_{scale\,up}}{h_{pilot}} = S^{1/9}S_N^{2/3} \quad \text{(S63)}$$

where S_N is the ratio of the rotational velocity of the agitator after and before the scale-up which is equal to $S^{-2/9}$ for the constant power per volume situation, as discussed above[16]. Therefore, the scaled-up heat transfer coefficient will increase by the factor of $S^{1/27}$ or the hA_{ext} will be scaled by the factor of $S^{17/27}$. Therefore, if we don't increase the ΔT, which *is the* difference between inlet and outlet of the cooling/heating liquid in the heat jacket, the heat accumulation/loss will decrease the yield proportionally. As an illustrative example, consider the industrially important reaction of acetic acid esterification at 100 °C [15]:

$$CH_3COOH + C_2H_5OH \rightleftharpoons CH_3COOC_2H_5 + H_2O \quad \text{(S64)}$$

with k, and k'being forward and backward reaction constants:

$$k = 4.76 \times 10^{-4} \frac{m^3}{kmol\,.\,min} \quad \text{(S65)}$$

$$k' = 1.63 \times 10^{-4} \frac{m^3}{kmol\,.\,min} \quad \text{(S66)}$$

This reaction is endothermic with $\Delta H^0_{rxn}\,(at\ 25\ \circ C) = 14.7\ KJ/mol$ and $\Delta H^0_{rxn}\,(at\ 100\ \circ C)\ \simeq 13.7\ KJ/mol$. Assuming this reaction is being run in a CSTR reactor with size 1 m^3 with the outlet and outlet stream that allows 37% of ester to be produced. Using equations S60 and S61 we can calculate the heat required for this reaction as $\dot{Q} = 3243\ KJ/min$ which has to be fed into the reactor. Now if we scale the volume up to 1000 m^3 (S=1000), and accordingly increase the throughput and output by the same scale (and assuming that the mixer parameters are selected appropriately to achieve the same mixing for large- and pilot- scale reactors), we will need 1000 time larger heat flow to achieve the same yield as before. However, linear scaling of the reactor dimensions including the heat jack, does not result increase in the heat flow by the same factor but, instead, the heat supply will fall short by the factor of $S^{17/27} \sim 77$ as described by Nauman et al.[16]. Hence, the provided heat flow will drop from required $\dot{Q} = 3243\ MJ/min$ to only 42116 KJ/min which is a dramatic change in the heat supply resulting in the yield markedly decreasing upon this scale-up. To solve this problem, the flow temperature in the heat jacket must be adjusted, or alternatively the larger reactor could have different shape compared to the pilot reaction or the impeller specification could be changed (note that the above estimation related to the constant power per volume for the agitator).

For scaling-up of a tubular reactor, the design is rather sensitive to the flow regime (deep laminar, laminar, or fully turbulent); in contrast. CSTR reactors often operate in the turbulent regime save for specific reactions (e.g., polymer synthesis, or reactions involving highly viscous materials, etc.). To evaluate the flow regime at PFRs one should quantify the Reynolds (Re), Prandtl (Pr), and Graetz (Gz) numbers. The scaling factor for each case is separately derived in

detail by Nauman et al.[16] It should also be mentioned that PFRs are typically scaled-up not only by increasing the volume of one tube but also by adding more reactors in series or in parallel each of which has a completely different scale-up formulation as elaborated by Nauman et al.[16]

S.8. Additional 200 test reactions from "*Current Syntheses*"

No.	REACTION SMILES	Solvent	T/°C	Y exp (range)	Avg Y exp	Yield CALC	Ref.
1	[H][Si](CC)(CC)CC.COC(CCC(Cl)=O)=O>>COC(CCC=O)=O.Cl[Si](CC)(CC)CC	Et3SiH/none	25	70	70	39	J. Org. Chem., 1994, 59, 3676
2	OO.OO.COC1=CC=CC(C=O)=C1O>>COC2=C(C(O)=CC=C2)O.O=C=O.O.O	water	40	68-80	74	72	Organic Syntheses, 1955, Coll.Vol.III,p. 759
3	O=N(C1=CC=CC=C1N=N#N)=O>>[O-][N+]2=C3C=CC=CC3=NO2.N#N	toluene	100	77-85	81	98	Organic Syntheses, 1963, Coll.Vol. IV, p. 74
4	NC1=CC=CC=C1Br.BrC2=CC=C(C=C2)C(Cl)=O>>BrC3=CC=C(C=C3)C(NC4=CC=CC=C4Br)=O.[H+].[Cl-]	THF	20	79	79	81	Organic Syntheses, 2012, vol. 88, p. 398
5	S[H].CC1=CC=C(N(=O)=O)C=C1>>NC2=CC=C(C=C2)C=O.O.[S]	water/ethanol 2:1	80	40-50	45	52	Organic Syntheses, 1963, Coll.Vol. IV, p. 31
6	OCCCCCO.CC(OC(C)=O)=O.CC(OC(C)=O)=O>>CC(OCCCCCOC(C)=O)=O.CC(O)=O.CC(O)=O	none	120	92-94	93	89	Organic Syntheses, 1958, vol. 38, p. 80
7	CC(OCCCCCOC(C)=O)=O>>CC(O)=O.C=CCC=C.OC(C)=O	none	575	63-71	67	100	Organic Syntheses, 1958, vol. 38, p. 80
8	CC(OCC)Cl.BrBr>>BrC(OCC)CBr.Cl	none	0	66-73	70	73	Organic Syntheses, 1958, vol. 38, p. 80
9	CCN(CC)CC.ClC(C1=CC=CC=C1I)=O.NC2=CC=CC=C2>>IC3=C(C(NC4=CC=CC=C4)=O)C=CC=C3.CC[NH+](CC)CC.[Cl-]	DCM	20	87-89	88	64	Organic Syntheses, 2013, vol. 90, p. 164
10	NN.ClS(=O)(C1=CC=C(C)C=C1)=O>>NNS(=O)(C2=CC=C(C)C=C2)=O.Cl	THF	15	91-94	92	66	Organic Syntheses, 1960, vol. 40, p. 93
11	C/C(C1=CC=CN=C1)=N\O.BrCC2=CC=CC=C2.[H-]>>C/C(C3=CC=CN=C3)=N\OCC4=CC=CC=C4.[H][H].[Br-]	DMF	-10	99	99	100	Organic Syntheses, 2010, vol. 87, p. 36
12	CC(C)([Si](C)(Cl)C)C.OCC=C.C1=NC=CN1>>CC(C)([Si](C)(OCC=C)C)C.[Cl-].C2=[NH+]C=CN2	DMF	4	71-73	72	72	Organic Syntheses, 2012, vol. 89, p. 243
13	CCC(Cl)=O.S=C1N[C@H](CS1)CC2=CC=CC=C2.C3=CC=CN=C3>>CCC(N4[C@H](CSC4=S)CC5=CC=CC=C5)=O.[Cl-].C6=CC=C[NH+]=C6	DCM	20	76-78	77	67	Organic Syntheses, 2012, vol. 88, p. 364
14	O=N[O-].NC1=CC=C(C=C1)N(=O)=O.CC2(C)OB(OC2(C)C)B3OC(C(O3)(C)C)(C)C>>CC4(C)OB(OC4(C)C)C5=CC=C(C=C5)N(=O)=O.CC6(C)OB(OC6(C)C)O.N#N.[OH-]	acetonitrile	22	72-74	73	66	Organic Syntheses, 2014, vol. 91, p. 106
15	[H][C@@]12CCC([C@]1(CC[C@]3([C@]2(CC=C4C[C@H](CC[C@]34C)O)[H])[H])C)=O.CC(C)([Si](C)(Cl)C)C.C5=NC=CN5>>[H][C@@]67CCC([C@]6(CC[C@]8([C@]7(CC=C9C[C@@H](O[Si](C)(C(C)(C)C)C)CC[C@]89C)[H])[H])C)=O.[Cl-].C%10=[NH+]C=CN%10	DMF	20	98	98	70	Organic Syntheses, 2012 , vol. 89, p. 19
16	CC(C)(C)OOC(C)(C)C.[H]C(C1=CC=CC=C1)=O.NCCC2=CC=CC=C2>>O=C(C3=CC=CC=C3)NCCC4=CC=CC=C4.OC(C)(C)C.OC(C)(C)C	acetonitrile	40	89	89	53	Organic Syntheses, 2012 , vol. 88, p. 14
17	[O-]C([O-])=O.NC1=CC=CC=C1C(O)=O.CC2=CC=C(S(Cl)(=O)=O)C=C2>>CC3=CC=C(S(=O)(NC4=C(C([O-])=O)C=CC=C4)=O)C=C3.O=C=O.[Cl-].O	water	70	88-91	90	53	Organic Syntheses, 1963, Coll.Vol. IV, p. 34
18	COC1=CC=C(C=C1)Br.CC(OC(C)(C)C)=O.C2(CCCCC2)[N-]C3CCCCC3>>COC4=CC=C(C=C4)CC(OC(C)(C)C)=O.C5(CCCCC5)NC6CCCCC6.[Br-]	Toluene	20	86	86	100	Organic Syntheses, 2012, vol. 88, p. 4

19	COC(C(C)C)=O.FC1=CC(Br)=CC=C1.C2(CCCCC2)[N-]C3CCCCC3>>COC(C(C)(C4=CC(F)=CC=C4)C)=O.C5(CCCCC5)NC6CCCCC6.[Br-]	Toluene	20	72	72	65	Organic Syntheses, 2012, vol. 88, p. 4
20	CC(=O)CCC1(CC=C)C(=O)CCCC1=O>>C=CC[C@]12CCC(=O)C=C1CCCC2=O.O	none	22	93-96	95	99	Organic Syntheses, 2012, vol. 88, p. 330
21	CC(C)[N-]C(C)C.[H][C@@]12CCCN1[C@@H](C(Cl)(Cl)Cl)OC2=O.BrCC=C>>ClC(Cl)([C@H]3OC([C@@]4(CCCN34)CC=C)=O)Cl.CC(C)NC(C)C.[Br-]	THF	-78/-40	80-82	81	72	Organic Syntheses, 2009, vol. 86, p. 262
22	[H][C@@]1(C(O)=O)CCCN1.OC(OCC)C(Cl)(Cl)Cl>>[H][C@@]23CCCN2[C@@H](C(Cl)(Cl)Cl)OC3=O.CCO.O	chloroform	60	62-65	63	50	Organic Syntheses, 2009, vol. 86, p. 262
23	BrBr.O=C1OCC2=C1C=CC=C2>>BrC3OC(C4=CC=CC=C34)=O.[H]Br	none	140	82-95	88	90	Organic Syntheses, 1943 , vol. 23, p. 74
24	CCNCC.CCNCC.COC1=CC(C(Cl)=O)=CC(OC)=C1OC>>COC2=CC(C(N(CC)CC)=O)=CC(OC)=C2OC.CC[NH2+]CC.[Cl-]	DCM	15	95	95	61	Organic Syntheses, 2012, vol. 88, p. 427
25	CC(OC(C)(C)C)=O.FC1=CC=C(C=C1)Br.C2(CCCCC2)[N-]C3CCCCC3>>CC(C)(OC(CC4=CC=C(C=C4)F)=O)C.C5(CCCCC5)NC6CCCCC6.[Br-]	Toluene	20	82	82	100	Organic Syntheses, 2012, vol. 88, p. 4
26	BrBr.CC(/C=C/C1=CC=CC=C1)=O>>BrC(C(C2=CC=CC=C2)Br)C(C)=O	CCl4	20	52-57	54	97	Organic Syntheses, 1955, Coll.Vol.III,p. 105
27	ClC1=CC=C(C=C1)Br.COC(C(C)C)=O.C2(CCCCC2)[N-]C3CCCCC3>>COC(C(C)(C4=CC=C(C=C4)Cl)C)=O.C5(CCCCC5)NC6CCCCC6.[Br-]	toluene	20	89	89	65	Organic Syntheses, 2012 , vol. 88, p. 4
28	COC(C1=CC=C(N(C(CCl)=O)C2CCN(C(OCC3=CC=CC=C3)=O)CC2)C=C1)=O.CCN(CC)CC>>COC(C4=CC=C(N(C(C5)=O)C6CCN(C(OCC7=CC=CC=C7)=O)CC6)C5=C4)=O.CC[NH+](CC)CC.[Cl-]	MeTHF/isopropanol 4:1	75	87	87	54	Organic Syntheses, 2013 , vol. 90, p. 74
29	ClC1=CC=C(C=C1)Br.COC(C(C)C)=O.C2(CCCCC2)[N-]C3CCCCC3>>COC(C(C)(C4=CC=C(C=C4)Cl)C)=O.C5(CCCCC5)NC6CCCCC6.[Br-]	Toluene	20	89	89	65	Organic Syntheses, 2012 , vol. 88, p. 4
30	COC1=CC=C(C=C1)Br.COC(C(C)C)=O.C2(CCCCC2)[N-]C3CCCCC3>>COC(C(C)(C4=CC=C(C=C4)OC)C)=O.C5(CCCCC5)NC6CCCCC6.[Br-]	Toluene	20	85	85	65	Organic Syntheses, 2012 , vol. 88, p. 4
31	C=CC#N.OCC1=CC=CC=C1>>C=CC(=O)NCC1=CC=CC=C1	sulfuric acid	20	59-62	62	36	Organic Syntheses, 1962, vol. 42, p. 18
32	ClP(Cl)(Cl)(Cl)Cl.NC(CC#N)=O>>N#CCC#N.ClP(Cl)(Cl)=O.[H]Cl.[H]Cl	water	100	67-80	74	94	Organic Syntheses, 1930, vol. 10, p. 66
33	O=C1CCCCCCC1.CCOC(OCC)=O.[H-]>>CCOC(/C2=C([O-])/CCCCCC2)=O.CCO.[H][H]	benzene	80	91-94	93	100	Organic Syntheses, 1973,Coll.Vol.V,p.198
34	CC(CCC1(C(CCCC1=O)=O)CC2=CC=CC=C2)=O>>O=C(C=C3CCCC4=O)CC[C@@]34CC5=CC=CC=C5.O	none	22	70	70	99	Organic Syntheses, 2012, vol. 88, p. 330
35	CC(OC(C)(C)C)=O.CC(C)(C1=CC=C(C=C1)Br)C.C2(CCCCC2)[N-]C3CCCCC3>>CC(C)(OC(CC4=CC=C(C(C)(C)C)C=C4)=O)C.C5(CCCCC5)NC6CCCCC6.[Br-]	Toluene	20	83	83	100	Organic Syntheses, 2012,vol. 88, p. 4
36	CS(=O)(Cl)=O.CCN(CC)CC.CC(C)([C@H](NC(C1=C(C=CC=C1)Br)=O)CO)C>>CC(C)([C@H]2COC(C3=C(C=CC=C3)Br)=N2)C.CS(=O)(O)=O.CC[NH+](CC)CC.[Cl-]	DCM	40	96	96	54	Organic Syntheses, 2009, vol.86, p. 181
37	COC(C(C)C)=O.CN(C1=CC=C(C=C1)Br)C.C2(CCCCC2)[N-]C3CCCCC3>>COC(C(C)(C4=CC=C(N(C)C)C=C4)C)=O.C5(CCCCC5)NC6CCCCC6.[Br-]	Toluene	20	88	88	65	Organic Syntheses, 2012, vol. 88, p. 4
38*	[H][H].[H]C(C1=CC=CC=C1)=O.CC2=C(C=CC=C2)N>>CC3=C(C=CC=C3)NCC4=CC=CC=C4.O	Diethyl ether	25	89-94	92	74	Organic Syntheses, 1941, vol. 21, p. 109
39	NC1=C(C=C(C=C1)N(=O)=O)N(=O)=O.S.S.S>>NC2=C(N)C=C(C=C2)N(=O)=O.O.O.[S].[S].[S]	ethanol	50	52-58	55	61	Organic Syntheses, 1941 , vol. 21, p. 20
40	BrC1=CSC=C1.COC(C(C)C)=O.C2(CCCCC2)[N-]C3CCCCC3>>COC(C(C)(C4=CSC=C4)C)=O.C5(CCCCC5)NC6CCCCC6.[Br-]	toluene	20	75	75	65	Organic Syntheses, 2012 , vol. 88, p. 4
41	CCN(CC)CC.CCN(CC)CC.O=C(C(F)(F)F)OC(C(F)(F)F)=O.O/N=C(CC1=CC=C(C=N1)C(F)(F)F)\C2=CC(Br)=CC=C2>>FC(C3=CC=C(N=C3)C4N=C4C5=CC(Br)=CC=C5)(F)F.[O-]C(C(F)(F)F)=O.[O-]C(C(F)(F)F)=O.CC[NH+](CC)CC.CC[NH+](CC)CC	DCM	20	76	76	78	Organic Syntheses, 2009, vol. 86, p. 18

42	O=C(CC1=CC=C(C=N1)C(F)(F)F)C2=CC(Br)=CC=C2.NO>>O/N=C(CC3=CC=C(C=N3)C(F)(F)F)\C4=CC(Br)=CC=C4.O	methanol	70	72	72	58	Organic Syntheses, 2009, vol. 86, p. 18
43	FC(F)(F)C1=CC=C(Cl)N=C1.O=C(C)C2=CC(Br)=CC=C2.[H-]>>O=C(CC3=CC=C(C=N3)C(F)(F)F)C4=CC(Br)=CC=C4.[Cl-].[HH]	THF	60	79	79	100	Organic Syntheses, 2009, vol. 86, p. 18
44	C1COCCN1.O=C1CCCCC1>>C1CCC(=CC1)N1CCOCC1.O	toluene	110	72-80	76	54	Organic Syntheses, 1973, Coll.Vol.V, p. 808
45	NC1=CC=CC=C1.C=C(O2)CC2=O>>O=C(CC(C)=O)NC3=CC=CC=C3	benzene	80	74	74	61	Organic Syntheses, 1941, vol. 21, p. 5
46	OO.CC1=CN=CC=C1>>CC2=CC=C[N+]([O-])=C2.O	Acetic acid	70	73-77	75	68	Organic Syntheses, 1963,Coll.Vol.IV, p. 654
47	O[N+]([O-])=O.CC1=CC=C[N+]([O-])=C1>>CC2=C([N+]([O-])=O)C=C[N+]([O-])=C2.O	Sulfuric acid	100	70-73	72	72	Organic Syntheses, 1963,Coll.Vol.IV, p. 654
48	BrBr.OC1=CC=CC=C1>>BrC2=CC=C(O)C=C2.[H]Br	CS2	5	80-84	82	64	Organic Syntheses, 1921, vol.1, p. 39
49	[O+]#[C-].[O+]#[C-].O=C(C1=CC=CC(N(=O)=O)=C1/C=C/C2=CC=CO2)N(C)C>>CN(C(C3=CC=CC4=C3C=C(C5=CC=CO5)N4)=O)C.O=C=O.O=C=O	DMF	80	91	91	66	Organic Syntheses, 2009, vol. 86, p. 92
50	[C-]#[O+].[C-]#[O+].FC1=CC=C(C=C1)/C=C/C2=C(N(=O)=O)C=CC=C2C(N3CCCC3)=O>>FC4=CC=C(C5=CC6=C(C=CC=C6C(N7CCCC7)=O)N5)C=C4.O=C=O.O=C=O	DMF	80	96	96	61	Organic Syntheses, 2009, vol. 86, p. 92
51	COC(C(C)C)=O.FC1=CC=C(C=C1)Br.C2(CCCCC2)[N-]C3CCCCC3>>COC(C(C)(C4=CC=C(C=C4)F)C)=O.C5(CCCCC5)NC6CCCCC6.[Br-]	toluene	20	85	85	65	Organic Syntheses, 2012 , vol. 88, p. 4
52	[C-]#[O+].[C-]#[O+].O=C(C1=CC=CC(N(=O)=O)=C1/C=C/C2=CC3=C(C=C2)OCO3)N4CCOCC4>>O=C(C5=CC=CC6=C5C=C(C7=CC8=C(C=C7)OCO8)N6)N9CCOCC9.O=C=O.O=C=O	DMF	80	92	92	61	Organic Syntheses, 2009, vol. 86, p. 92
53	N.N.ClC1=CC(Cl)=C(C=C1N(=O)=O)N(=O)=O>>NC1=CC(N)=C(C=C1N(=O)=O)N(=O)=O.Cl.Cl	ethylene glycol	140	88-96	92	62	Organic Syntheses, 1960, vol. 40, p. 96
54	OC(=O)C(O)(C1=CC=CC=C1)C1=CC=CC=C1>>OC(=O)C1C2=C(C=CC=C2)C2=C1C=CC=C2.O	benzene	80	93-97	95	87	Organic Syntheses,1963, Coll.Vol.IV, p. 482
55	ON=O.NNC1=CC=CC=C1>>N#N=NC2=CC=CC=C2.O.O	water	0	65-68	67	61	Organic Syntheses, 1955,Coll.Vol.III, p. 710
56	BrBr.CC1=CC(C)=CC(C)=C1>>CC2=CC(C)=C(C(C)=C2)Br.Br	CCl4	15	79-82	81	100	Organic Syntheses, 1931 , vol. 11, p. 24
57	CC([O-])=O.CC1(C)OI(Cl)C2=CC=CC=C12.C[Si](C(F)(F)F)(C)C>>CC3(C)OI(C4=C3C=CC=C4)C(F)(F)F.C[Si](OC(C)=O)(C)C.[Cl-]	acetonitrile	-15	85	85	60	Organic Syntheses, 2012, vol. 88, p. 168
58	O=NO.NC1=C(C=CC=C1)N>>C23=CC=CC=C2NN=N3.O.O	Water/Acetic acid 2:1	80	75-81	78	86	Organic Syntheses, 1940, vol. 20, p. 16
59	[C-]#[O+].[C-]#[O+].CS(=O)(N1CCN(C(C2=CC=CC(N(=O)=O)=C2/C=C/C3=CC(Cl)=CC(F)=C3)=O)CC1)=O>>CS(=O)(N4CCN(C(C5=CC=CC6=C5C=C(C7=CC(Cl)=CC(F)=C7)N6)=O)CC4)=O.O=C=O.O=C=O	DMF	80	94	94	61	Organic Syntheses, 2009, vol. 86, p. 92
60	O=N(O)=O.O=N(O)=O.CC1=C(C=C(C(C)=C1)C)C>>CC2=C(C([N+]([O-])=O)=C(C(C)=C2[N+]([O-])=O)C)C.O.O	chloroform	50	92-94	93	100	Organic Syntheses, vol. 10, p. 40
61	[C-]#[O+].[C-]#[O+].CC(C)(OC(N1CCN(C(C2=CC=CC(N(=O)=O)=C2/C=C/C3=CC4=C(C=C3)OCO4)=O)CC1)=O)C>>CC(C)(OC(N5CCN(C(C6=CC=CC7=C6C=C(C8=CC9=C(C=C8)OCO9)N7)=O)CC5)=O)C.O=C=O.O=C=O	DMF	80	95	95	61	Organic Syntheses, 2009, vol. 86, p. 92

62	BrBr.CC1=C(C(OC)=O)C=CC=C1[N+]([O-])=O>>O=C(OC)C2=C(CBr)C([N+]([O-])=O)=CC=C2.[Br]	CCl4	77	96	96	98	Org. Synth. 2003, 80, 75
63	CCN(CC)CC.CCSCO.CC(C)([Si](C)(Cl)C)C>>CCSCO[Si](C)(C(C)(C)C)C.CC[NH+](CC)CC.[Cl-]	DCM	5	78-80	79	73	Organic Syntheses, 2013, vol. 90, p. 10
64	CCOC(=O)C1CCC2=C(C=CC=C2)C1=O.C#C>>CCOC(=O)C1(CCC2=CC=CC=C2C1=O)C=C	toluene	100	100	100	92	Organic Syntheses, 2009, vol. 86, p. 325
65	NC(N)=O.CCOC(CC(OCC)=O)=O>>O=C1CC(NC(N1)=O)=O.CCO.CCO	Ethanol	80	72-78	75	65	Organic Syntheses, 1938, vol. 18, p. 8
66	[O-]C([O-])=O.CCCC#CCCC.CNC1=CC=CC=C1I>>CCCC2=C(C3=C(N2C)C=CC=C3)CCC.OC([O-])=O.[I-]	DMF	100	71	71	40	Organic Syntheses, 2012, vol. 88, p. 377
67	[O-]C([O-])=O.CC(C#CC(C)(O)C)=C.NC1=C(C=CC=C1)I>>CC(C2=C(C(C)(O)C)NC3=C2C=CC=C3)=C.OC([O-])=O.[I-]	DMF	100	70	70	41	Organic Syntheses, 2012, vol. 88, p. 377
68	CC([O-])=O.OC[C@@H](C1=C(C=C(C=C1Br)OCC2=CC=CC=C2)Br)CNCC3=CC=CC=C3>>OC[C@@H]4CN(C5=C4C(Br)=CC(OCC6=CC=CC=C6)=C5)CC7=CC=CC=C7.CC(O)=O.[Br-]	DMSO	20	67	67	99	Organic Syntheses, 2012, vol. 88, p. 388
69	O=CC1=CC=CC=C1N.C2CCCN2>>C34=CC=CC=C3CN5C(CCC5)N4.O	n-butanol	120	72-77	75	61	Organic Syntheses, 2012, vol. 89, p. 274
70	[O-]C([O-])=O.CCCC(C=O)(C1=C(C=CC(OCC2=CC=C(C#N)C=C2)=C1)Br)CCC>>CCCC3(C(C4=CC=C(C=C34)OCC5=CC=C(C#N)C=C5)=O)CCC.OC([O-])=O.[Br-]	dioxane	100	70	70	92	Organic Syntheses, 2012, vol. 89, p. 159
71	CC1=CC=C(C=C1)C1=NC=CC=C1.CC1=CC=C(C=C1)S(=O)(=O)N=[N]#N>>CC1=CC=C(C=C1)S(=O)(=O)NC1=C(C=CC(C)=C1)C1=NC=CC=C1.N#N	1,2-Dichloroethane	80	86	86	86	Organic Syntheses, 2014, vol. 91, p. 52
72	NO.CCOC(C1=CC=CC=C1)=O>>ONC(C2=CC=CC=C2)=O.CCO	methanol	20	57-65	61	54	Organic Syntheses, 1939, vol. 19, p. 15
73	O=C(C(C1=CC=CC=C1)=O)C2=CC=CC=C2.NN.NN>>N/N=C(/C(C3=CC=CC=C3)=N/N)C4=CC=CC=C4.O.O	isopropanol	83	83-89	86	89	Organic Syntheses, 1963,Coll.Vol.IV, p. 377
74	O=NO.CCOC(CN)=O>>CCOC(C=N#N)=O.O.O	water	0	85	85	64	Organic Syntheses,1955, Coll.Vol.III, p. 392
75	CCN(CC)CC.CS(=O)(Cl)=O.NC1=CC=CC=C1/C(CC2=CC=CC=C2)=N\O>>CC[NH+](CC)CC.CS(=O)(O)=O.C3(CC4=CC=CC=C4)=NNC5=CC=CC=C35.[Cl-]	DCM	25	68	68	61	Organic Syntheses, 2012 , vol. 88, p. 33
76	ON=O.CCCCO>>CCCCON=O.O	water	0	81-85	83	51	Organic Syntheses, 1936 , vol. 16, p. 7
77	NC1=CC=CC=C1N.O=CC=O>>C23=CC=CC=C2N=CC=N3.O.O	water	80	85-90	87	76	Organic Syntheses, 1950 , vol. 30, p. 88
78	ClC(=O)CCCCC(Cl)=O.C1=CC=CC=C1.C1=CC=CC=C1>>Cl.Cl.O=C(CCCCC(=O)C1=CC=CC=C1)C1=CC=CC=C1	benzene	20	75-81	78	64	Organic Syntheses, 1933 , vol. 13, p. 32
79	CCCCC(CC)OSc1ccccc1>>CCC(O)CCC(C)Sc1ccccc1	benzene	20	53	53	51	Organic Syntheses, vol. 81, p. 244
80	[C-]#[O+].[C-]#[O+].FC1=CC=C(C=C1)/C=C/C2=C(N(=O)=O)C=CC=C2C(N3CCC(N4C(NC5=C4C=CC=C5)=O)CC3)=O>>FC6=CC=C(C7=CC8=C(C=CC=C8C(N9CCC(N%10C(NC%11=C%10C=CC=C%11)=O)CC9)=O)N7)C=C6.O=C=O.O=C=O	DMF	80	94	94	61	Organic Syntheses, 2009, vol. 86, p. 92
81	CCN(CC)CC.ClSc1ccccc1.CCCCC(O)CC>>CCCCC(OSc2ccccc2)CC.CC[NH+](CC)CC.[Cl-]	DCM	-70	89	89	55	Organic Syntheses, vol. 81, p. 244
82	NCC=C.NCC=C.COC(Cl)=O>>COC(NCC=C)=O.[NH3+]CC=C.[Cl-]	CH2Cl2	0	97-98	98	81	Organic Syntheses, vol. 84, p. 88
83	COC(NCC=C)=O.CCCCCCC#CC#CBr.C1=CC=CC=N1>>CCCCCCC#CC#CN(CC=C)C(OC)=O.[Br-].[NH+]2=CC=CC=C2	THF/pyridine	20	80	80	54	Organic Syntheses, vol. 84, p. 088
84	CC(C)[N-]C(C)C.CCCCCCC#CC(Cl)=C>>CCCCCCC#CC#C.CC(C)NC(C)C.[Cl-]	THF	-78	69-73	71	84	Organic Syntheses, vol. 84, p. 077

85	CCN(CC)CC.CC1CCCCC1=O.CC([Si](Cl)(C)C)(C)C>>CC2=C(CCCC2)O[Si](C(C)(C)C)(C)C.CC[NH+](CC)CC.[Cl-]	acetonitrile	20	95	95	86	Organic Syntheses, vol. 83, p. 193
86	OC(CC(c1ccccc1)C=C)=O.II>>IC[C@@H]2OC(C[C@H]2c3ccccc3)=O.[H+].[I-]	acetonitrile	0	85-91	88	100	Organic Syntheses, vol. 64, p. 175
87	C=C(C(S(c1ccccc1)=O)=C)S(c2ccccc2)=O.OO.OO>>C=C(S(=O)(c3ccccc3)=O)C(S(=O)(c4ccccc4)=O)=C.O.O	AcOH	90	70	70	54	Organic Syntheses, vol. 74, p. 147
88	O=Cc1ccccc1.FS(F)(c2ccccc2)F>>FC(c3ccccc3)F.FS(c4ccccc4)=O	none	100	71-91	81	98	Organic Syntheses, vol. 44, p. 039
89	O=S1(CC=CC1)=O.CCOC(/C=C/C(OCC)=O)=O>>CCOC([C@@H]2CC=CC[C@@H]2C(OCC)=O)=O.O=S=O	ethanol	110	66-73	70	54	Organic Syntheses, vol. 50, p. 43
90	CCN(CC)CC.BrC1CCOC1=O>>O=C2OCC=C2.CC[N+](CC)([H])CC.[Br-]	diethyl ether	35	60	60	55	Organic Syntheses, vol. 45, p. 22
91	CCOC(C1CC(C(C(OCC)=O)CC1=O)=O)=O.O.O>>O=C(CC2)CCC2=O.CCO.CCO.O=C=O.O=C=O	water	190	81-89	85	95	Organic Syntheses, vol. 45, p. 25
92	CCN.CCN.BrCC(Br)=C>>CCNCC(Br)=C.CC[NH3+].[Br-]	none	20	70-78	74	64	Organic Syntheses, vol. 44, p. 7
93	C1C2C=CC1C=C2.CC(C)(C)OOC(=O)c1ccccc1>>CC(C)(C)OC1C2C=CC1C=C2.OC(=O)c1ccccc1	benzene	80	20-25	23	44	Organic Syntheses, vol. 44, p. 12
94	CC(OC(C)=O)=O.CC(OC(C)=O)=O.O=Cc1ccccc1.OC(CCC(c2ccccc2)=O)=O>>O=C3OC(c4ccccc4)=C/C3=C\c5ccccc5.CC(O)=O.CC(O)=O.CC(O)=O.CC(O)=O	acetic anhydride	95	45-50	48	88	Organic Syntheses, vol. 43, p. 3
95	II.C1=CC=CC=C1>>IC2=CC=CC=C2.[H+].[I-]	Nitric acid	80	86-87	87	100	Organic Syntheses, vol. 9, p. 46
96	NO.CCCCCCC([H])=O>>CCCCCC/C([H])=N/O.O	water	40	81-93	87	59	Organic Syntheses, vol. 11, p. 54
97	O=CC1=CC=CC=C1.O=CC2=CC=CC=C2.NN>>C3(/C=N/N=C/C4=CC=CC=C4)=CC=CC=C3.O.O	water	20	91-94	93	67	Organic Syntheses, vol. 16, p. 51
98	O=CC1=CC=CC=C1.N#CCC2=CC=CC=C2>>N#C/C(C3=CC=CC=C3)=C\C4=CC=CC=C4.O	ethanol	30	83-91	87	66	Organic Syntheses, vol. 29, p. 83
99	C=C(O[Si](C)(C)C(C)(C)C)/C=C/N(C)C.C=CC(OC)=O>>CN(C)C1C=C(O[Si](C)(C)C(C)(C)C)CCC1C(OC)=O	Diethyl Ether	25	98	98	94	Organic Syntheses, vol. 78, p. 160
100	OC(C1=CC=CC2=C1OC3=C(C=CC=C32)C(O)=O)=O.ClS(Cl)=O.ClS(Cl)=O>>ClC(C4=CC=CC5=C4OC6=C(C=CC=C65)C(Cl)=O)=O.O=S=O.O=S=O.[H+].[Cl-].[H+].[Cl-]	chloroform	68	90	90	99	Organic Syntheses, vol. 80, p. 046
101	CC1=C(NC(C)C(C)=O)C(C)=CC(C)=C1.CC(OC=O)=O>>CC2=C(N(C=O)C(C)C(C)=O)C(C)=CC(C)=C2.CC(O)=O	THF	20	77-80	79	78	Organic Syntheses, vol. 85, p. 34
102	CCN(CC)CC.C[Si](C)(OC1=CCCC1)C.ClC(C(Cl)=O)Cl>>O=C2C(Cl)(C3(C2CCC3)O[Si](C)(C)C)Cl.CC[NH+](CC)CC.[Cl-]	hexane	25	84	84	100	Organic Syntheses, vol. 85, p. 138
103	NCC1=CC=CC=C1.NCC2=CC=CC=C2.ClC[Si](C)(C)C>>C[Si](C)(C)CNCC3=CC=CC=C3.[NH3+]CC4=CC=CC=C4.[Cl-]	acetonitrile	84	54-67	60	60	Organic Syntheses, vol. 91, p. 162
104	COC1=CC=CC=C1C#CC2=CC=CC=C2.II>>IC(C3=CC=CC=C3O4)=C4C5=CC=CC=C5.CI	DCM	20	87	87	100	Organic Syntheses, vol. 91, p. 283
105	ClP(Cl)(Cl)=O.O=C1C2=C(NC3=C1C=CC=C3)C=CC=C2>>ClC4=C5C=CC=CC5=NC6=C4C=CC=C6.OP(Cl)(Cl)=O	Trichlorophosphate	105	99	99	68	Organic Syntheses, 1942, vol. 22, p. 5
106	OC(C1=CC=CC=C1)C2=CC=CC=C2.OCCCl>>OCCOC(C3=CC=CC=C3)C4=CC=CC=C4.Cl	benzene	80	81-88	84	63	Organic Syntheses, 1963, Coll.Vol.IV, p. 72
107	CCN(CC)CC.O=C1CCC2(OCCO2)CC1.C[Si](C)(Cl)C>>C[Si](C)(OC3=CCC4(OCCO4)CC3)C.CC[NH+](CC)CC.[Cl-]	DMF	80	85-93	89	66	Organic Syntheses, 2012, vol. 88, p. 121
108	CC1CO1.BrBr.[H][C@@]23CCC([C@]2(CC[C@]4([C@]3(CCC5=CC(CC[C@]45C)=O)[H])[H])C)=O>>[H][C@@]67CCC([C@]6(CC[C@]8([C@]7(CCC9=C(C(CC[C@]89C)=O)Br)[H])[H])C)=O.CC(O)CBr	Acetic acid/propylene oxide	0	63	63	58	Organic Syntheses, 2010, vol. 87, p. 126
109	CC(OC(C)(C)C)=O.COC(C1=CC=C(C#N)C=C1)=O.O>>COC(C2=CC=C(C(NC(C)(C)C)=O)C=C2)=O.CC(O)=O	Acetic acid	20	95	95	100	Organic Syntheses, 2012, vol. 89, p. 519

110	CC(C)(OO)C.O=CC1=CC2=CC=CC=C2C=C1.C3CCNC3>>O=C(C4=CC5=C(C=C4)C=CC=C5)N6CCCC6.CC(C)(O)C.O	acetonitrile	70	85	85	51	Organic Syntheses, 2010, vol. 87, p. 1
111	CC(C)([O-])C.FC1=CC=C(C=C1)NC(C2=CC=CC=C2I)=O>>FC3=CC4=C(C=C3)NC(C5=C4C=CC=C5)=O.CC(C)(O)C.[I-]	Benzene	80	85	85	58	Organic Syntheses, 2013, vol. 90, p. 164
112	CC(C)OC(=S)SC(C)C(C)=O>>CC1=C(C)SC(=O)S1.CC(C)O	water	20	82	82	78	Organic Syntheses, 2009, vol. 86, p. 333
113	CCN(CC)CC.CS(=O)(Cl)=O.C/C(C1=CC=CC=C1NCC=C)=N\O>>CC2=NN(C3=C2C=CC=C3)CC=C.CC[NH+](CC)CC.CS(=O)(O)=O.[Cl-]	DCM	23	70	70	60	Organic Syntheses, 2012, vol. 88, p. 33
114	NC1=CC=CC2=C1C(N)=CC=C2.SC(C3=CC=CC=C3)(C4=CC=CC=C4)C5=CC=CC=C5.CC(C)([Si](C)(OCCl)C)C>>CC(C)([Si](C)(OCSC(C6=CC=CC=C6)(C7=CC=CC=C7)C8=CC=CC=C8)C)C.[NH3+]C9=CC=CC%10=C9C(N)=CC=C%10.[Cl-]	DMF	20	77	77	91	Organic Syntheses, 2013, vol. 90, p. 10
115	CC1=CC=CC(C)=N1.CC(C)([Si](C)(OCCl)C)C.CCCCCCCCCCCCCS>>CCCCCCCCCCCCCSCO[Si](C)(C(C)(C)C)C.CC2=[NH+]C(C)=CC=C2.[Cl-]	DMF	20	80	80	54	Organic Syntheses, 2013, vol. 90, p. 10
116	NC(N)=S.CC(CCl)=O.[OH-]>>CC1=CSC(N)=N1.[Cl-].O.O	water	100	70-75	66	51	Organic Syntheses, 1939, vol. 19, p. 10
117	O=C(OCC)/C=C/C1=CC=CC=C1.BrBr>>O=C(OCC)C(Br)C(Br)C2=CC=CC=C2	CCl4	0	83-85	84	100	Organic Syntheses, vol. 12, p. 36
118	ON=O.CC1=C(N)C=CC(N(=O)=O)=C1>>O=N(C2=CC3=C(NN=C3)C=C2)=O.O.O	Acetic acid	20	80-96	88	61	Organic Syntheses, 1940, vol. 20, p. 73
119	ClS(Cl)=O.NCCC1=C(C=C(C=C1)Cl)CO.[OH-].[OH-]>>ClC2=CC=C3CCNCC3=C2.O=S=O.O.O.[Cl-].[Cl-]	DME/water	60	79	79	82	Organic Syntheses, 2013, vol. 90, p. 251
120	NC(CO)C1=CC=CC=C1.ClS(Cl)=O>>[NH3+]C(CCl)C2=CC=CC=C2.O=S=O.[Cl-]	DME	20	99	99	96	Organic Syntheses, 2013, vol. 90, p. 251
121	CC(OC(C)(C)C)=O.ClC1=C(C=CC(C#N)=C1)OC2=C(C=CC=C2)Br.O>>CC(C)(C)NC(C3=CC(Cl)=C(C=C3)OC4=C(C=CC=C4)Br)=O.CC(O)=O	Acetic acid	20	85	85	100	Organic Syntheses, 2012, vol. 89, p. 519
122	CC(OC(C)(C)C)=O.N#CCCC1=CC=CC=C1.O>>CC(O)=O.CC(C)(NC(CCC2=CC=CC=C2)=O)C	Acetic acid	20	88	88	100	Organic Syntheses, 2012, vol. 89, p. 519
123	CC(OC(C)(C)C)=O.CC(NC1=CC=C(C#N)C=C1)=O.O>>CC(NC2=CC=C(C(NC(C)(C)C)=O)C=C2)=O.CC(O)=O	Acetic acid	20	95	95	100	Organic Syntheses, 2012, vol. 89, p. 519
124	[O-]C([O-])=O.CCCCCC(NC1=CC=CC=C1Br)=O>>CCCCCC2=NC3=C(C=CC=C3)O2.OC([O-])=O.[Br-]	DMSO	110	73	73	43	Organic Syntheses, 2012, vol. 88, p. 398
125	O=C1OC(C=C1)=O.NC2=CC=CC=C2>>OC(/C=C\C(NC3=CC=CC=C3)=O)=O	Diethyl ether	20	97-98	98	79	Organic Syntheses, 1961, vol. 41, p. 93
126	O=C(OC(C)=O)C.OC(/C=C\C(NC1=CC=CC=C1)=O)=O>>O=C2N(C3=CC=CC=C3)C(C=C2)=O.O=C(O)C.O=C(O)C	Acetic anhydride	100	75-80	78	86	Organic Syntheses, 1961, vol. 41, p. 93
127	CC1=CC=CC(C)=N1.SCCCSC(C2=CC=CC=C2)(C3=CC=CC=C3)C4=CC=CC=C4.CC(C)([Si](C)(OCCl)C)C>>CC(C)([Si](C)(OCSCCCSC(C5=CC=CC=C5)(C6=CC=CC=C6)C7=CC=CC=C7)C)C.CC8=[NH+]C(C)=CC=C8.[Cl-]	DMF	55	61	61	53	Organic Syntheses, 2013, vol. 90, p. 10
128	CC(C)([O-])C.CC(C)([Si](C)(OCCl)C)C.OCCS>>CC(C)([Si](C)(OCSCCO)C)C.CC(C)(O)C.[Cl-]	DMF	20	63	63	40	Organic Syntheses, 2013, vol. 90, p. 10
129	ClS(Cl)=O.CS(=O)(O)=O>>CS(=O)(Cl)=O.O=S=O.[H]Cl	none	95	71-83	77	97	Organic Syntheses, 1950, vol. 30, p. 58
130	CC(=C)C(C)=C.O=C1C=CC(=O)C2=C1C=CC=C2>>[H][C@@]12CC(C)=C(C)C[C@]1([H])C(=O)C1=CC=CC=C1C2=O	Ethanol	80	96	96	83	Organic Syntheses, 1955, Coll.Vol.III, p.310
131	O=C(C)CC(OC)OC.CNC>>O=C(C)/C=C/N(C)C.OC.OC	Methanol	20	90	90	78	Organic Syntheses, 2002, vol. 78, p. 152
132	N#CCC1=CC(OC)=C(OC)C=C1.CNC>>N=C(N(C)C)CC2=CC(OC)=C(OC)C=C2	Ethanol	70	93-96	95	79	Organic Syntheses, 1999, vol. 76, p. 133

133	CC(C)[N-]C(C)C.CCCCCCCCCC(OCC)=O.Cl[Si](C1=CC=CC=C1)(C)C2=CC=CC=C2>>CC(C)NC(C)C.CCCCCCCCC([Si](C3=CC=CC=C3)(C)C4=CC=CC=C4)C(OCC)=O.[Cl-]	THF	20	93-94	94	66	Organic Syntheses, 1989, vol. 67, p. 125
134	OC1=C(C2=C(C=CC=C3)C3=CC=C2O)C4=C(C=CC=C4)C=C1.ClP(Cl)Cl.C5(C=CC=C6)=C6C=CC(C=CC=C7)=C7[N-]5>>C8(C=CC=C9)=C9C(C%10=C(C=CC=C%11)C%11=CC=C%10OP(N%12C%13=C(C=CC=C%13)C=CC%14=C%12C=CC=C%14)O%15)=C%15C=C8.Cl.Cl.[Cl-]	THF	20	73-75	74	83	Organic Syntheses, 2015, vol. 92 p. 1
135	C[Si](C)(C)C1=CC=CC=C1OS(=O)(C(F)(F)F)=O.O=C(CC(OC)=O)C.[F-]>>O=C(C)C2=CC=CC=C2CC(OC)=O.C[Si](C)(F)C.[O-]S(=O)(C(F)(F)F)=O	acetonitrile	80	67	67	75	Organic Syntheses, 2006, vol. 86 p. 161
136	F/C(F)=C(CCCC)\C1=CC=CC=C1NS(C2=CC=C(C)C=C2)(=O)=O.[H-]>>FC3=C(CCCC)C4=CC=CC=C4N3S(C5=CC=C(C)C=C5)(=O)=O.[H][H].[F-]	DMF	80	80	80	100	Organic Syntheses, 2006, vol. 83 p. 111
137	BrBr.CC12C(C)(O2)CC=CC1>>CC34C(C)(O4)CC(Br)C(Br)C3	Chloroform/DCM 1:1	-60	80-86	83	70	Organic Syntheses, 1969, vol. 49 p. 62
138	CC(CCC1(C(CCCC1=O)=O)CCCOCC2=CC=CC=C2)=O>>O=C3CC[C@@]4(C(CCCC4=C3)=O)CCCOCC5=CC=CC=C5.O	none	22	78	78	99	Organic Syntheses, 2012, vol. 88, p. 330
139	CN.ClC1=CC=C(CC(O)=O)C=C1.ClS(Cl)=O>>ClC2=CC=C(CC(NC)=O)C=C2.O=S=O.[H]Cl.[H]Cl	toluene	40	95-96	96	88	Organic Syntheses, 2012, vol. 89, p. 44
140	CCCC(CCC=C(C)C)(C=O)C1=C(Br)C=CC=C1>>CCCC1(CCC=C(C)C)C(=O)C2=CC=CC=C12.Br	dioxane	100	82	82	100	Organic Syntheses, 2012, vol. 89, p. 159
141	O=C(OC)C1OC1(C)CC(OC)OC>>O=C(OC)C2=C(C)C=CO2.OC.OC	none	160	65-70	67	65	*Org. Synth.* **1959**, *39*, 49
142	NC1=C(I)C=CC=C1.COC2OC(OC)CC2>>IC3=CC=CC=C3N4C=CC=C4.CO.CO.O	Acetic acid	118	64-66	65	71	*Org. Synth. 2006, 83, 103*
143	IC1=CC=CC=C1N2C=CC=C2.C#CC>>CC#CC3=CC=CC=C3N4C=CC=C4.I	toluene	25	75-78	77	58	*Org. Synth.* **2006**, *83*, 103
144	CC(OC(N1CCCCC1)=O)(C)C.BrC2=CC(OC)=CC=C2>>COC3=CC=CC(C4CN(C(OC(C)(C)C)=O)CCC4)=C3.Br	toluene	60	58-60	59	54	*Org. Synth.* **2015**, *92*, 76
145	C#C[Si](C)(C)C.C=CC(CC)=O>>O=C(CC)CCC#C[Si](C)(C)C	acetone	60	49	49	42	*Org. Synth.* **2014**, *91*, 72
146	O=C(OC)C1=CC=CC=C1.CCCCCCN>>O=C(NCCCCCC)C2=CC=CC=C2.OC	toluene	55	82-84	83	62	*Org. Synth.* **2014**, *91*, 201
147	NCC1=CC=CC=C1.O=C(OC)[C@H](CC(C)C)NC(OC(C)(C)C)=O>>O=C(NCC2=CC=CC=C2)[C@H](CC(C)C)NC(OC(C)(C)C)=O.CO	toluene	70	82-85	83	60	*Org. Synth.* **2014**, *91*, 201
148	O=C(O)C1=CC2=CC=CC=C2C=C1O.CI.CI>>O=C(OC)C3=CC4=CC=CC=C4C=C3OC.I.I	DMF	40	96	96	68	*Org. Synth.* **2014**, *91*, 260
149	OCC1=CC=CO1.O=C(C2=CC=CC(Cl)=C2)OO>>O=C3COC(C=C3)O.OC(C4=CC=CC(Cl)=C4)=O	DCM	20	62-63	63	37	*Org. Synth.* **2014**, *91*, 293
150	C[C@H](N)[C@H](O)C1=CC=CC=C1.BrCCCCBr>>C[C@H](N2CCCC2)[C@H](O)C3=CC=CC=C3.Br.Br	toluene	118	90	90	75	*Org. Synth.* **2000**, *77*, 12
151	C1=CC=CC1.C2C=C2>>C34C=CC(C5C4C5)C3	pentane	20	72	72	56	*Org. Synth.* **2000**, *77*, 254
152	O=CC1=CC=CC=C1.NC2=CC=C(OC)C=C2>>COC3=CC=C(/N=C/C4=CC=CC=C4)C=C3.O	DCM	20	88	88	71	*Org. Synth.* **2003**, *80*, 160
153	CSC1=CC=CC=C1.OO>>CS(C2=CC=CC=C2)=O.O	trifluoroethanol	20	91	91	66	*Org. Synth.* **2003**, *80*, 184
154	C[C@H](N)C1=CC=CC=C1.BrCC#N>>C[C@H](NCC#N)C2=CC=CC=C2.Br	acetonitrile	25	100	100	67	*Org. Synth.* **2003**, *80*, 207
155	COC1=C(OC)C(C=CC=C2)=C2C3=C1C=CC=C3.CC(Cl)=O.CC(Cl)=O>>COC4=C(OC)C(C=CC(C(C)=O)=C5)=C5C6=C4C=CC(C(C)=O)=C6.Cl.Cl	DCM	20	77	77	68	*Org. Synth.* **2003**, *80*, 227

156	O=C1OC(C2CC=CCC21)=O.OCC.CCO>>O=C(OCC)C3C(C(OCC)=O)CC=CC3.O	Ethanol/toluene	80	83-86	85	54	*Org. Synth.* **1950**, *30*, 29
157**	O=C(OCC)C1C(C(OCC)=O)CC=CC1.[H][H]>>O=C(OCC)C2C(C(OCC)=O)CCCC2	Ethanol	25	94-96	95	72	*Org. Synth.* **1950**, *30*, 29
158	NC1=CC=CC=C1.OC(C2=CC=CC=C2)(C3=CC=CC=C3)C4=CC=CC=C4>>NC5=CC=C(C(C6=CC=CC=C6)(C7=CC=CC=C7)C8=CC=CC=C8)C=C5.O	Acetic acid	118	70-74	72	48	*Org. Synth. 1950, 30, 5*
159	OC(C1=CC=CC=C1)(C2=CC=CC=C2)C(C3=CC=CC=C3)(O)C4=CC=CC=C4>>C5(C(C=CC=C6)=C6C(C7=CC=CC=C7)=C8C9=CC=CC=C9)=C8C=CC=C5.O.O	toluene	20	90	90	94	*Org. Synth.* **1999**, *76*, 294
160	C=C/C=C/OC(C)=O.C1=CC=CCC=C1>>[H][C@@]23C=CC=C[C@@](CC=C[C@@H]3OC(C)=O)([H])C2	hexane	20	86	86	96	*Org. Synth.* **2000**, *77*, 121
161	NC1=CC=C(I)C=C1F.CC(CCC(C)=O)=O>>FC2=CC(I)=CC=C2N3C(C)=CC=C3C.O.O	toluene	110	100	100	89	*Org. Synth.* **2002**, *78*, 63
162	C=C(C)CCN(C(OC(C)(C)C)=O)C1=CC=CO1>>O=C2CC3(C)C(N(C(OC(C)(C)C)=O)CC3)=CC2	toluene	160	70-75	73	76	*Org. Synth.* **2002**, *78*, 202
163	CC1(CO)COC1.CC2=CC=C(S(=O)(Cl)=O)C=C2>>CC3(COS(C4=CC=C(C)C=C4)(=O)=O)COC3.Cl	pyridine	20	78-87	83	63	*Org. Synth.* **2002**, *79*, 216
164	BrC1=CC=CC=C1.O=C(CC2)OC2=O>>BrC3=CC=C(C(CCC(O)=O)=O)C=C3	none	20	94	94	60	Org. Synth. 2002, 79, 204
165	O=C1N(C(OC)C)C(C2=CC=CC=C2)C(C3=CC=CC=C3)O1>>O=C4N(C=C)C(C5=CC=CC=C5)C(C6=CC=CC=C6)O4.OC	none	150	84	84	50	*Org. Synth.* **1998**, *75*, 45
166	O=C1C=CCCC1.II>>O=C2C(I)=CCCC2.I[H]	Et2O/pyridine	20	80	80	100	*Org. Synth.* **1998**, *75*, 69
167	CC=O.C=CC(OC)=O>>CC(O)C(C(OC)=O)=C	Methyl acrylate	25	79	79	69	*Org. Synth.* **1998**, *75*, 106
168	N[C@@H](CO)C1=CC=CC=C1.O=CCCCC=O.[H]C#N>>N#C[C@H]2N3C(COC3CCC2)C4=CC=CC=C4.O.O	DCM/water	25	65	65	50	*Org. Synth.* **1992**, *70*, 54
169	C[Si](C)(C)O/C=C1CCCCCCCCCCC\1(CC(C)=C)O>>CC2=CC=C3CCCCCCCCCCC3=C2.O[Si](C)(C)C.O	toluene	80	86	86	85	*Org. Synth.* **1993**, *71*, 158
170	O=C(/C(Br)=C/Br)OCC.C#C[Si](C)(C)C>>O=C(/C(Br)=C/C#C[Si](C)(C)C)OCC.[H]Br	DMF	0	48-56	52	57	*Org. Synth.* **1995**, *72*, 104
171	OCCOCCN.CC(OC(C)=O)=O>>CC(O)=O.OCCOCCNC(C)=O	ethanol	40	92-97	94	95	*Org. Synth.* **1992**, *70*, 129
172	SCCS.ClC(CCl)=O>>O=C1CSCCS1.[H]Cl.[H]Cl	DCM	20	60	60	74	*Org. Synth.* **1995**, *72*, 265
173	NC(C[C@@H](C(O)=O)N)=O.CC(C)(C)C=O.ClC(OC)=O>>O=C1N[C@H](C(C)(C)C)N(C(OC)=O)[C@H](C(O)=O)C1.Cl.O	water	20	72-79	75	63	*Org. Synth.* **1996**, *73*, 201
174	CCC(Cl)=O.C1COCCN1>>O=C(CC)N2CCOCC2.Cl	DCM	20	86	86	66	*Org. Synth.* **2007**, *84*, 22
175	CCCCCC#C.ClC(Cl)=C>>ClC(C#CCCCCC)=C.Cl	toluene	40	69-73	71	37	*Org. Synth.* **2007**, *84*, 77
176	CC1=CC=C(S(OC2CCC2)(=O)=O)C=C1>>CC3=CC=C(S(=O)(O)=O)C=C3.C4CC=C4	DMSO	70	70-84	77	65	*Org. Synth.* **1986**, *64*, 50
177	CC(C1=CC=NC=C1)=O.NO>>C/C(C2=CC=NC=C2)=N/O.O	water	0	66-69	67	57	*Org. Synth.* **1986**, *64*, 19
178	C/C(C1=CC=NC=C1)=N/O.CC2=CC=C(S(=O)(Cl)=O)C=C2>>C/C(C3=CC=NC=C3)=N/OS(C4=CC=C(C)C=C4)(=O)=O.Cl	pyridine	25	95	95	69	*Org. Synth.* **1986**, *64*, 19
179	CCCCCCC(C=C)OC(CC(C)=O)=O>>CCCCCC/C=C/CC(C(C)=O)C(O)=O	THF	25	97	97	100	*Org. Synth.* **1990**, *68*, 210

180	COCN(C[Si](C)(C)C)CC1=CC=CC=C1.O=C(C=C2)N(C3=CC=CC=C3)C2=O>>O=C(C4C5CN(CC6=CC=CC=C6)C4)N(C7=CC=CC=C7)C5=O.C[Si](C)(OC)C	acetonitrile	25	72-75	73	80	*Org. Synth.* **1989**, *67*, 133
181	OC(C=C1)=CC=C1C2=CC=CC=C2.ClC3=NN=NN3C4=CC=CC=C4>>C5(C6=CC=C(OC7=NN=NN7C8=CC=CC=C8)C=C6)=CC=CC=C5.Cl	acetone	60	86-89	87	67	*Org. Synth.* **1971**, *51*, 82
182	CC1CCCCC1=O.CC(OC(C)=O)=O>>CC2=C(OC(C)=O)CCCC2.CC(O)=O	CCl4	25	87-92	90	78	*Org. Synth.* **1972**, *52*, 39
183	O=C1C(Br)=CC(Br)(Br)C=C1Br.CN(C)C2=CC(C(F)(F)F)=CC=C2>>OC3=C(Br)C=C(Br)C=C3Br.CN(C)C4=CC(C(F)(F)F)=C(Br)C=C4	DCM	20	82-90	86	85	*Org. Synth.* **1976**, *55*, 20
184	CC(C(C)C)=O.BrBr>>O=C(C(C)C)CBr.Br	methanol	20	95	95	77	*Org. Synth.* **1976**, *55*, 24
185	O=C(CC1)N(Br)C1=O.C2(/C=C/C3=CC=CC=C3)=CC=CC=C2.O>>Br[C@@]([H])(C4=CC=CC=C4)[C@@]([H])(C5=CC=CC=C5)O.O=C(CC6)N([H])C6=O	DMSO	50	80-90	85	100	*Org. Synth.* **1979**, *59*, 16
186	O=C1C=C(O)CCC1.CCO>>O=C2C=C(OCC)CCC2.O	benzene	75	70-75	72	52	*Org. Synth.* **1960**, *40*, 41
187	O=C1C(C)=CC[C@@H](C(C)=C)C1.C=C=C([Si](C)(C)C)C>>C[C@@]23[C@@](CC([Si](C)(C)C)=C3C)([H])C[C@@H](C(C)=C)CC2=O	DCM	0	82	82	80	*Org. Synth.* **1988**, *66*, 8
188	CI.C12=CC=CC=C1NC=C2>>CN3C=CC4=CC=CC=C43.I	amonnia	-40	85-95	90	63	*Org. Synth.* **1960**, *40*, 68
189	O=CC1=CC=CO1.N#CCC(O)=O>>N#C/C=C/C2=CC=CO2.O=C=O.O	Toluene/pyridine	118	75-78	76	94	*Org. Synth.* **1960**, *40*, 46
190	O=CC1=CC=CC=C1.O=CC2=CC=CC=C2.CC(C)=O>>O=C(/C=C/C3=CC=CC=C3)/C=C/C4=CC=CC=C4.O.O	ethanol	20	90-94	92	64	*Org. Synth.* **1932**, *12*, 22
191	CCCC/C=C/C.O=C(CC1)N(Br)C1=O>>CCCC(Br)/C=C/C.O=C(CC2)NC2=O	CCl4	77	58-64	61	38	*Org. Synth.* **1958**, *38*, 8
192	C=CC=O.CCOC(OCC)OCC>>C=CC(OCC)OCC.CCOC=O	Triethyl orthoformate	25	72-80	76	57	*Org. Synth.* **1952**, *32*, 5
193	O=C(OC)OC>>CC(C#N)C1=CC=CC=C1.CO.O=C=O	Dimethyl carbonate	180	93	93	100	*Org. Synth.* **1999**, *76*, 169
194	O=C(C1=CC=CC=C1)NCC(O)=O.CC(OC(C)=O)=O>>O=C2CN=C(C3=CC=CC=C3)O2.OC(C)=O.OC(C)=O	Acetic anhdride	80	66-68	67	90	*Org. Synth.* **1967**, *47*, 101
195	O=C(CC1=CC=CC=C1)CC2=CC=CC=C2.OC(C)=O.OC(C)=O>>O=C3C(C4=CC=CC=C4)=C(C)OC(C)=C3C5=CC=CC=C5.O.O.O	Acetic acid	130	39-45	42	49	*Org. Synth.* **1967**, *47*, 54
196	ClC(C1CCCCC1)=O.ClC(C2CCCCC2)=O>>O=C3C4(CCCCC4)C(C35CCCCC5)=O.Cl.Cl	benzene	80	49-58	54	83	*Org. Synth.* **1967**, *47*, 34
197	BrBr.C12=CC=CC3=C1C(C(C=C3)=CC=C4)=C4C=C2>>BrC5=CC=C(C=C6)C7=C5C=CC8=CC=CC6=C87.Br	CCl4	25	78-86	82	63	*Org. Synth.* **1968**, *48*, 30
198	CCCCCC/C=C\CCCCCC(OCC1OC(C)(C)OC1)=O.O>>CCCCCC/C=C\CCCCCC(OCC(CO)O)=O.CC(C)=O	methanol	25	71	71	41	*Org. Synth.* **2012**, *89*, 183
199	ClC1=CC(Cl)=C(C=CC=C2)C2=N1.N[C@H]3[C@H](N)CCCC3.ClC4=CC(Cl)=C(C=CC=C5)C5=N4>>ClC6=C(C=CC=C7)C7=NC(N[C@@H]8CCCC[C@@H]8NC9=CC(Cl)=C(C=CC=C%10)C%10=N9)=C6.Cl.Cl	toluene	80	73-77	75	65	*Org. Synth.* **2012**, *89*, 380
200	O=C/C=C/C1=CC=CC=C1.NN>>N#N.C2(C3CC3)=CC=CC=C2.O	ethanol	250	45-56	51	67	*Org. Synth.* **1967**, *47*, 98

* Reaction has been performed at 1000 psi

** Reaction has been performed at 30 psi

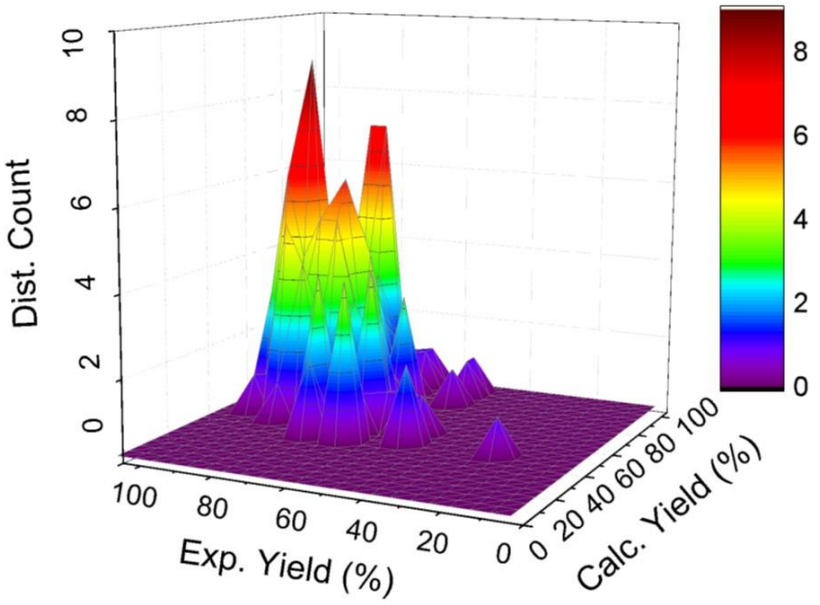
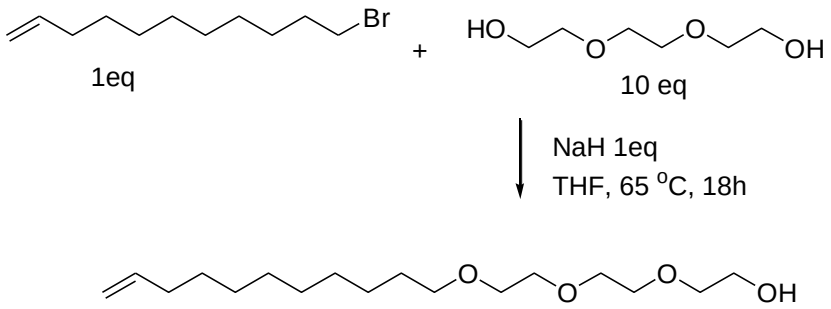
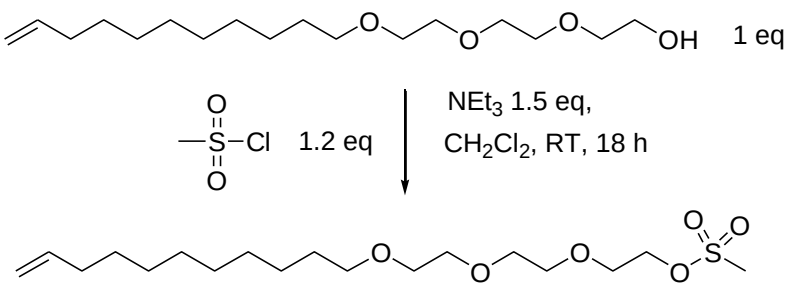


Figure S.14. Histogram of the numbers of the 200 additional "Organic Syntheses" reactions characterized by given values of calculated and experimental yields, (ξ_{exp}, ξ_{pred}). Data is scattered within ~16 % of the (0,0) to (100,100) diagonal corresponding to perfect prediction.

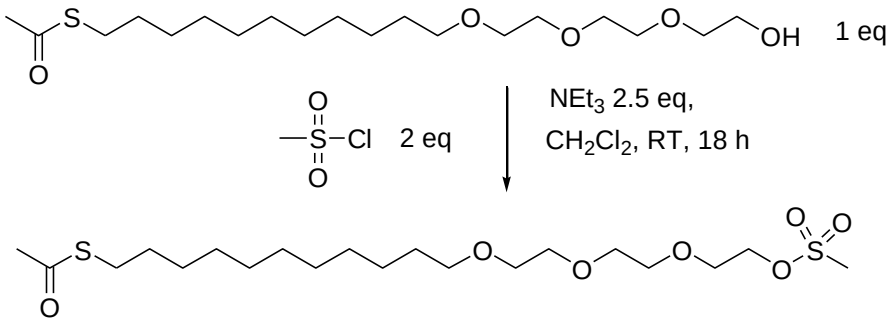
S.9. Variability of experimental yields illustrated by additional test reactions repeated multiple times in industry (courtesy of ProChimia Surfaces)



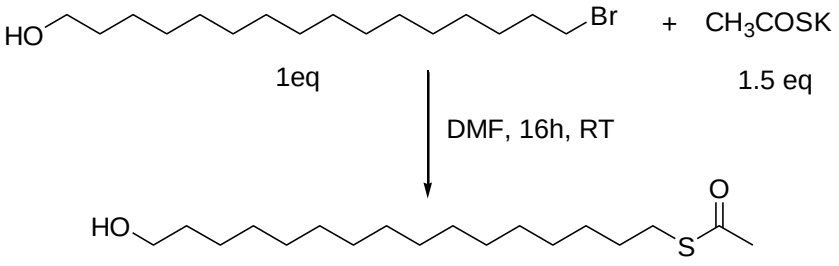
entry	Scale (mmmol of substrate)	Yield (%)	Remarks
1	85.7	51.7	
2	85.7	58.2	
3	115	54.0	
4	57.5	72.5	Time of reaction: weekend
5	115	54.6	
6	115	64.7	
7	115	54.6	
8	115	66.1	
9	115	67.2	
Predicted Yield		70.94	



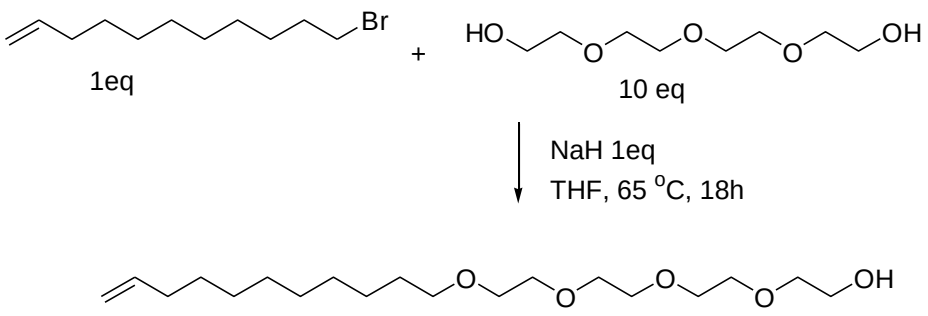
entry	Scale (mmol of substrate)	Yield (%)	remarks
1	52	62.6	
2	33	66.9	
3	33	75.5	Reaction time: weekend
4	49.9	65.7	Reaction time: weekend
Predicted Yield		48.34	



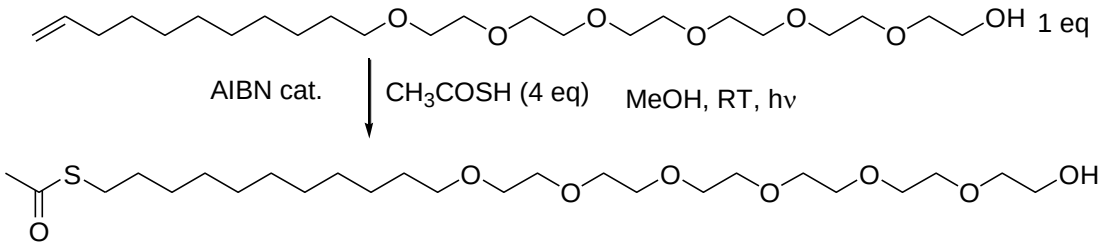
entry	Scale (mmol of substrate)	Yield (%)	remarks
1	26	66.7	
2	18.5	88.6	
3	19	92.1	
Predicted Yield		67.49	



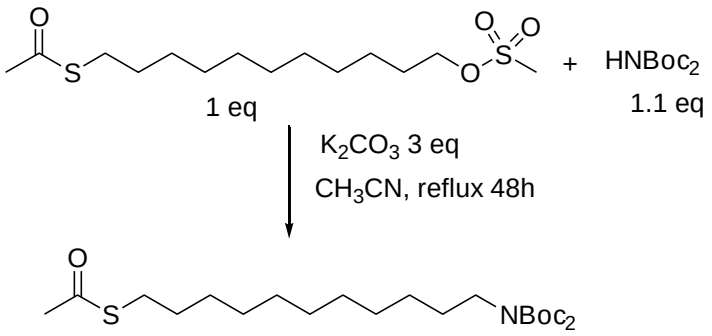
entry	Scale (mmol of substrate)	Yield (%)	remarks
1	6.22	60.9	
2	3.73	55.0	
Predicted Yield		63.28	



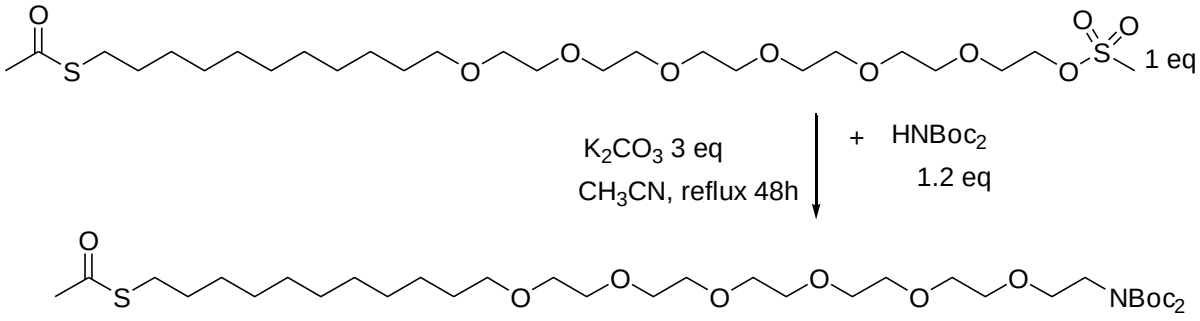
entry	Scale (mmol of substrate)	Yield (%)	remarks
1	34	63.7	
2	241.4	64.5	
3	115	60.3	
Predicted Yield		77.9	



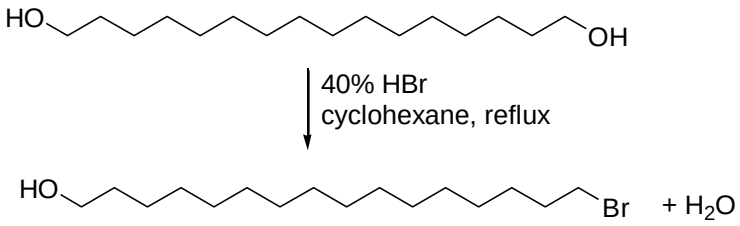
entry	Scale (mmmol of substrate)	Yield (%)	Remarks
1	22.14	75.4	Irradiation time: 1h for each 1g of substrate
2	36	50	
3	32	70.3	
4	35	64.9	
5	23	72.4	
Predicted Yield		32.88	



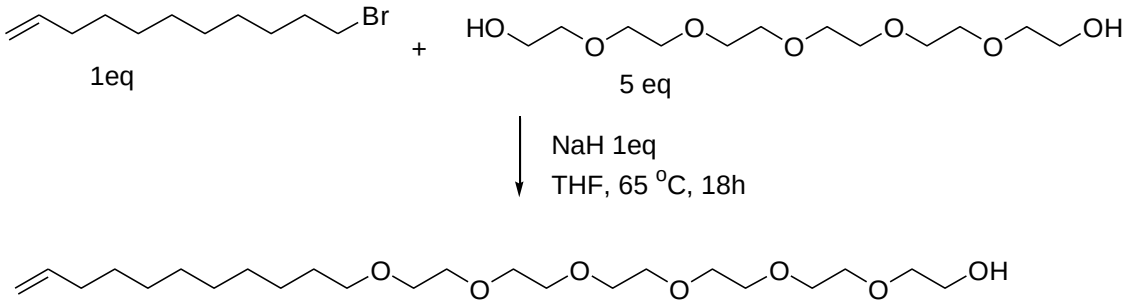
entry	Scale (mmmol of substrate)	Yield (%)	Remarks
1	21	76.9	
2	21	81.2	
3	21	78.2	
Predicted Yield		47.29	



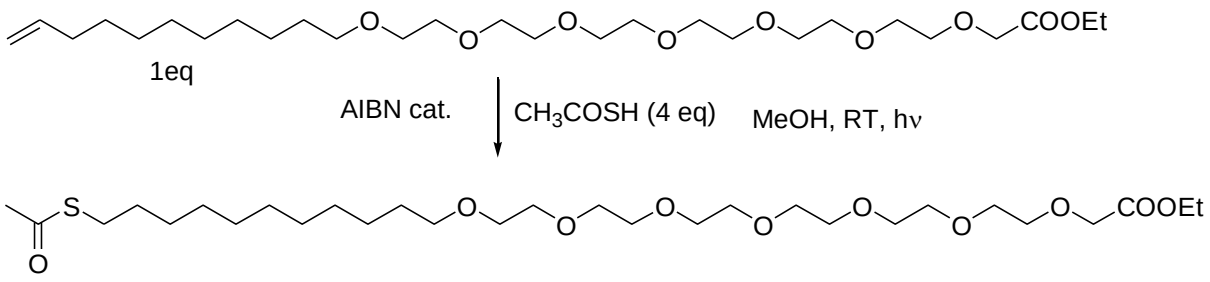
entry	Scale (mmmol of substrate)	Yield (%)	Remarks
1	5.4	51.9	
2	9.3	37.6	
3	10.5	59.0	
Predicted Yield		47.29	



entry	Scale (mmmol of substrate)	Yield (%)	Remarks
1	19.34	22.5	
2	30.95	15.7	
3	40.6	53.6	
Predicted Yield		42.8	



entry	Scale (mmmol of substrate)	Yield (%)	Remarks
1	16.29	62.5	
2	22.66	65.2	
3	70	79.8	
4	80	83.4	
5	50	74.4	
6	50	78.2	
7	50	80.6	
Predicted Yield		73.06	



entry	Scale (mmmol of substrate)	Yield (%)	Remarks
1	10.7	55.0	Irradiation time: 1h for each 1g of substrate
2	24	52.9	
3	26	58.1	
4	21	79.5	
5	11.5	51.3	
Predicted Yield		32.88	

S.10. Computer codes used in the calculations

- **Function to Decompose Molecules Into Functional Groups Using RDKIT package (Python)**

```
# Dictionary that contains all the functional groups and their SMARTS

from rdkit import Chem
groups =\
[['name_group_a', 'group_a_SMART'],
['name_group_b', 'group_b_SMART'],
['name_group_c', 'group_c_SMART'],
.
..
…
['name_group_n', 'group_n_SMART']]
group_smart_mols = [Chem.MolFromSmarts(s) for _,s in groups]


def smart_vector(Molecule_SMARTS):
    from dictionaries import group_smart_mols
    No_atoms = mol.GetNumAtoms()
    atom_flag = [0] * No_atoms
    matches = [0] * len(group_smart_mols)
    for i, s in enumerate(group_smart_mols):

        # Finding substructures of functional groups in the molecule having the
        # SMART Molecule_SMARTS
        y = Molecule_SMARTS.GetSubstructMatches(s)
        # Eliminate the identified groups with overlapping atoms
        if len(y):
            for ix in y:
                overlap = 0
                for iy in ix:
                    if (atom_flag[iy] == 1):
                        overlap = 1
                        break
                if (overlap == 0):
                    for iy in ix:
                        atom_flag[iy] = 1
                    matches[i] += 1
    if(check_if_mol_decomposed(mol, matches) == 0):
        idecompsed=0
    return matches
```

- **Free Energy Formation Calculation of Molecules Contributing in Reactions (c++)**

```
Double G_K_298_func ( int reaction_index){
    int i= reaction_index;
    for (int j=0; j< reaction[i].No_compound; j++){
        reaction[i].compound_Ggc[j]=0;
        for (int k=0; k < No_Grp ; k++){
            reaction[i].compound_Ggc[j] += reaction[i].compound_RptGrp[j][k] *
                    Grp_Gf[k];
        }
        reaction[i].rxn_Ggc_R298 += reaction[i].compound_nu[j] *
             reaction[i].compound_Ggc[j] * 1000/rGas/298.15;

    }
   double reaction[i].Krxn298 = exp(-1*reaction[i].rxn_Ggc_R298 );
   return reaction[i].Krxn298;
}
```

- **Temperature Effect**

```
double  sum_prod_func (int * multiplier, double * group_prop) {
    double molec_prop =0;
    for (int i = 0 ; i< num_groups; i++){
        molec_prop += multiplier[i] * group_prop[i]
     }
     return molec_prop;
}

double cp_func(double tK)  {
       double cp_RXN=0;
       for (int i=0; NumSpecies; i++){
          Cp_a0 = 105.94/1000; // J/mol
          Cp_b0 = -5.4/1000;   // J/(mol.K)
          Cp_c0= -7.24/1000;   // J/(mol.K^2)
          double t0 = 298.15 ;
          double Teta0 = 298.15/100;
          double Teta = tK/100 ;
                    Cp_a[i]= sum_prod(func_groups[i], Cp_ai);
                    Cp_b[i]= sum_prod(func_groups[i], Cp_bi);
                    Cp_c[i]= sum_prod(func_groups[i], Cp_ci);
                    CA[i] = Cp_a[i] + Cp_a0;
                    CB[i] = (Cp_b[i]+Cp_b0) * Teta;
                    CC[i] = (Cp_c[i]+Cp_c0) * SQR(Teta);
```

```
                Cp[i] =  CA[i] + CB [i] + CC [i];
                cp_RXN += Cp[i]*StoiCoeff[i];
              }
              return cp_RXN;
}

 delHrxn_func(double tk){
     double del_Hrxn298=0;
     CpTot=0;
     for (int i=0; NumSpecies; i++){
        hf298[i]= sum_prod(func_groups[i], hfi);
        del_Hrxn298 += StoiCoeff[i]*hf298[i];
     }

     int binNum=500;
     double Tmin=298.15;
     double Tmax=tK;
     double delT=(Tmax-Tmin)/binNum;
     double Tbin=Tmin;
     double CpIntegral=0;
     for (int i=0; i<binNum; i++){
       CpIntegral += Tbin * cp_func(Tbin+delT/2);
       Tbin += delT
     }
     double del_Hrxn= del_Hrxn298 + CpIntegral;

  return del_Hrxn
}

double lnK_T_func(double tK){
     double del_G298=0;
     CpTot=0;
     for (int i=0; NumSpecies; i++){
        gf298[i]= sum_prod(func_groups[i], gi);
        del_G298 += StoiCoeff[i]*gf298[i];
     }

     double Krxn_T
     double ln_Krxn298 = -1*del_G298*1000/rGas/298.15
     double lnKT_lnK298= 0;

     int binNum=500;
     double Tmin=298.15;
     double Tmax=tK;
     double delT=(Tmax-Tmin)/binNum;
     double Tbin=Tmin;
```

```
    for (int i=0; i<binNum; i++){
      double delH_RT2 =delHrxn_func(Tbin+delT/2);
      delH_RT2 /= (rGas*(Tbin+delT/2)*(Tbin+delT/2));
      LnK_Integral += Tbin * delH_RT2
      Tbin += delT
    }
    Krxn_T = exp(LnK_Integral + ln_Krxn298 );
    return Krxn_T;
}
```

- **Fugacity Calculation**

```
double fug_fun(int i, int j, double tK, double Pbar, double s_mShape, double s_sigma,
double s_eok, double s_eHB_k, double s_kab, int s_Nd)) {

// In order to use PCSAFT Equation of State to solve for fugacity coefficients, we require
// to obtain the EOS parameters. PCSAFT parameters can be found by Emami et al.[11]
// method using Table S.3 or by any other available group-contribution approaches.
// Parameters can also be fitted on thermo-physical data which is specially useful and
// solvents and small molecules.

// State point calculations at T/K and P/bar of reaction should be performed to obtain
// fugacities of every substance in the reaction with solvent. Since we assume the
// solution is at infinite dilution, we will not need to count for interactions between solute
// molecules, otherwise the calculations would be more complicated

        double pi=3.141592654;
        double kBoltzman= 8.314/6.022d23;

        /* the definition of many of parameters are skipped only for brevity*/

        x(0)=0.999; //Compound 0 is the solvent. Infinite dilution assumption
        for (int i=1; i<icomp; i++){
              x(i)=0.001;
        }

        density = density_func(tKelvin,Pbar);
           //This function iterates on density to obtain Pcal=Pbar at given tKelvin.
           // This calculation is standard thermodynamic calculation and the definition
           // of this function is not addressed here for brevity.

        mShape[0] = s_mShape[i];
        mShape[1] = s_mShape[j];

        sigma[0] = s_sigma[i];
        sigma[1] = s_sigma[j];
```

```
eok[0] = s_eok[i];
eok[1] = s_eok[j];

eHB_k[0] = s_eHB_k[i];
eHB_k[1] = s_eHB_k[j];

kab [0] = s_kab [i];
kab [1] = s_kab[j];

Nd[0] = s_Nd[i];
Nd[1] = s_Nd[j];

Xitha0=0.0
Xitha1=0.0
Xitha2=0.0
Xitha3=0.0
m_mean=0.0


for (int i=0; i<icomp; i++){

        dEhs(i)=sigma(i)*(1-0.12*EXP(-3*eok(i)/tKelvin));
        md(i)=mShape(i)*dEhs(i);
        md2(i)=mShape(i)*dEhs(i)*dEhs(i);
        mReverse(i)=1/mShape(i);
        mdReverse(i)=mReverse(i)/dEhs(i);
        md2Reverse(i)=mdReverse(i)/dEhs(i);
        md3Reverse(i)=md2Reverse(i)/dEhs(i);

        Xitha0 +=  pi/6*rho*x(i) * mShape(i) ;
        Xitha1 +=  pi/6*rho*x(i) * mShape(i) *dEhs(i);
        Xitha2 += pi/6*rho*x(i) * mShape(i) * dEhs(i) * dEhs(i);
        Xitha3 += pi/6*rho*x(i) * mShape(i) * dEhs(i) * dEhs(i)*dEhs(i);
        eta = Xitha3;
        m_mean += x(i) * mShape(i)

}

ap(1,1)=  0.91056314451539;
ap(1,2)= -0.30840169182720;
ap(1,3)= -0.09061483509767;
.
..
…
ap(7,1)=  91.2977740839123;
ap(7,2)= -33.7469229297323;
```

```
ap(7,3)= -8.67284703679646;

bp(1,1)=  0.72409469413165;
bp(1,2)= -0.57554980753450;
bp(1,3)=  0.09768831158356;
.
..
…
bp(7,1)= -50.8003365888685  *7;
bp(7,2)= -23.6010990650801  *7;
bp(7,3)= -4.23812936930675  *7;

am(1)=ap(1,1) + (m_mean -1)/ m_mean *ap(1,2) +
     (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *ap(1,3);
am(2)=ap(2,1) + (m_mean -1)/ m_mean *ap(2,2) +
     (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *ap(2,3);
am(3)=ap(3,1)+ (m_mean -1)/ m_mean *ap(3,2) +
    (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *ap(3,3);
am(4)=ap(4,1) + (m_mean -1)/ m_mean *ap(4,2) +
    (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *ap(4,3);
am(5)=ap(5,1) + (m_mean -1)/ m_mean *ap(5,2) +
    (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *ap(5,3);
am(6)=ap(6,1) + (m_mean -1)/ m_mean *ap(6,2) +
   (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *ap(6,3);
am(7)=ap(7,1) + (m_mean -1)/ m_mean *ap(7,2) +
   (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *ap(7,3);

bm(1)=bp(1,1) + (m_mean -1)/ m_mean *bp(1,2) +
   (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *bp(1,3);
bm(2)=bp(2,1) + (m_mean -1)/ m_mean *bp(2,2) +
   (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *bp(2,3);
bm(3)=bp(3,1) + (m_mean -1)/ m_mean *bp(3,2) +
   (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *bp(3,3);
bm(4)=bp(4,1)+ (m_mean -1)/ m_mean *bp(4,2) +
   (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *bp(4,3);
bm(5)=bp(5,1) + (m_mean -1)/ m_mean *bp(5,2) +
   (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *bp(5,3);
bm(6)=bp(6,1) + (m_mean -1)/ m_mean *bp(6,2) +
   (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *bp(6,3);
bm(7)=bp(7,1) + (m_mean -1)/ m_mean *bp(7,2) +
    (m_mean -1)/ m_mean *( m_mean -2)/ m_mean *bp(7,3);

order1=0.0;
order2=0.0;

for (int i=0; i< icomp; i++){
```

```
        for (int j=0; j<icomp; j++){
                order1 += x(i) * x(j) * mShape(i) * mShape(i)
                    *sigma(i)**3*eok(i)/tKelvin;
                order2 += x(i) * x(j) * mShape(i) * mShape(i)
                    *sigma(i)**3*(eok(i)/tKelvin)**2;
        }
}

voidFrac=1-eta;
voidFrac2=voidFrac*voidFrac;
voidFrac3=voidFrac*voidFrac2;
voidFrac4=voidFrac*voidFrac3;
voidFrac5=voidFrac*voidFrac4;
voidFrac6=voidFrac*voidFrac5;

I1=am(1);
I2=bm(1);
dEtaI1_dEta=am(1);
dEtaI2_dEta=bm(1);
etaProd=eta;

for (int iCoeff=2; iCoeff<8; iCoeff++){
        I1=I1+etaProd*am(iCoeff);
        I2=I2+etaProd*bm(iCoeff);
        dEtaI1_dEta=dEtaI1_dEta + iCoeff*am(iCoeff)*etaProd ;
        dEtaI2_dEta=dEtaI2_dEta + iCoeff*bm(iCoeff)*etaProd ;
        etaProd=etaProd*eta;
}

denom=voidFrac*(2-eta);
denom2=denom*denom;
denom3=denom*denom2;

if (voidFrac<0) cerr << "Check! density is not feasible...\n";

zHs=eta/(1-eta) + 3*Xitha1*Xitha2/Xitha0/(1-eta)**2
        +(3*Xitha2**3-eta*Xitha2**3)/Xitha0/(1-eta)**3;
zHs=4*eta*(1-eta/2)/voidfrac3;
aHs=1/Xitha0* (3*Xitha1*Xitha2/voidFrac +
        Xitha2**3/eta/(1-eta)**2+(Xitha2**3/(eta*eta)-
            Xitha0)*DLOG(voidFrac));
gHs=1/(1-eta)+(dEhs/2)*3*Xitha2/(1-
     eta)**2+(dEhs/2)**2*2*Xitha2**2/voidFrac3;
gHs=(1-eta/2)/voidfrac3;
rhogHs_rho=eta*(2.5-eta)/voidfrac4;
rhogHs_rho=eta/(1-eta)**2+dEhs/2*(3*Xitha2/(1-eta)**2+6*Xitha2
```

```
        *eta/(1-eta)**3)+(dEhs/2)**2*(4*Xitha2**2/
       (1-eta)**3+6*Xitha2**2*eta/(1-eta)**4);
zHc=mShape*zHs - ((mShape-1)/gHs*rhogHs_rho);
if (gHs<0) cerr << "gHs<0\n";
aHc=mShape*aHs - (mShape-1) * log(gHs);
C1inv=1+mShape*(8*eta-2*eta*eta)/voidFrac4 +
        (1-mShape)*eta*(  20+eta*(-27+eta*(12-2*eta) )  )/denom2;
C1=1/C1inv;
C2=-(C1*C1)* (mShape*(-4*eta**2+20*eta+8)/voidFrac5+
         (1-mShape)*  (2*eta**3+12*eta**2-48*eta+40)/denom3   );
zDisp=-2*pi*rho*dEtaI1_dEta*order1-pi*rho*mShape*
        (C1*dEtaI2_dEta+C2*eta*I2)*order2;
aDisp=-2*pi*rho*I1*order1-pi*rho*mshape*C1*I2*order2;
Sig2=sigma*sigma;
Sig3=Sigma*Sig2;
Yhb=exp(eHB_k/tkelvin)-1;
Del=gHs*Yhb*Sig3*kab;
dDel_deta=Yhb*Sig3*Kab*(-0.5*voidFrac+3*(1-eta/2))/voidFrac4;

Factor=1+4*rho*Del;
if (Factor<0) cerr << "Error! divided by zero\n ";
Xa=(-1+sqrt(Factor))/(2*rho*Del+1e-10);
dXa_deta=(Del+eta*dDel_deta)*(1/sqrt(Factor)-Xa)/eta/
            (Del+1e-10);
zAssoc=eta*Nd*(2/(Xa+1e-10)-1)*dXa_deta;
aAssoc=2*Nd* log(XA+1e-10)+Nd*(1-XA);
zFactor=1+zHc+zDisp+zAssoc;
aRes=aHc+aDisp+aAssoc;


LnPhi=aRes+zFactor-1-log(zFactor);

// calculate the derivative of Helmholtz with respect to molfraction x(i)
nComp=2;
for (int m=0; m<nComp; m++{
        for (int n=0; n<nComp; n++){
                Dij(m,n) = dEhs(m)* dEhs(n)/( dEhs(m) + dEhs(n))
        }
}

for (int i=0; i<nComp; i++) {

        zeta0_xi = pi/6 * rho *mShape(i);
        zeta1_xi = pi/6 * rho *mShape(i)*dEhs(i);
        zeta2_xi = pi/6 * rho *mShape(i)*dEhs(i) *dEhs(i);
        zeta3_xi = pi/6 * rho *mShape(i)*dEhs(i) *dEhs(i)* dEhs(i);
```

```
m_xi(i) = (mShape (i) - m_mean ) / rho;

// hard sphere contribution

mu_hs(i) =  (  3.0*( zeta1_xi * zeta2+ zeta1* zeta2_xi)/ voidFrac +
          3.0*z1* zeta2* zeta3_xi / voidFrac / voidFrac + 3.0* zeta2* zeta2*
          zeta2_xi/ zeta3/ voidFrac / voidFrac + zeta2**3 * zeta3_xi *(3.0*
          zeta3-1.0)/ zeta3/ zeta3/ voidFrac **3 + ((3.0* zeta2*
          zeta2*z2_rk*z3-2.0* zeta2**3 * zeta3_xi)/z3**3 – zeta0_xi)
         *log(voidFrac) + (zeta0- zeta2**3 / zeta3/ zeta3)* zeta3_xi /
          voidFrac);

// hard chain contribution
for (int m=0; m<nComp; m++){
        for (int n=0; n<nComp; n++){
                gij(m,n) = 1.0/ voidFrac + 3.0* Dij(m,n) (m,n)* zeta2/
                      voidFrac / voidFrac + 2.0*( Dij(m,n) * zeta2)**2 /
                      voidFrac **3;

                gij_rx(i,j) = zeta3_xi / voidFrac / voidFrac +
                        3.0*dij_ab(i,j)*( zeta2_xi +2.0* zeta2* zeta3_xi /
                        voidFrac)/ voidFrac / voidFrac + Dij(m,n) **2 * zeta2/
                        voidFrac **3*(4.0* zeta2_xi +6.0* zeta2* zeta3_xi /
                        voidFrac);
        }
}

mu_ahc(i) = 0.0;
for (int m=0; m<nComp; m++){
        mu_hc(i) += x(m)*rho * (1.0-mShape(m)) / gij(m,m) * gij_rx(m,m);
}

mu_ahc(i) += ( 1.0- mShape (i)) * log( gij(i,i) );

// dispersion contribution

for (int m=0; m<6; m++) {
        ap_xi(i,m) = m_xi (i)/m_mean**2 * ( ap(m,2) +
                (3.0 -4.0/m_mean) *ap(m,3) );

        bp_xi(i,m) = m_xi(i)/m_mean**2 * ( bp(m,2) +
                (3.0 -4.0/m_mean) *bp(m,3) );
}

I1_xi = 0.0
```

```
I2_xi = 0.0
for (int m=0; m<6; m++) {
        I1_xi += am(m)*double(m)*eta**double(m-1)* zeta3_xi +
                ap_xi(i,m)*eta**double(m);

        I2_xi += bp(m)*double (m)*eta**double(m-1)* zeta3_xi +
                bp_xi(i,m)*eta**double(m);
}

ord1_xi  = 0.0;
ord2_xi  = 0.0;

for (int m=0; m<nComp; m++){
        k_im=0; //this could be added to the model to correct for
                segment-segment interaction
        sigij(i,m) =(sig(m)+sig(i))/2;
        eokij(i,m) = sqrt( eok(i) * eok(j) ) *(1- k_im);
        order1_xi += 2.0*mShape(i)*rho*x(m)*mseg(m)*
                sigij(m, i)**3 *eokij(m, i)/tKelvin;

        order2_xi +=  2.0*mShape(i)*rho*x(m)*mseg(m)*sig_ij(m,i)**3
                *(eokij(m,i)/tKelvin)**2;
}

C1_xi = C2* zeta3_xi - C1*C1* m_xi (i) * ( (8.0*zeta3-2.0*zeta3*zeta3)/
        voidFrac **4 - (-2.0*zeta3**4 +12.0*zeta3**3 -27.0*zeta3*
        zeta3+20.0*zeta3) / (voidFrac *(2.0-zeta3))**2  );

mu_dsp(i) = -2.0* pi * ( order1*rho*rho*I1_xi + order1_xi*I1 ) - pi*
         C1*m_mean * ( order2*rho*rho*I2_xi + order2_xi*I2 ) - pi*
         (C1*m_xi(i) + C1_xi*m_mean ) * order2*rho*rho*I2;



//Association Part
mu_hbon(i) = 0.0
ass_s2  = 0.0
for (int l=0, l< nhb(i); m++){
        ass_s2  += nhb_no(i,l) * log(m_x(i,l));
}

mu_hbon(i) = ass_s2;

for (int m=0; m<nComp; m++) {
        for (int n=0; n<nComp; n++) {
                mu_hbon(i) +=  -1* rho*rho/2.0*x(m)*x(n)
```

```
                                *mx(m)*mx(n) *nhb_no(m)*nhb_no(n) * gij_xi(m,n)
                                * ass_d(m,n);
                        }
                }

                mu_res(i) = mu_hs(i) +mu_hc(i) +mu_dsp(i) +mu_hbon(i);

        }

        for (int i=0; i<nComp; i++){
                LnPhi(i) = ares_xi(i) – log (zFactor);
                Fug(i) = exp (LnPhi(i));

        return Fug;

}

double * gamma_func (double tK) {
   double rho;

   double phi_solvent=0;
   double phi_dilute_solute=0;

   double my_fug[2];

   for (int i=1; i< NumSpecies){
       my_fug = fug_fun(i, 0 , tK, Pbar, mShape, sigma, eok, eHB_k, kab, Nd));
                    //specie 0 is the solvent
      phi_solvent = my_fug [0];
      phi_dilute_solute = my_fug [1];
      gamma[i] = phi_pure_solvent/phi_pure_solute;
   }

   return gamma;
}
```

- **Yield Calculation**

```
double yield_func ( double tk, int reaction_index){

   int i = reaction_index;
   double K_t = lnK_T_func(tK);
   double * gammas;
   gammas = new double [NumSpecies];
   gammas = gamma_func(tK);
```

```
    //The simplest way to perform analytical calculations is to
    //arrange the reaction in a way that all the stoichiometric coefficients are
    //equal to 1. e.g. for 2A->B, we can consider A+A->B (ireact=2 & iprod =1)

    if (ireact==1 && iprod==1) { //A->B
        double gamma_a = gammas[0];
        double gamma_b = gammas[1];
        double Krxn_T_ = Krxn_T * gamma_a/ gamma_b;
        reaction[i].rxn_YieldGC = Krxn_T_ /(1 + Krxn_T_);
    }else if (ireact==2 && iprod==1) { //A+B->C

        double gamma_a = gammas[0];
        double gamma_b = gammas[1];
        double gamma_c = gammas[2];
        double Krxn_T_ = Krxn_T * gamma_a* gamma_b/ gamma_c;
        reaction[i].rxn_YieldGC  = (2+ Krxn_T_-sqrt(4+Krxn_T_*Krxn_T_)) /2 ;

    }else if (ireact==2 && iprod==2) {//A+B->C+D
        double gamma_a = gammas[0];
        double gamma_b = gammas[1];
        double gamma_c = gammas[2];
        double gamma_d = gammas[3];
        double Krxn_T_ = Krxn_T * gamma_a * gamma_b/ gamma_c/ gamma_d;
        reaction[index].rxn_YieldGC = sqrt(Krxn_T_ )/(1+sqrt(Krxn_T_)) ;
    }else if (ireact==n && iprod==m) { // put the formula corresponding to number of
reactant and product in the reaction
    .
    ..
    …
return reaction[i].rxn_YieldGC
}
```

Supplementary References:

[1] K. J. M. Bishop, R. Klajn, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2006**, *45*, 5548-5554; C. M. Gothard, S. Soh, N. A. Gothard, B. Kowalczyk, Y. H. Wei, B. Baytekin, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2012**, *51*, 7922-7927; M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski, K. J. M. Bishop, *Angew. Chem. Int. Ed.* **2012**, *51*, 7928-7932; M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2005**, *44*, 7263-7269.

[2] A. S. Hukkerikar, B. Sarup, A. Ten Kate, J. Abildskov, G. Sin, R. Gani, *Fluid Phase Equilib.* **2012**, *321*, 25-43.

[3] M. Kleiner, J. Gross, *AIChE J.* **2006**, *52*, 1951-1961.

[4] J. Vrabec, J. Gross, *J. Phys. Chem. B.* **2008**, *112*, 51-60.

[5] J. Gross, *AIChE J.* **2005**, *51*, 2556-2568.

[6] K. P. Shukla, W. G. Chapman, *Mol. Phys.* **1997**, *91*, 1075-1081.

[7] W. G. Chapman, G. Jackson, K. E. Gubbins, *Molecular Phys.* **1988**, *65*, 1057-1079.

[8] G. A. Mansoori, N. F. Carnahan, K. E. Starling, T. W. Leland, *J. Chem. Phys.* **1971**, *54*, 1523-&.

[9] J. Gross, O. Spuhl, F. Tumakaka, G. Sadowski, *Ind. Eng. Chem. Res.* **2003**, *42*, 1266-1274.

[10] M. S. Wertheim, *J. Chem. Phys.* **1987**, *87*, 7323-7331.

[11] F. S. Emami, A. Vahid, J. R. Elliott, Jr., F. Feyzi, *Ind. Eng. Chem. Res.* **2008**, *47*, 8401-8411.

[12] G. C. Pimentel, A. L. McClellan, *The Hydrogen Bond*, 1st ed., Reinhold Publishing Corp., New York, NY., **1960**.

[13] T.-B. Nguyen, J.-C. de Hemptinne, B. Creton, G. M. Kontogeorgis, *Ind. Eng. Chem. Res.* **2013**, *52*, 7014-7029.

[14] Z. Kolska, J. Kukal, M. Zabransk, V. Ruzicka, *Ind. Eng. Chem. Res.* **2008**, *47*, 2075-2085.

[15] S. I. Sandler, *Chemical, Biological, and Engineering Thermodynamics*, 4th ed., Wiley, New York, **2006**.

[16] E. B. Nauman, *Chemical Reactor Design, Optimization, and Scaleup*, 2nd ed., Wiley, New Jersey, **2008**.

[17] K. Inagaki, T. Ohta, K. Nozaki, H. Takaya, *J. Organomet. Chem.* **1997**, *531*, 159-163.

[18] H. Stamm, V. Gailius, *Chem. Ber.-Recl.* **1981**, *114*, 3599-3608.

[19] (Ed.: N. I. S. R. Institute), **2007**.

[20] T. Ohkuma, C. A. Sandoval, R. Srinivasan, Q. Lin, Y. Wei, K. Muniz, R. Noyori, *J. Am. Chem. Soc.* **2005**, *127*, 8288-8289.

[21] T. Yamamura, H. Nakatsuka, S. Tanaka, M. Kitamura, *Angew. Chem. Int. Ed.* **2013**, *52*, 9313-9315.

[22] A. Kamal, V. Devaiah, K. L. Reddy, N. Shankaraiah, *Adv. Synth. Catal.* **2006**, *348*, 249-254.

[23] Y. Endo, K. Shudo, T. Okamoto, *Synthesis-Stuttgart* **1983**, 471-472.

[24] S.-F. Hsu, B. Plietker, *Chemcatchem* **2013**, *5*, 126-129.

[25] (Ed.: C. o. S. a. I. Research), **2006**.

[26] A. A. Kirchanov, A. S. Zanina, I. L. Kotlyarevskii, ***B. Acad. Sci. USSR CH+*** **1981**, *30*, 1579-1580.

14. Statements of contribution

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES

01-224 WARSAW
ul. KASPRZAKA 44/52
Phone: + 48 (22) 631 87 88
Fax: + 48 (22) 632 66 81
E-mail: icho-s@icho.edu.pl

July 22, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the papers:

1. Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I contributed to the development of methods described in the paper, especially in formalizing the higher order chemical logic and inputting new reactions (ca 7 000 reactions since 2016 and 15 000 in total). I queried Chematica to produce pathways that were subsequently validated experimentally. I also participated in the writing of the manuscript.

2. Michał Bajczyk, Piotr Dittwald, Agnieszka Wołos, Sara Szymkuć and Bartosz A. Grzybowski "Discovery and Enumeration of Organic-Chemical and Biomimetic Reaction Cycles within the Network of Chemistry" *Angew. Chem. Int. Ed.* **2018**, 57, 2367- 2371

I co-performed analyses of chemical cycles within Cyclorg network. I also helped with preparing the dataset of reactions. I also participated in the writing of the manuscript.

3. Bartosz A. Grzybowski, Sara Szymkuć, Karol Molga, Ewa P. Gajewska, Agnieszka Wołos "Synthetic design with the Chematica program – the importance of accurate rules and of higher-order logic" *CHIMIA* **2017**, 71, 512

I contributed to the development of methods described in the paper.

4. Grzegorz Skoraczyński, Piotr Dittwald, Błażej Miasojedow, Sara Szymkuć, Ewa P. Gajewska, Bartosz A. Grzybowski, Anna Gambin "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?" *Sci. Rep.* **2017**, 7, 3582

I provided chemical examples and validated chemical conclusions. I also participated in the writing of the manuscript.

Sara Szymkuć

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES

01-224 WARSAW
ul. KASPRZAKA 44/52
Phone: + 48 (22) 631 87 88
Fax: + 48 (22) 632 66 81
E-mail: icho-s@icho.edu.pl

5. Sara Szymkuć, Ewa Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk and Bartosz A. Grzybowski "Computer-assisted synthetic planning: The end of the beginning" *Angew. Chem. Int. Ed.* **2016** , 55, 5904-5937

I designed Chematica's knowledge base and input ca 8 000 reactions (out of 15 000 up till 2018). I proposed dual scoring functions and their variables. I contributed to the development of other modules and methods described in the paper. I also participated in the writing of the manuscript.

6. Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, Sara Szymkuć, Piotr Dittwald, Karol Molga and Prof. Bartosz A. Grzybowski "A Priori Estimation of Organic Reaction Yields", *Angew. Chem. Int. Ed.* **2015**, 54, 10797-10801

I performed statistical analysis of organic reactions under thermodynamic vs. kinetic control.

mgr inż. Sara Szymkuć

Sara Szymkuć

July 22, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the papers:

1. Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I conceived the project, supervised research and paper writing.

2. Michał Bajczyk, Piotr Dittwald, Agnieszka Wołos, Sara Szymkuć and Bartosz A. Grzybowski "Discovery and Enumeration of Organic-Chemical and Biomimetic Reaction Cycles within the Network of Chemistry" *Angew. Chem. Int. Ed.* **2018**, 57, 2367- 2371

I conceived the project, supervised research and paper writing.

3. Bartosz A. Grzybowski, Sara Szymkuć, Karol Molga, Ewa P. Gajewska, Agnieszka Wołos "Synthetic design with the Chematica program – the importance of accurate rules and of higher-order logic" *CHIMIA* **2017**, 71, 512

I conceived the project, supervised research and paper writing.

4. Grzegorz Skoraczyński, Piotr Dittwald, Błażej Miasojedow, Sara Szymkuć, Ewa P. Gajewska, Bartosz A. Grzybowski, Anna Gambin "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?" *Sci. Rep.* **2017**, 7, 3582

I conceived the project, supervised research and paper writing.

5. Sara Szymkuć, Ewa Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk and Bartosz A. Grzybowski "Computer-assisted synthetic planning: The end of the beginning" *Angew. Chem. Int. Ed.* **2016** , 55, 5904-5937

I conceived the project, supervised research and paper writing.

6. Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, Sara Szymkuć, Piotr Dittwald, Karol Molga and Prof. Bartosz A. Grzybowski "A Priori Estimation of Organic Reaction Yields", *Angew. Chem. Int. Ed.* **2015**, 54, 10797-10801

I conceived the project, supervised research and paper writing.

Prof. Bartosz A. Grzybowski

July 20, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the papers:

1. Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I was one of the key developers of the search algorithms in Chematica.

2. Michał Bajczyk, Piotr Dittwald, Agnieszka Wołos, Sara Szymkuć and Bartosz A. Grzybowski "Discovery and Enumeration of Organic-Chemical and Biomimetic Reaction Cycles within the Network of Chemistry" *Angew. Chem. Int. Ed.* **2018**, 57, 2367- 2371

I implemented search algorithm for cycles and developed Cyclorg webservice.

3. Grzegorz Skoraczyński, Piotr Dittwald, Błażej Miasojedow, Sara Szymkuć, Ewa P. Gajewska, Bartosz A. Grzybowski, Anna Gambin "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?" *Sci. Rep.* **2017**, 7, 3582

I designed the models and performed calculations.

4. Sara Szymkuć, Ewa Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk and Bartosz A. Grzybowski "Computer-assisted synthetic planning: The end of the beginning" *Angew. Chem. Int. Ed.* **2016** , 55, 5904-5937

I was one of the key developers of the search algorithms in Chematica.

5. Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, Sara Szymkuć, Piotr Dittwald, Karol Molga and Prof. Bartosz A. Grzybowski "A Priori Estimation of Organic Reaction Yields", *Angew. Chem. Int. Ed.* **2015**, 54, 10797-10801

I performed additional data analysis.

Dr. Piotr Dittwald

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES

01-224 WARSAW
ul. KASPRZAKA 44/52
Phone: + 48 (22) 631 87 88
Fax: + 48 (22) 632 66 81
E-mail: icho-s@icho.edu.pl

July 22, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the papers:

1. Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I contributed to the development of methods described in the paper. I also participated in the writing of the manuscript.

2. Bartosz A. Grzybowski, Sara Szymkuć, Karol Molga, Ewa P. Gajewska, Agnieszka Wołos "Synthetic design with the Chematica program – the importance of accurate rules and of higher-order logic" *CHIMIA* **2017**, 71, 512

I contributed to the development of methods described in the paper.

3. Grzegorz Skoraczyński, Piotr Dittwald, Błażej Miasojedow, Sara Szymkuć, Ewa P. Gajewska, Bartosz A. Grzybowski, Anna Gambin "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?" *Sci. Rep.* **2017**, 7, 3582

I provided chemical examples and validated chemical conclusions.

4. Sara Szymkuć, Ewa Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk and Bartosz A. Grzybowski "Computer-assisted synthetic planning: The end of the beginning" *Angew. Chem. Int. Ed.* **2016** , 55, 5904-5937

I helped with the development of Chematica's knowledge base and input ca. 4 000 reactions according to specifications and computer routines developed by Ms Szymkuć. I also participated in the writing of the manuscript.

Ewa Gajewska

mgr inż. Ewa Gajewska

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES

01-224 WARSAW
ul. KASPRZAKA 44/52
Phone: + 48 (22) 631 87 88
Fax: + 48 (22) 632 66 81
E-mail: icho-s@icho.edu.pl

July 22, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the papers:

1. Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I contributed to the development of methods described in the paper. I also participated in the writing of the manuscript.

2. Bartosz A. Grzybowski, Sara Szymkuć, Karol Molga, Ewa P. Gajewska, Agnieszka Wołos "Synthetic design with the Chematica program – the importance of accurate rules and of higher-order logic" *CHIMIA* **2017**, 71, 512

I contributed to the development of methods described in the paper.

3. Sara Szymkuć, Ewa Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk and Bartosz A. Grzybowski "Computer-assisted synthetic planning: The end of the beginning" *Angew. Chem. Int. Ed.* **2016** , 55, 5904-5937

I helped with the development of Chematica's knowledge base and input ca. 4 000 reactions according to specifications and computer routines developed by Ms Szymkuć. I also participated in the writing of the manuscript.

4. Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, Sara Szymkuć, Piotr Dittwald, Karol Molga and Prof. Bartosz A. Grzybowski "A Priori Estimation of Organic Reaction Yields", *Angew. Chem. Int. Ed.* **2015**, 54, 10797-10801

I helped with the statistical analysis of organic reactions under thermodynamic vs. kinetic control.

Molga

mgr inż. Karol Molga

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES

01-224 WARSAW
ul. KASPRZAKA 44/52
Phone: + 48 (22) 631 87 88
Fax: + 48 (22) 632 66 81
E-mail: icho-s@icho.edu.pl

July 20, 2018
TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the papers:

1. Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I synthesized 5β/6β-hydroxylurasidone, dronedarone, and (*S*)-4hydroxyduloxetine. I also helped with the development of Chematica's knowledge base and input ca. 3 000 reactions according to specifications and computer routines developed by Ms Szymkuć.

2. Sara Szymkuć, Ewa Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk and Bartosz A. Grzybowski "Computer-assisted synthetic planning: The end of the beginning" *Angew. Chem. Int. Ed.* **2016** , 55, 5904-5937

I helped with the development of Chematica's knowledge base and input ca. 3 000 reactions according to specifications and computer routines developed by Ms Szymkuć.

mgr inż. Tomasz Klucznik

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES

01-224 WARSAW
ul. KASPRZAKA 44/52
Phone: + 48 (22) 631 87 88
Fax: + 48 (22) 632 66 81
E-mail: icho-s@icho.edu.pl

July 20, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the papers:

1. Michał Bajczyk, Piotr Dittwald, Agnieszka Wołos, Sara Szymkuć and Bartosz A. Grzybowski "Discovery and Enumeration of Organic-Chemical and Biomimetic Reaction Cycles within the Network of Chemistry" *Angew. Chem. Int. Ed.* **2018**, 57, 2367- 2371

I co-developed algorithms and methods for the discovery of organic chemical cycles.

2. Sara Szymkuć, Ewa Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk and Bartosz A. Grzybowski "Computer-assisted synthetic planning: The end of the beginning" *Angew. Chem. Int. Ed.* **2016** , 55, 5904-5937

I participated in user-evaluation of Chematica's network-module routines.

MBajczyk

mgr Michał Bajczyk

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES

01-224 WARSAW
ul. KASPRZAKA 44/52
Phone: + 48 (22) 631 87 88
Fax: + 48 (22) 632 66 81
E-mail: icho-s@icho.edu.pl

July 20, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the papers:

1. Michał Bajczyk, Piotr Dittwald, Agnieszka Wołos, Sara Szymkuć and Bartosz A. Grzybowski "Discovery and Enumeration of Organic-Chemical and Biomimetic Reaction Cycles within the Network of Chemistry" *Angew. Chem. Int. Ed.* **2018**, 57, 2367- 2371

I co-performed analyses of chemical cycles within Cyclorg network. I also helped with the design of the Cyclorg webservice.

2. Bartosz A. Grzybowski, Sara Szymkuć, Karol Molga, Ewa P. Gajewska, Agnieszka Wołos "Synthetic design with the Chematica program – the importance of accurate rules and of higher-order logic" *CHIMIA* **2017**, 71, 512

I helped with the development of Chematica's knowledge base and input ca. 3 000 reactions according to specifications and computer routines developed by Ms Szymkuć.

mgr inż. Agnieszka Wołos

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES

01-224 WARSAW
ul. KASPRZAKA 44/52
Phone: + 48 (22) 631 87 88
Fax: + 48 (22) 632 66 81
E-mail: icho-s@icho.edu.pl

July 20, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I synthesized 5β/6β-hydroxylurasidone, dronedarone, and (*S*)-4hydroxyduloxetine.

Barbara Mikulak-Klucznik

mgr inż. Barbara Mikulak-Klucznik

INSTITUTE OF ORGANIC CHEMISTRY
POLISH ACADEMY OF SCIENCES

01-224 WARSAW
ul. KASPRZAKA 44/52
Phone: + 48 (22) 631 87 88
Fax: + 48 (22) 632 66 81
E-mail: icho-s@icho.edu.pl

July 26, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I was working on the Chematica's filter for electrophilic aromatic substitution reactions.

Rafał Roszak

Dr. Rafał Roszak

July 23, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I helped in the synthesis of (*S*)-4-hydroxyduloxetine and also helped prepare substrates for the synthesis of dronedarone .

Bianka Sieredzińska

Bianka Sieredzińska

July 19, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I helped in the synthesis of (*S*)-4-hydroxyduloxetine.

Dr Ariel Adamski

Ariel Adamski

July 3, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Grzegorz Skoraczyński, Piotr Dittwald, Błażej Miasojedow, Sara Szymkuć, Ewa P. Gajewska, Bartosz A. Grzybowski, Anna Gambin "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?" *Sci. Rep.* **2017**, 7, 3582.

I conceived the project and supervised research.

Prof. Anna Gambin

July 3, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

1. Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I developed Synturus' search algorithm.

2. Sara Szymkuć, Ewa Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk and Bartosz A. Grzybowski "Computer-assisted synthetic planning: The end of the beginning" *Angew. Chem. Int. Ed.* **2016** , 55, 5904-5937

I developed Synturus' search algorithm.

Startek

Dr. Michał Startek

July 3, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Grzegorz Skoraczyński, Piotr Dittwald, Błażej Miasojedow, Sara Szymkuć, Ewa P. Gajewska, Bartosz A. Grzybowski, Anna Gambin "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?" *Sci. Rep.* **2017**, 7, 3582

I designed the models and performed calculations.

Dr. Błażej Miasojedow

July 3, 2018

TO WHOM IT MAY CONCERN

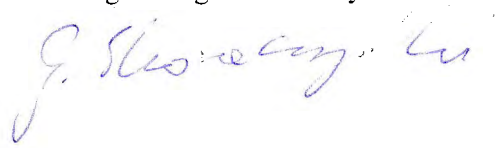
Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Grzegorz Skoraczyński, Piotr Dittwald, Błażej Miasojedow, Sara Szymkuć, Ewa P. Gajewska, Bartosz A. Grzybowski, Anna Gambin "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?" *Sci. Rep.* **2017**, 7, 3582.

I designed the models and performed calculations.

mgr Grzegorz Skoraczyński



August 12, 2018

Institute of Organic Chemistry of the Polish Academy of Sciences:

This letter describes my contribution in the paper "A Priori Estimation of Organic Reaction Yields"; Angew. Chem. Int. Ed. 2015, 54, 10797-10801. This letter is written to clarify the co-author contributions required for the graduation of Sara Szymkuć from her PhD program. My main contribution was to program the method using c++ and python programing languages. A large data-base of chemical reactions was used for training the model and for predicting the yield of reactions; I had a contribution in proofreading the applied database. The thermodynamic model and the mathematical algorithms were developed and chosen by Amir Vahid (first co-author).

Fateme Sadat Emami, PhD

Fateme S. Emami

08/12/2018

NORTHWESTERN
UNIVERSITY

Milan Mrksich
Henry Wade Rogers Professor
Departments of Chemistry,
Biomedical Engineering, and
Cell and Molecular Biology

milan.mrksich@northwestern.edu
Phone 847-467-0472

2145 Sheridan Rd
Technological Institute
Willens Laboratory, B490
Evanston, Illinois 60208

Assistant: Yael Mayer
yael.mayer@northwestern.edu
Phone: 847-467-0710

July 5, 2018

Intytut Chemii Organicznej
Polskieg Akademii Nauk
Ul. Kasprzaka 44/52
01-224 Warsaw
Poland

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I supervised the synthesis of engelheptanoxide C.

Sincerely,

Milan Mrksich

July 3, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I conceived and supervised the technology evaluation project on behalf of Millipore-Sigma.

Dr. Sarah Trice

July 3, 2018

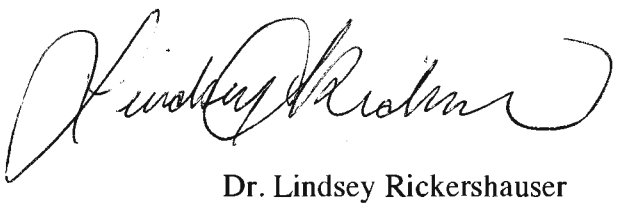
TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I provided advice on the synthesis of (*S*)-4-hydroxyduloxetine.



Dr. Lindsey Rickershauser

July 2, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I synthesized the engelheptanoxide C.

Dr. Yubai Zhou

July 3, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I synthesized ATR kinase inhibitor.

Dr. Alexei Toutchkine

July 3, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I synthesized the anti-leukemia drug candidate.

Dr. Michael McCormack

July 3, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I synthesized α-hydroxyetizolam.

Dr. Heather Lima

03 July 2018

July 3, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I provided advice on the synthesis of (*S*)-4-hydroxyduloxetine.

Tu, 03.2018

Dr. Gregory Kirkovits

July 3, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L. J. Trice, Bartosz A. Grzybowski "Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory" *Chem*, **2018**, 4, 522-532

I synthesized the BRD7/9 inhibitor.

Manishabrata Bhowmick

Dr. Manishabrata Bhowmick 7/2/2018

July 3, 2018

TO WHOM IT MAY CONCERN

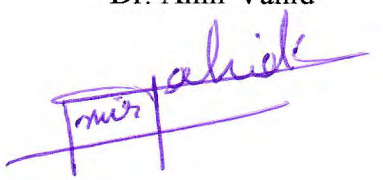
Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, Sara Szymkuć, Piotr Dittwald, Karol Molga and Prof. Bartosz A. Grzybowski "A Priori Estimation of Organic Reaction Yields", *Angew. Chem. Int. Ed.* **2015**, 54, 10797-10801

I developed thermodynamic model for the prediction of reaction yields.

Dr. Amir Vahid



July 23, 2018

TO WHOM IT MAY CONCERN

Statement of contribution

Hereby, I would like to claim that my contribution to the paper:

Fateme S. Emami, Amir Vahid, Elizabeth K. Wylie, Sara Szymkuć, Piotr Dittwald, Karol Molga and Prof. Bartosz A. Grzybowski "A Priori Estimation of Organic Reaction Yields", *Angew. Chem. Int. Ed.* **2015**, 54, 10797-10801

I co-developed thermodynamic model for the prediction of reaction yields.

Elizabeth Wylie