Shenvi Lab Group Meeting
October 2nd, 2017

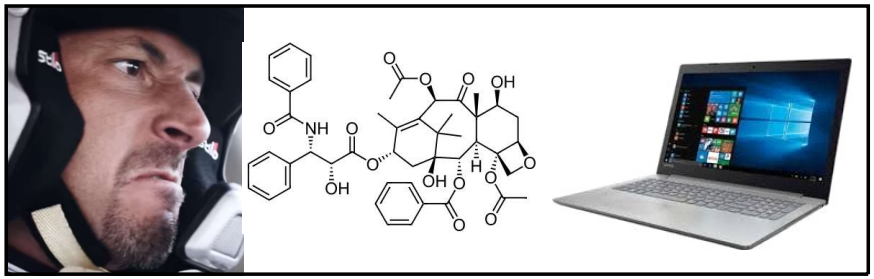
Computer Aided Retrosynthesis

Tucker Huffman

Man vs. Machine:

1997, Chess Grandmaster Garry Kasparov is defeated by IBM's Deep Blue.

2017, World Go Champion Ke Jie is defeated by Google's AlphaGo, 20 years later.



Is synthetic organic chemistry next? Will it take another 20 years?

Overview:

This talk will cover the basics of computer aided retrosynthesis, the history of its development and use, and the current state of the art.

While significant effort has been put into the related problems of computational reaction and mechanism prediction and database searching such as SciFinder and Reaxys, the focus here will be devoted towards software that given a target structure will produce (either automatically or with manual guidance) a retrosynthesis allowing the molecule's construction.

Why?

"That there is a need for such an application is made apparent by the fact that a complete, logic-centered synthetic analysis of a complex organic structure often requires so much time, even of the most skilled chemist, as to endanger or remove the feasibility of this approach." -E.J. Corey, 1969

"...it could be argued that one might be able to recognize these motifs using human intuition. But this is like arguing that we could, using paper and pencil, "eventually" divide two ten-digit numbers to the precision of ten decimal places–why do so if we have a pocket calculator available?" -Bartosz Gryzbowski, 2016

Goals:

"The intent ... is not to replace art in organic synthesis but to show where real art lies." -Jim Hendrickson

A computer aided retrosynthesis tool should assist a chemist plan a synthetic route towards a given target molecule. There have been different specific models by which different pieces of software have aimed to make this happen.

An important point: a retrosynthetic tool does not need to be perfect or even better than humans to be useful, it just needs to accelerate or otherwise improve the retrosynthetic process.

Fully Automated Retrosynthesis Tools:

Perhaps the easiest to understand and hardest to execute approach; a fully automated retrosynthesis tool would take a target molecule as input and give a retrosynthesis as an output.

Manual Retrosynthesis Tools:

More in line with Gryzbowski's calculator analogy, a large body of the work in this area has been towards software where the user actively directs the retrosynthesis that it produces.

Target Analysis Tools:

Some software has gone after the more simple aim of helping a chemist analyze a given target rather than explicitly handle any reactions. While less common, it does loosely fit under the same umbrella as retrosynthesis generating software.

Some key problems in computer aided retrosynthesis:

1) The reaction identification problem
2) The scoring problem
3) The tree-search problem

The Reaction Identification Problem

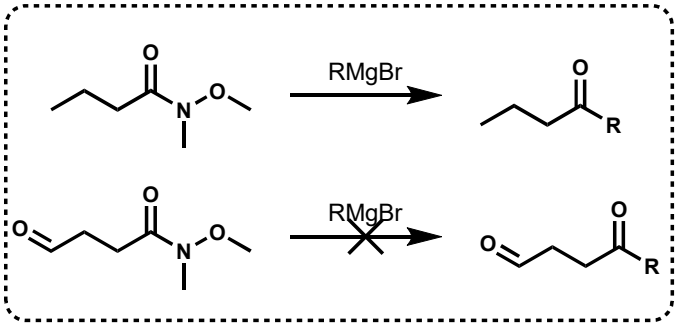
Tucker Huffman

For a computer to make disconnections retrosynthetically, it must know what reactions can be carried out in the forward direction.

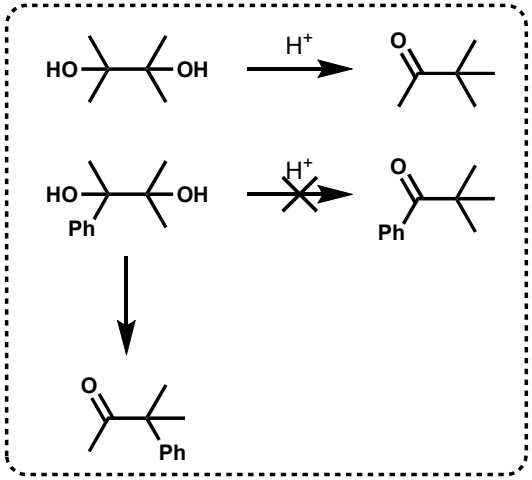
Overgeneralization:

Molecular context is highly important in a wide variety of reactions. If a piece of software is too loose with it's reactions, it can easily start applying transforms that are not feasible.

Functional Group Incompatibility: Requires Protection



Substituent Change Alters Reactivity



Szymkuc, S *et. al. Angew. Chem. Int. Ed.* **2016,** *55,* 5904.

Lack of Coverage:

Ideally, a retrosynthesis tool will be able to identify every synthetically useful transform that can be applied to a target molecule. This could include rare reactions for which scope is unknown, or recently published methods that might be left out.

The Scoring Problem

For a computer to identify a good retrosynthesis among many possibilities, it must be able to score disconnections/substructures.

Heuristics: Applying a number to a disconnection

While the heuristics traditionally used to aid retrosynthesis provide a useful qualitative tool for humans to evaluate routes, computers ultimately rely on quantitative, numerical values in order to rank things. Translation of human derived heuristics to numerical values can be done, but is fairly arbitrary. Data-driven heuristic generation is challenging and assumes that the literature contains largely good syntheses for models to be built off of.

What synthesis is the best synthesis?

It is essentially impossible to objectively rank the quality of a synthesis. Factors such as step count, overall yield, convergency/divergency, and "elegance" all matter differently depending on the purpose of the route.

The Tree-Search Problem

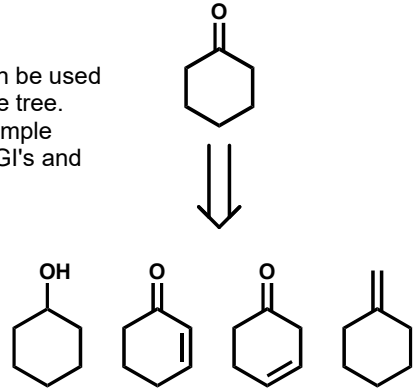
Tucker Huffman

For a computer to explore the vast space of possible disconnections, it must have an efficient means of searching synthetic trees.

This is a general computer science problem for which numerous algorithms have been made. There is no objective best choice of method to use for this problem.

Branching Factor:

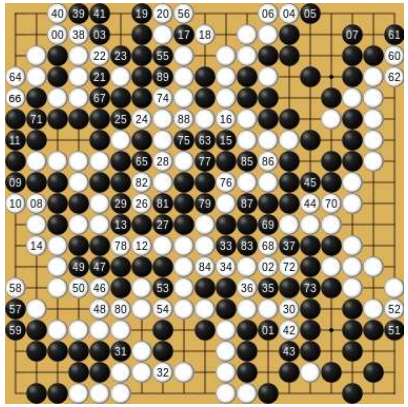
-Describes the number of reactions that can be used to reach the target or any intermediatein the tree.
-This number can be quite large even for simple structures, particularly when considering FGI's and sequences that remove functionality.



How many more possible precursors are there?

Search Depth:

-Describes the number of steps out from the target molecular that are examined.
-Every incremental increase in search depth requires the number of options examined be multiplied by the branching factor.
i.e. Nodes=Branching Factor^{Depth}
-For games like Chess and Go, the search depth can be quite large. Thankfully, in synthesis it is desireable to have shorter routes, allowing the searches to be truncated (e.g at ~20 steps).



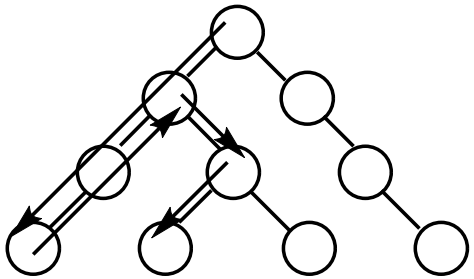
Final board after 289 moves, victory by half a point AlphaGo vs. Ke Jie

The number of transforms that can be applied to an arbitrary non-trivial target has been estimated to be ~80 on average, meaning that for 10 steps there are ~10^{19} synthetic possibilities.

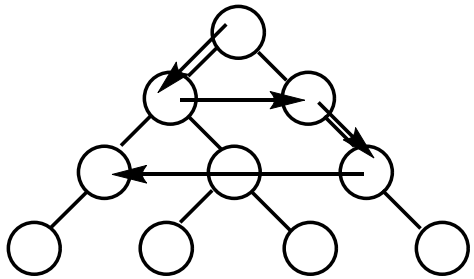
⟹

The search space is too large to exhaustively explore in a reasonable time frame even for modern computers. Searches must be prioritized.

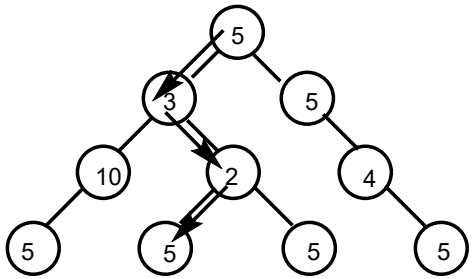
Depth First Search



Breadth First Search



Best First Search

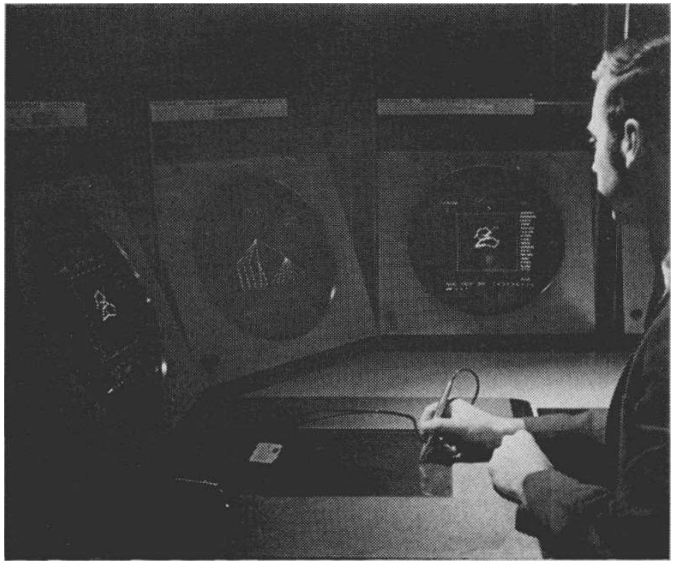


Szymkuc, S *et. al. Angew. Chem. Int. Ed.* **2016,** *55,* 5904.

The Beginning: Corey's OCSS and LHASA

Tucker Huffman

1957: Vlèduts and Finn present idea of "*information machine for chemistry*"
1965: The Dentral project at Stanford attempts to computationally predict structure from spectroscopic data and use AI to plan synthetic routes.
1969: Corey and Wipke publish a paper describing their system for computer aided retrosynthesis titled OCSS (Organic Chemical Simulation of Synthesis.)



Corey, E. J. *et. al. Science*, **1969,** *166,* 178.

Corey's Perspective on Computer Aided Retrosynthesis:

"...it is important to test our understanding of synthetic analysis by writing and evaluating such machine programs."

In Corey and Wipke's seminal paper, several "general requirements" are laid out that can be summarized as follows:

-The chemist has active control over the system at every stage of the process, allowing them to guide the retrosynthesis as it is produced.
-The computer can produce on it's own a tree small enough so that each option can be evaluated by the chemist.

A completely automated synthesis plan was intentionally avoided to allow the chemist to do their own evaluation of produced structures.

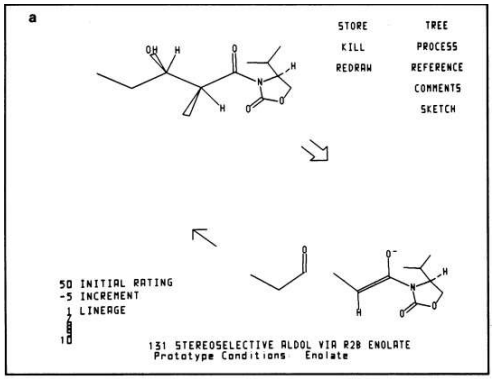
"These requirements limit the task to be performed by computer to the 'logic-centered' part of the synthesis and leave to the chemist the complex, ill-defined, and 'information-centered part, which is at present beyond the scope of computation."

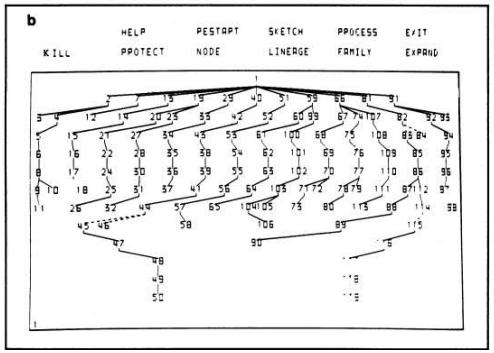
OCSS becomes LHASA:

Improvements to OCSS carried out by the Corey group became LHASA (Logic and Heuristics Applied to Synthetic Analysis), which is likely the most well known effort in the field. Wipke split off to develop SECS in a similar vein.

General Design:

-LHASA used an expert coded knowledge base and would generate a synthesis tree by comparing the input molecule against the transforms stored in it's database.
-Users could specify certain strategies that the software would use to generate the search tree, such as a transform-based strategy around a desired reaction or a stereochemistry-based strategy to reduce the number of stereocenters.
-The software relied on vaguely described heuristics to deterimine what disconnections should be made and to rank reactions.
-Functional group incompatibilites would be identified and flagged in suggeted routes. These routes would still be suggested, but would essentially leave it up to the chemist to plan protection/deprotection steps.





Corey, E. J. *et. al. Science,* **1985,** *228,* 408.

Downsides:

LHASA was inherently limitted by the quality of the transform knowledge base. Additionally, the tree-search problem was mostly dodged by asking the chemist to pick a strategy from the beginning, which while certainly limitting the search space also removes the possibility that the user finds a synthetic strategy that they didn't already have in mind from the beginning.

"The task of expanding the database is monumental, not only because of the number of organic reactions in the literature but also because the LHASA database contains more than just simple descriptions of reactants, products, and reaction conditions."

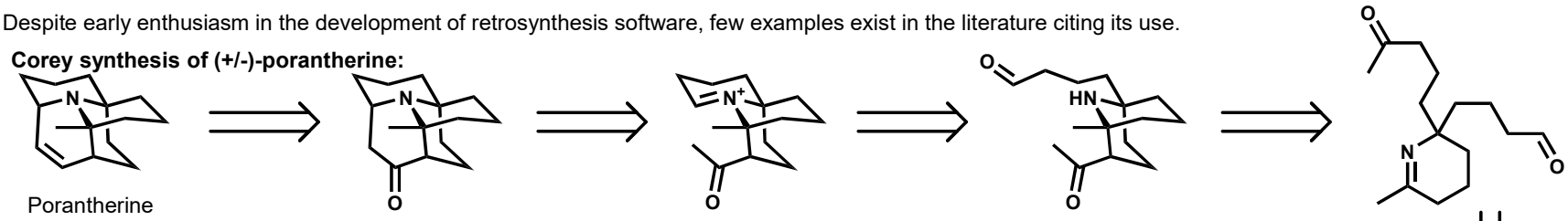
Szymkuc, S *et. al. Angew. Chem. Int. Ed.* **2016,** *55,* 5904.
Hanessian, S. *Curr. Opin. Drug Discov. & Devel.* **2005,** *8*, 798.

LHASA in Practice

Tucker Huffman

Despite early enthusiasm in the development of retrosynthesis software, few examples exist in the literature citing its use.

Corey synthesis of (+/-)-porantherine:



Even from Corey's lab, LHASA use is only ever described as occuring after the fact. While this demonstrates that LHASA is capable of producing realistic retrosyntheses, it doesn't show the software fulfilling the task for which it was designed, namely to help a chemist come up with a new synthetic plan.

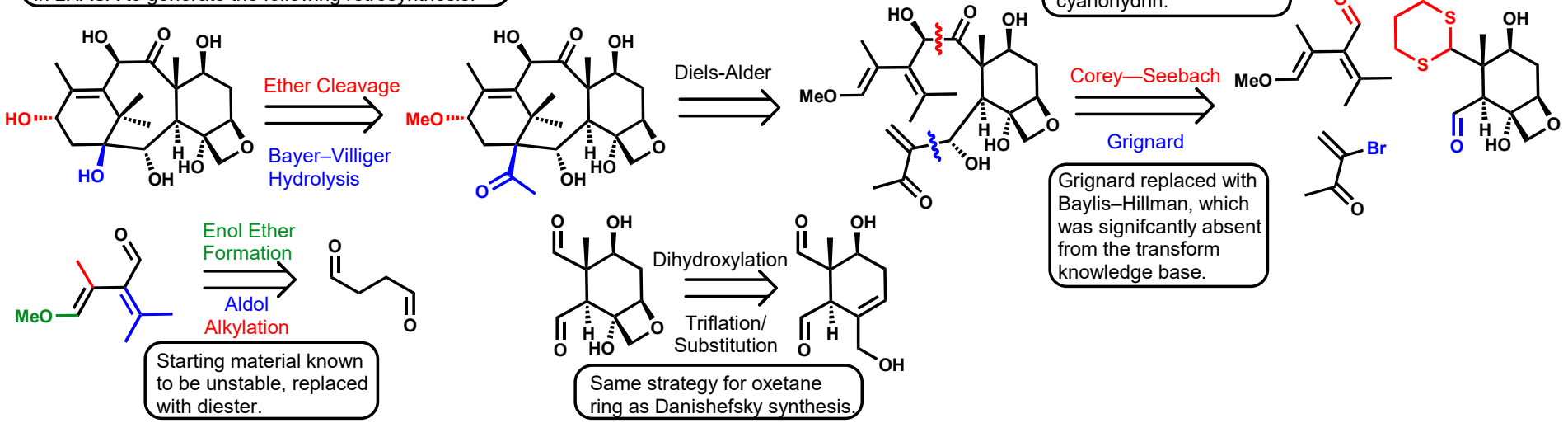
LHASA is only briefly mentioned in the last sentence of the paper:
"*The synthetic approach outlined in Scheme I (among others) was also suggested by LHASA-10, the Harvard* program for computer-assisted synthetic analysis."

Corey, E. J. et. al. *J. Am. Chem. Soc.* **1974,** *96*, 6516.

Retrosynthesis of Taxol using LHASA:

"To our knowledge, no reports have been published about the use of the LHASA program for designing syntheses of complex molecules, prior to practical application."

The user specified the Long-range (transform-based) Diels-Alder and Topological (disconnective) strategies in LHASA to generate the following retrosynthesis.



Dithiane failed in practice, replaced with protected cyanohydrin.

Grignard replaced with Baylis–Hillman, which was signifcantly absent from the transform knowledge base.

Starting material known to be unstable, replaced with diester.

Same strategy for oxetane ring as Danishefsky synthesis.

"...it is virtually impossible to account for all possible applications of a reaction and sometimes a doubtful application is shown. The main conclusions which can be drawn here are that the user should critically evaluate the suggestions and that the LHASA program is only useful in the hands of a skilled chemist."

Rozendaal, E. V. Ph.D. Dissertation, Radboud University Nijmegen, 1994.

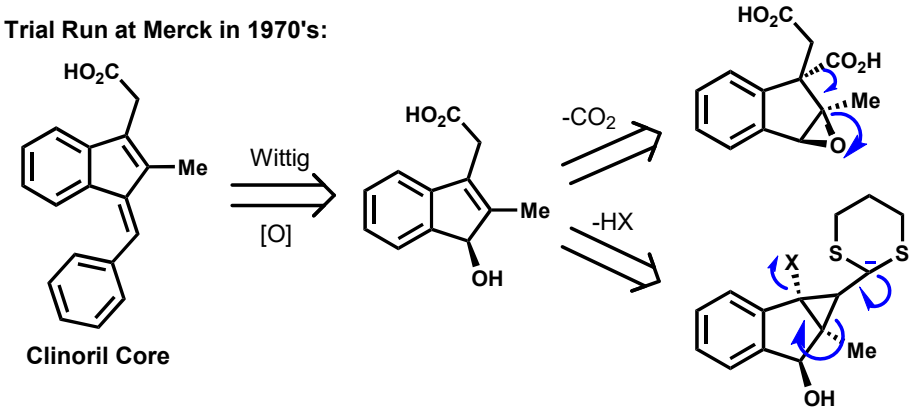
Other Early Retrosynthesis Software

Tucker Huffman

SECS (Simulation and Evaluation of Chemical Synthesis):

SECS was structurally very similar to LHASA. Differences include improvments in stereochemistry and topology handling and introduction of some rudimentary energy minimization.

Trial Run at Merck in 1970's:



Clinoril Core

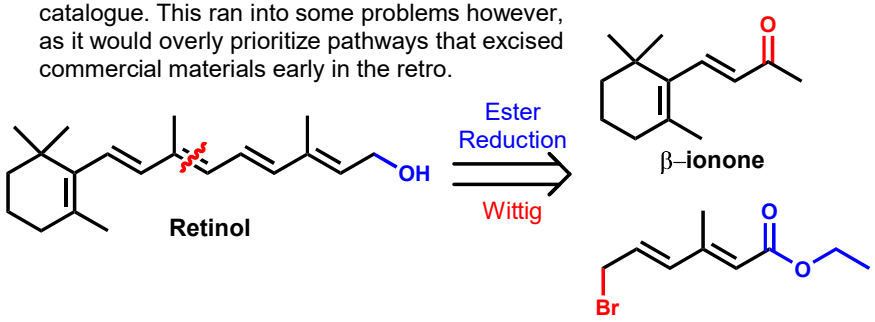
"...the suggestion of [these structures] by the synthesis program incited chemists working on the problem to think about fragmentation processes as a route to [the Clinoril nucleus]... Although these proposals did not, in fact, lead to a significant new synthesis... this example illustrates the program's ability to induce chemists to think about new approaches to familiar synthetic problems."

Andose, J. D. *et. al. J. Chem. Inf. Comput. Sci.* **1980,** *20*, 88.

Despite good reviews and apparent enthusiasm by Merck and Pfizer, the project was ended and SECS vanished.

SYNCHEM:

Project led by Gelernter at Stanford/Stony Brook, beginning near LHASA's initial publication. SYNCHEM was the first system to due more proper tree-searches (best first searches) to identify commercial starting materials off of the Sigma catalogue. This ran into some problems however, as it would overly prioritize pathways that excised commercial materials early in the retro.



Retinol — **β–ionone**

SYNLMA:

Escapade by P.Y. Johnson's group that tried a different strategy in knowlege base and reasoning construction, but was rendered unusable by lack of tree pruning and efficient search. Dissolved in 1989.

SYNGEN:

Developed by Jim Hendrikson at Brandeis University. The general approach was to build routes around convergent skeletal construction, essentially ignoring functionality using a heuristics approach to prioritize searches around routes that combined fragments of similar molecular weight to maximize convergency.

"Since syntheses generally involve making larger molecules from smaller ones, only the construction steps are truly obligatory. Hence the shortest and most economical synthesis plan should be one with no refunctionalization..."

Hendrickson, J. B. *J. Am. Chem. Soc.* **1977,** *99*, 5439.

While this approach simplified the search space, it naturally struggled in cases where refunctionality was required by the synthesis. While an additional program was planned to fix these problems, it was largely unsuccessful.

IGOR (Interactive Generation of Organic Reactions):

Ivar Ugi's approach using bond-electron and reaction matrices to describe reactions in the place of expert coded reactions. While mostly used for reaction prediction, retrosynthetic capabilities existed that ultimately were limitted by poor search functions and slower calculations on matrices as apposed to those on character strings.

WODCA (Workbench for the Organization of Data for Chemical Applications):

Johann Gasteiger's program using heuristically defined "breakable bonds" to generate reactions, giving smaller trees than IGOR. Ran best first searches using calculated bond enthalpies to guide retrosyntheses.

SST (Starting Material Selection Strategies):

Pattern recognition approach to tie target structure to similar starting materials from the Aldrich catalogue, essentially only suggests starting materials.

SESAM (The Search for Starting Material):

Similar to SST, identifies skeletal overlap between target and potential starting materials, although ignores functionality.

CHIRON:

Hanessian's strategy of identifying chiral substructures within target molecules that are present in commercially available materials. Again, like SST and SESAM finds its strengths in identifying starting materials and less so in producing full synthetic routes.

SYNSUP-MB/KOSP/LILITH/TRESOR:

Over the years, the graveyard of software grew quite large...

Todd, M. H. *Chem. Soc. Rev.* **2005,** *34*, 247.

Szymkuc, S *et. al. Angew. Chem. Int. Ed.* **2016,** *55,* 5904.

Hanessian, S. *Curr. Opin. Drug Discov. & Devel.* **2005,** *8*, 798.

Trees, Networks, and Searches

Tucker Huffman

A Deeper Look at SYNCHEM's Tree-Searching:

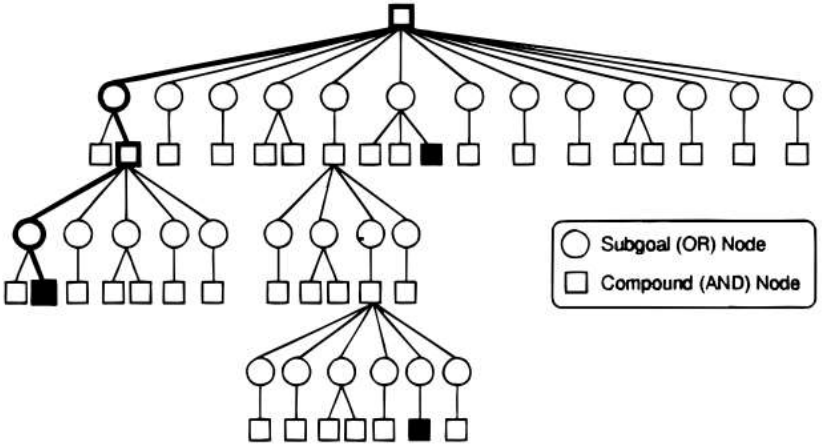


Figure 1. A problem solving graph (PSG) with the current best pathway indicated with bold lines.

Updated versions of SYNCHEM made improvements to the way that it navigated the synthetic tree it generated (referred to as a problem solving graph). While still following the best first search approach, the data structure separates out compounds (squares) from reactions (circles) and the search algorithm proceeds with a best-first search of the reactions and a worst-first search of the compounds. This is to allow the software to "fail fast" and more quickly exclude disconnections to intractable intermediates.

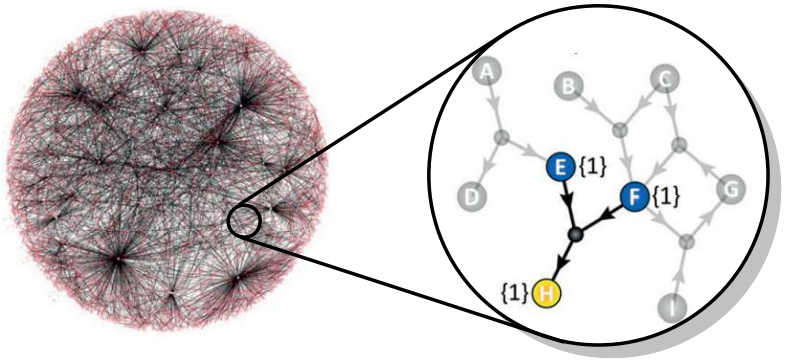
Selection:
-The algorithm finds the reaction pathway with the highest current score, then explores the compound beneath it with the worst score. The search works its way down the tree until it finds a compound that hasn't been expanded yet.

Expansion:
-The knowledge base is exhaustively screened for possible reactions leading to the selected target, expanding the tree. The reactions and compounds are then scored heuristically (e.g. Baldwin's rules for reactions, Bredt's rules for compounds.)

Update:
-The scores of nodes higher up the tree are updated using the information from the new expansion and the algorithm is then repeated from the top.

Network of Organic Chemistry (NOC) and Chematica:

In the early 2000's, Bartosz Grzybowski and his group began working on building a large network of known chemical reactions, essentially tying the chemistry literature into a giant web.



A 5500 node portion of Grzybowski's NOC, with a diagram showing the underlying structure of the system. The black dots represent reactions, the labelled circles are compounds, and the arrows show their connections in the synthetic (forward) direction.

NOC Searches:

For sufficiently short searches (e.g. 3 steps deep) Chematica runs a simple breadth first search to exhaustively examine the options and then picks the path for which the user specified cost function (e.g. sum of costs of starting materials on Sigma catalogue) is minimized.

For longer searches, the software rapidly calculates optimal short pathways for many compounds, thus building a "library of end-games" that will allow it to truncate searches whenever it reaches specific intermediates. This approach borrows off of a chess computer strategy for reducing search depth.

The NOC only gives information on known materials, and can only suggest the exact routes that were carried out in the literature or chimeric routes that minimize cost/steps/etc.

Krebsbach, D *et. al. J. Chem. Inf. Comput. Sci.* **1998,** *38,* 595.

Szymkuc, S *et. al. Angew. Chem. Int. Ed.* **2016,** *55,* 5904.

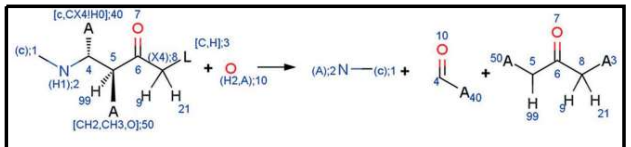
Syntaurus: The Expert Expert-System

Tucker Huffman

A Modernerized Retrosynthesis Tool:

Syntaurus, Grzybowski's *de novo* retrosynthetic design software within the Chematica package, largely borrows the overall strategy of earlier software (manually coded reations, heuristics guided searches) with major improvements and a more thorough execution.

Manually Coded Reactions:



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

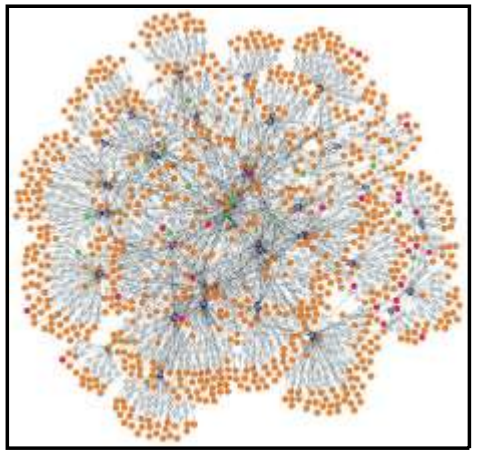
Like many of its predecessors, Syntaurus takes the approach of manully encoding reactions as is shown above. The general transform as well as protection requirements, functional group incompatibilities, reaction conditions, and references are all manually encoded into a database containing ~20,000 reactions.

Customizeable Heurisics

The software allows the user to choose from a set of options what herustics to use and what weighting to give each one. Options include pentalties for protection groups and preferences towards disconnecting rings and stereocenters. This allows users to tune the seach strategy towards the desired synthesis.

Searches:

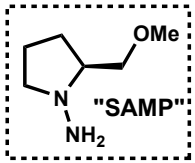
Syntaurus can run either manual, step-by-step searches or fully automatic searches. In manual searches, the possible disconnections are ranked according to the set heuristics and the user navigates in the retrosynthetic reaction one step at a time. In automatic searches, a SYNCHEM like algorithm engages in a best first search but reverts back to the top of the tree when a given route become more costly than the higher level alternatives.

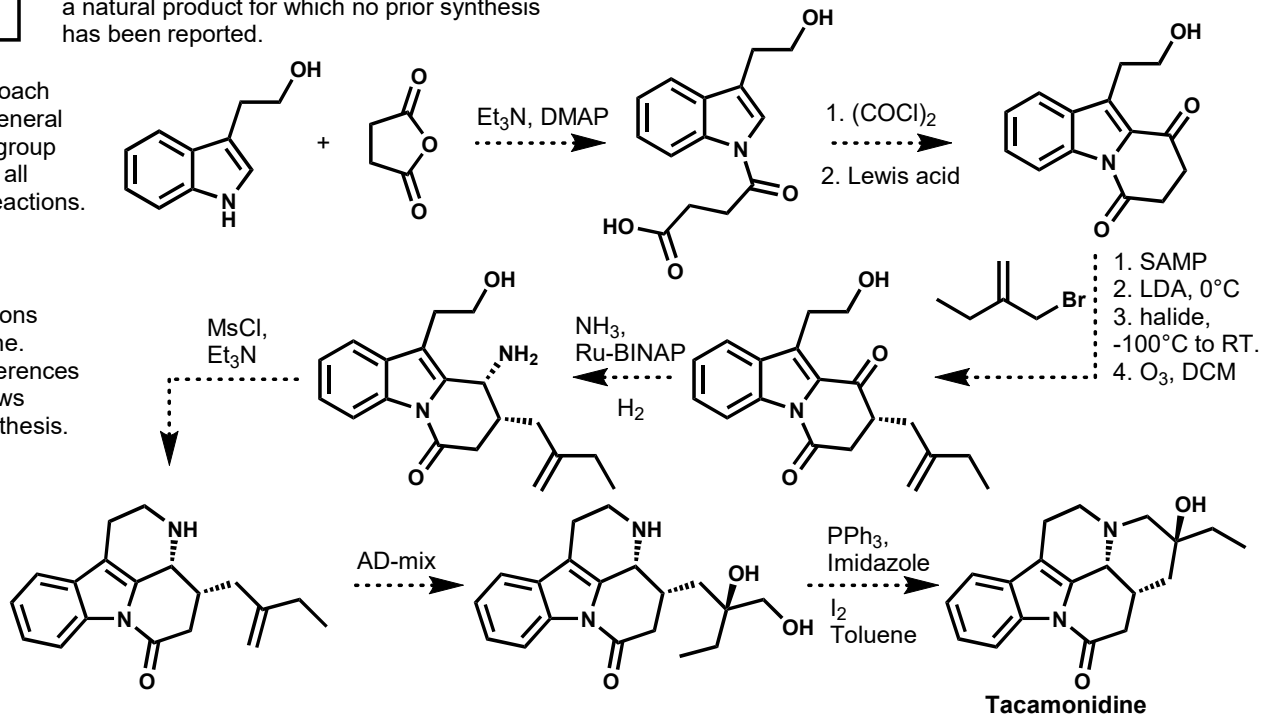


Example of the branches explored during an automatic search over 2 min. with 35 expansions.

Tacamonidine test:

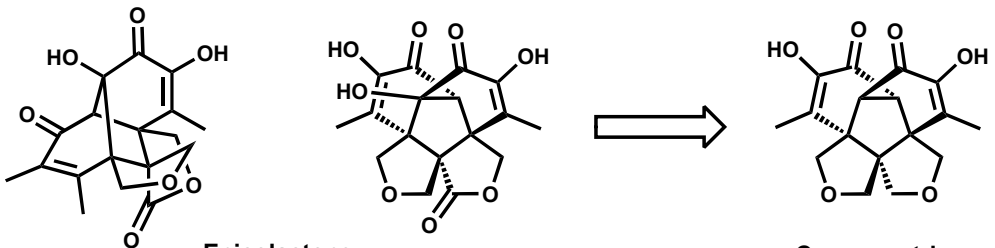
Grzybowski tested the software with a fully automatic search towards tacamonidine, a natural product for which no prior synthesis has been reported.





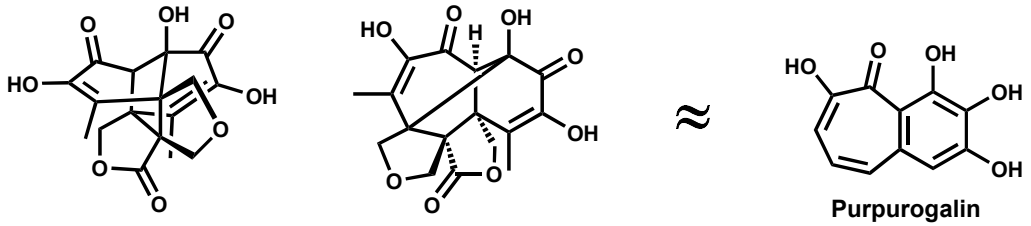
Epicolactone: Syntaurus vs. Trauner

Tucker Huffman

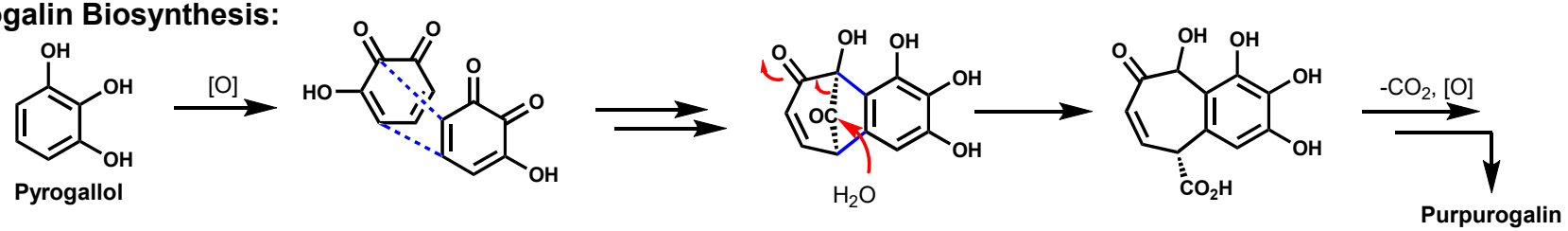


Careful analysis of the structure of epicolactone allowed Trauner and coworkers to draw an analogy to purpurogalin, a natural product with an established biosynthesis via an oxidative coupling

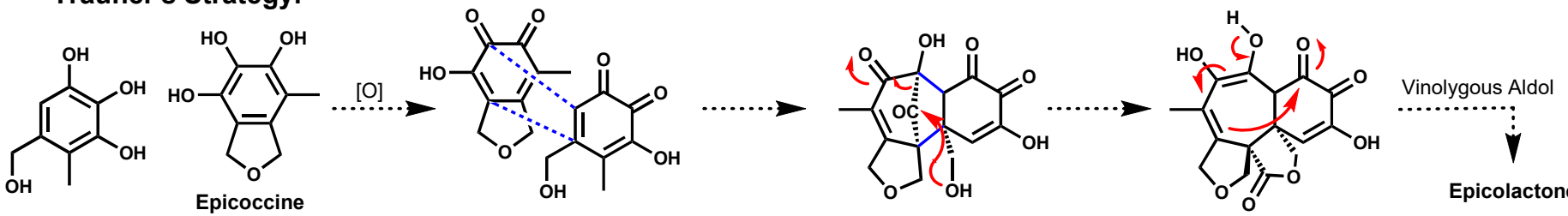
"The systematic disconnection of bonds, however, usually leads to an exponential increase in potential precursor molecules and reactions. These are difficult to evaluate, even in an age of fast computation and advanced theory. An alternative approach involves pattern recognition. This does not require a systematic procedure and is especially attuned to the human brain... Although humans have difficulties with scanning through a database of millions of reactions within seconds, they are very good at drawing analogies and recognizing similarities..."



Purpurogalin Biosynthesis:



Trauner's Strategy:



Ellerbrock, P. *et. al. Nat. Chem.* **2015,** *7*, 879.

Trauner and coworkers had previously hypothesized this epicolactone cascade during the total synthesis of related natural product dibefurin (see: Ellerbrock, P. *et. al. Angew. Chem. Int. Ed.* **2014,** *53*, 13414.)

Some time later, at Syntaurus Headquarters...

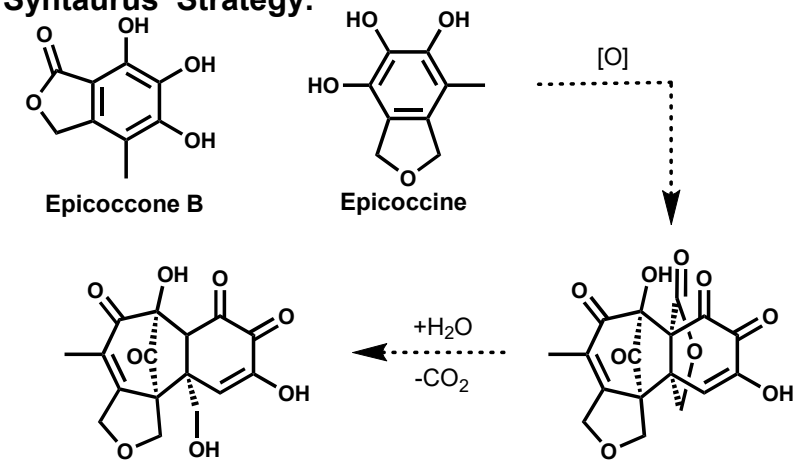
"...we illustrate one such example, suggested to us by Prof. Dirk Trauner, of an Epicoccum nigrum metabolite called epicolactone, which was isolated only in 2012 and for which only a plausible biosynthetic pathway (though not an actual total synthesis) has been published."

Szymkuc, S *et. al. Angew. Chem. Int. Ed.* **2016,** *55,* 5904.

Epicolactone: Syntaurus vs. Trauner

Tucker Huffman

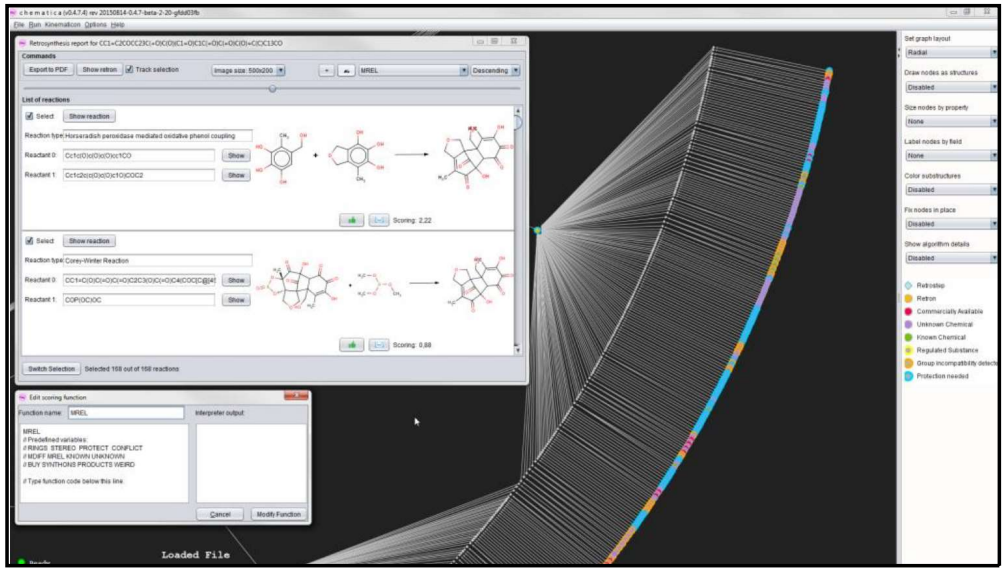
Syntaurus' Strategy:



Using Syntaurus under manual control, a Syntaurus user was able to find a route using the same coupling strategy in a few hours.

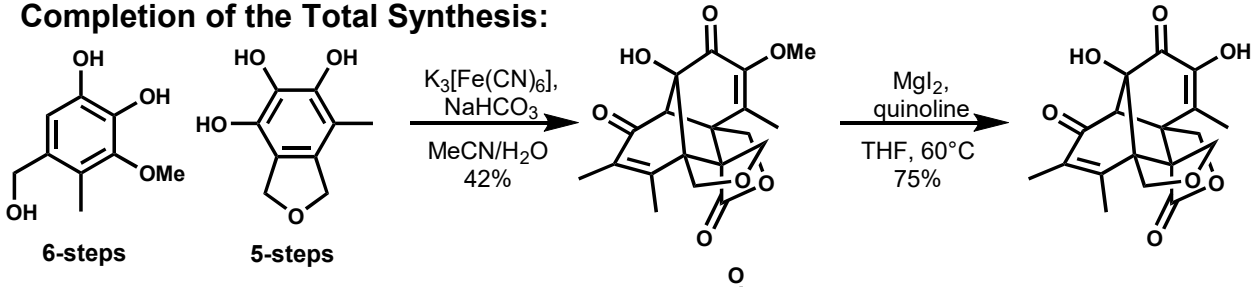
"We wish to stress that while the process required a 'chemically savvy user,' it was 'blind' as we did not know the 'correct' pathway beforehand – it was only after we completed the synthetic design that we shared our synthetic solution to Prof. Trauner who then disclosed to us his own approach to the problem."

However, the disclosure in the literature of the epicolactone cascade by Trauner 2 years prior makes this claim of supposed independence questionable.

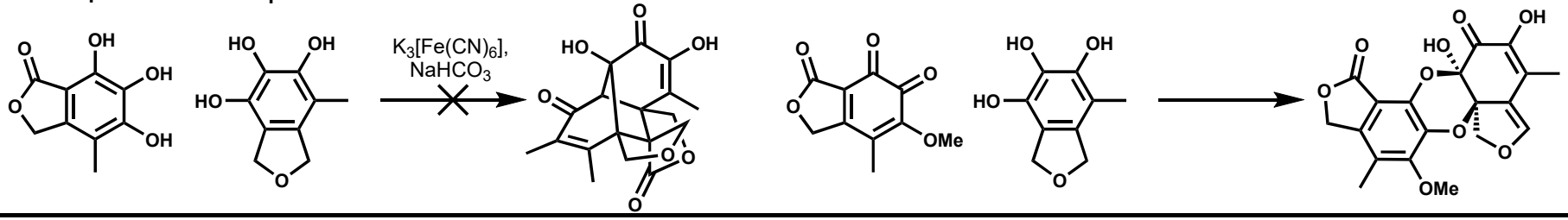


The first path explored by the user actually found the same starting material and strategy as the total synthesis that was completed. However, they reasoned that the two coupling partners wouldn't be sufficiently differentiated electronically to favor heterocoupling over nonspecific homocoupling.

Completion of the Total Synthesis:



The specific route found using Syntaurus failed in practice, although the same general approach that was arrived at by human pattern recongition ultimately was able to result in the synthesis of epicolactone.



Szymkuc, S *et. al. Angew. Chem. Int. Ed.* **2016,** *55,* 5904.
Ellerbrock, P. *et. al. Nat. Chem.* **2015,** *7*, 879.

Modern Approaches: Machine Extraction

Tucker Huffman

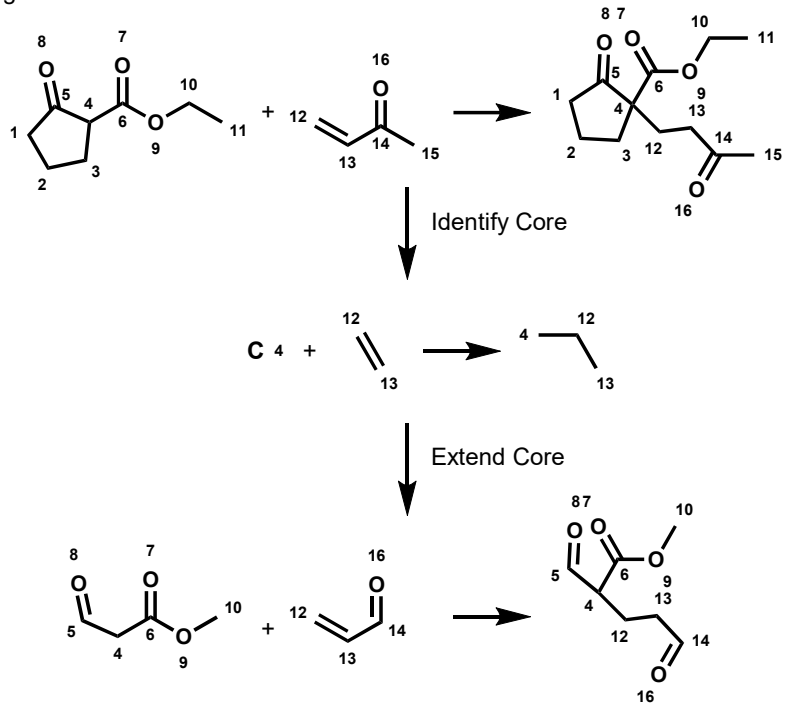
Machine learning approaches attempt to solve the reaction identification problem by extracting reactions from databases rather than laborious manual encoding of reactions into knowledge bases.

Early Work: SYNCHEM and KOSP

-Work was done towards expanding SYNCHEM's knowledge base by extracting reaction information from early databases using existing machine learning software.
-KOSP (Knowledge base-Oriented system for Synthesis Planning) was a retrosynthesis software developed around using database extracted reactons.
-Neither attempt was particularly successful, but developed the idea for later efforts.

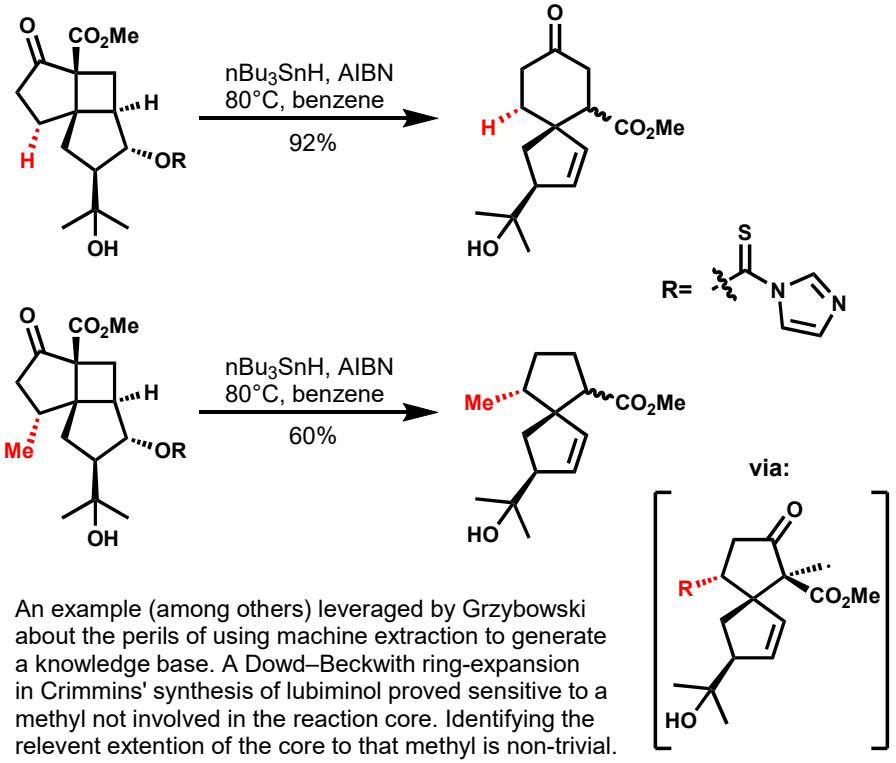
General Idea:

1) Identify what atoms are changing over the course of a reaction to define the reaction core. Requires accurate mapping of atoms between starting material and products.
2) Extend the reacion core to include required nearby functionality and ideally exclude any irrelevant passenger groups.
3) Group extended reaction cores from several reactions together to generate a generalized reaction.



Law, J. *et. al. J. Chem. Inf. Model.* **2009,** *49,* 593.

Problematic Reactions for Machine Learning:



An example (among others) leveraged by Grzybowski about the perils of using machine extraction to generate a knowledge base. A Dowd–Beckwith ring-expansion in Crimmins' synthesis of lubiminol proved sensitive to a methyl not involved in the reaction core. Identifying the relevent extention of the core to that methyl is non-trivial.

Crimmins, M. T. *et. al. J. Am. Chem. Soc.* **1998,** *120*, 1747.
A review with similarly problematic cases in total synthesis:
Sierra, M. A. *et. al. Angew. Chem. Int. Ed.* **2000,** *39*, 1538.

While this does point out a challenge in machine extracting reactions from databases, it is also indicates a more general limitation in the entire field of computer aided retrosynthesis: Many reactions, including several synthetically useful and powerful transformations, are very hard to generalize and reaction outcomes can be unpredictable. The question, then, is whether or not to neglect these reactions entirely, or to include some things that have a reasonable probability of failing in the lab.

ChemPlanner and IC*Synth*

Tucker Huffman

Two large retrosynthesis software packages exist today that operate using reactions that are machine extracted from databases, ChemPlanner and IC*Synth*. These programs both function strategically very similar, and differ mostly in the reaction extraction process, specifically in core extension. They also use separate databases for reaction extraction.

Route Designer/ARChem/ChemPlanner:

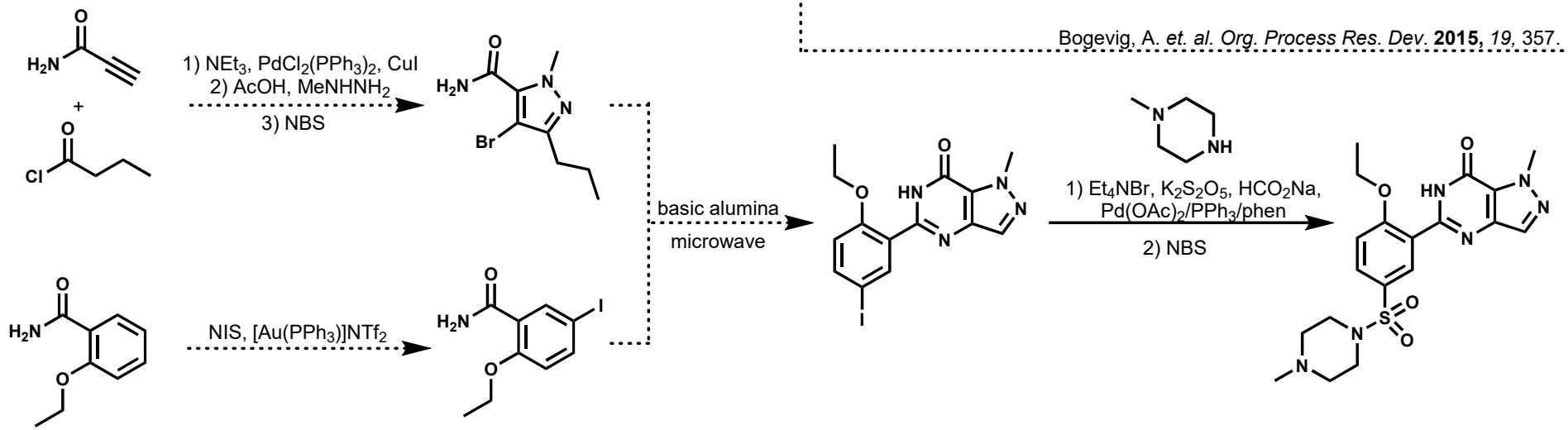
This project was originally produced under the name Route Designer, then was renamed to ARChem. Simbiosys, the company that produced ARChem, was aquired in 2014 by Wiley, where the software appears to have been merged into ChemPlanner.

ARChem uses a series of rules to determine how to extend the reaction core, such as automatically extending to contain whole aromatic systems but truncating after consecutive sp^3 carbons.

Law, J. *et. al. J. Chem. Inf. Model.* **2009,** *49,* 593.

ChemPlanner Case Study: Sildenafil

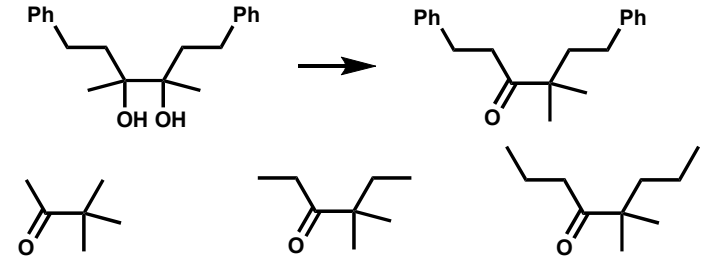
ChemPlanner was tested to see if it could provide an improved route towards sildenafil. Like most tests, there is no indication that it was ever attempted in the lab, but it did produce a route. (Last 2 steps are known)



Wiley ChemPlanner. Synthesis of Sidenafil Case Study. "http://www.chemplanner.com/resource-center/" (Accessed Oct. 1, 2017).

IC*Synth*:

IC*Synth* conducts core extension in a series of 3 "shells" of various size. Each increse in shell extends the core incrementally by all adjacent atoms. This system allows the user to choose the precision of the transforms used in the specific search.



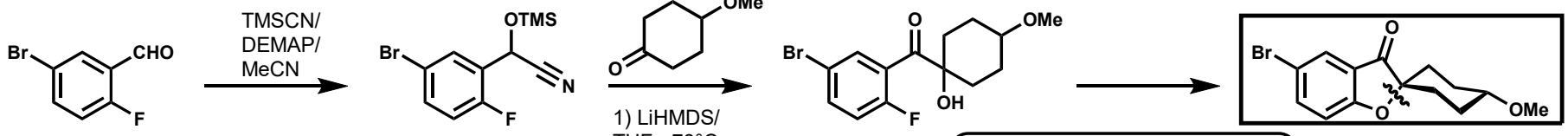
Bogevig, A. *et. al. Org. Process Res. Dev.* **2015,** *19,* 357.

IC*Synth* & AstraZeneca

Tucker Huffman

Similarly to SECS being taken into Merck in the 1970's, IC*Synth* was implemented at AstraZeneca, but in this case actully resulting in some application.

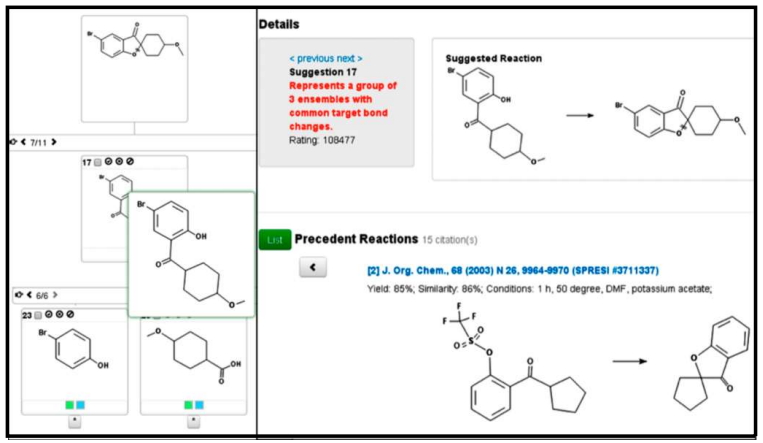
Medicinal chemistry route:



Yields proved to be irreproducible (~10% overall) TMSCN and LiHMDS undesireable on scale, and tedious chromatography was required.

IC*SYNTH* identified the possibility of an unanticipated disconnection and pulled relevant literature precedent justifying the approach.

Developed Route based on IC*SYNTH Prediction:*



"The major conclusion is that in all cases the computer has been able to indentify new ideas for defining routes to synthetic targets that go beyond known chemist-derived suggestions. However, we emphasize that this result in no way detracts from the continuing central importance of the chemist, both in their own generation of new route options (which may well go beyond what the computer suggests) and in evaluation of the computer's suggestions. In fact, we also find that a computer-derived new idea can lead the open-minded chemist to further new ideas of his/her own."

Bogevig, A. *et. al. Org. Process Res. Dev.* **2015,** *19,* 357.

Modern Approaches: Deep Learning

Tucker Huffman

What is Deep Learning?

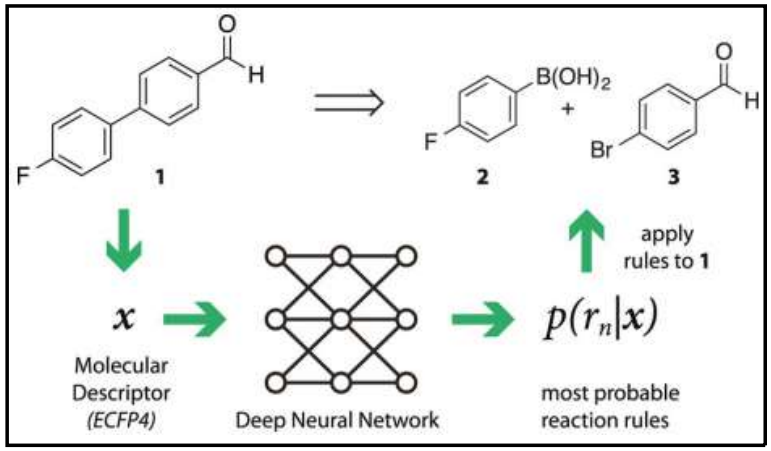
Deep learning software attempts to **learn** how to solve a given problem rather than be explicitly programmed how to do so. The program is initially trained on a data set where it is allowed to modify the way it processes the input data to allow it to accurately produce a right answer. This method allows software to make predictions and judgements where the relationships in the data are very complicated.

Can a computer learn chemical intuition?

The chemist's intuition comes from years of observations and associations between observed results and the conditions that led to them. In principle, given the same set of information, a computer could pick out the same set of intuitive rules as a human, although this is not trivial. A large complication is the absence of all of the failed reactions that have corrected overgeneralizations.

Learning to apply reaction rules:

A neural network approach has been taken to build a system that more accurately applies reaction rules generated by other means. The system was trained on database reactions labelled with the reaction rule used to carry out the transform. The resulting neural network was able to assign the probability that a given reaction rule would be applicable to a queried target molecule.

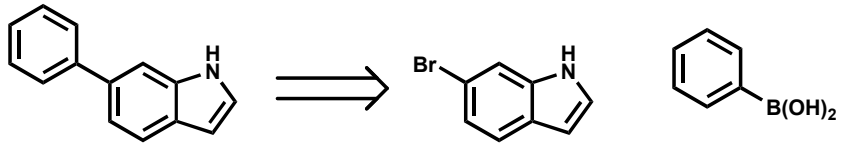


Segler, M. H. S. *et. al. Chem. Eur. J.* **2017,** *23,* 5966.

Predicting Correct Retrosyntheses:

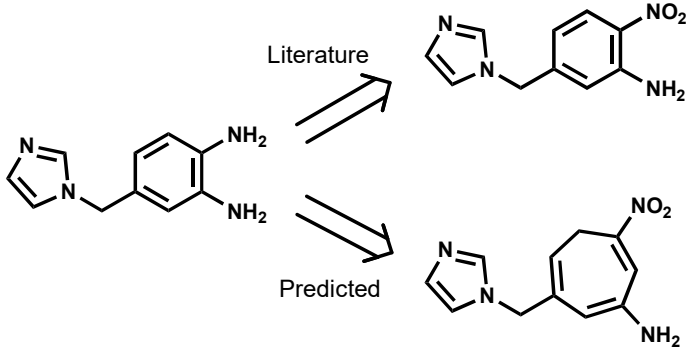
One recent approach attempted to train a neural network to predict the starting materials from a product using the SMILES codes for the materials alone. This approach divorces itself entirely from any preconception of chemistry and lets the software derive everything from the database used. While the neural network was able to produce starting materials from a given product, there were frequently errors arrising from the fragility in the SMILES codes.

Some reactions returned invalid SMILES codes, the rate ranging from 12 to 30% depending on search parameters.



Literature: Brc1ccc2cc[nH]c2c1.OB(O)c1ccccc1
Prediction: Brc1ccc2cc[nH]c2c2.OB(O)c1ccccc1

Some errors came from returning chemically nonsensical reactions.



Liu, B. *et. al. ACS Cent. Sci.* [Online early access] DOI: 10.1021/acscentsci.7b00303

"AlphaChem"

Tucker Huffman

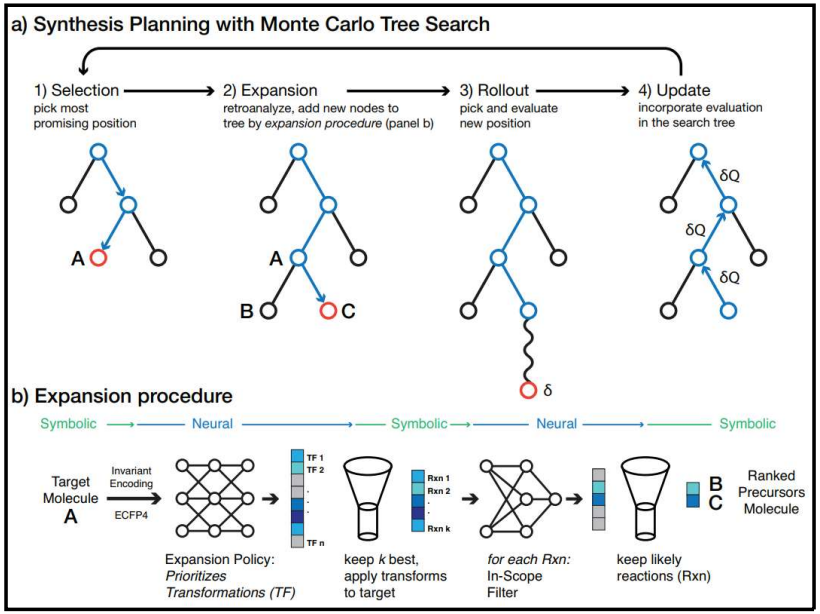
Monte Carlo Tree Search:

Google DeepMind's AlphaGo software uses a modified version of a Monte Carlo tree search (MCTS) to explore the vast combinatorial space of moves ahead of it. Mark Waller's group adapted this strategy towards making a retrosynthesis software.

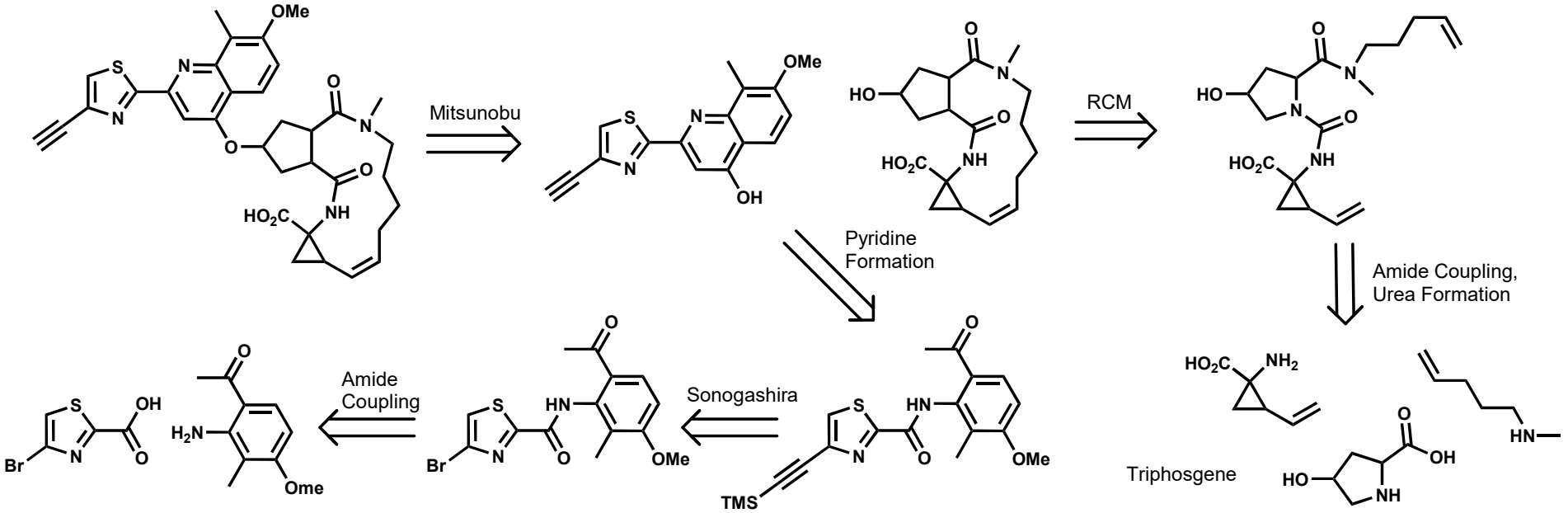
As opposed to the modified best first search strategies like those employed by SYNCHEM and Syntaurus, a MCTS introduces a randomization component andpursues a route to completion before updating. Compared to Grzybowski's search algorithm, the MCTS evaluates fewer molecules per seconds but successfully completes retrosyntheses at a higher rate.

The system was shown to be capable of solving over 80% of the test molecules given (largely drug like molecules, no treatment of stereochemistry) within 5 seconds. With an expanded time frame of 300 seconds ~95% could be solved. The produced routes couldn't be distinguished from the literature reports in blind tests.

This system is entirely driven directly from database reactions. All reaction identification, scoring, and searching is done via neural policy networks. While this removes the need of any manual coding or heuristics, it also makes it an automatic process by necessity.



Representative Retrosynthesis:



Segler, M. H. S. *et. al. arXiv,* **2017,** 1708.04202.

Conclusions

Tucker Huffman

1) Despite advances in hardware and software, computer aided retrosynthesis is still a very hard problem with some major barriers left to overcome.

2) The success of computer aided retrosynthesis programs is hard to accurately assess because of the lack of reported real world useage. Producing chemically reasonable or literature routes is not convincing of general utility, nor is lack of useage necessarily an indication of uselessness.

3) Even if these tools are shown to be useful and are widely implemented, I am skeptical that their benefit would be more than marginal for many applications.