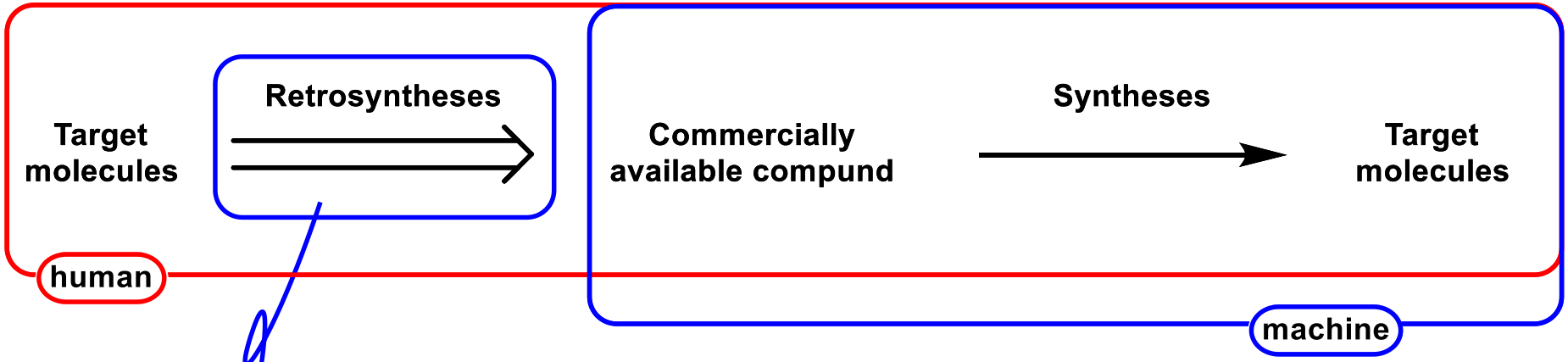
Computer-Assisted Synthetic Planning The End of the Beginning

2016.09.16. Yuki Nakagawa

Szymkuc, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald,P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904.

Human vs. Machine in Synthetic Organic Chemistry



eg. Solid-phase peptide synthesis
HPLC, flow reactor

We are helped by database when we consider retrosyntheses.
However, *"useful"* retrosyntheses programs do not exist yet.

Prof. Grzybowski, B. A. and Chematica

Prof. Bartoszm A. Grzybowski

Profile:
1995: B.S. Yale University, New haven, CT
2000: Ph.D. Harvard University, Cambridge, MA
2000-2003: Post-Doc. Harvard University, Cambridge, MA
2003-2007: Assistant Professor, Department of Chemical and Biological Engineering and Department of Chemistry, Northwestern University, Evanston, IL
2007-2014: Associate Professor at the same university
2009-2014: Director, Non-Equilibrium Energy Research Center at the same univ.
2002-present: Chief Scientific Officer, ProChimia Surfaces, Ltd.
2009-present: President, GSI L.L.C.
2014-present: Distinguished Professor, UNIST, Ulsan

Research area: Nanoscience, Nanomaterials, Chemical networks, Prgrammable reactions

Chematica: A Machine that thinks like a Chemist! (http://chematica.net/)
Synthetic plannning software

- **Network Module**

- **Retrosynthesis Module: Syntaurus**

Network Module

Network module works off of a graph network of about 10 million chemical substances that are "connected" by a similar number of reactions from the chemical literature.

With network algorithms, network module scrutinizes labeling the molecules and reactions with desirable attributes (molecular masses, solubilities, yields, etc.) to allow for user-specified criteria/constraints to be imposed on the search results.

Retrosynthesis Module: Syntaurus

Syntaurus combines ~20,000 reaction rules taught to the computer by expert organic chemists with advanced, chess-like algorithms to score synthetic positions during synthetic planning. Each of the rules accounts fully for the possible substituents, for stereo- and regio-chemistry, for protection group requirements, and for potential reactivity conflicts.

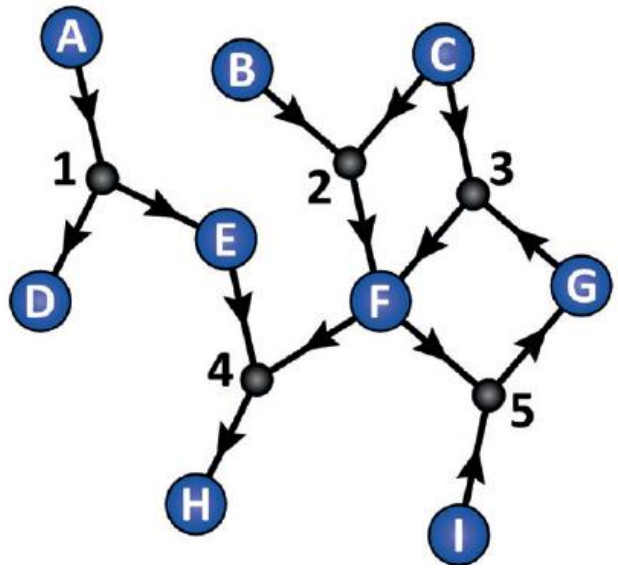
The search algorithms, in turn, codify "chemical intuition" and can intelligently back-track from unpromising synthetic pathways.

Syntaurus can construct hundreds to thousands of synthetic pathways per minute and can rank them according to synthetic viability.

The Network of Organic Chemistry (NOC)

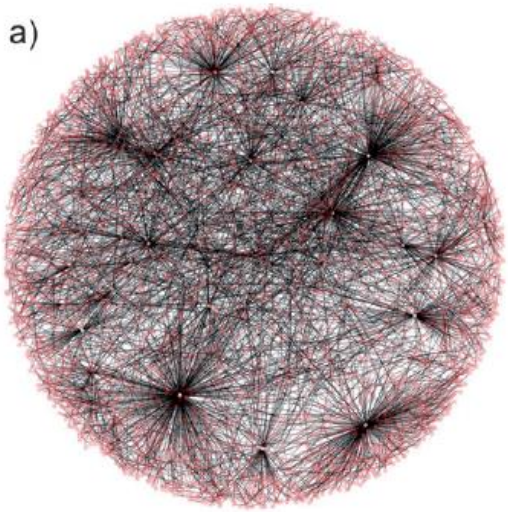
Reactions in databases

$$A \longrightarrow D + E$$
$$B + C \longrightarrow F$$
$$C + G \longrightarrow F$$
$$E + F \longrightarrow H$$
$$F + I \longrightarrow G$$

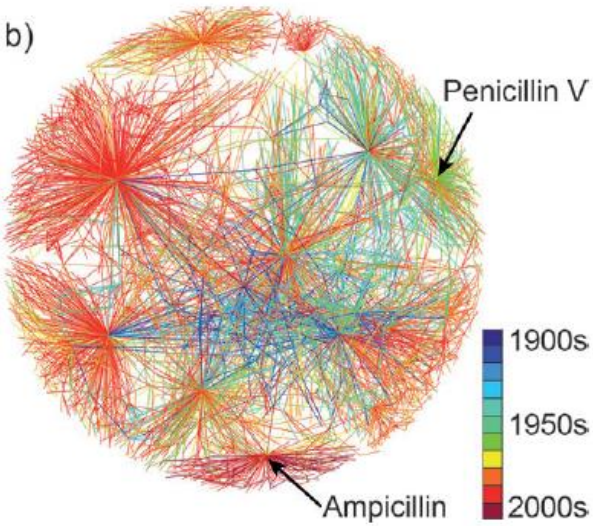


Network of chemical reactions

blue nodes: compounds
black nodes: reaction conditions

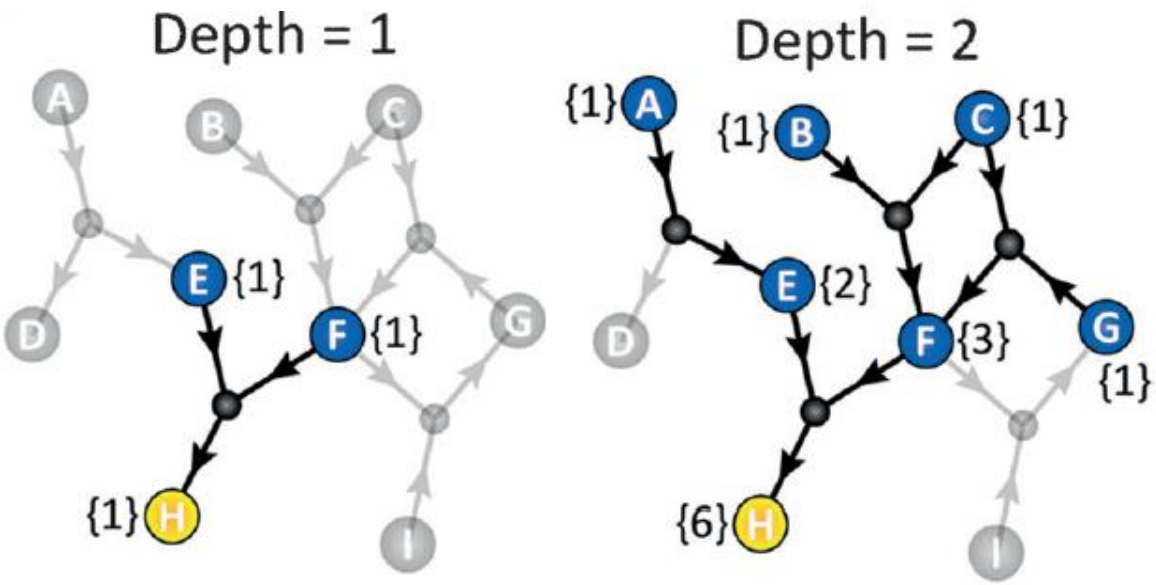


ca. 5500 compounds and reaction conditions (ca. 0.1% of the total) fragment of the NOC



reaction arrows colored by the times these reactions were first reported

Complexity of the Network of Chemistry



Possible reaction route in depth=1

E + F → H

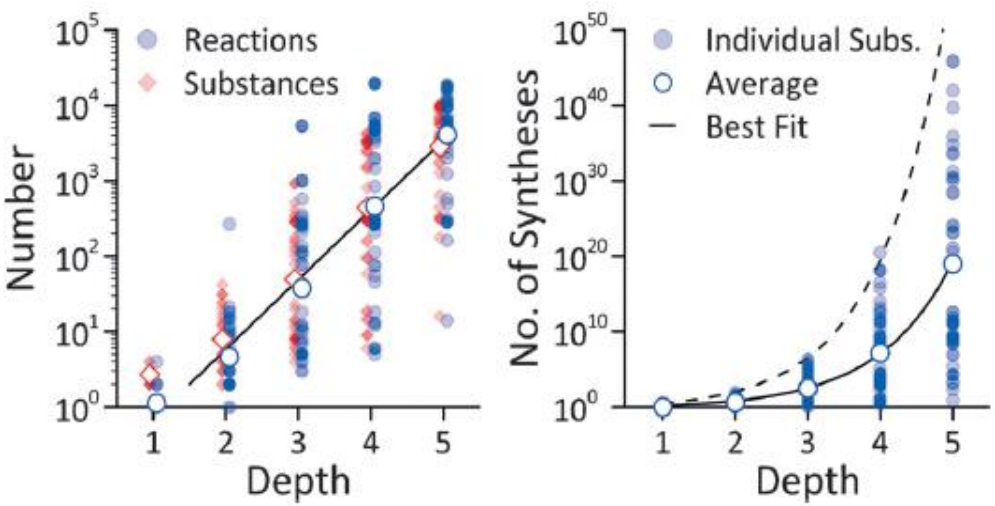
Possible reaction routes in depth=2 (E and F are commercially available)

A → E
E

B + C → F
C + G → F
F

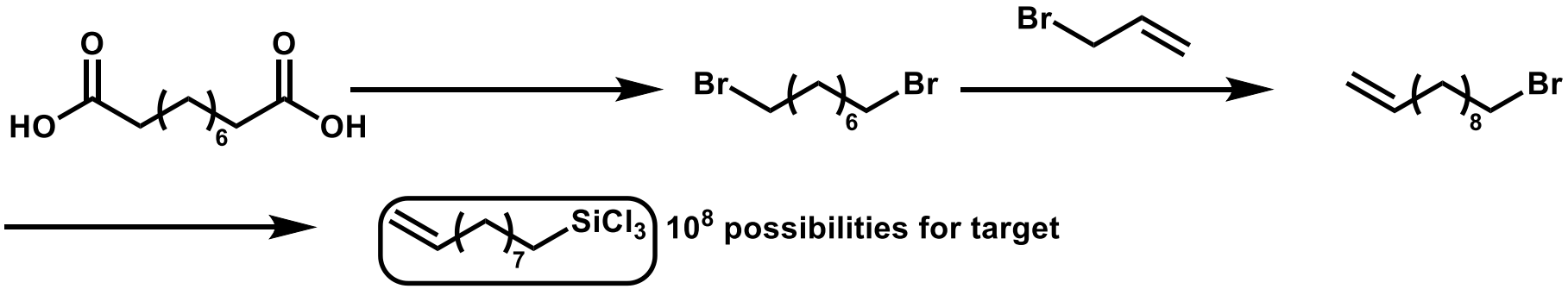
→ H

six possible routes

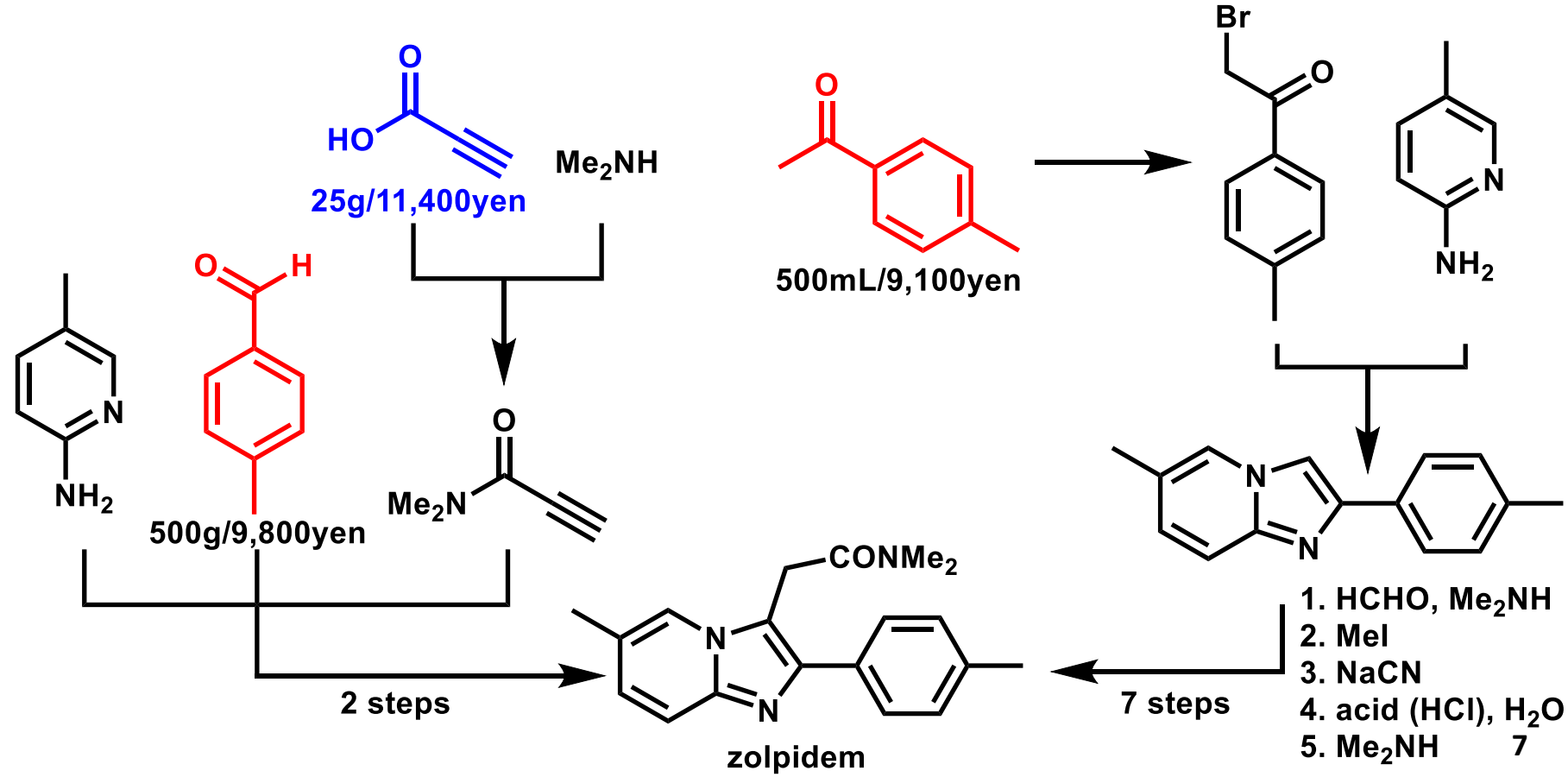


Based on network searches in the vicinity of 51 different target substances

"Optimal" Pathways



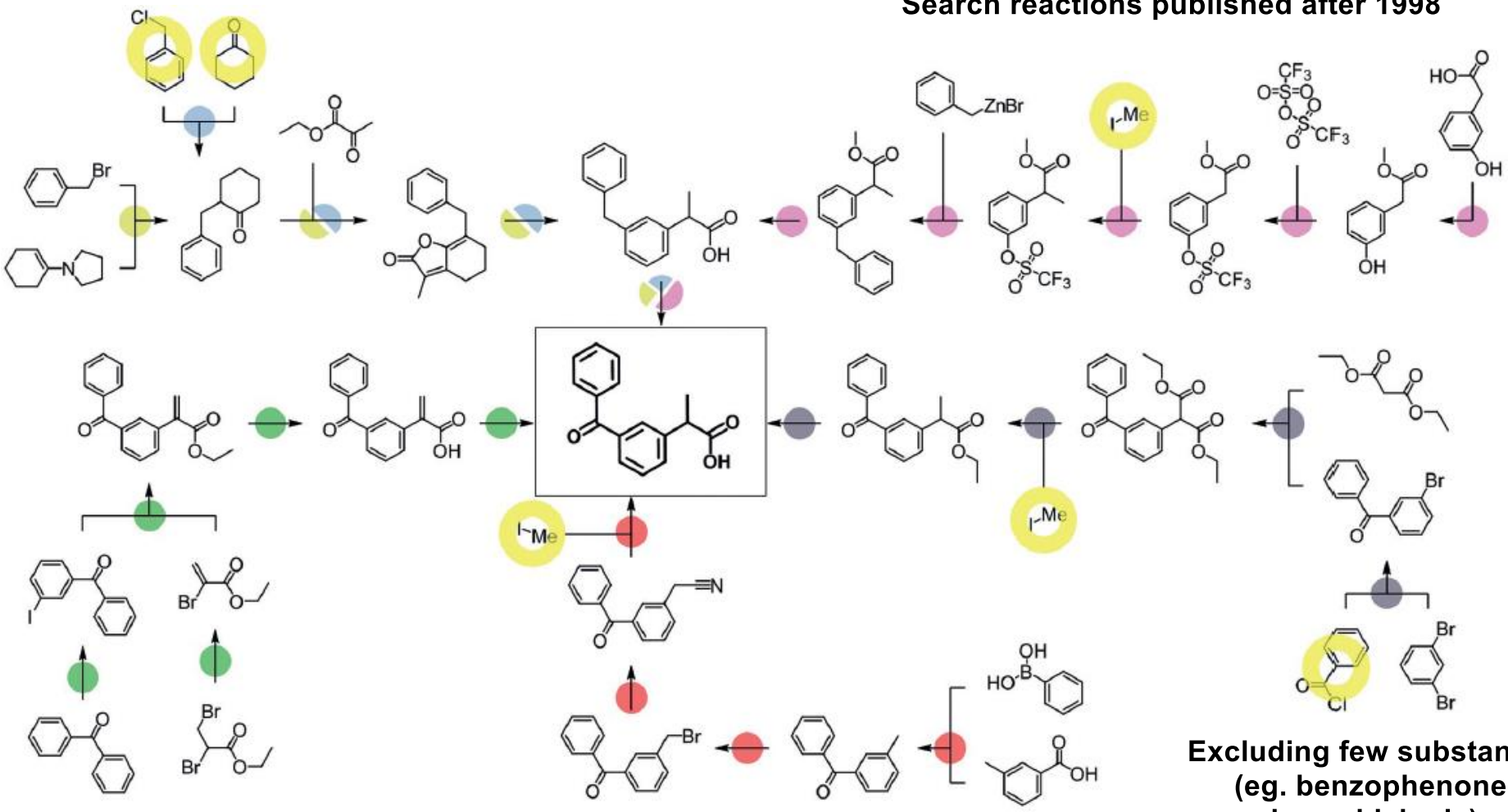
10^8 possibilities for target



Cost-optimized Syntheses of ketoprofen

Priority for atom economy

Search reactions published after 1998

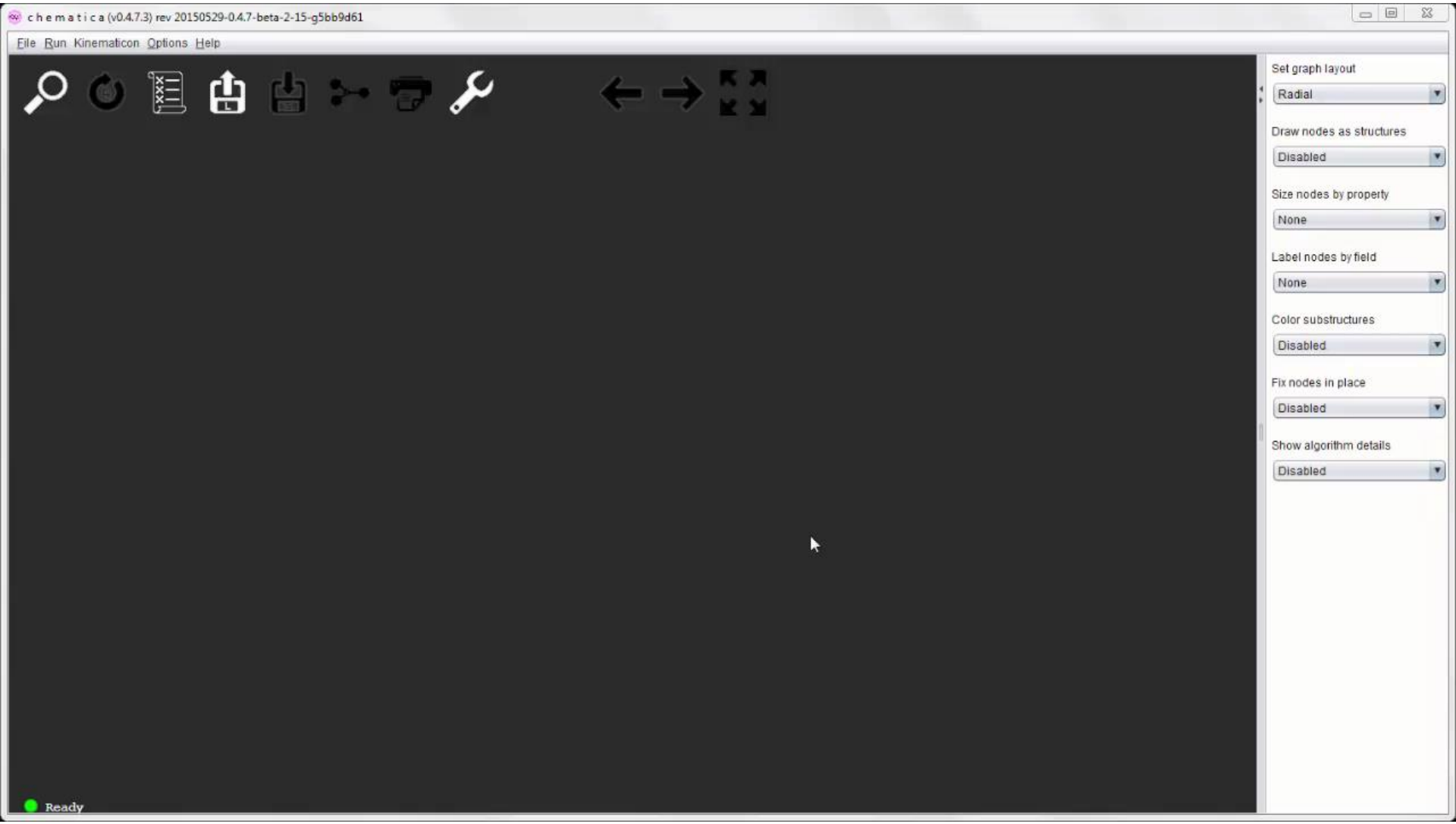


No toxic and regulated compounds

Most cheapest route

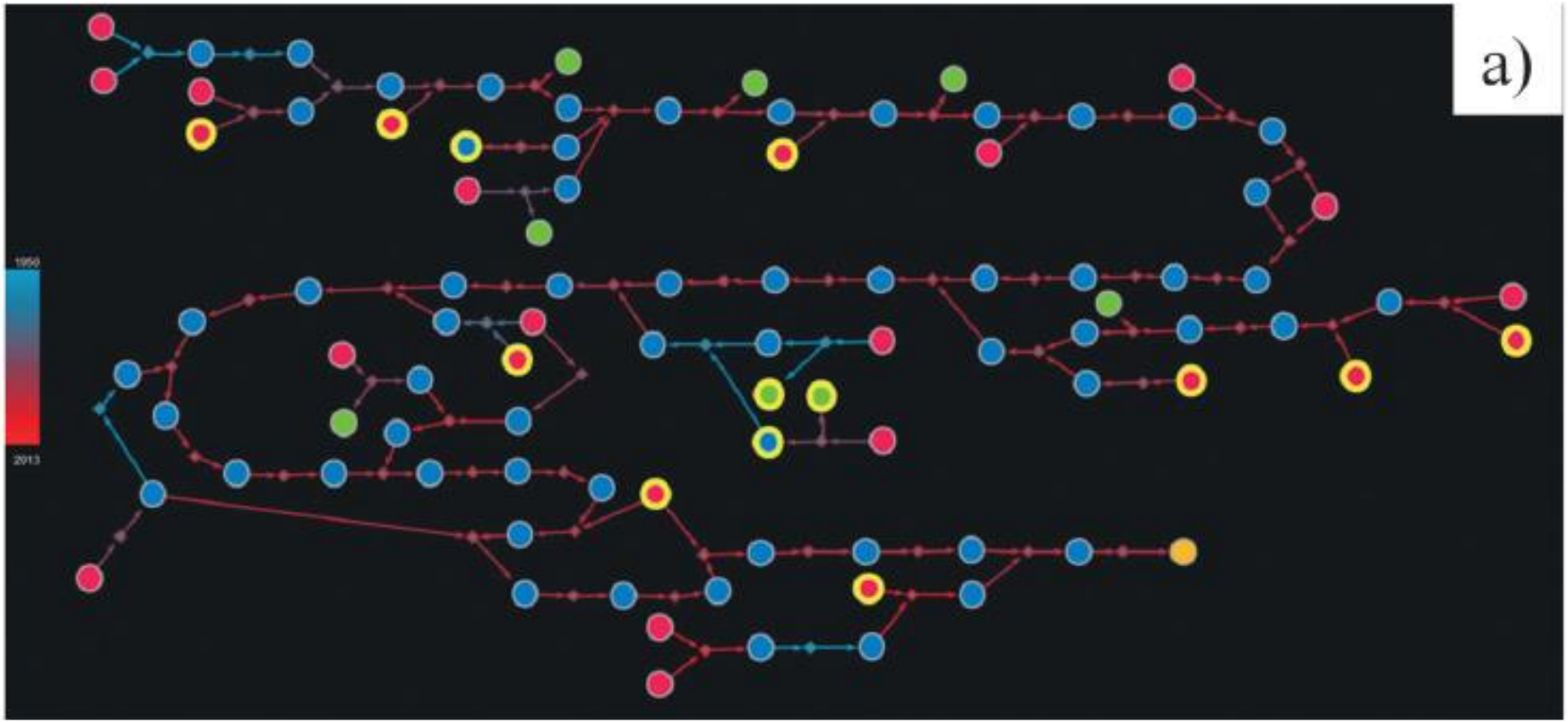
Excluding few substances (eg. benzophenone, benzaldehyde)

Cost-optimized Syntheses of Taxol



345 million sequences of possible steps and over 400 million combinations of participating substances

Cost-optimized Syntheses of Taxol



Node representation of the lowest-cost synthesis of taxol limited to 50 steps.
orange: target, red: commercially availble, blue: intermediates, green: minor/side product, yellow halos: regulated substances

Most of the pathway shown (red arrows) is based on Danishfsky's synthesis from 1995.

Network Module

Network module works off of a graph network of about 10 million chemical substances that are "connected" by a similar number of reactions from the chemical literature.

With network algorithms, network module scrutinizes labeling the molecules and reactions with desirable attributes (molecular masses, solubilities, yields, etc.) to allow for user-specified criteria/constraints to be imposed on the search results.

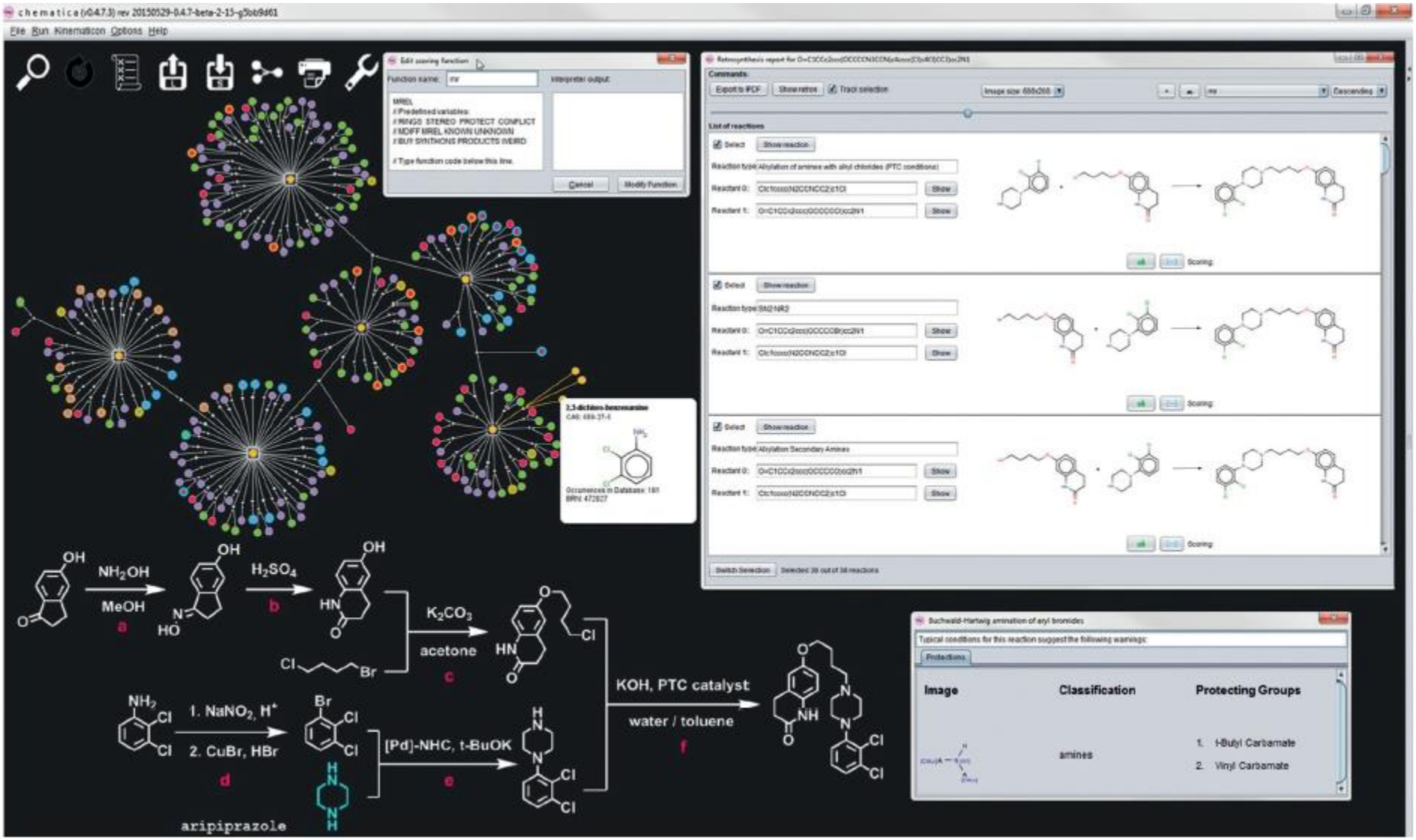
Retrosynthesis Module: Syntaurus

Syntaurus combines ~20,000 reaction rules taught to the computer by expert organic chemists with advanced, chess-like algorithms to score synthetic positions during synthetic planning. Each of the rules accounts fully for the possible substituents, for stereo- and regio-chemistry, for protection group requirements, and for potential reactivity conflicts.

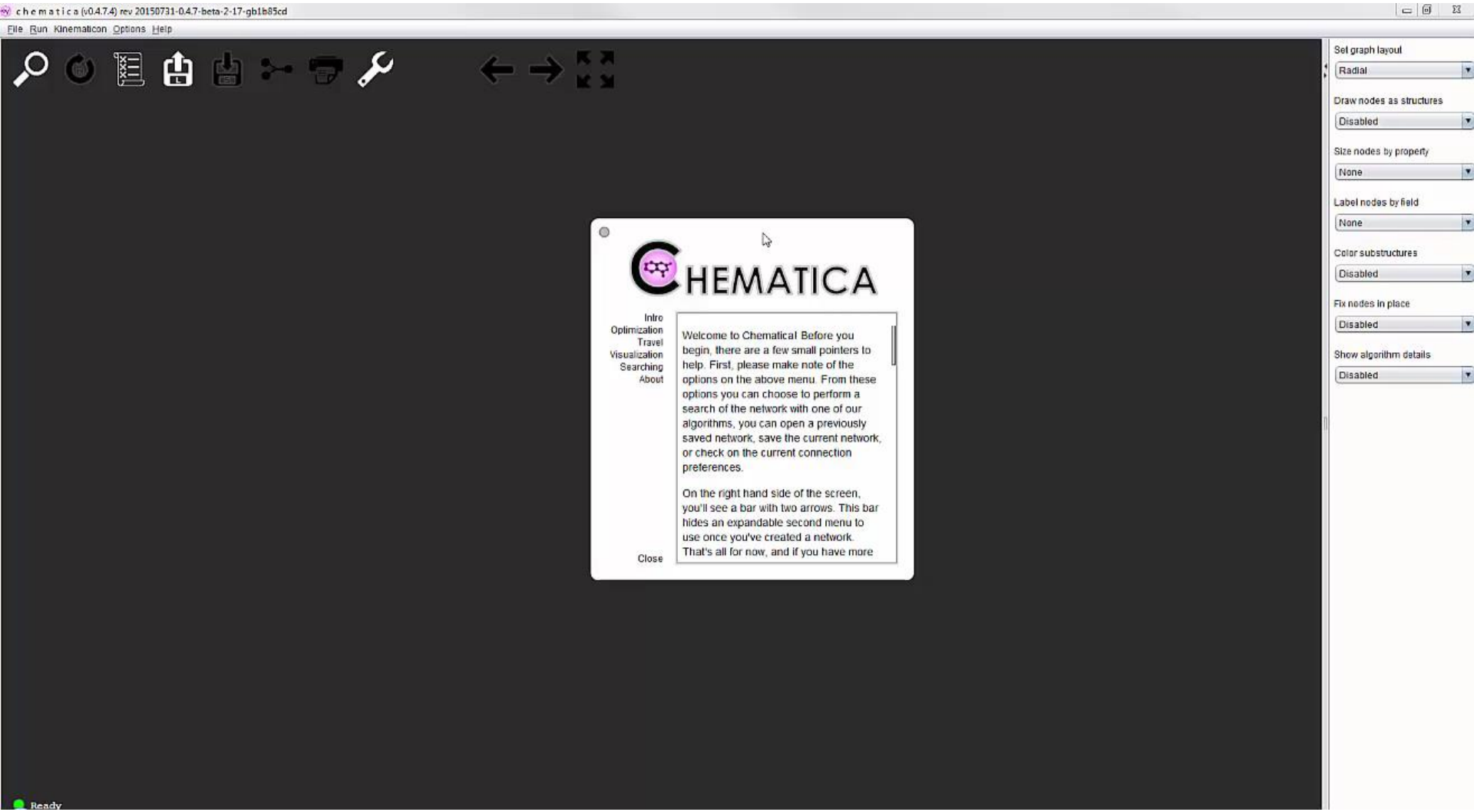
The search algorithms, in turn, codify "chemical intuition" and can intelligently back-track from unpromising synthetic pathways.

Syntaurus can construct hundreds to thousands of synthetic pathways per minute and can rank them according to synthetic viability.

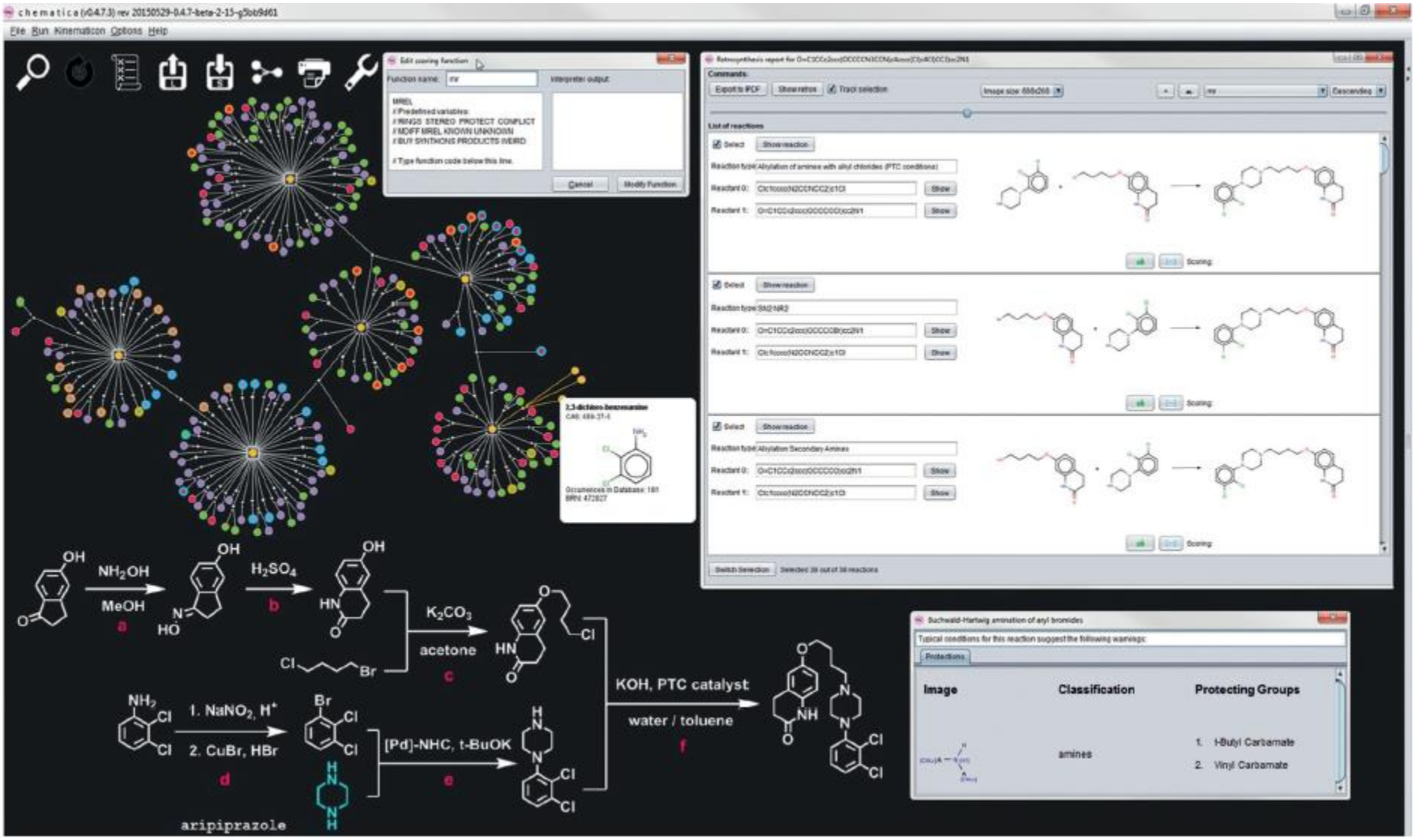
Synthesis Design of Aripiprazole - step by step -



Synthesis Design of Aripiprazole - step by step -



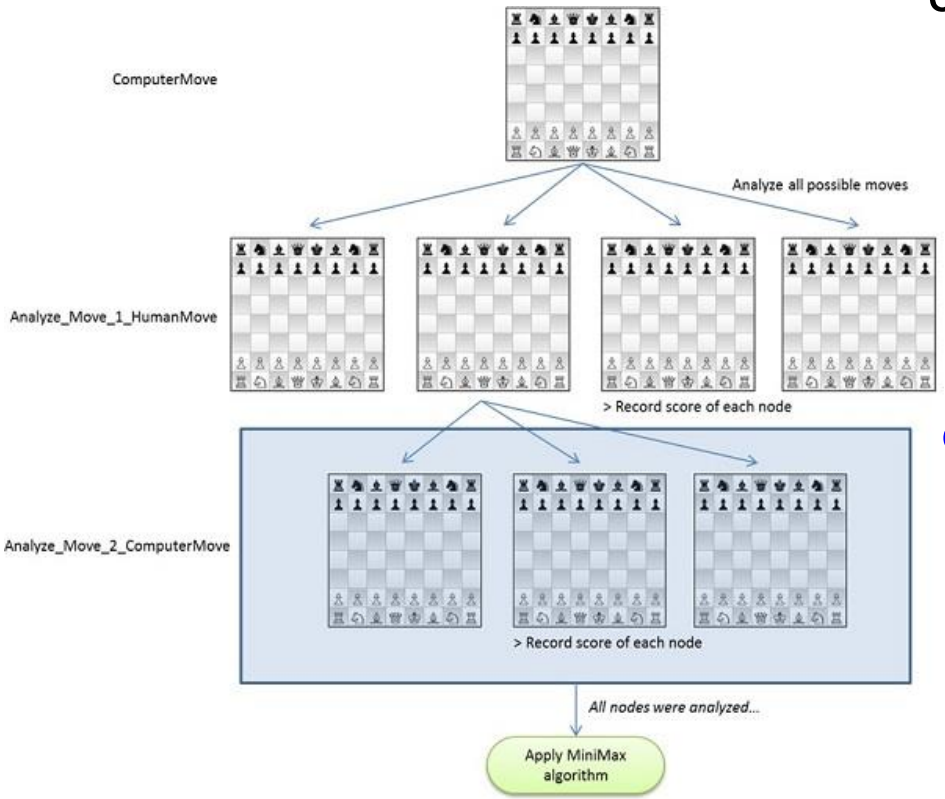
Synthesis Design of Aripiprazole - step by step -



Step by step method is similar to the method of LHASA which programmed by Corey.

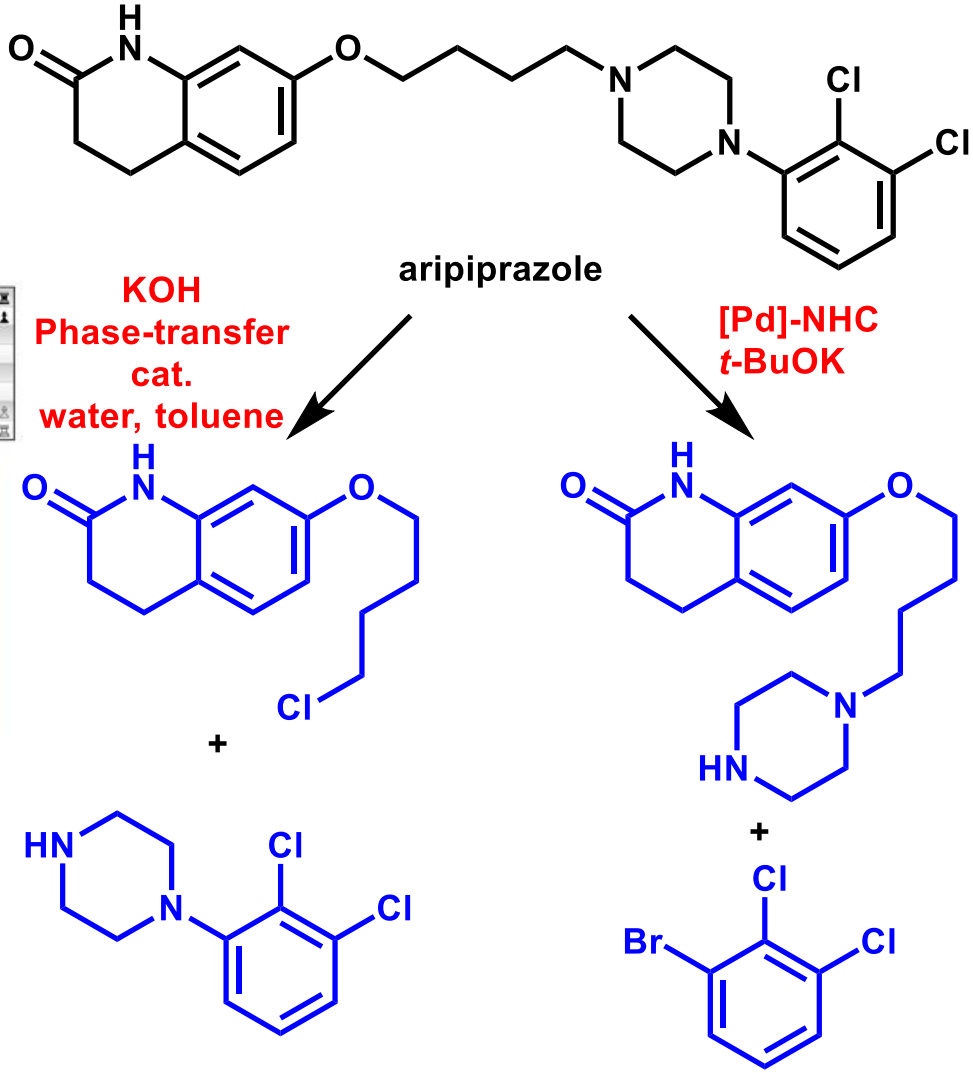
Defining "Synthetic Positions"

Posistions in the chess programs



http://www.codeproject.com/Articles/20736/
C-C-CLI-Micro-Chess-Huo-Chess

Posistions in the synthetic routes



Score both the reactions and the sets of substrates created in each retrosynthetic steps.

Chemical's and Reaction Scoring Functions

Chematica has two scoring functions: the Chemical's Scoring Function (CSF)
the Reaction Scoring Function (RSF)

CSF evaluates structual features of molecules including *Rings, Stereo, Known, Buy and Mass*.

e.g.)
Rings: CSF = RINGS = 2 mean that only one substrate with two rings.

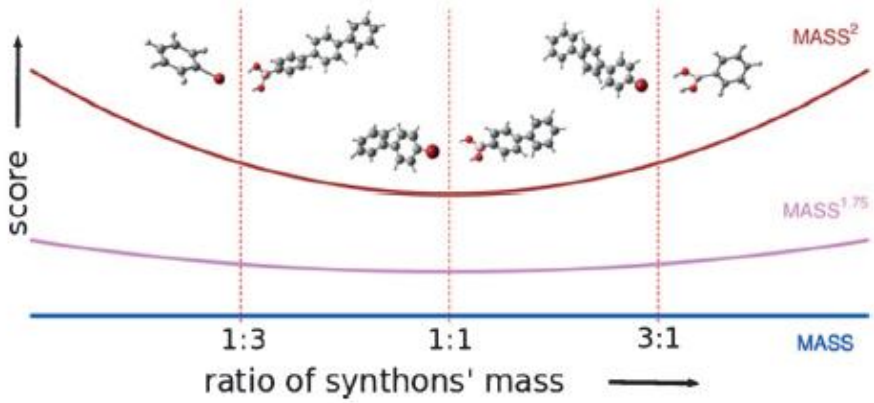
: CSF = RINGS = 1+0=1 mean that there are two substrates, one of which has one ring.

Known: +1 if molecule is not commercially available but known in the NOC
+0 otherwise

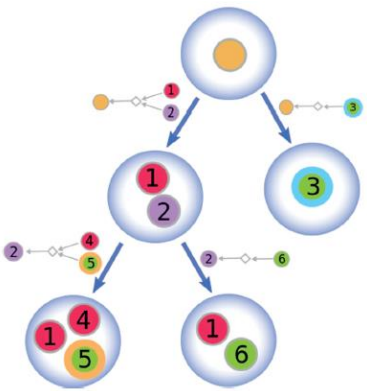
Mass: CSF = MASS = 200 + 200 or 100 + 300 = 400 mean that target of MW 400 is cut into two smaller substrates.
CSF = $MASS^2$ ($200^2 + 200^2 < 300^2 + 100^2$)

RSF assaigns some constant cost of perfoming a reaction step and a combination of *Protect* (a set penalty for each group that needs to be protected) or *Conflict* (penalty for each group incompatibility)

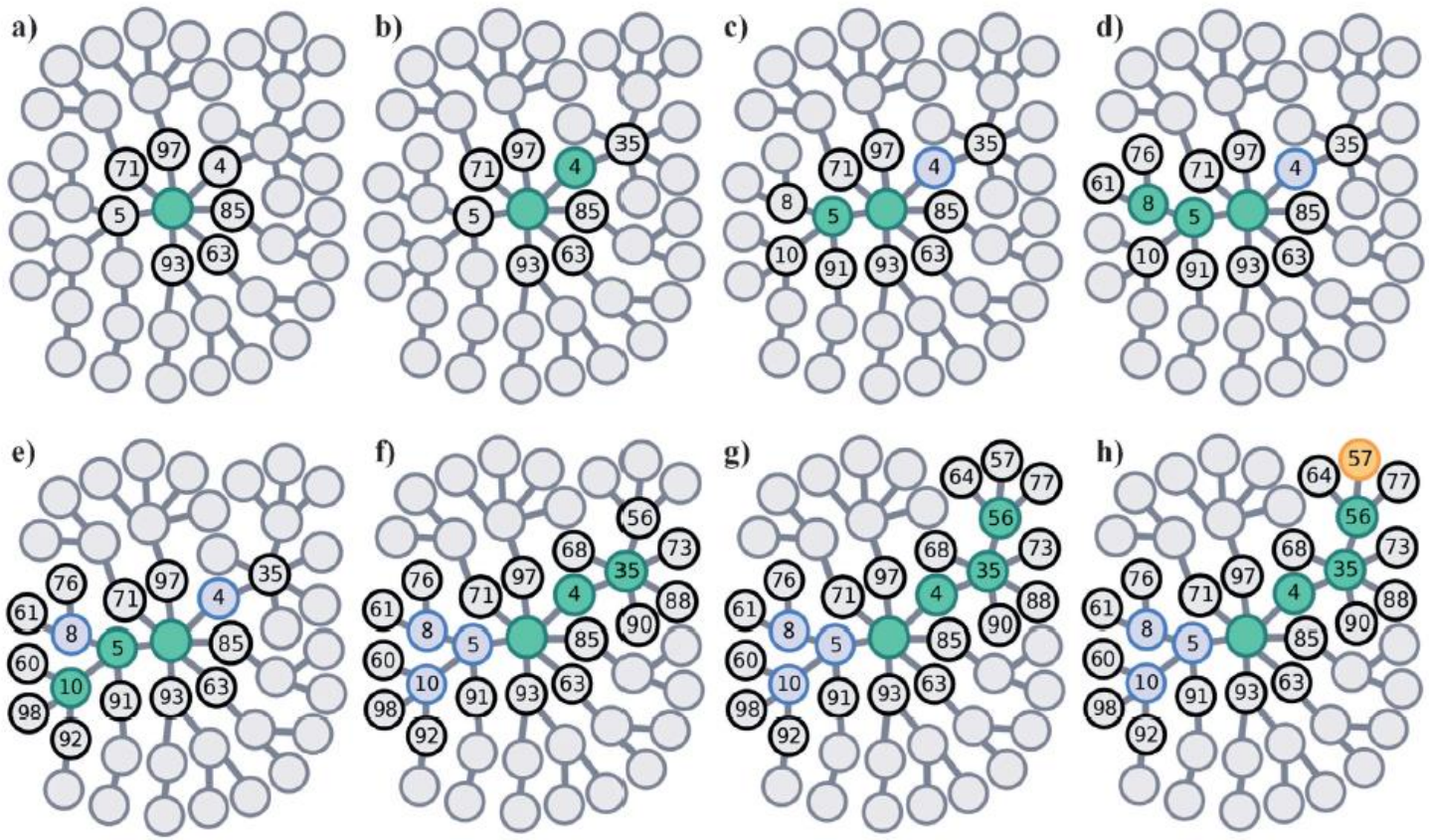
e.g.)
Protect: RSF = 30 + 10000xPROTECT
Protection-Free synthesis



Automated Searches

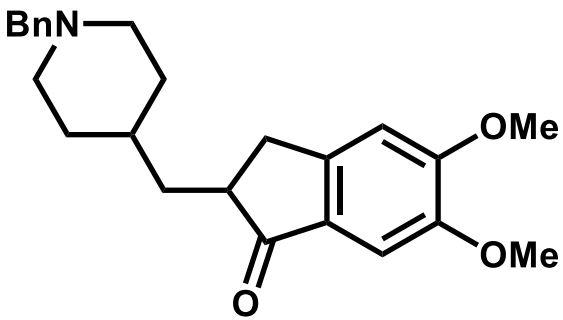


A dual-graph representat-ion underlying Syntaurus' automated searches.



A simplified scheme of syntaurus's retrosynthetic graph-search algorithm. Each node represents a collection of substrates generated in each reaction "move". node 57 can be bought or is known.

Retrosyntheses of Donepezil



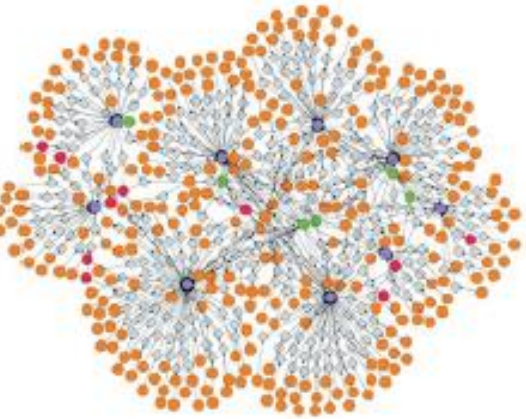
Donepezil

Treatment for alzheimer's disease

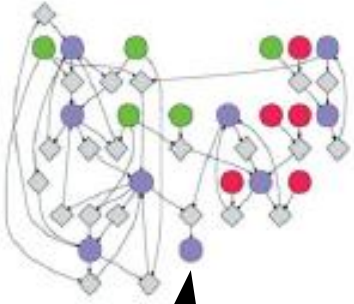
Eisai: Aricept®

Seraching for syntheses after 8 expansions

main-network 8 expansions



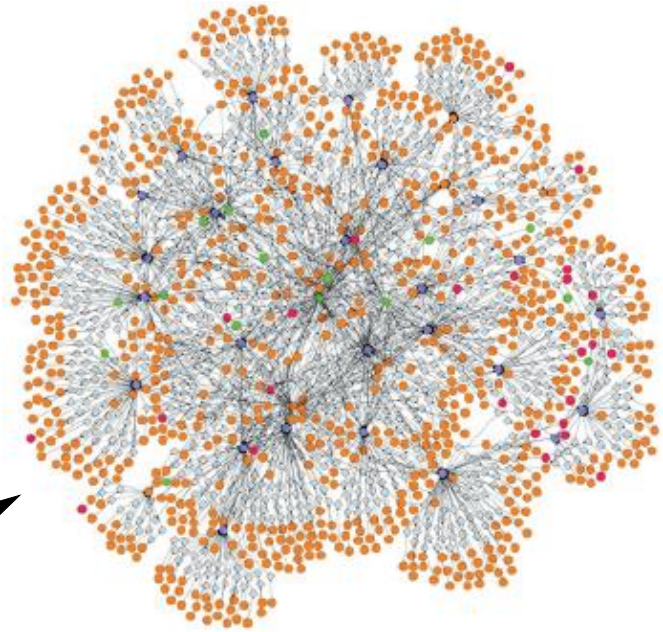
sub-network 8 expansions (viable syntheses)



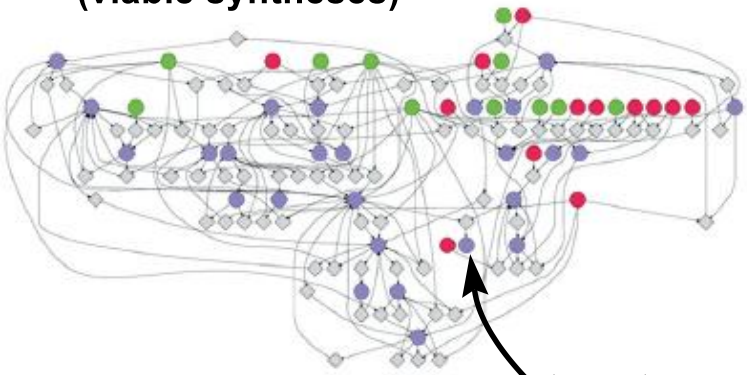
target

Seraching for syntheses after 35 expansions

main-network 35 expansions



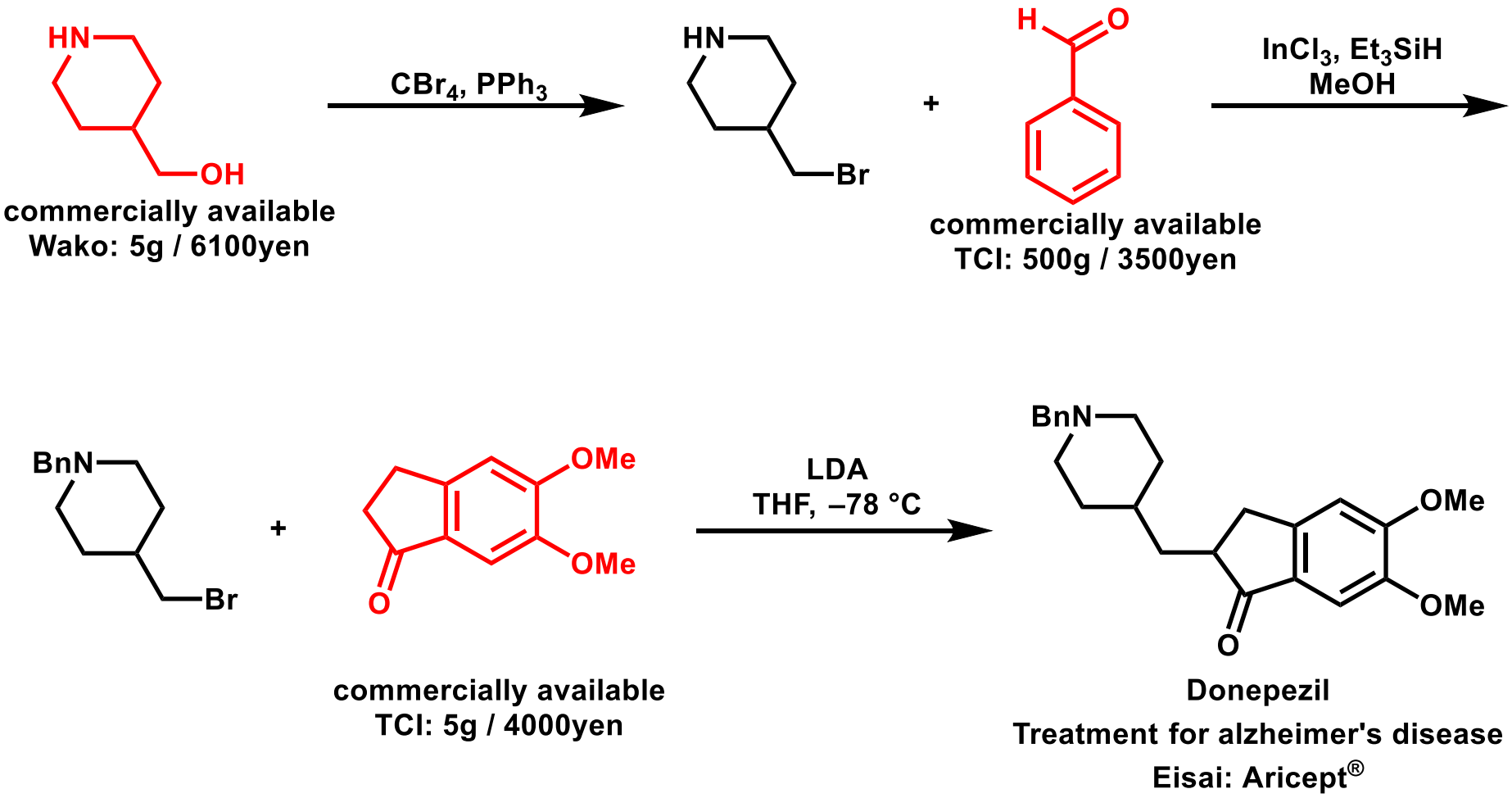
sub-network 35 expansions (viable syntheses)



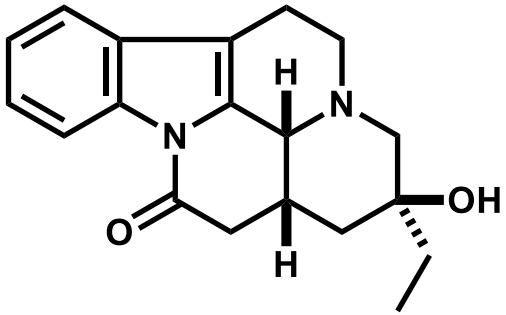
target

red: commercially available, green: known in the NOC

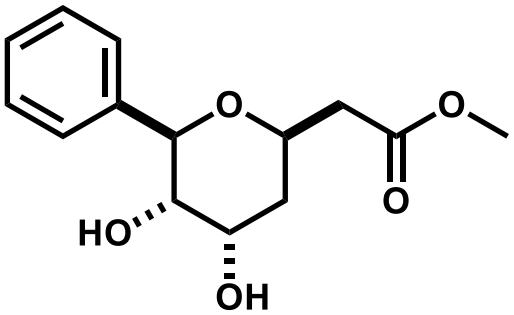
Retrosynthesis of Donepezil



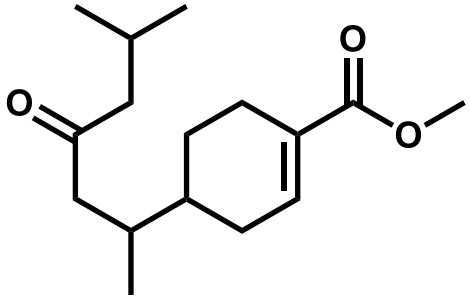
Synthetic Route of Recently Isolated Natural Product



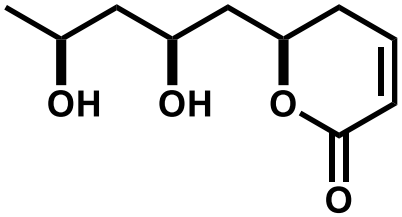
tacamonidine



goniothalesdiol A

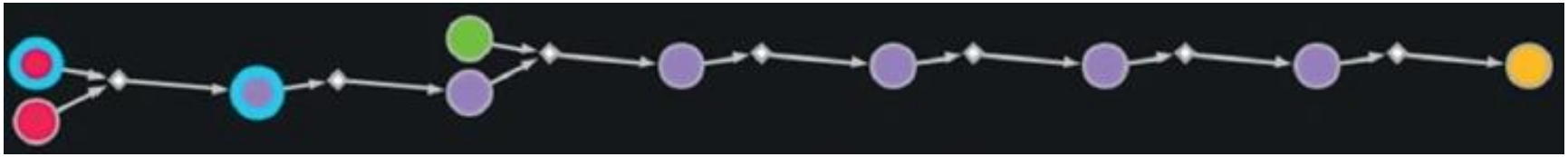


juvabione

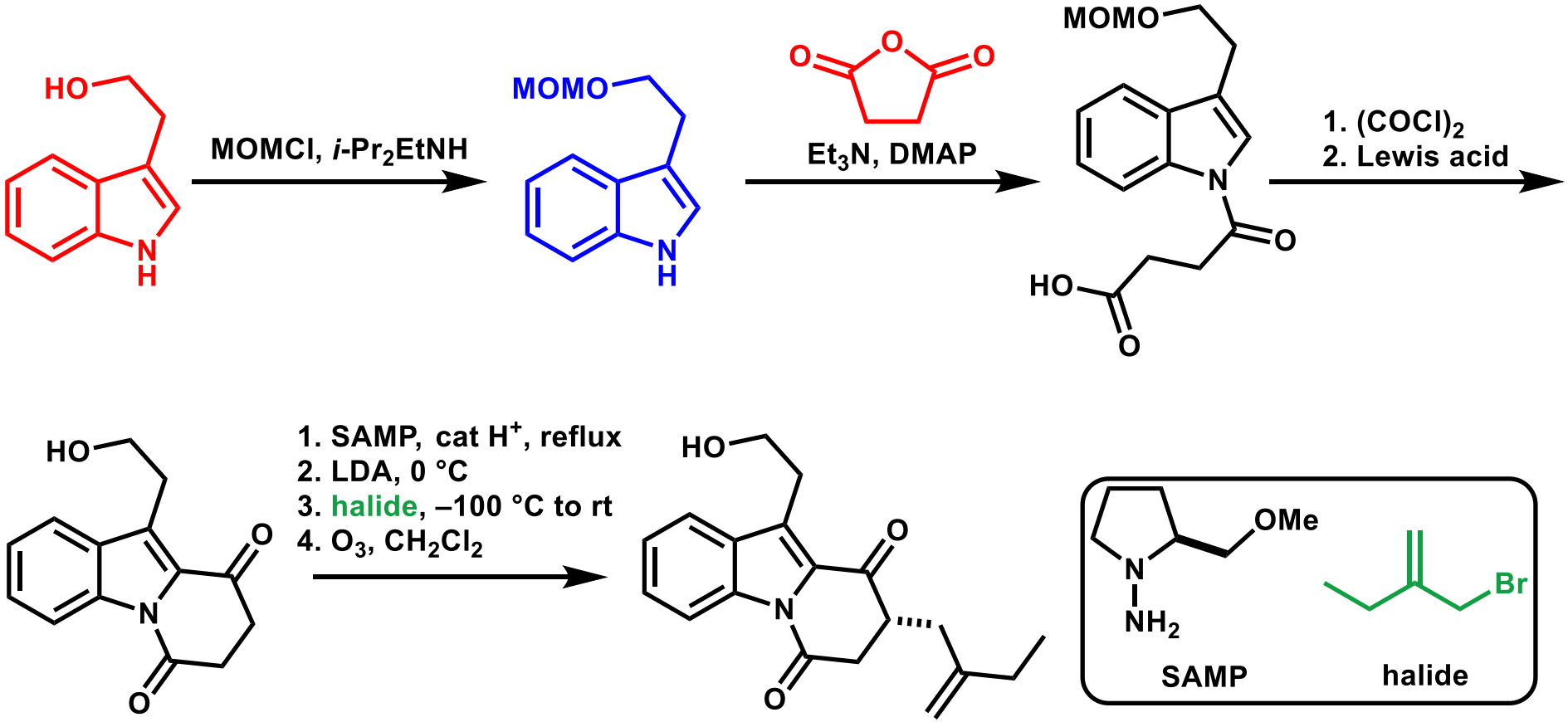


polyhydroxylated natural product isolated from *Cryptocarya latifolia*

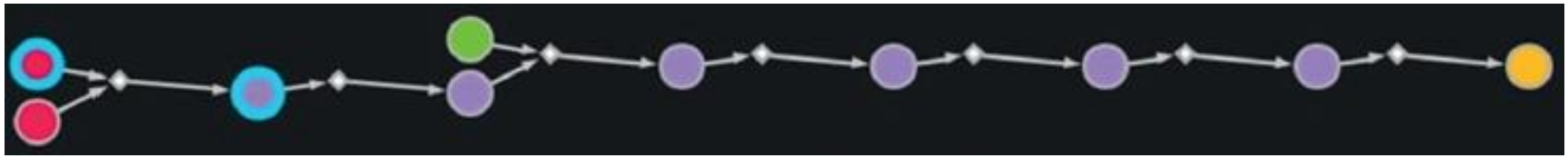
Synthetic Route of Tacamonidine -1



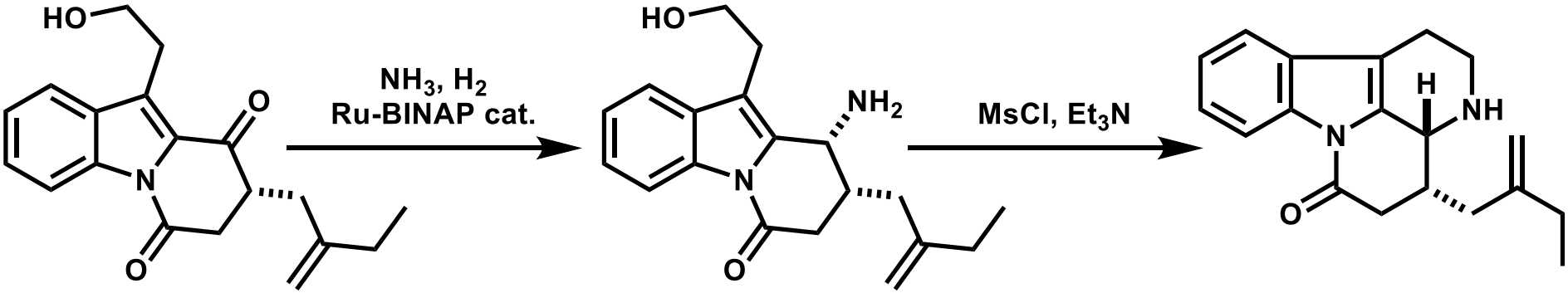
red: commercially available, green: known in the NOC, violet: unknown, yellow: target
blue halos: protection required

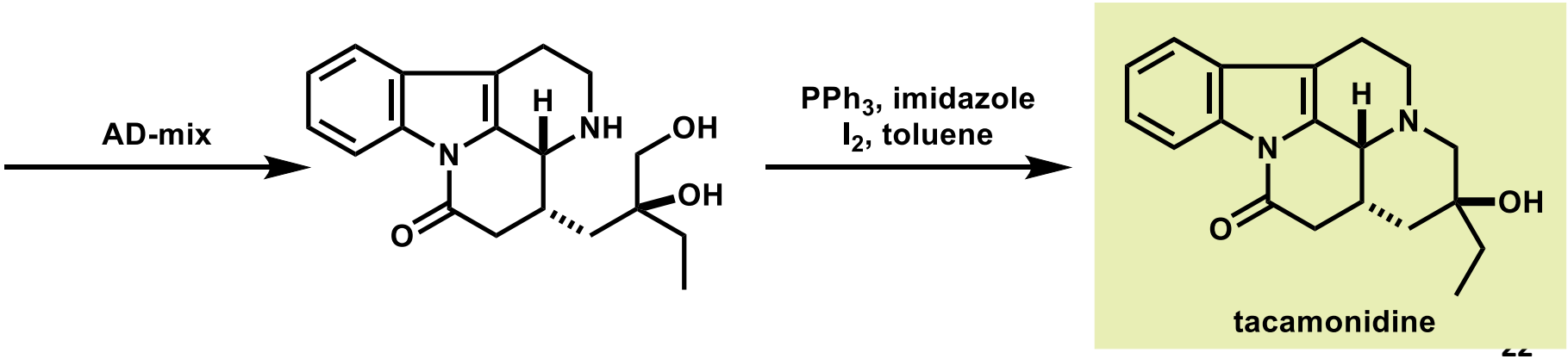


Synthetic Route of Tacamonidine -2

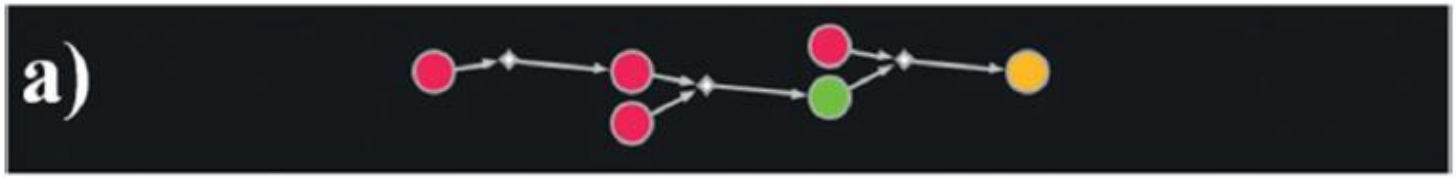


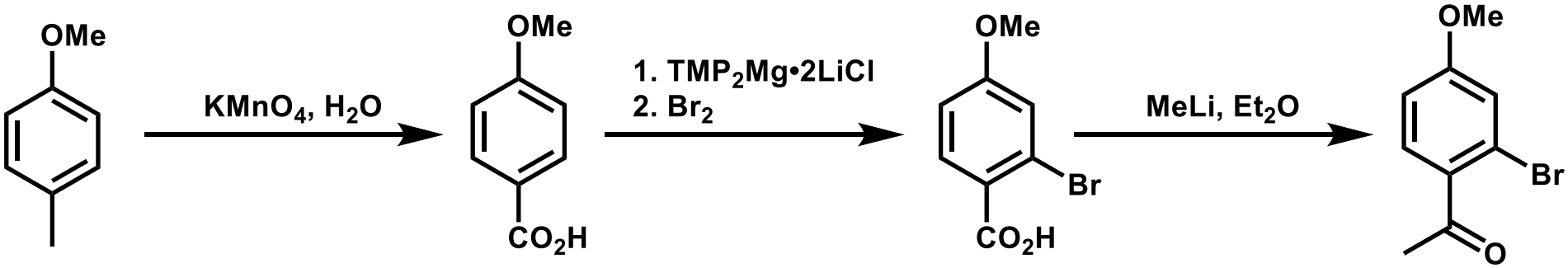
red: commercially available, green: known in the NOC, violet: unknown, yellow: target
blue halos: protection required

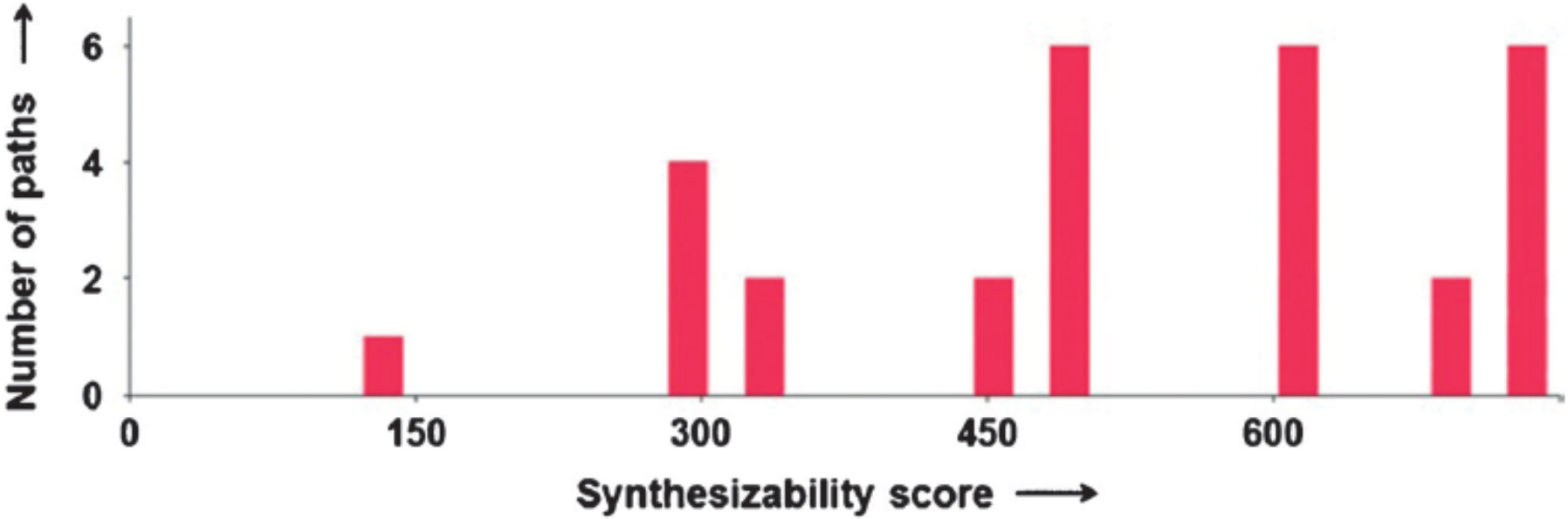




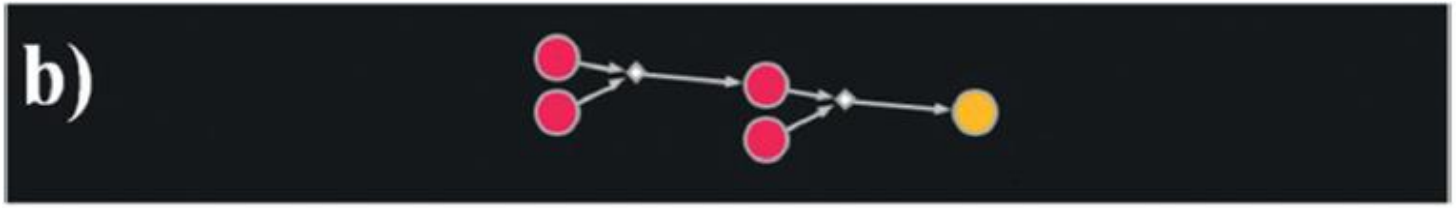
"Synthesizability" -1

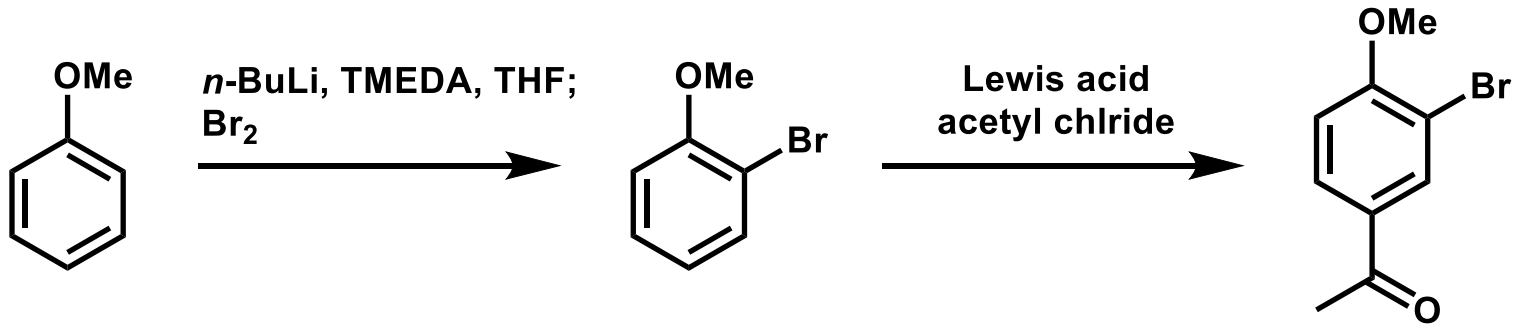


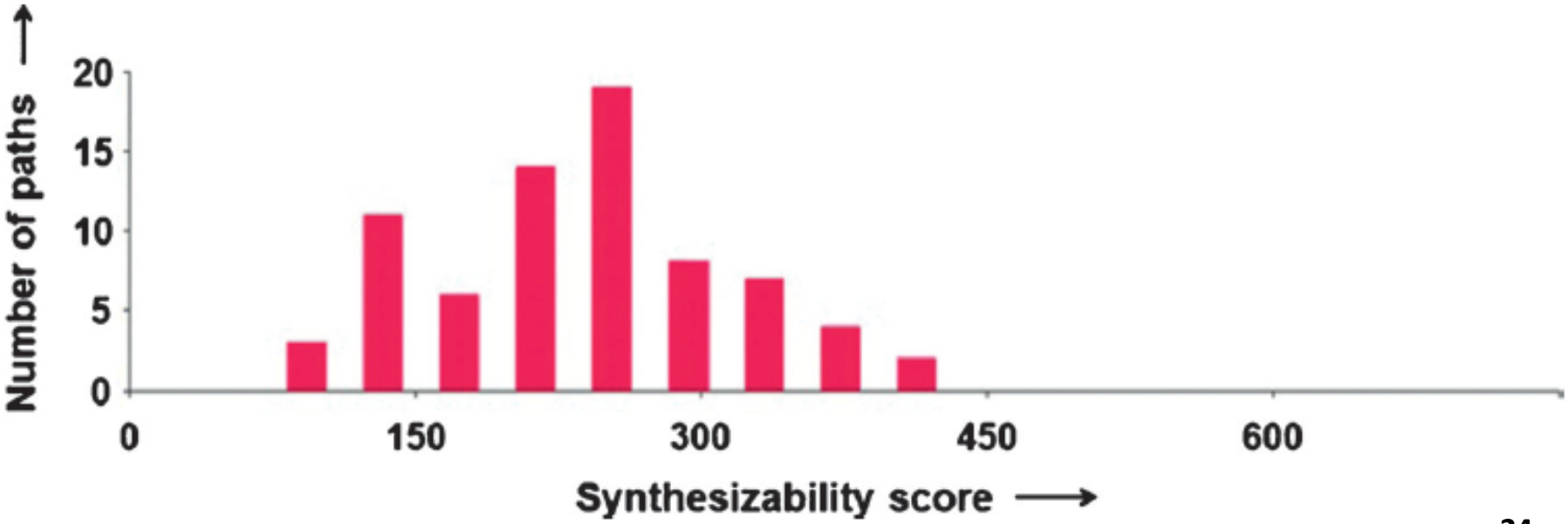




"Synthesizability" -2

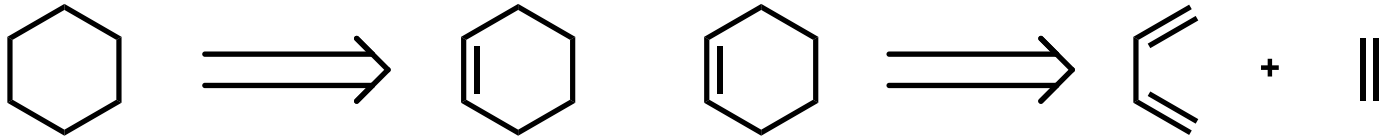
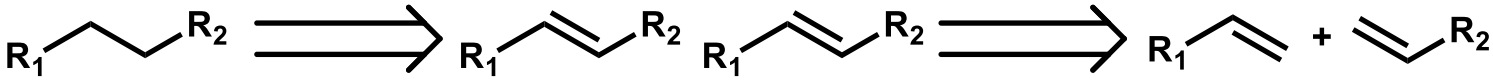


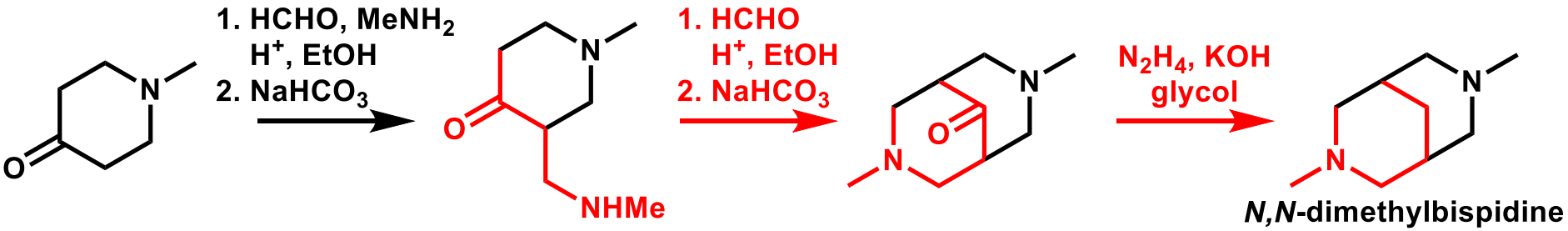




Search Speeds vs. Diversity of synthetic routes

Two-step strategies

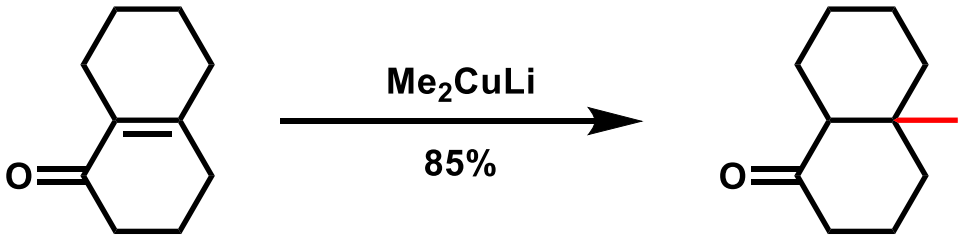
Strategy	Step 1	Step 2
Diels-Alder /alkene reduction		
Alkene metathesis /alkene reduction		
Mannich reaction /Ketone -> CH_2 reduction		

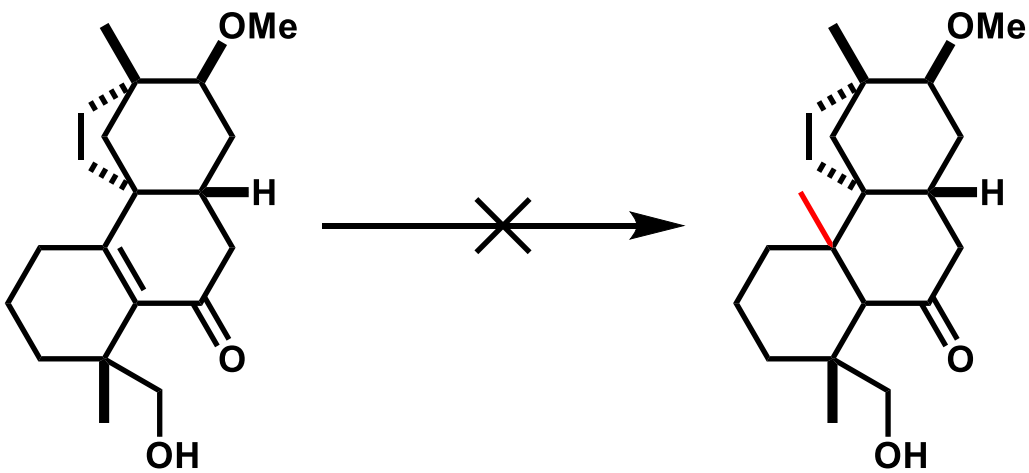


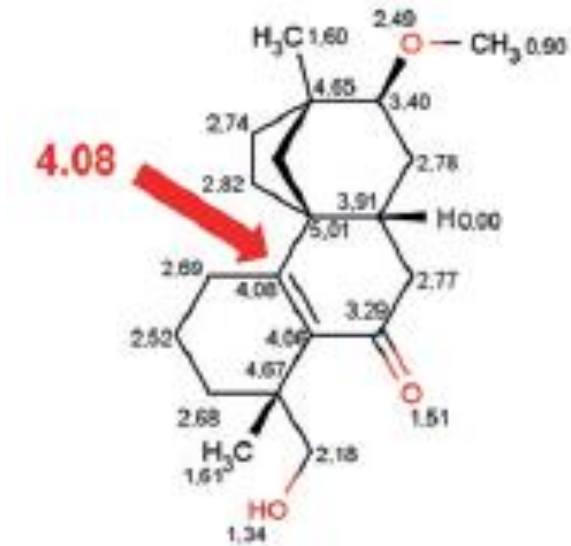
Use of one of in-built strategies reduced the number of search iterations from 151 to 48.

Introduction of too many strategies may excessively bias the searches into certain branches of synthetic possibilities thus limiting the diversity of pathways generated.

Problems - Steric Effect -1 -



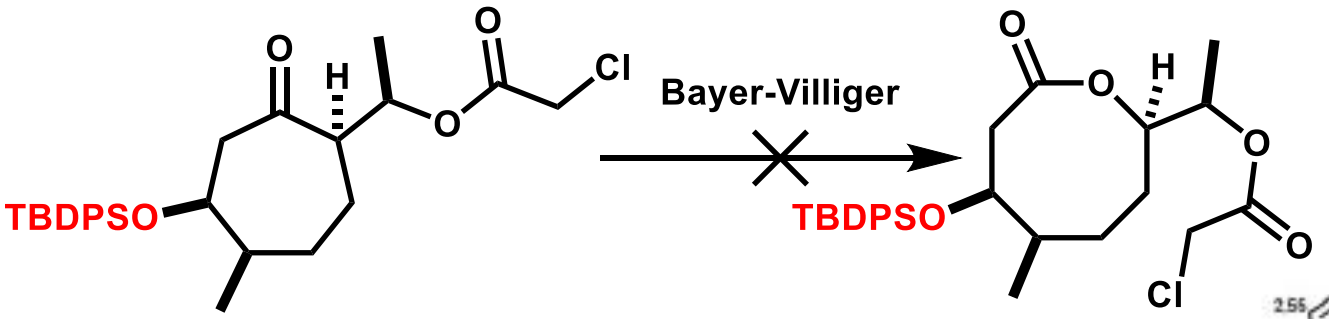


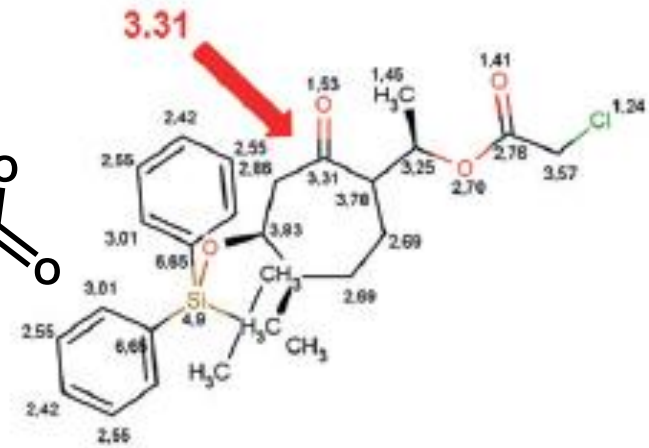


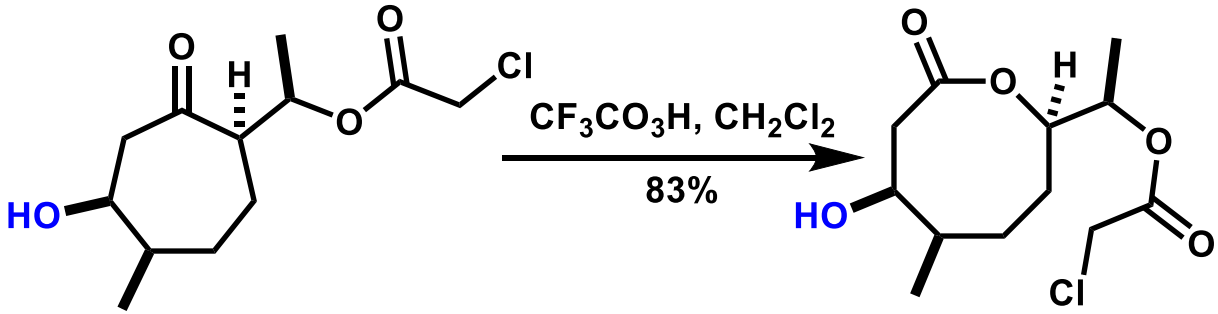
The numbers in the right pictures are the values of the TSEI steric crowding index.

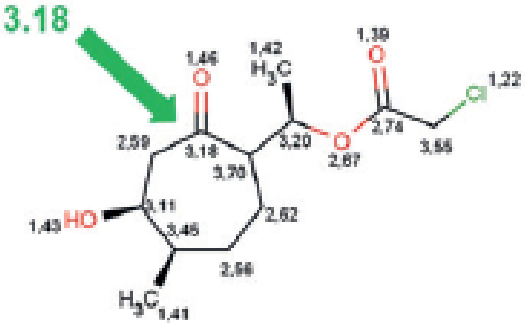
Overman, L. E.; Ricca, D. J.; Tran, V. D. *J. Am. Chem. Soc.* **1997**, *119*, 12031.

Problems - Steric Effect -2 -









The numbers in the right pictures are the values of the TSEI steric crowding index.

McWilliams, J. C.; Clardy, J. *J. Am. Chem. Soc.* **1994**, *116*, 8378.

Summary

Chematica - Network module (the Network of Oraganic Chemistry)
- Retrosynthesis module (Syntaurus)

	Network module	Retrosynthesis module
advantage	Search for syntheses in the NOC with desired search constraints	Search for syntheses including the reactions which don't reported yet
disadvantage	Unable to suggest any novel synthetic strategies and /or pathways leading to targets that have not yet been synthesized	How to evalutate the gained synthetic routes How to incorporate steric effects and protection method in synthetic routes

Author said comments in the review.

The machines are not yet likely to match the creativity of top level total-synthesis masters.

As we stressed in the title of this Review, it is only "the end of the beginning." and there are still many challenges to be overcome.

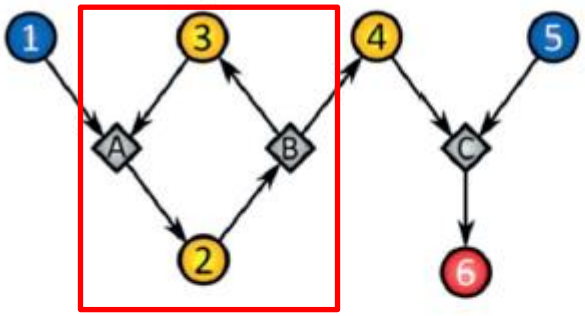
Appendix

Search Algorithms

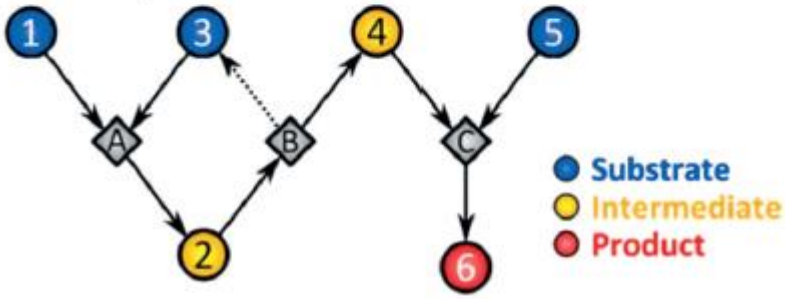
Contradiction: Intermediate 3 is synthesized from intermediate 2. On the other hand, Intermediate is synthesized from intermediate 3.

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

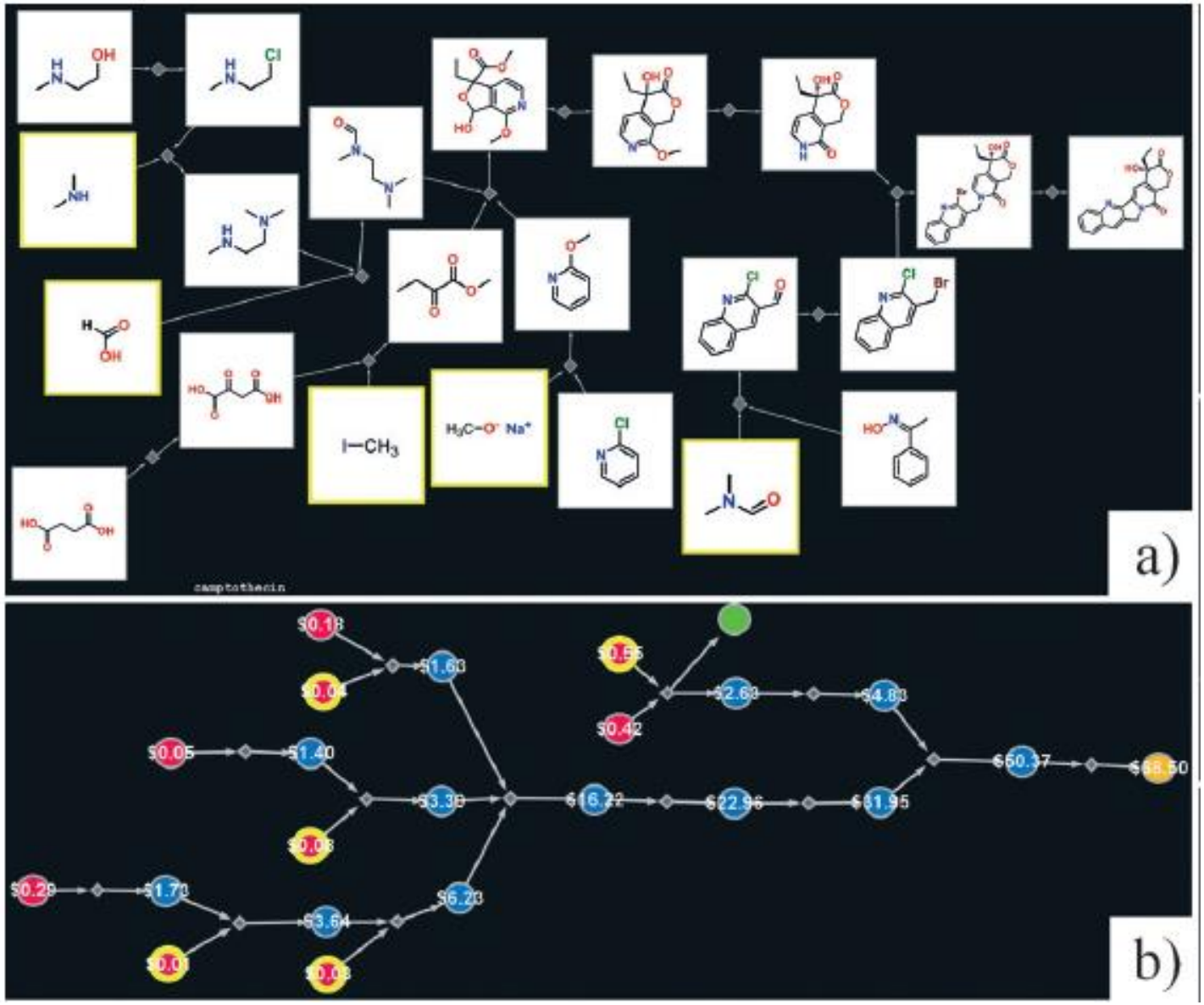
Nonviable Synthesis Tree



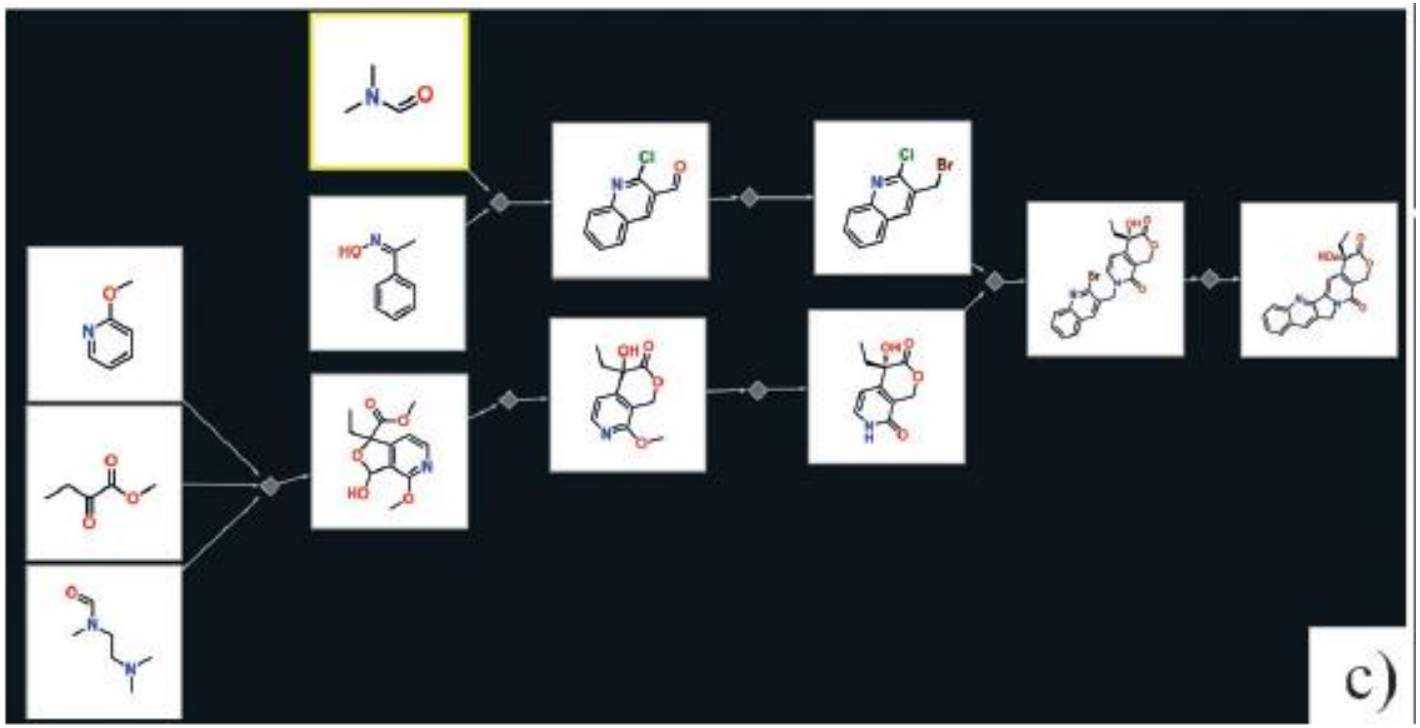
Viable Synthesis Tree



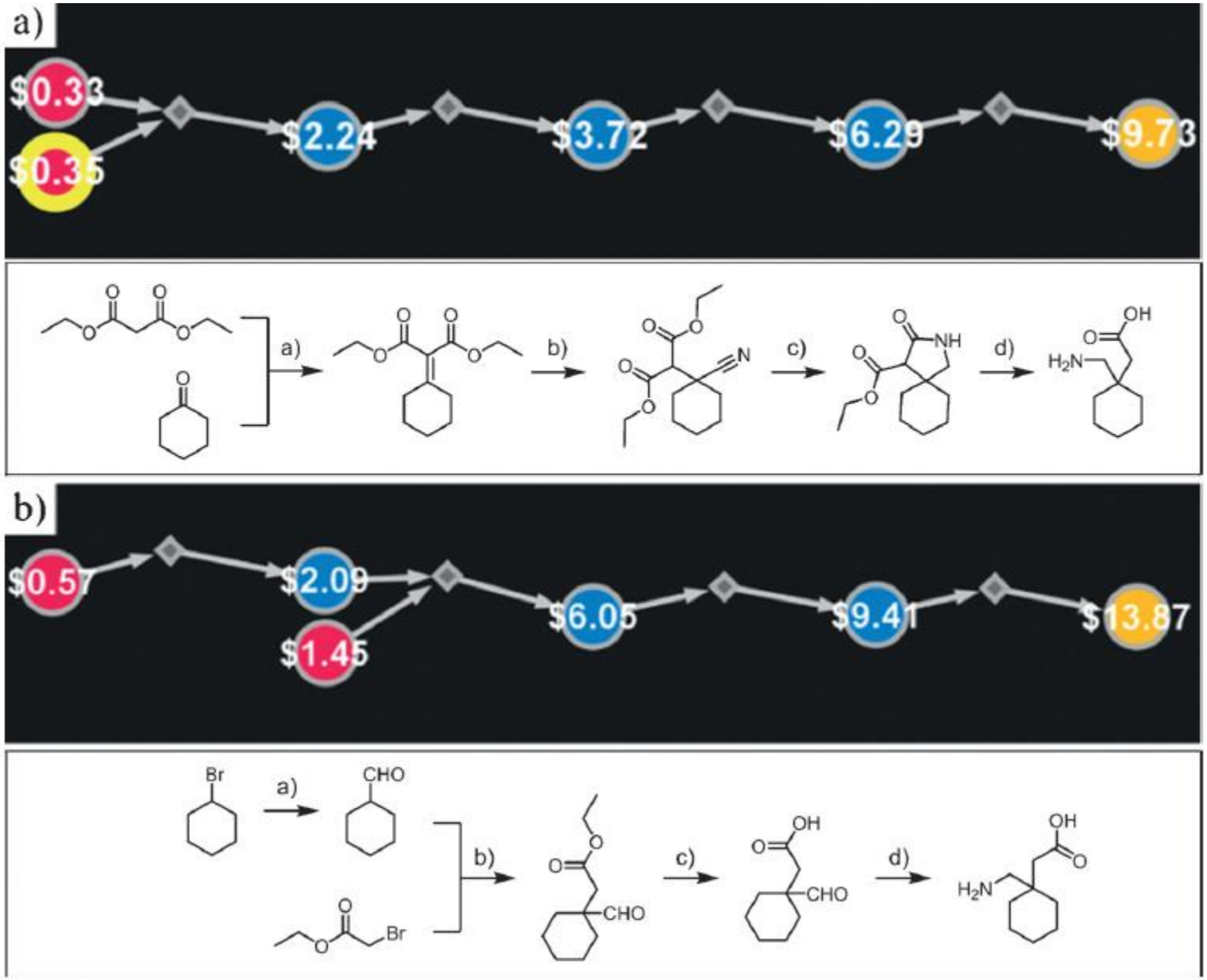
Cost-optimized Syntheses of camptothecin -1



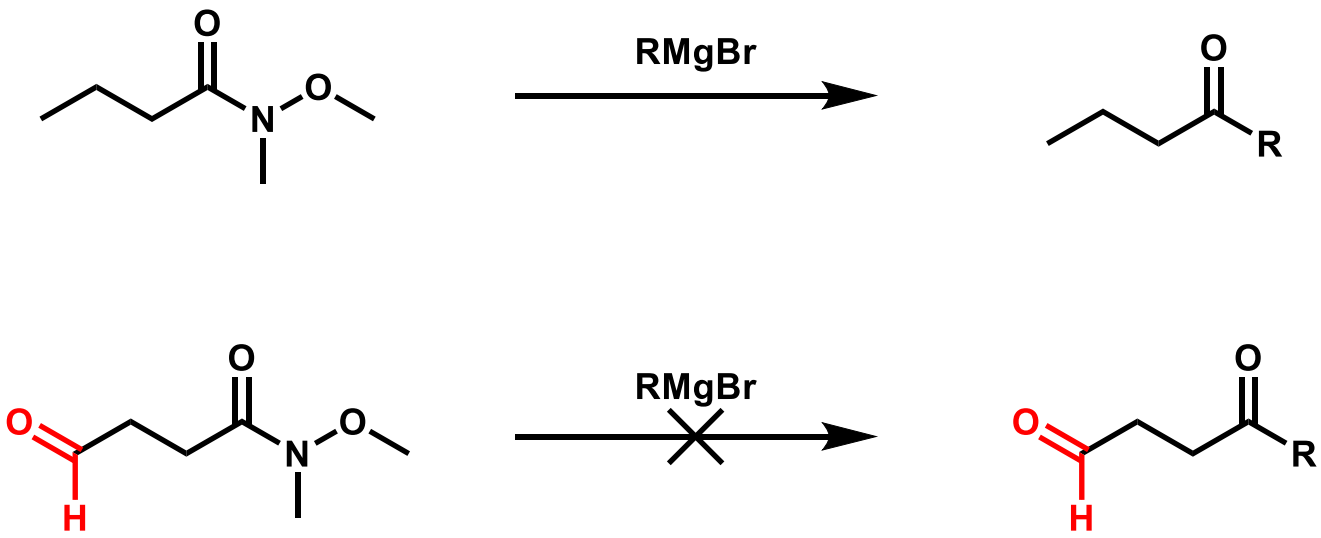
Cost-optimized Syntheses of camptothecin -2



Cost-optimized Syntheses of gabapentin

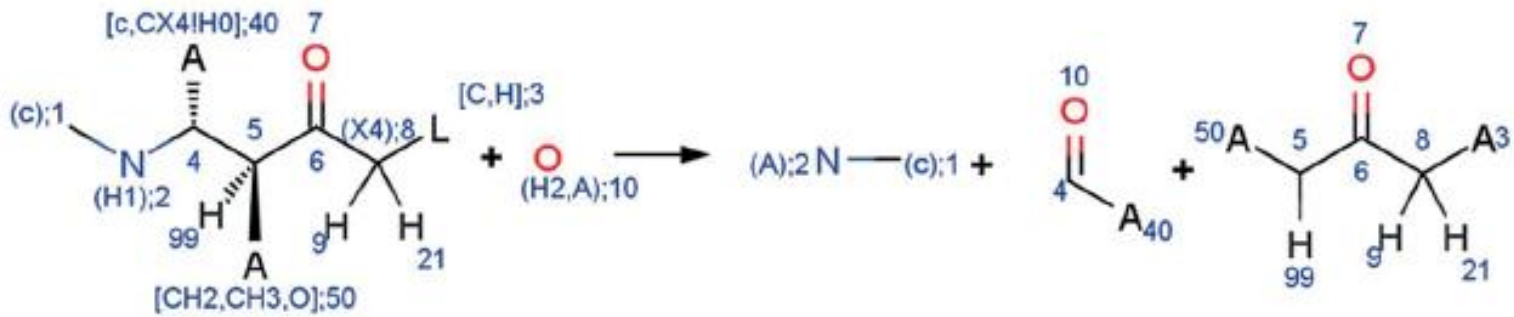


One Example of Reaction Rule



"Reaction X can proceed if groups Y, Z are not present" or "reaction X can proceed if groups Y, Z are apprpriately protected

Mannich Reaction as Coded into Syntaurus



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"]

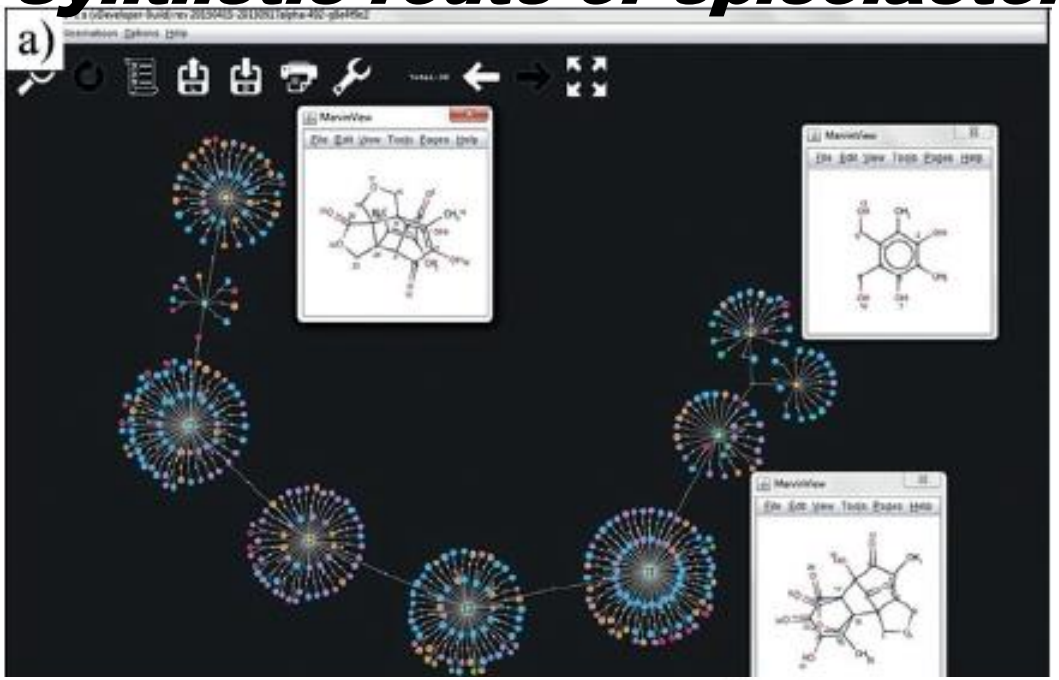
protection_conditions_code: ["NNB1", "EA12"]

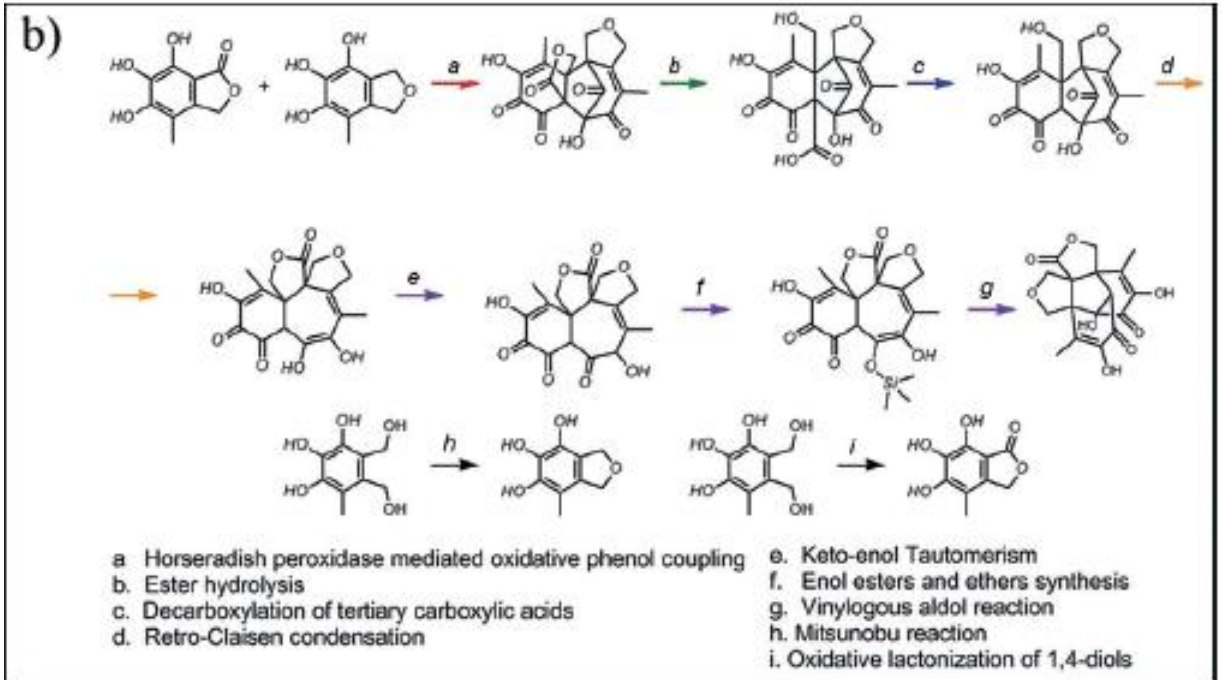
incompatible_groups: ["[#6]O[OH]", "c[N+]#[N]", "[NX2]=[NX2]", "[#6]OO[#6]", "[#6]C(=[O])OC(=[O])[#6]", "[#6]N=C=[O,S]", "[#6][N+]#[C-]", "[#6]C(=O)[Cl,Br,I]", "[CX3]=[NX2][*!O]", "[#6]C(=[SX1])[#6]", "[#6][CH]=[SX1]", "[#6][SX3](=O)[OH]", "[CX4]1[O,N][CX4]1", "[#6]=[N+]=[N-]","[CX3]=[NX2][O]"]

typical reaction conditions: "(S)-proline. Solvent, e.g., DMSO",

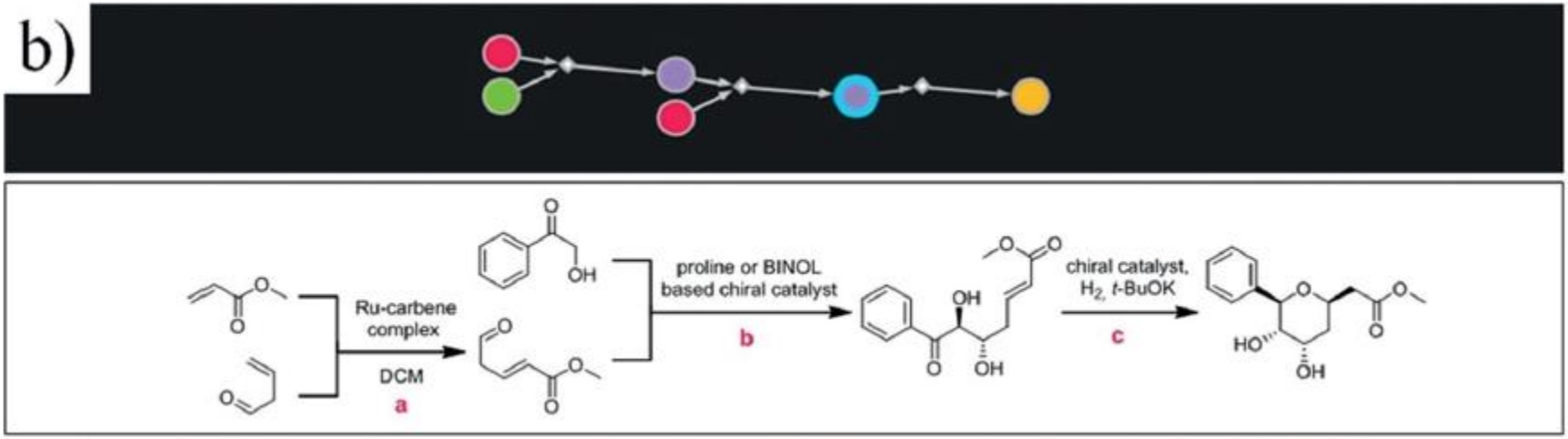
general references: "DOI: 10.1021/ja001923x or DOI: 10.1021/cr0684016 or DOI: 10.1021/ja0174231 or DOI: 10.1016/S0040-4020(02)00516-1"

Synthetic route of epicolactone

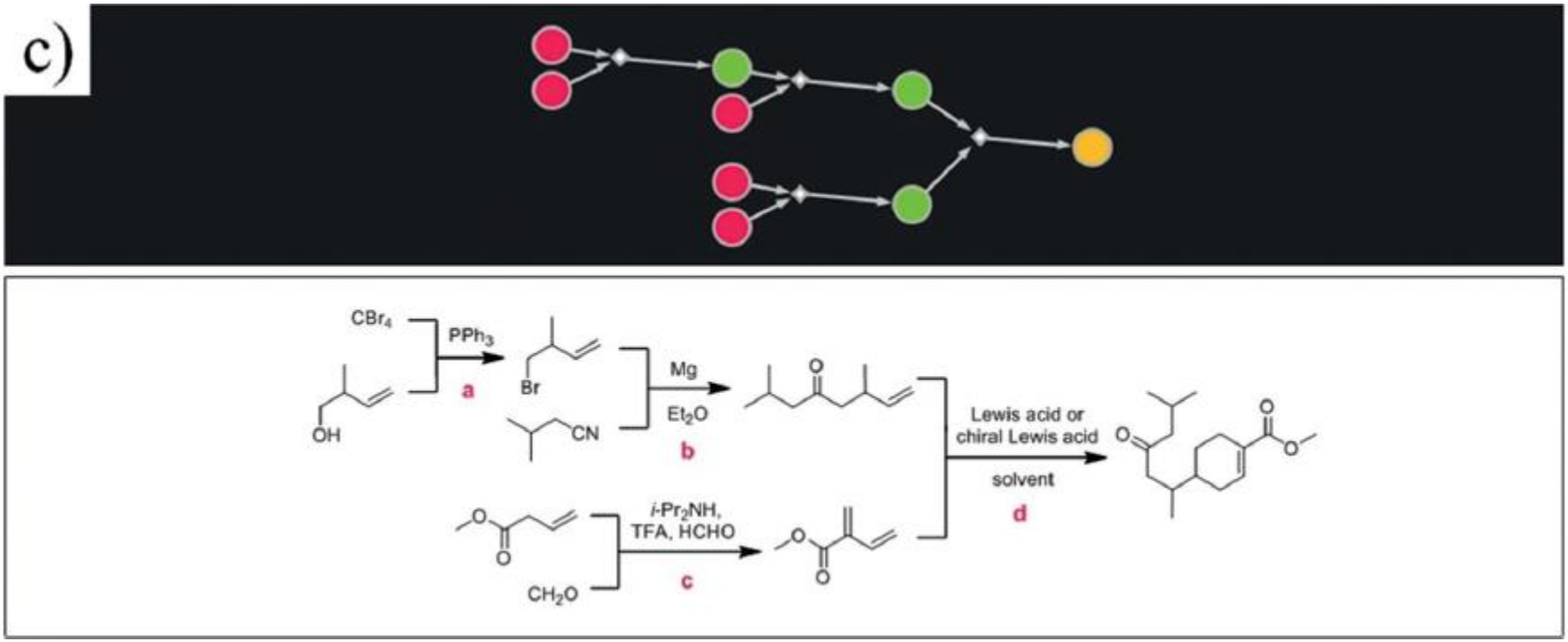




Synthetic route of geniothalesdiol A



Synthetic route of juvabione



Synthetic route of polyhydroxylated natural product