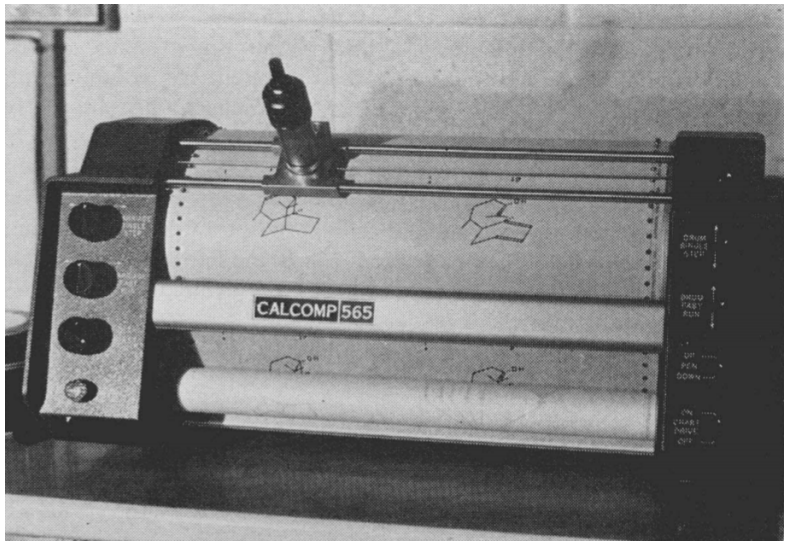
Planning Organic Synthesis:

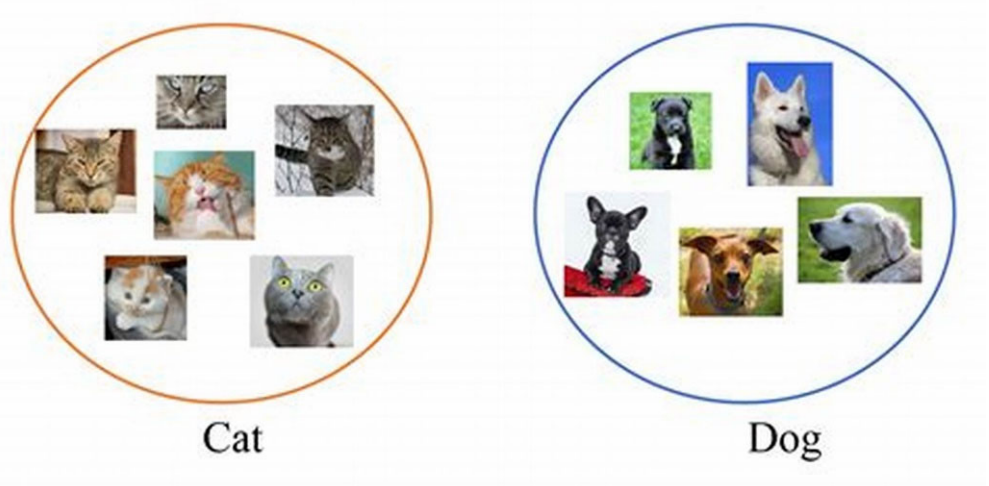
From Programmed Logic to Artificial Intelligence

Sicong Chen
College of Chemistry and Molecular Engineering
Peking University
Dec. 2019

Contents

- Introduction
- Rule-based expert system
  - LHASA
  - Chematica
- Machine learning strategy
  - Machine learning and deep learning
  - Predicting reaction type by neural network
  - Template free translation
  - MCTS-3N space search
- Summary





Organic Synthesis is a Strategic Game

	Chess	Rubik's cube	Chemical synthesis
			
Number of players	Two	One	One
Movements	Small set of moves defined for each piece, some moves may not be allowed for some positions	Rotation of cube's single layer; always the same number of moves allowed	Very large (>10000) number of possible moves (i.e., reaction rules); applicable moves depend on the structure of the molecule; database of moves can grow as chemistry advances
Start position	Always the same initial arrangement of pieces on the board; "white" player starts	(Random) rearrangement of the cube	Target that needs to be synthesized
Position	Current configuration of the pieces on the board	Configuration of the cube	Set of substrates/synthons at each step
End position	Check-mate or exceeding allowed time; draws also possible	Each of the six faces of the cube composed of one color	All substrates for target's synthesis judged as "available"
Score of the game	Won/lost/drawn/not finished	Solved/not solved; in addition, the time or the number of moves might be counted (less moves = better score)	Viable synthesis found/not found; viability ultimately confirmed by experimental execution; in addition to "hard" criteria (number of steps, yield) soft criteria such as "elegance" might be applied during evaluation

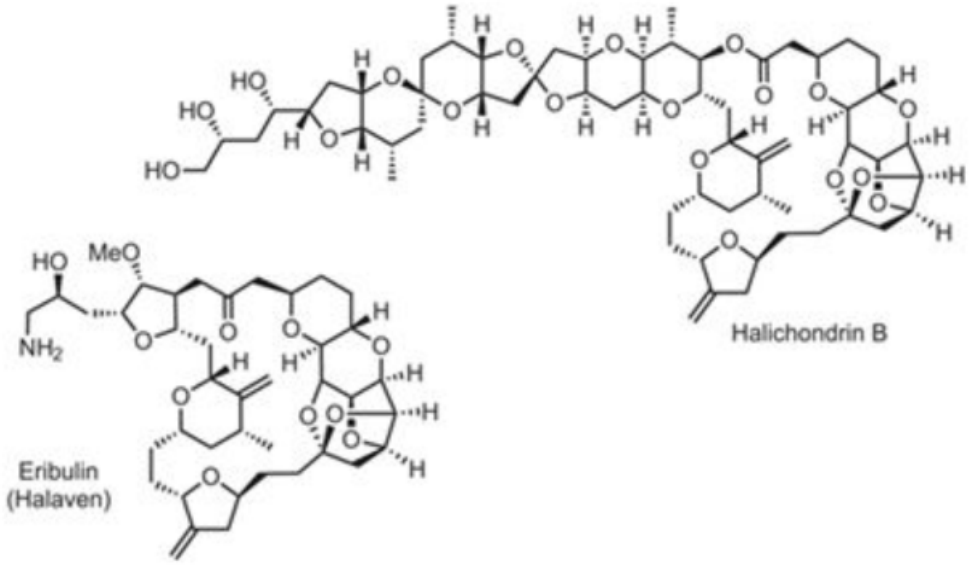
Angew. Chem. Int. Ed., **2016**, *55*, 5904

Organic Synthesis is a Strategic Game

Chess

Rubik's cube

Chemical synthesis



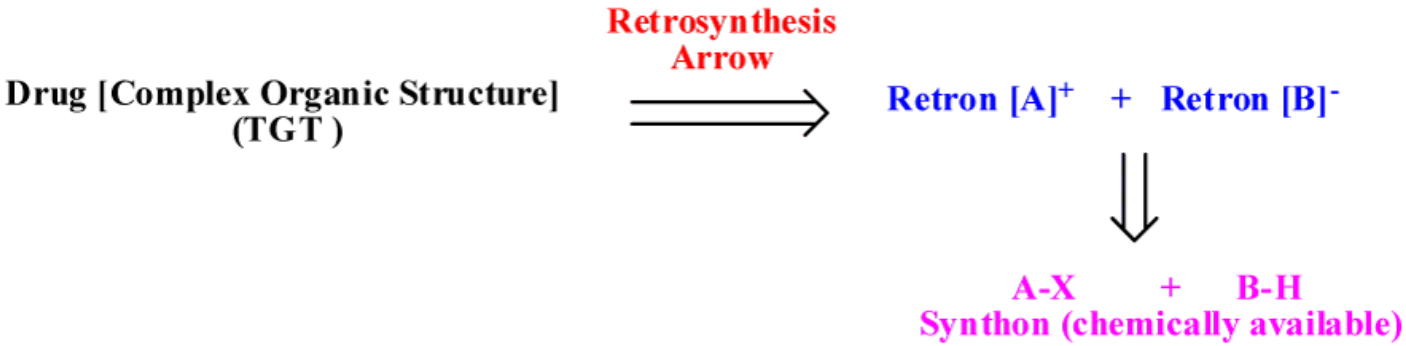
On average, 80.2 distinct reactions can be applied to a non-trivial retron,[13] translating into $\approx 3.5 \times 10^{28}$ possible 15-step pathways and $\approx 1.2 \times 10^{57}$ possible 30-step pathways

Commercially available Halaven is made in 62 synthesis steps[15] which seems to be an upper bound for industrially relevant syntheses

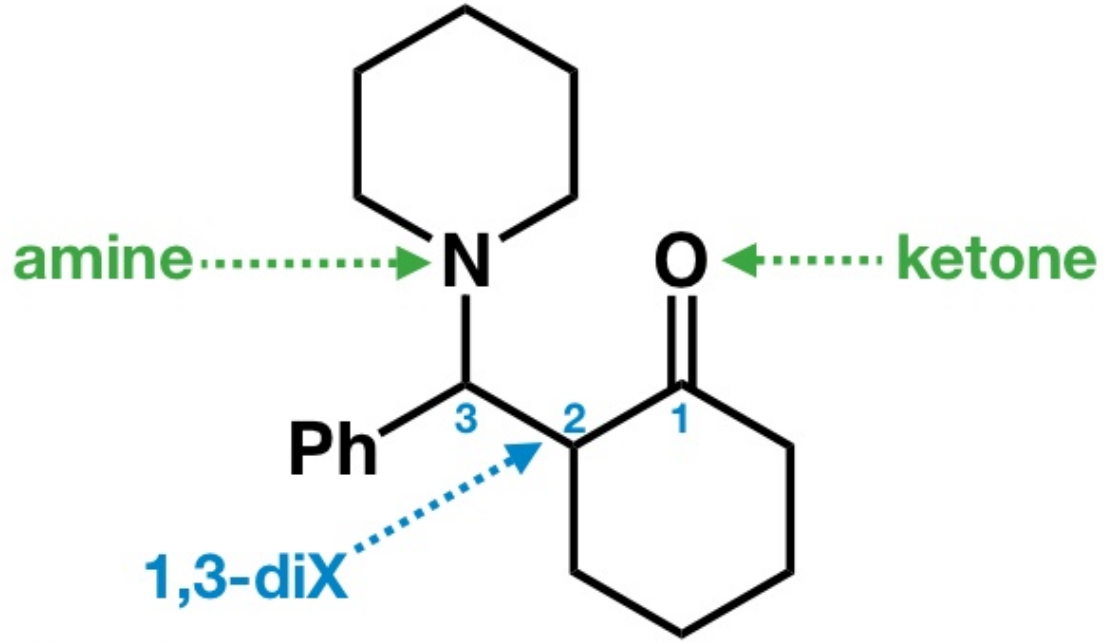
In general, no single solution can be objectively deemed as "optimal" as it depends on available substrates and/or the criteria applied (e.g., minimal number of steps, green conditions, no protection groups, etc.)

Angew. Chem. Int. Ed., **2016**, *55*, 5904

Retrosynthesis Analysis: Disconnection Approach

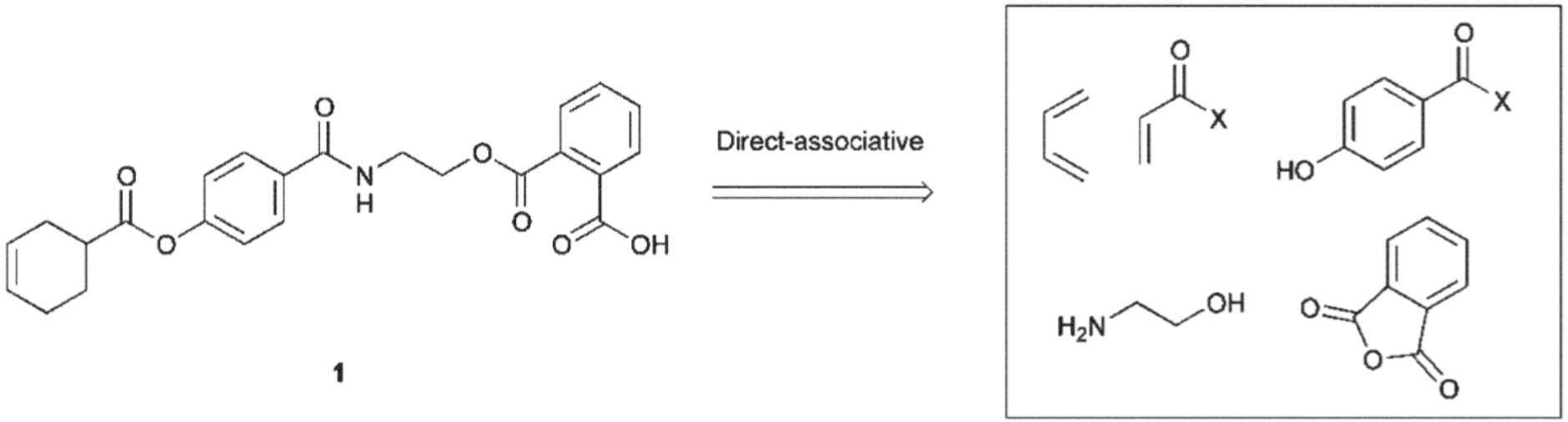


We have learned this since undergraduate class!



Synthetic Target Classification

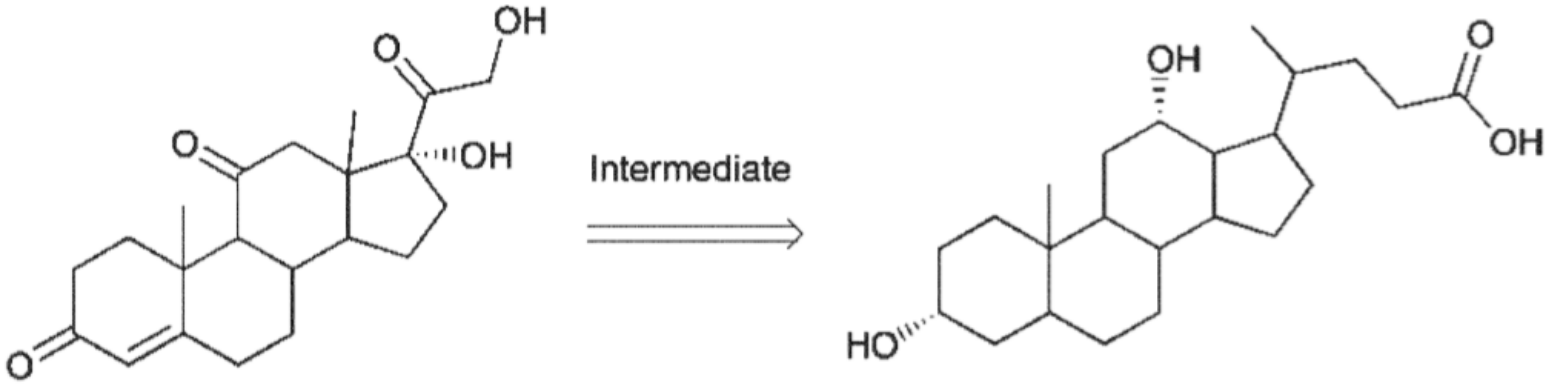
- **Direct associative**, where the synthetic target is a simple collection of ‘undisguised’ subunits, and where a minimal and uncontroversial analysis reveals the required starting materials.



Chem. Soc. Rev., **2005**, *34*, 247

Synthetic Target Classification

- **Intermediate**, where a complex synthetic target bears a close resemblance to another, but synthetically accessible, molecule, and the problem becomes finding the appropriate sequence of reactions for their interconversion

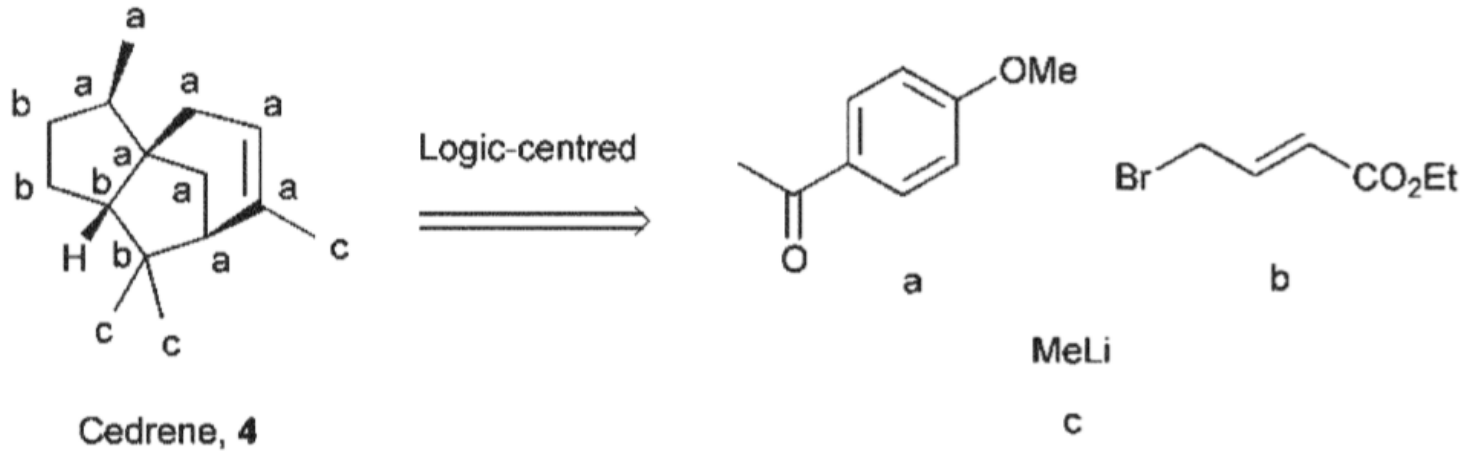


Cortisone, **2** Deoxycholic acid, **3**

Chem. Soc. Rev., **2005**, *34*, 247

Synthetic Target Classification

- **Logic-centered**, where a logical analysis generates a synthetic tree without any assumptions as to the starting materials required.

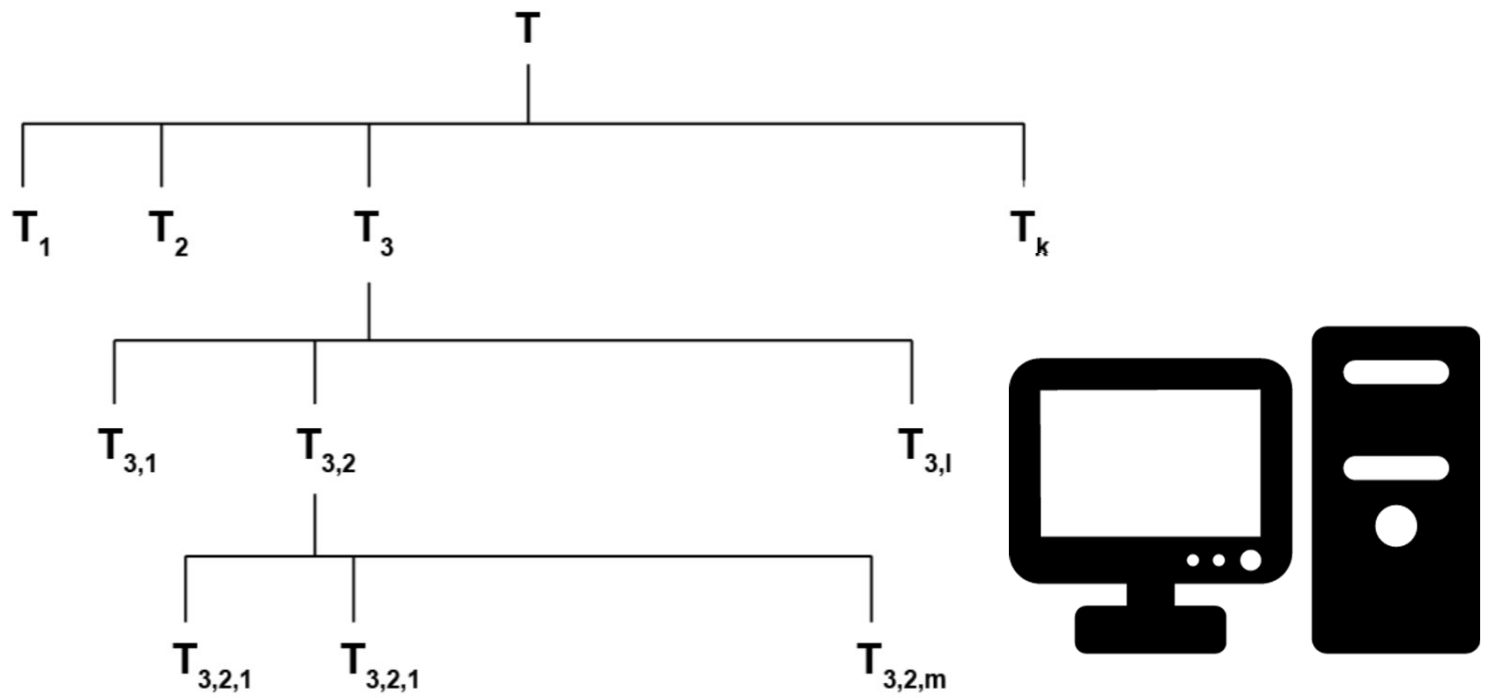


Cedrene, **4**

- In the subject of logic-centered complex molecular synthesis, at the other end of the spectrum, we encounter a **methodology limited** only by the **frontiers of chemistry** and the **power of human intelligence and creativity**. (Corey, 1969)

Chem. Soc. Rev., **2005**, *34*, 247

Computer-Assisted Synthesis Planning (CASP)



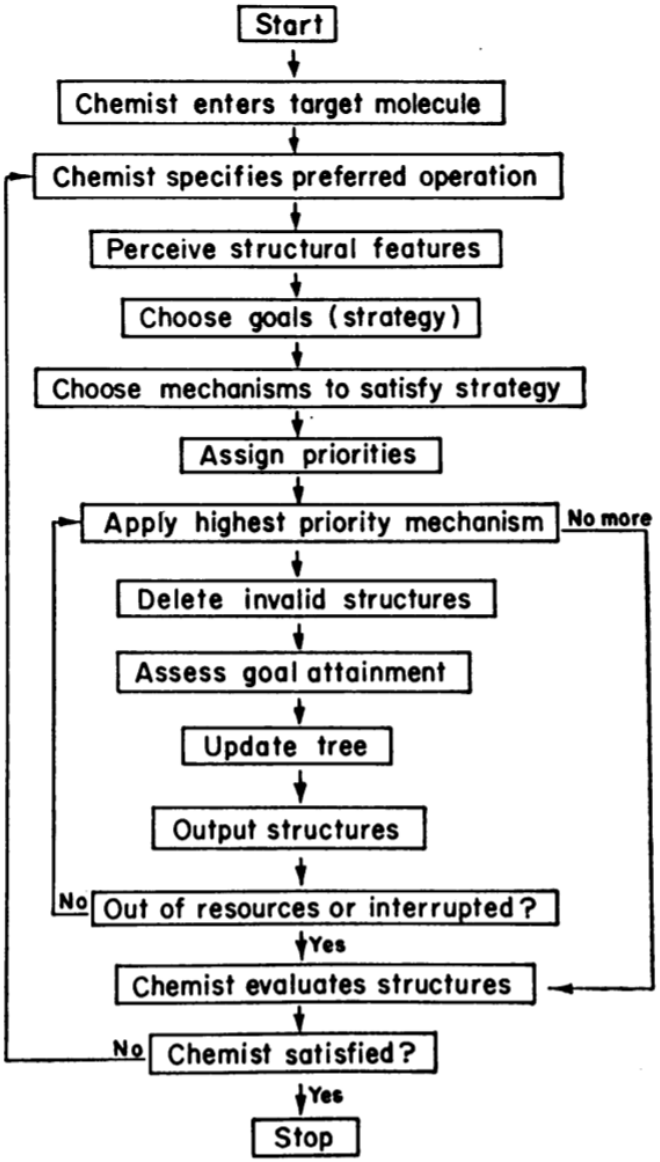
How to build the synthetic tree of the target **T**?

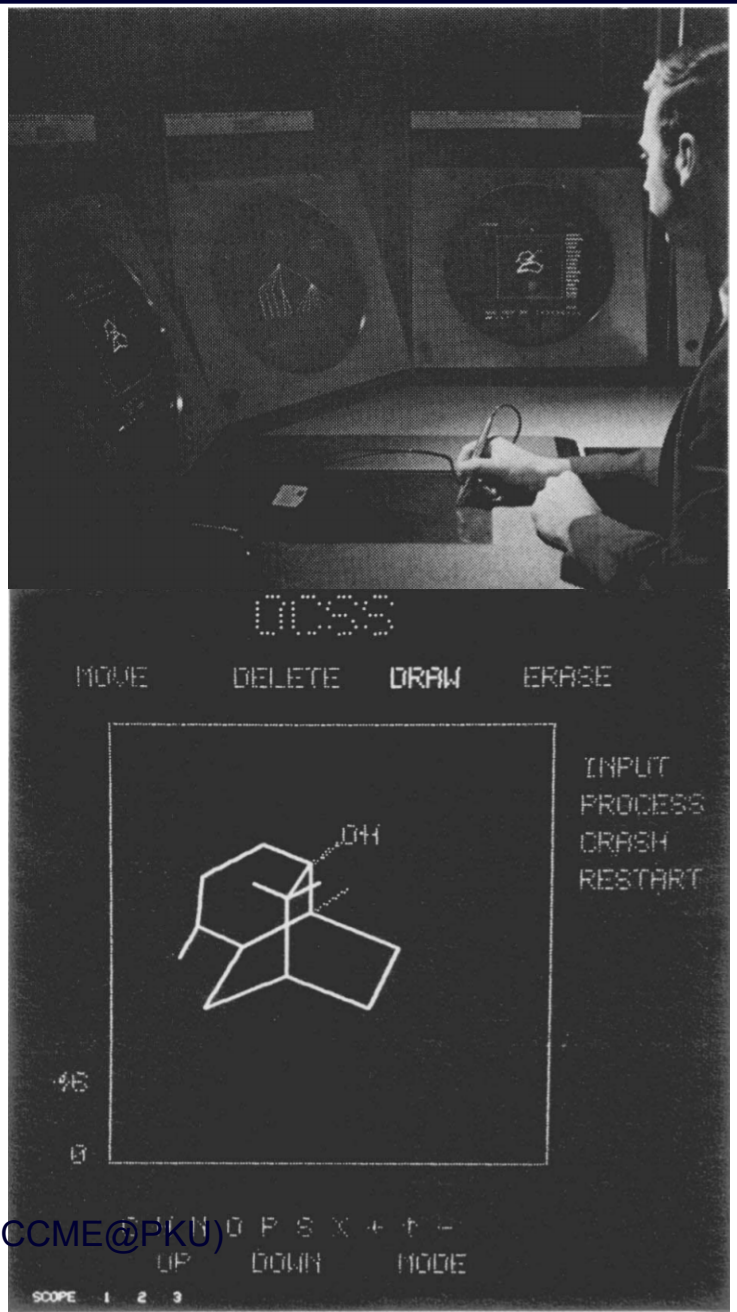
Whether a computer can do human 'logic-centered' analysis?

Organic Chemical Synthesis Simulation (**OCSS**) -
Logic and Heuristics Applied to Synthetic Analysis (**LHASA**)

Science, **1969**, *166*, 178

Workflow of OCSS-LHASA





Science, **1969**, *166*, 178

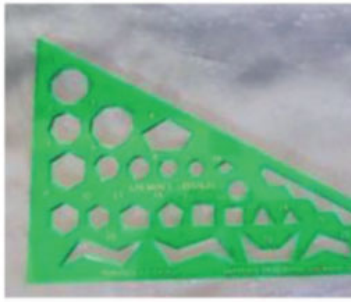
DOI: 10.1002/anie.201405820

History of the Harvard ChemDraw Project

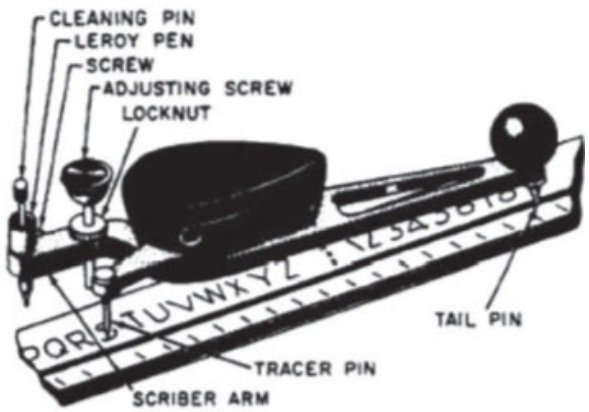
*David A. Evans**

a)

c)



b)



In 1985 Stewart was an NSF-sponsored graduate student pursuing his PhD degree as a student with Professor E. J. Corey. Stewart was integrated into that group as a member of the LHASA[14] effort. In September 1984 Stewart had also purchased a Mac, and he frequently visited our labs down the hall. He soon became interested in Sally's activities in slide preparation. She was producing india ink drawings with a Leroy Lettering Set (Figure 1). One of her frustrations was associated with the effort expended in creating a complex structure only to start over to draw the next structure in the reaction sequence where only minor chemical modifications had been made. On more than one occasion she suggested that I might consider working on structures less complex than vancomycin!

One afternoon during this period, Sally vented her frustration with the comment to Stewart: "How would you like to save my marriage?" In fact, this was a loaded question as she and I had already looked at the MacDraw software that appeared to be a reasonable starting point for a Mac-based structure-drawing program. This interaction culminated in a meeting between Sally, Stewart, and myself to discuss the possibility of creating such a program. Stewart stated that he would have a go at this challenge. Within several weeks he reported back with a rudimentary program that could handle many of the templates found on the Fieser Triangle. In his first rendition, the length of the line or the size of the ring was determined by the magnitude of the "drag" of the mouse from the "touchdown point." Sally then informed him that, while this feature was terrific, she also wanted to have the option of making all of the bonds the same length! Stew's response: "Why would you want to do that?" So began our meaningful collaboration into what was needed for a chemical drawing

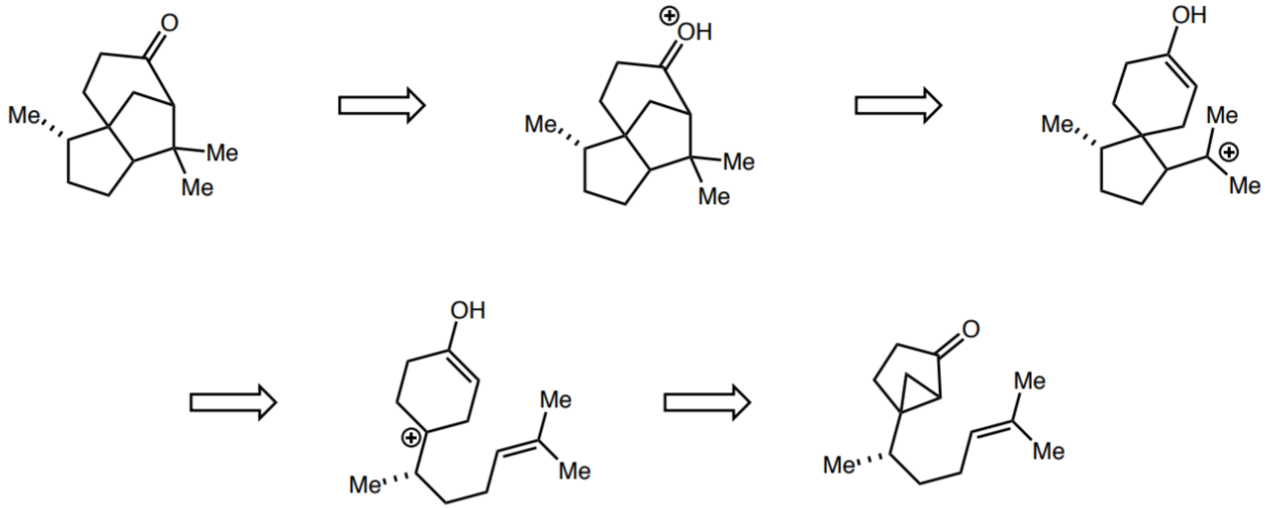
Angew. Chem. Int. Ed., **2014**, *53*, 11140

Five Retrosynthetic Strategies in LHASA

- **Transform-based strategies**
  - *Name Reaction* based transformations: powerful reaction template

Carbo Diels-Alder	Internal nucleophilic acylation
Quinone Diels-Alder	Internal ene reaction
Hetero Diels-Alder	Internal cycloaddition:
Robinson annulation	[4 + 2], [3 + 2], [4 + 3], [2 + 2], [2 + 1]
Position-selective partial aromatic reduction	Pericyclic cation or anion closure
Cation π-cyclization	Sigmatropic rearrangements
Radical π-cyclization	Photocyclizations
Aldol cyclization	Enantioselective π-addition
Sila-acyloin cyclization	Diastereoselective π-addition
Internal S_N2 cyclization	Fischer indole, Knorr pyrrole, and so on

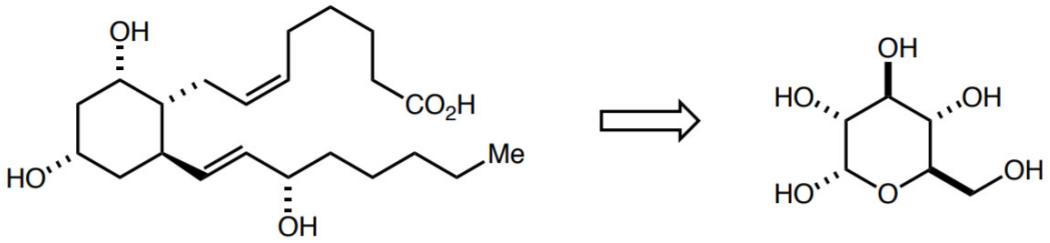
  - Mechanistic transforms: to a reactive intermediate



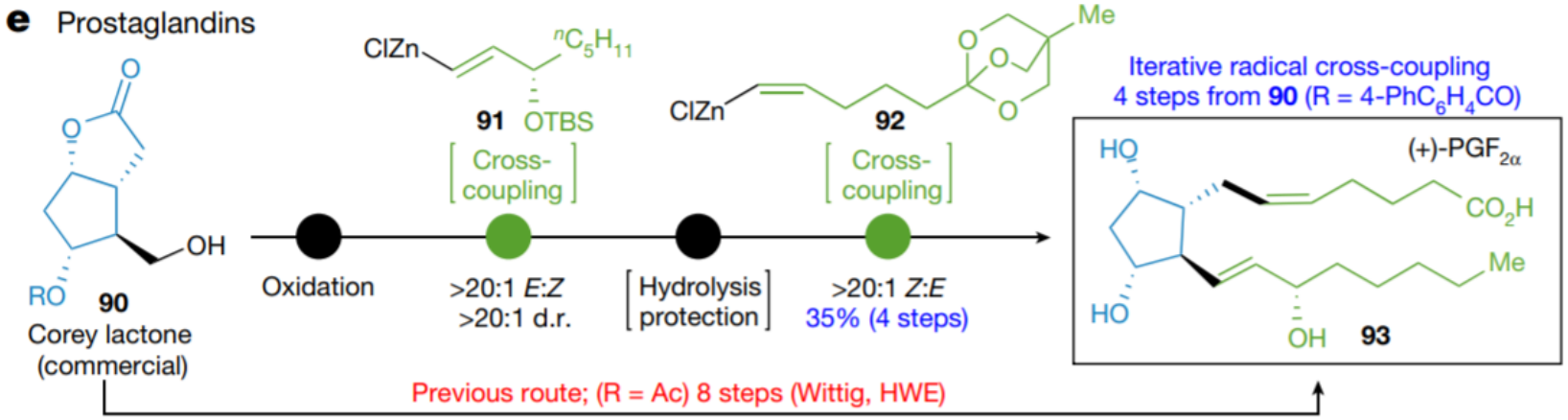
Science, **1985**, *228*, 408

Five Retrosynthetic Strategies in LHASA

- **Structure-goal (S-goal) strategies:** identification of a potential starting material, building block, retron-containing subunit or initiating chiral element



Baran's decarboxylative alkenylation

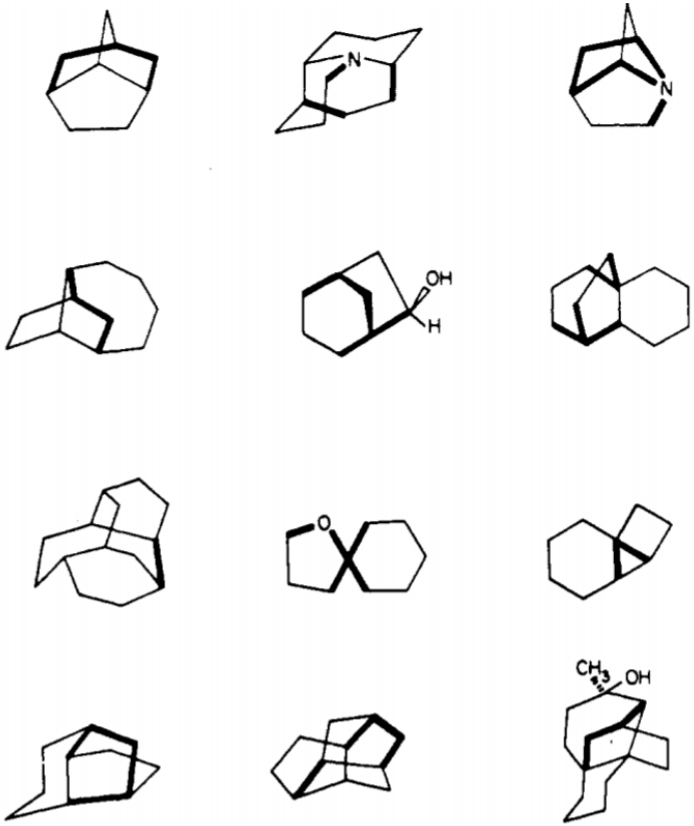


Science, **1985**, *228*, 408; *Nature,* **2017**, *545*, 213

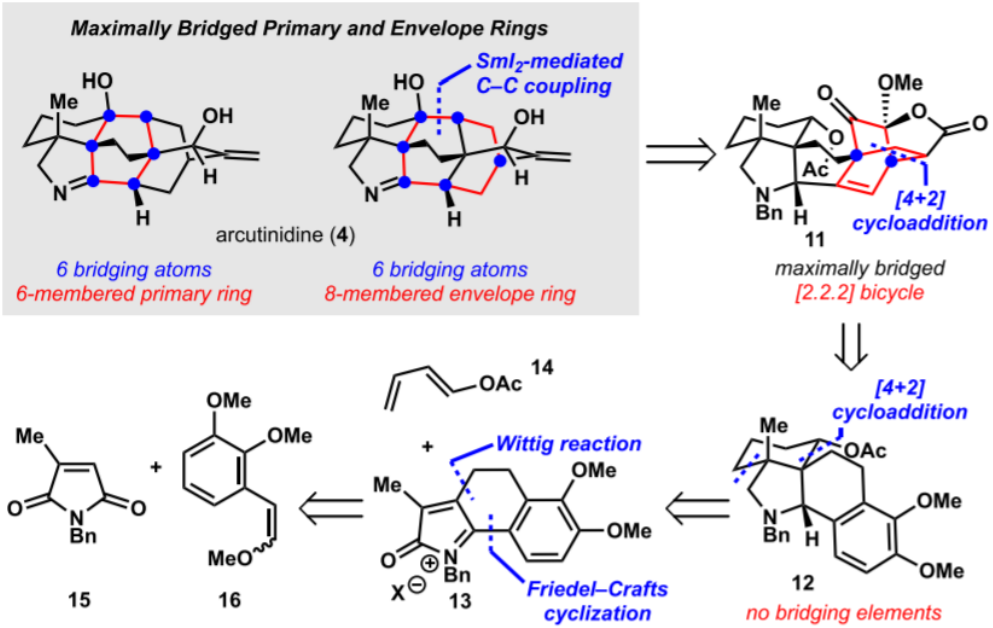
Five Retrosynthetic Strategies in LHASA

- **Topological strategies:** identify one or more bonds whose disconnection can lead to major molecular simplification

Preferred strategic bonds



Sarpong's retrosynthesis of Arcutinidine



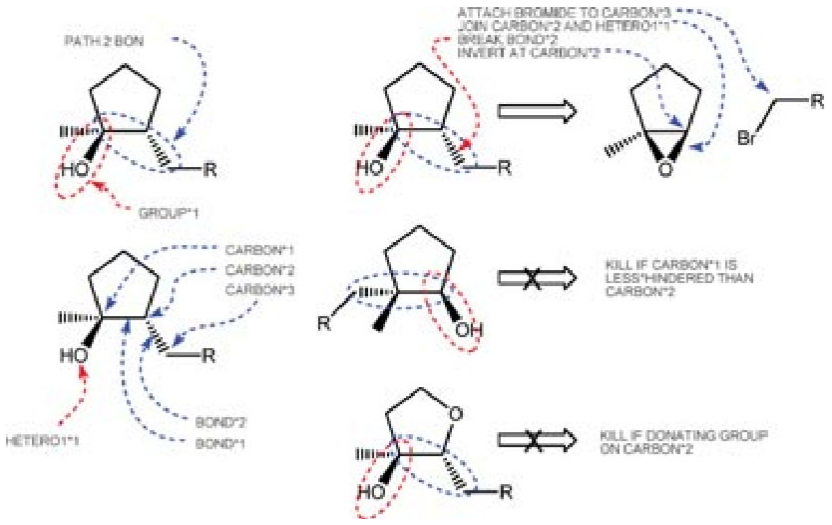
For analysis ring topology online:
http://www.maxbridge.org/

Science, **1985**, *228*, 408; *JACS*, **1975**, *97*, 6116; *JACS*, **2019**, *141*, 13713

Five Retrosynthetic Strategies in LHASA

- **Stereochemical strategies:** Stereoselective reactions, or steric based arguments used to reduce stereocomplexity
- **Functional group-oriented strategies**: Functional group interconversions and determining logical disconnections based off of functional group arrangement(s)

```
TRANSFORM 569
NAME GRIGNARD OPENING OF EPOXIDE
...PATH 2 BONDS
RATING 35
GROUP*1 MUST BE ALCOHOL
...
KILL IF DONATING GROUP ON CARBON*2  ...unstable epoxide
....
   ATTACH BROMIDE TO CARBON*3
   JOIN CARBON*2 AND HETERO1*1
   BREAK BOND*2
   INVERT AT CARBON*2
....
KILL IF CARBON*1 IS LESS*HINDERED THAN CARBON*2 ... undesired attack
```

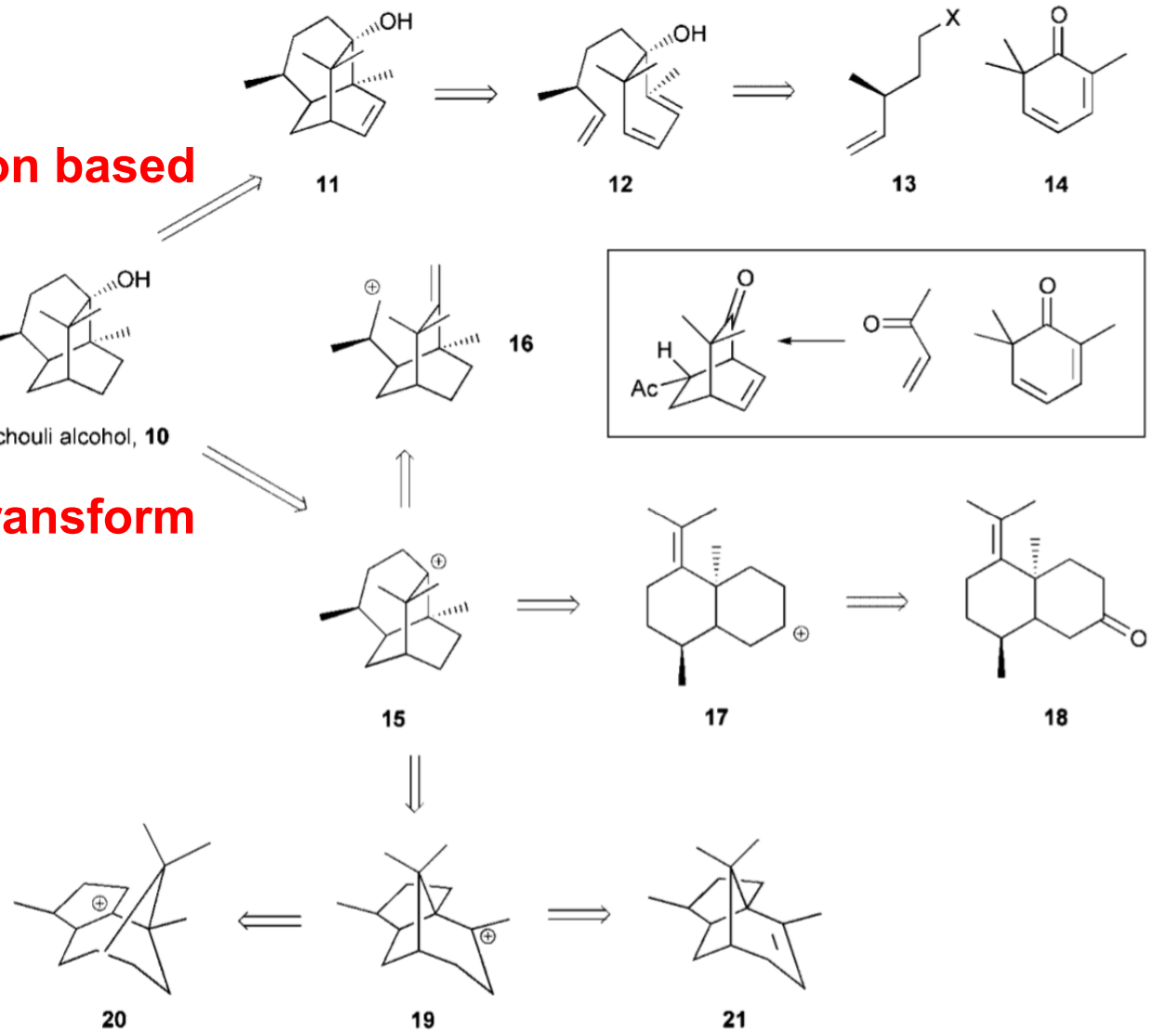


Science, **1985**, *228*, 408

Example of LHASA's Retrosynthetic Analysis

Name reaction based

Mechanistic transform



Review of LHASA

Year	1967–1997?
Author	E. J. Corey (Harvard)
Database	Template, 2100+ reactions
Language gramma	CHMTRN (**CH**e**M**istry **TR**a**N**slation)
Interactive	Yes
Score	Chemist
Retrosynthesis	5 Strategies guided
Retrosynthesis depth	~15 to key transformation
Predict new reaction	No
Perspective	A large expert system; Time-consuming but suitable for sophisticated organic chemist

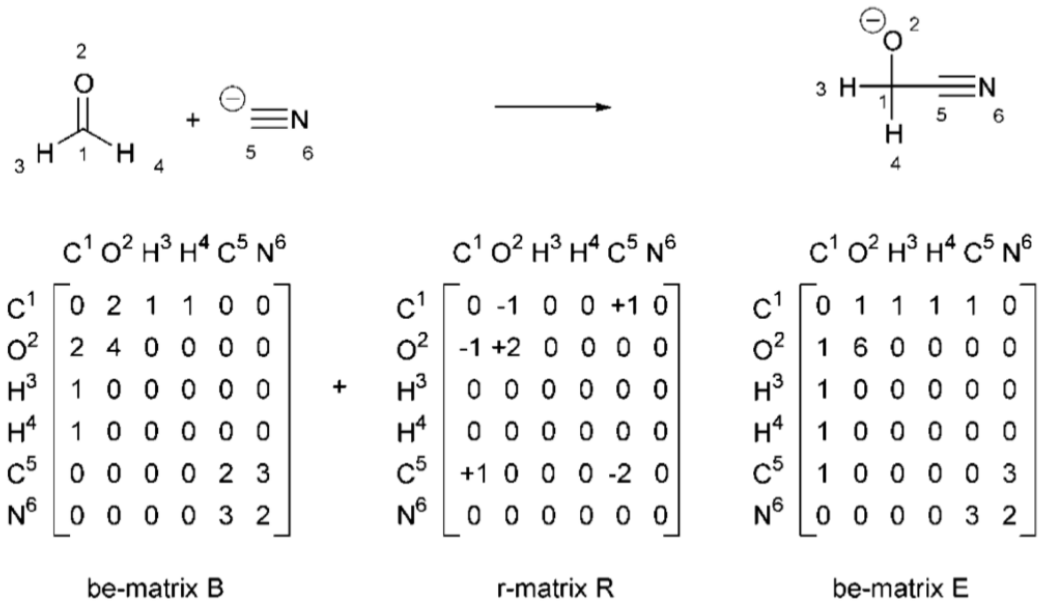
Chem. Soc. Rev., **2005**, *34*, 247

CASP Softwares Later

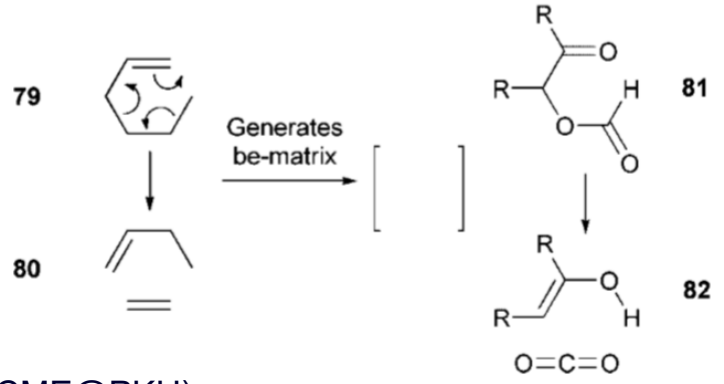
- SECS (Wipke, 1976)
- SYNCHEM (Gelernter, 1977)
- SYNLMA (Johnson, 1989)
- SYNGEN (Hendrickson, 1989)
- CHIRON (Hanessian, 1990)
- IGOR (Ugi, 1993)
- WODCA (Gasteiger, 1995)
- Etc.

Limitations:

- Too simplified rule sets
- Incompatible synthetic routes
- Poor computing power



Discovering Unprecedented Reaction



Chem. Soc. Rev., **2005**, *34*, 247

Network of Organic Chemistry (NOC)

a)

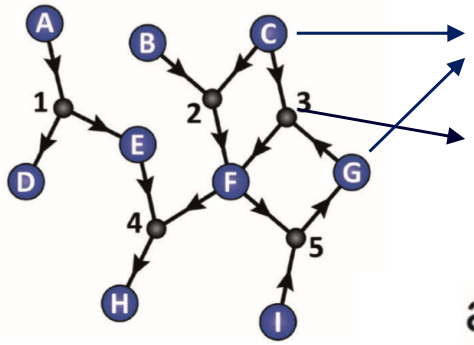
Cities		
Gdańsk	–	Warszawa
Gdańsk	–	Poznań
Szczecin	–	Poznań
Warszawa	–	Wrocław
Warszawa	–	Katowice
Warszawa	–	Kraków
Poznań	–	Wrocław
Poznań	–	Warszawa

- Classical Reaction Databases:
  C → F, G → F
- Network of Organic Chemistry:
  C + G –*3*→ F

b)

Reactions

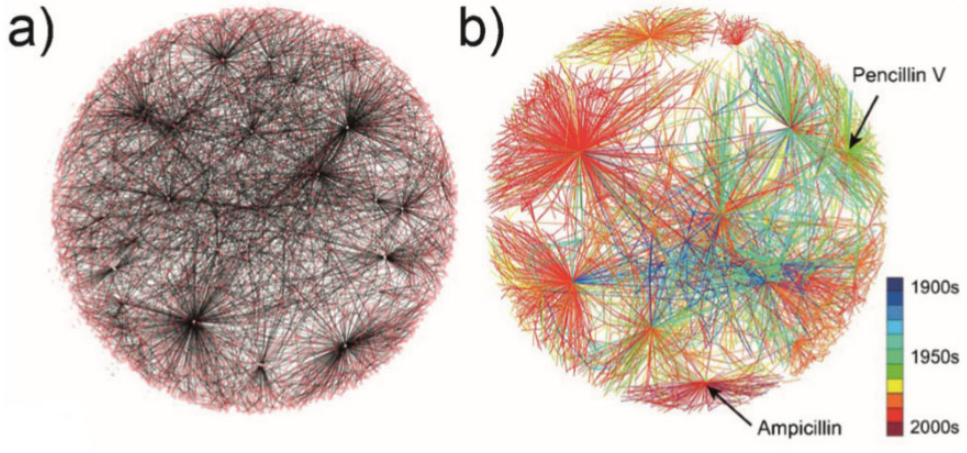
1) A → D + E
2) B + C → F
3) C + G → F
4) E + F → H
5) F + I → G



substrate/product molecule

reaction operations

The complete network of chemistry in Chematica is more than 1000 times larger than human metabolic network!



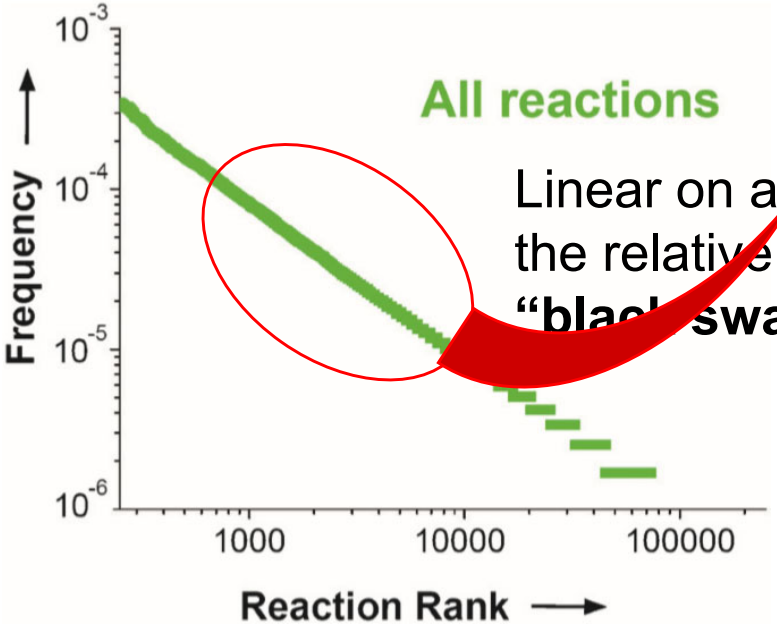
Angew. Chem. Int. Ed., **2016**, *55*, 5904

Key Challenges for Expert Systems

- …we know that certain complex types of computational problems cannot be simplified too much… Synthetic planning cannot be done by teaching computer few hundreds of general rules or working by analogy to literature-reported reactions. Computer has to be taught an enormous [illegible] trained how to use them, and be able to explore [illegible] s before its true power manifests itself.

> 10000 rules required

- Analysis of 1.2 million literature [illegible] ported reactions randomly in NOC:

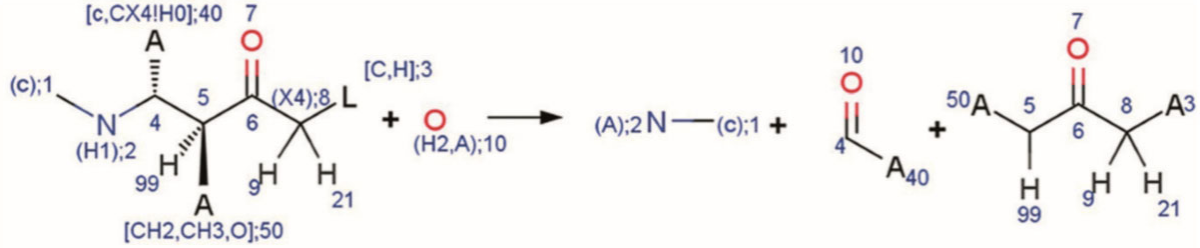


Linear on a doubly-logarithmic scale:
the relative importance of reactions that occur infrequently;
"black swan" events

Angew. Chem. Int. Ed., **2016**, *55*, 5904

Syntaurus in Chematica

- ARChem: Machine extract > 100k reaction rules from literature precedents?
- **Synthesis planned by analogy, meaningless.**



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"

Angew. Chem. Int. Ed., **2016**, *55*, 5904

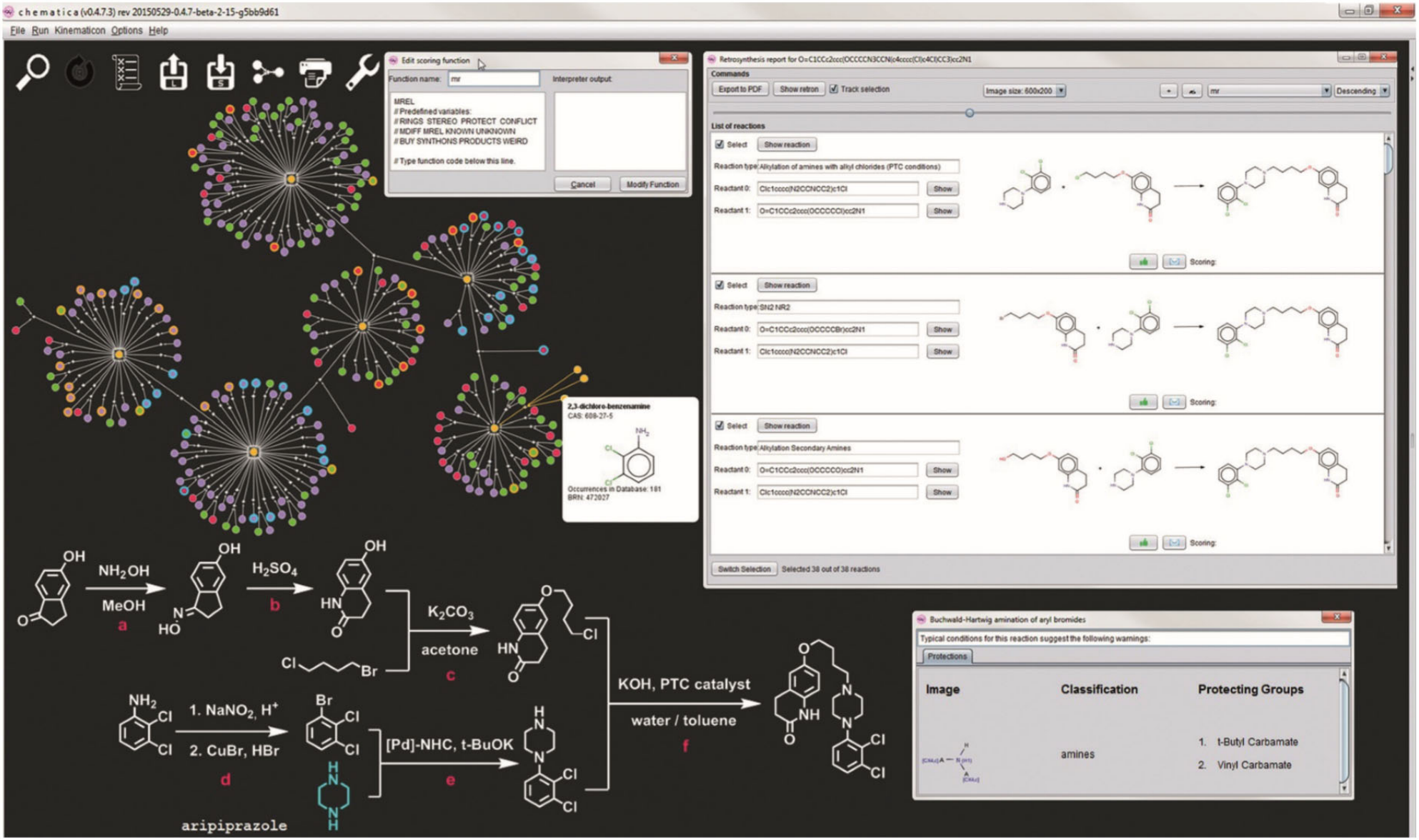
Chemicals' and Reaction Scoring Functions

- Chemical's Scoring Function (CSF), "synthetic positions":
  - Number of rings/stereocenter:

    Creating as many rings/stereocenter as possible
  - Synthon's mass:

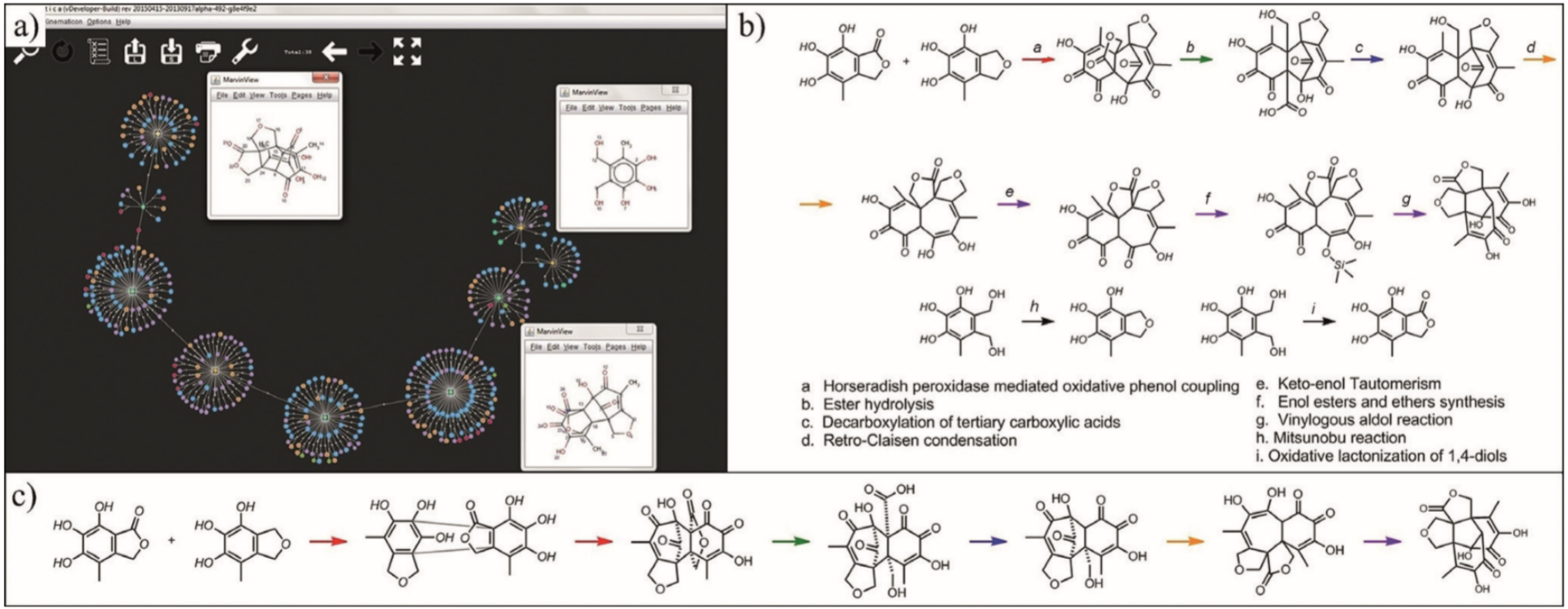
    Splitting into equivalent mass portions is encouraged
- Reaction Scoring Function(RSF), "synthetic moves":
  - Necessity of protection
  - Group incompatibility
  - Theoretically estimated yields

Interface of Chematica

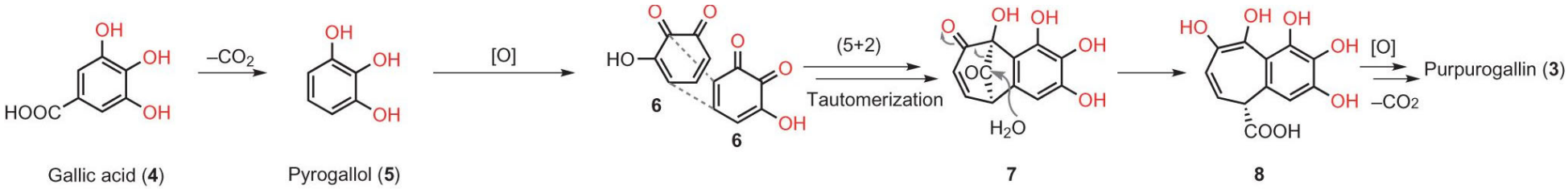


Angew. Chem. Int. Ed., **2016**, *55*, 5904 Luo Group Meeting (CCME@PKU)

Example of Chematica's Synthetic Route



Trauner's total synthesis route:



Angew. Chem. Int. Ed., **2016**, *55*, 5904; *Nature Chemistry*, **2015**, *7*, 879

Review of *Chematica*

Year	2001–present
Author	Bartosz A. Grzybowski (Ulsan National Institute of Science and Technology)
Database	*ca.* 20k hand-coded rules
Language gramma	*Syntaurus* (SMILES/SMART)
Interactive	No
Score	Scoring functions
Retrosynthesis	*Syntaurus*
Retrosynthesis depth	Full Automated
Predict new reaction	No
Perspective	A giant expert system; Automated, conformation not considered

Another Way…

Machine Learning

- Definition:

  A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with the experience **E**.

- Tasks:

  Regression *vs* Classification

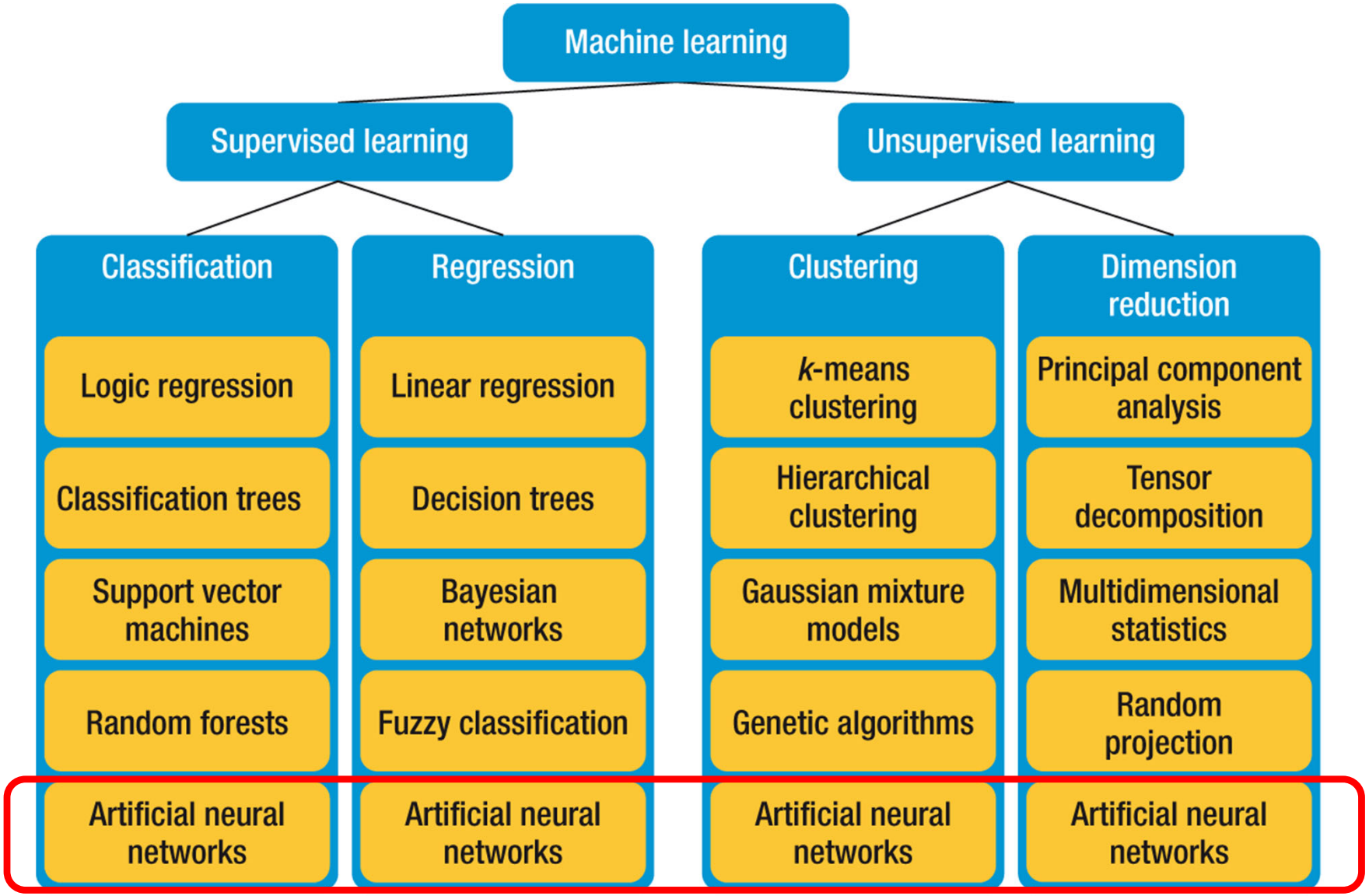
- Algorithms:

  Supervised learning (with labeled answer)

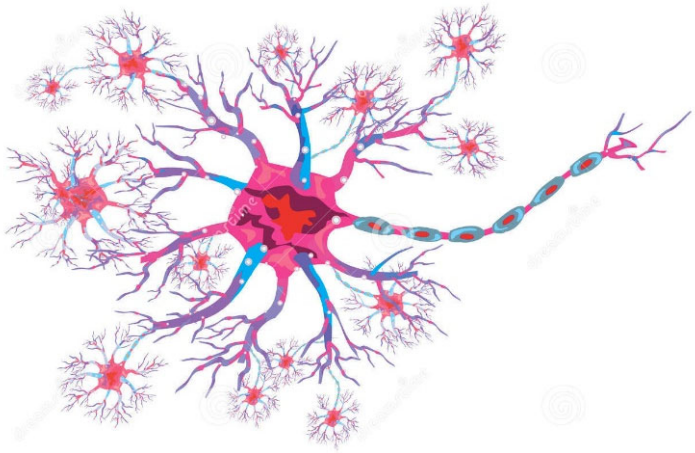
  Unsupervised learning (data mining from unlabeled data)

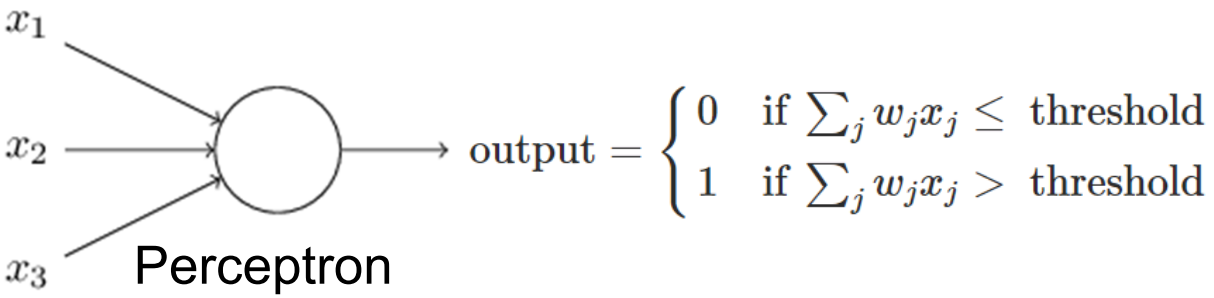
  Reinforcement leaning (maximizing rewards, *e.g.* AlphaGO)

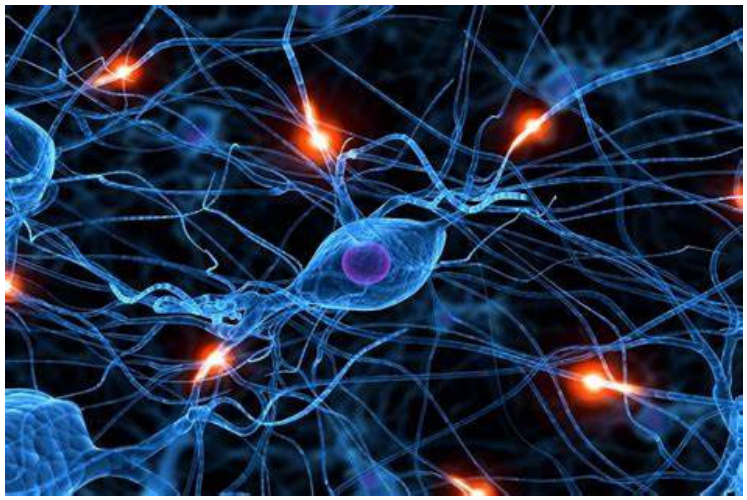
Machine Learning Algorithms

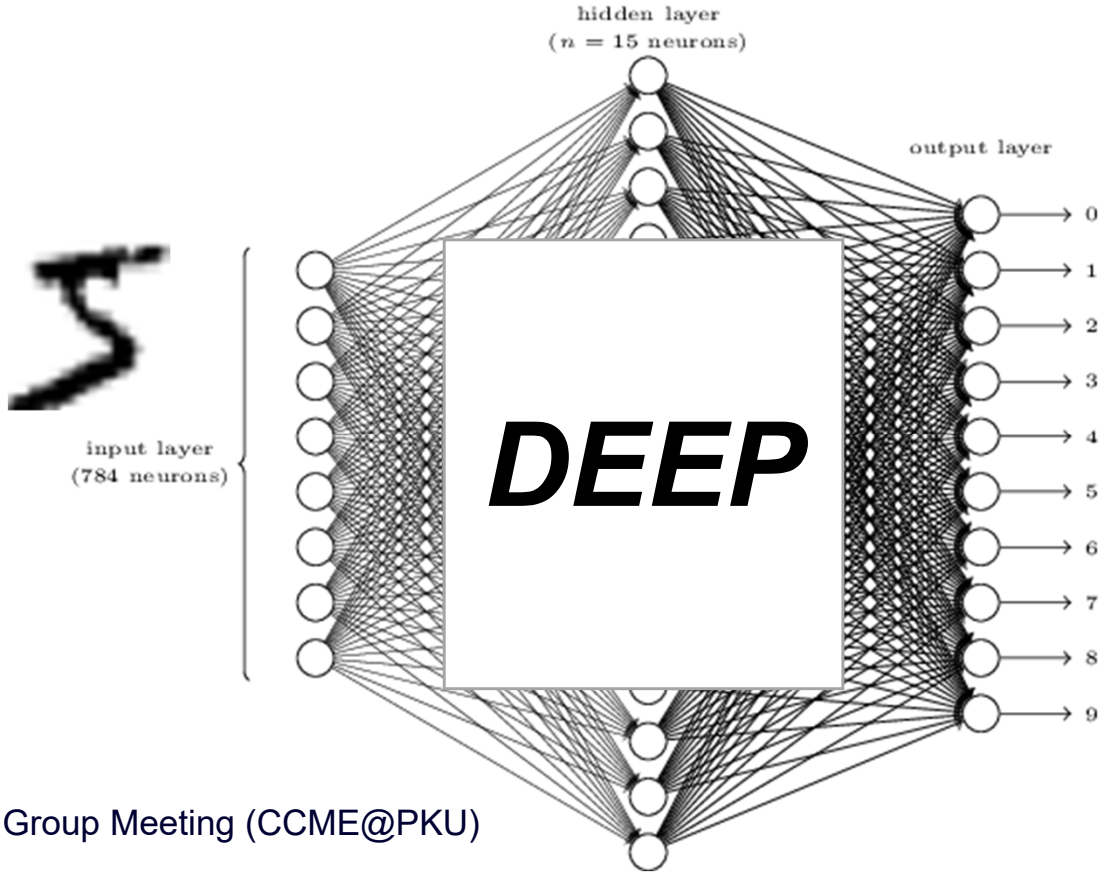


Neural Network and Deep Learning



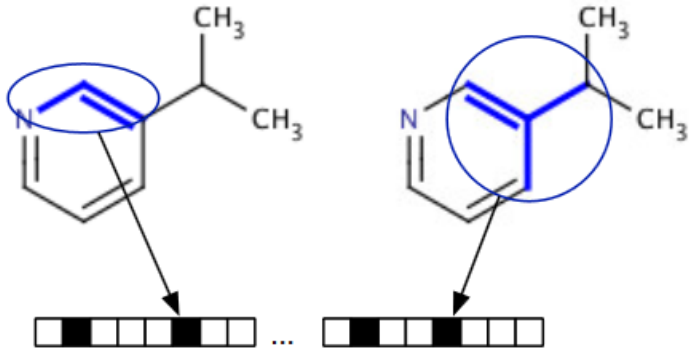






Machine Learning Approaches from Fingerprints

- Molecular descriptors (fingerprints) character similarity between molecules in chemical informatics.
- Physical descriptors
  - Molecular weight
  - Number of rings
  - Partial charge
  - …
- Predicted properties
  - Morgan fingerprint
  - Coulomb matrix
  - Radial distribution functions
  - …

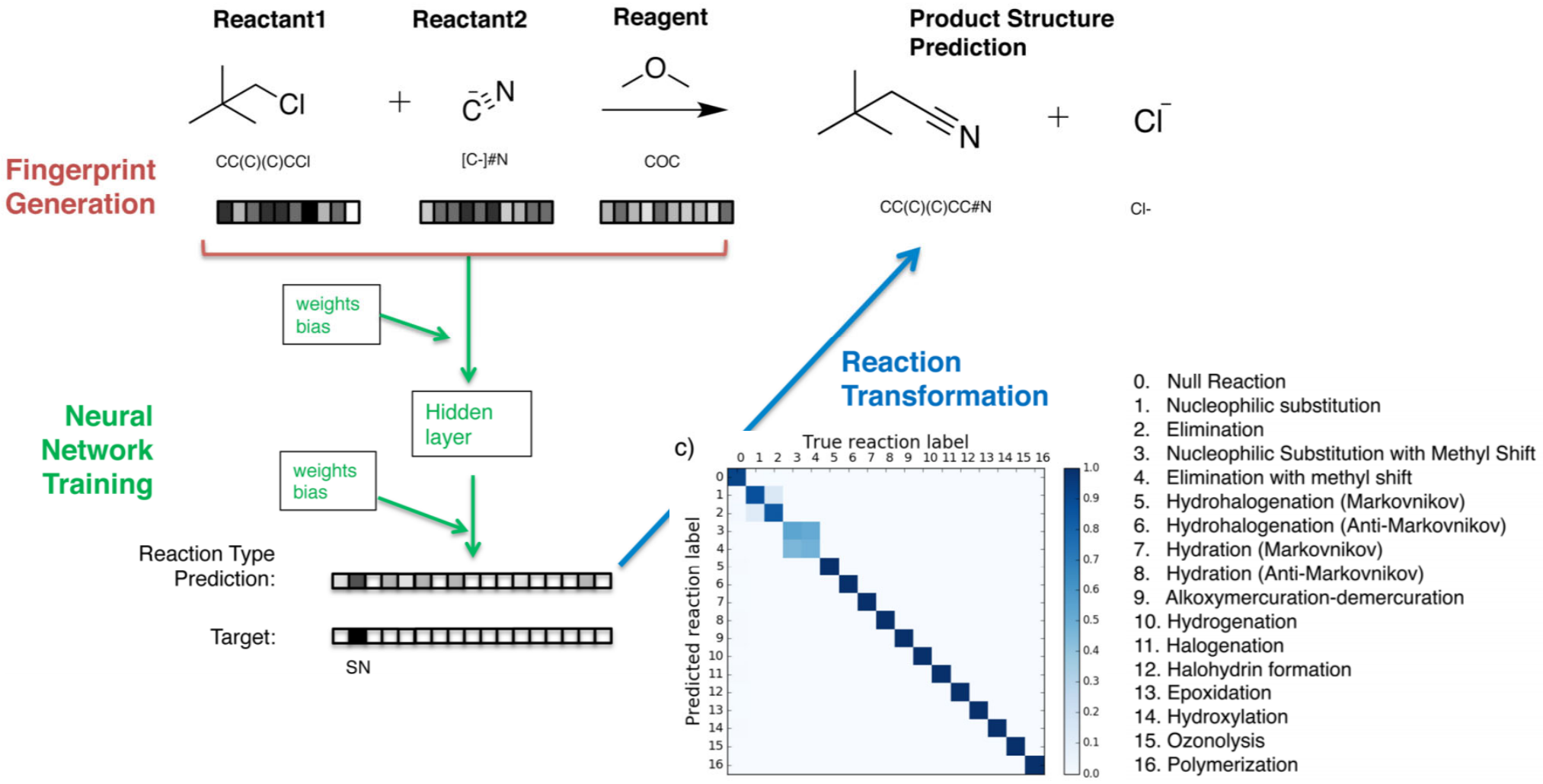


Predict molecular properties by comparation

HOMO-LUMO gaps
Ligand binding affinity
Binding Affinity
Reactivities

ACS Cent. Sci., **2016**, *2*, 725

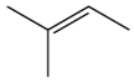
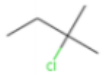
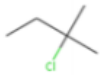
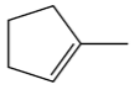
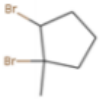
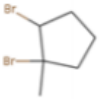
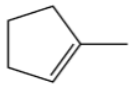
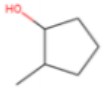
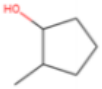
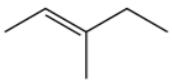
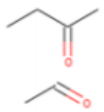
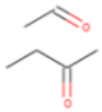
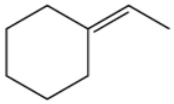
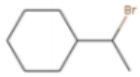
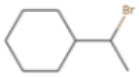
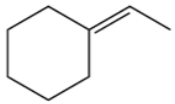
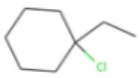
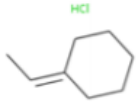
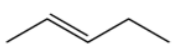
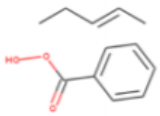
Prediction of Organic Chemistry Reactions



ACS Cent. Sci., **2016**, *2*, 725

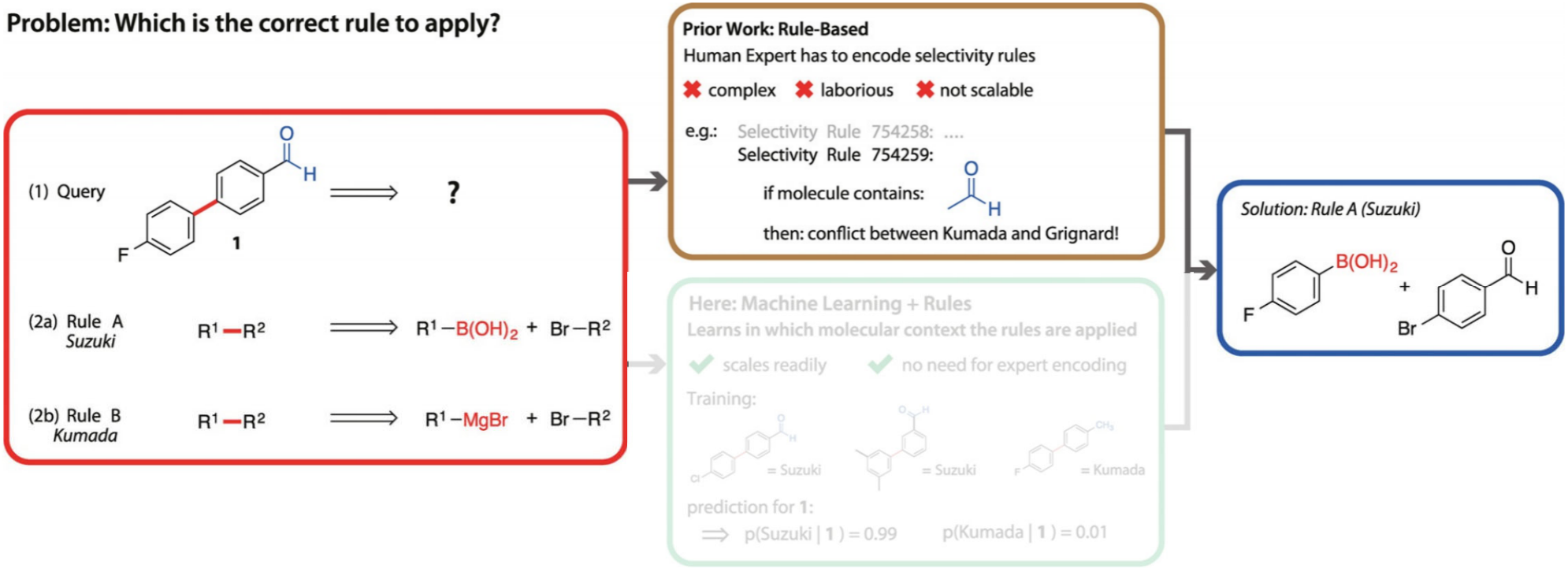
Attempts to Solve Textbook problems

- Problems from Wade's *Organic chemistry*, 6th ed.

		True Product	Major Predicted Product	Morgan Weighted Tanimoto Score	Estimated probability of correct reaction type
a)	 HCl →			0.9998	1.000
b)	 Br_2, CCl_4 →			0.8863	0.791
c)	 BH_3 THF → H_2O_2 OH-			0.8554	0.748
d)	 O_3 → $(CH_3)_2S$			0.9999	1.000
e)	 HBr → H_3COOCH_3			0.9999	1.000
f)	 HCl → H_3COOCH_3			0.3540	0.001
g)	 $PhCO_3H$ →			0.4296	0.073

ACS Cent. Sci., **2016**, *2*, 725

Retrosynthesis based on ML and Rules



Chem. Eur. J., **2017**, *23*, 5966

Retrosynthesis based on ML and Rules

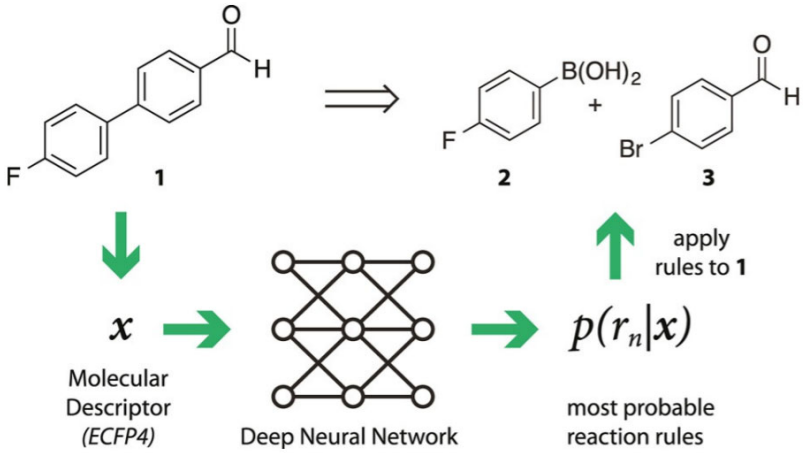
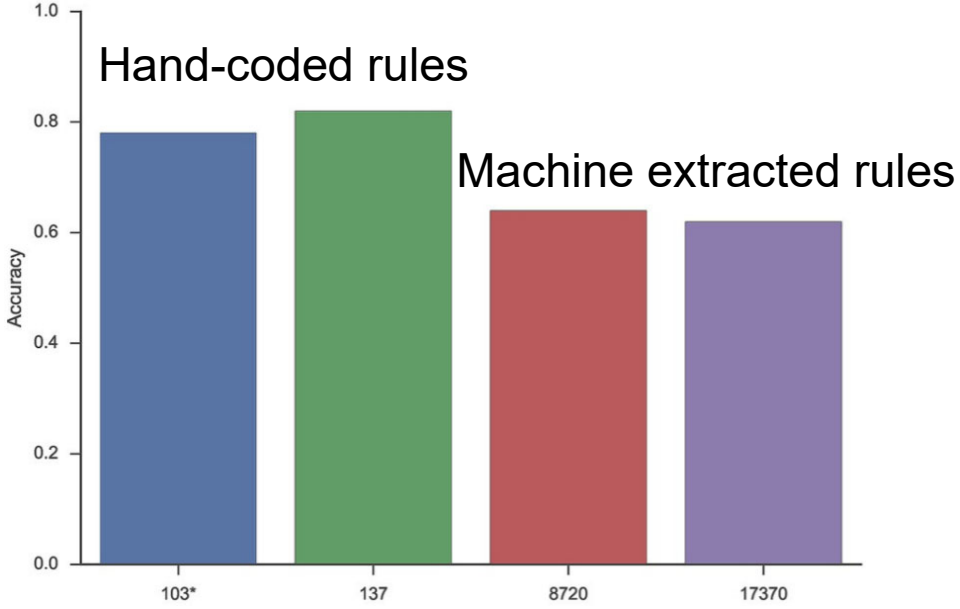
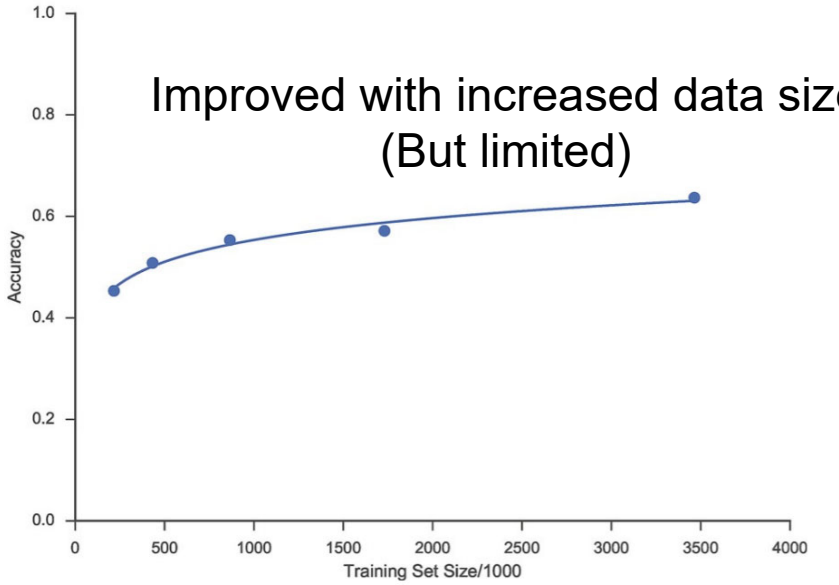


Table 1. Results for the study on 103 hand coded rules.

Task/Model	Acc	Top 3-Acc	MRR	W. Prec.
Reaction prediction				
random	0.03	0.12	0.04	0.03
expert system	0.07	0.33	0.12	0.46
logistic regression	0.86	0.97	0.91	0.86
highway network	0.92	0.99	0.96	0.92
FC512 ELU	0.92	0.99	0.96	0.92
Retrosynthesis				
random	0.03	0.12	0.04	0.03
expert system	0.05	0.30	0.06	0.11
logistic regression	0.64	0.95	0.77	0.62
highway network	0.77	0.98	0.86	0.77
FC512 ELU	0.78	0.98	0.87	0.78

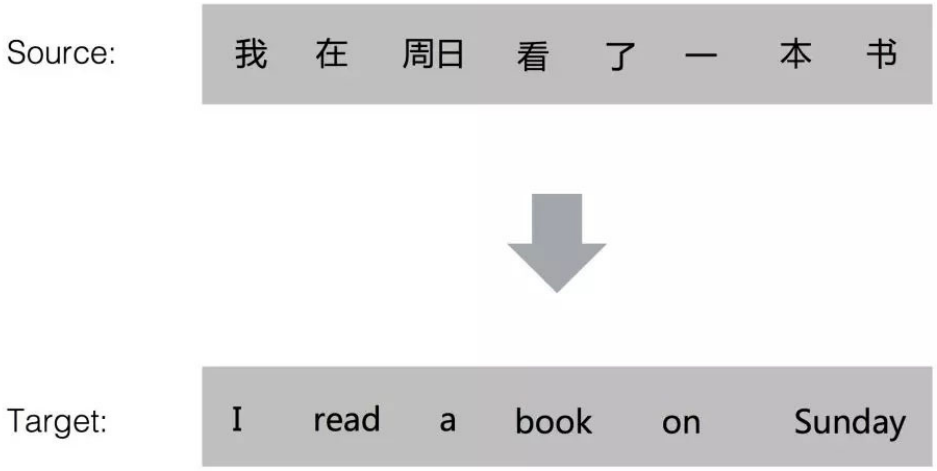




Chem. Eur. J., **2017**, *23*, 5966

Retrosynthesis based on Translation Model

- Natural Language Processing (NLP):



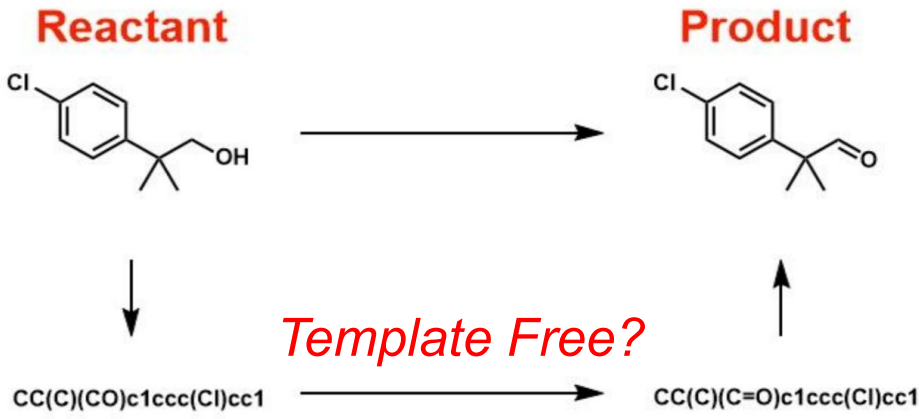
Classical machine translation:
Based on Rules (Dictionary)

"我": → "I", <*noun*>, subject

…

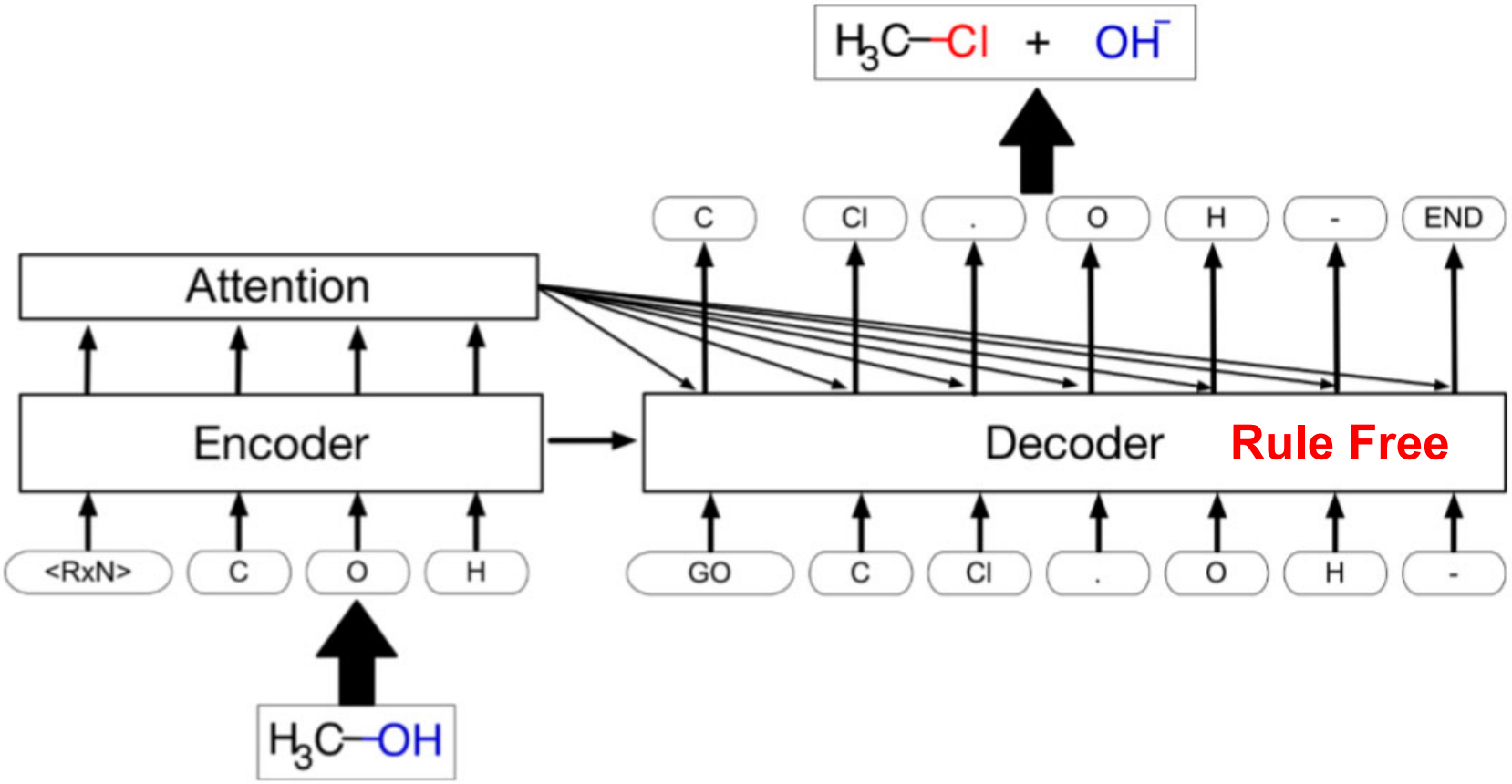
Modern machine translation:
Based on Statistics (Template Free)

$$P(y \mid x; \theta) = \sum_z \frac{\exp(\theta \cdot \phi(x,y,z))}{\sum_{y'} \sum_{z'} \exp(\theta \cdot \phi(x,y',z'))}$$



Sequence to Sequence (seq2seq):
Encoder–decoder Model

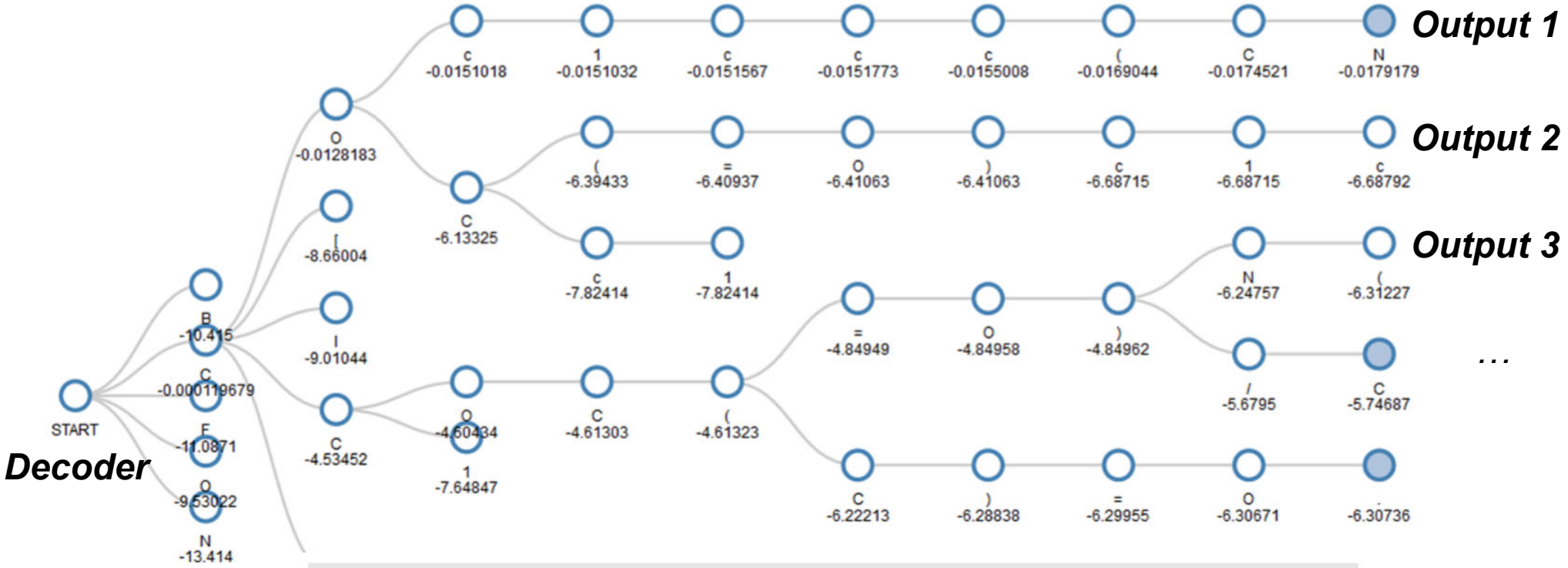
Seq2seq Model in Retrosynthesis



ACS Cent. Sci., **2017**, *3*, 1103

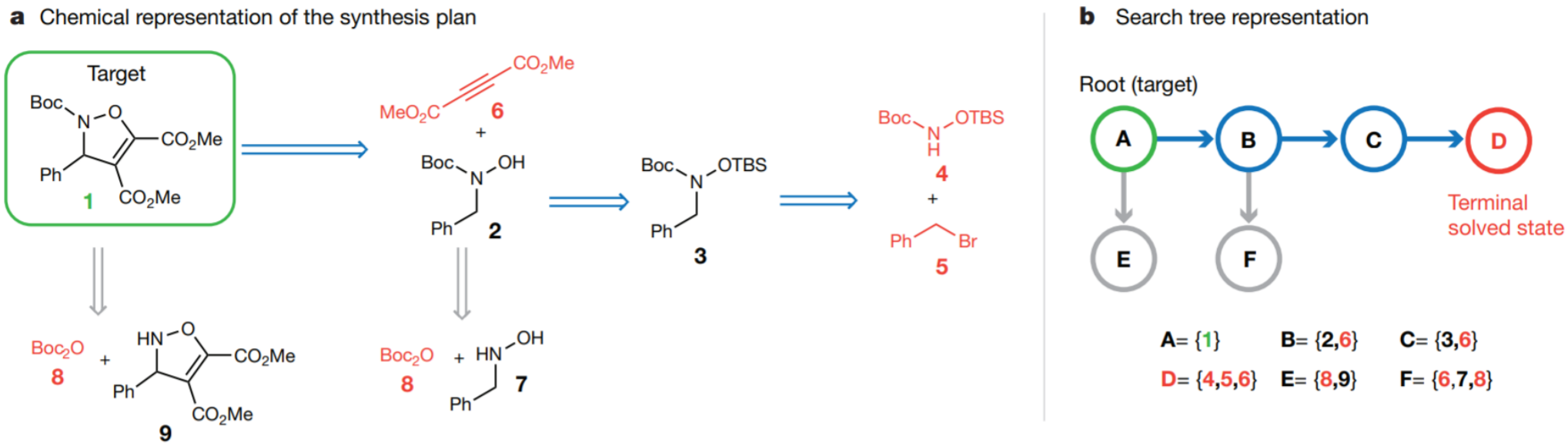
Seq2seq Model in Retrosynthesis

- **A partially completed beam search procedure**



		top-N accuracy (%)					
	model	top-1	top-3	top-5	top-10	top-20	top-50
Machine extracted rules	baseline	35.4	52.3	59.1	65.1	68.6	69.5
Template free	seq2seq	37.4	52.4	57.0	61.7	65.9	70.7

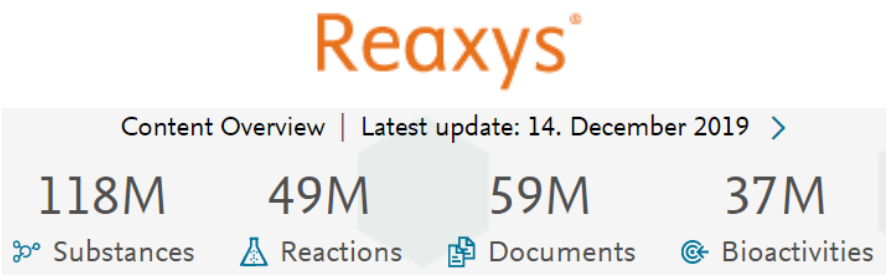
Searching Space in Retrosynthesis



Locating disconnecting position:

- (Previous) Heuristic best first search (BFS), difficult in: **Large searching space**
  - Chemists tend to **disagree** on what constitutes a good position
  - Temporarily **increase complexity** by the use of protecting or directing groups
  - The position value depends highly on the availability of suitable **precursors**
- Monte Carlo tree search (MCTS): **Reinforcement leaning – MCTS-3N**
  - Random steps
  - Accept policies: $p(t|s)$; t: transformation, s: position
  - Trained to predict the winning move

Training the Policies and Filter Network



12.4 million single step reaction

- **Rollout set: (17,134 rules)**

  Contain the atoms and bonds that changed in the course of the reaction and the first-degree neighboring atoms.

  Only rules that occurred at least 50 times in reactions published before 2015 were kept.
- **Expansion set: (301,671 rules)**

  Only the reaction center was extracted (more general)

  Rules occurring at least three times were kept
- **In-scope filter network: (classifier of unsuccessful reaction)**

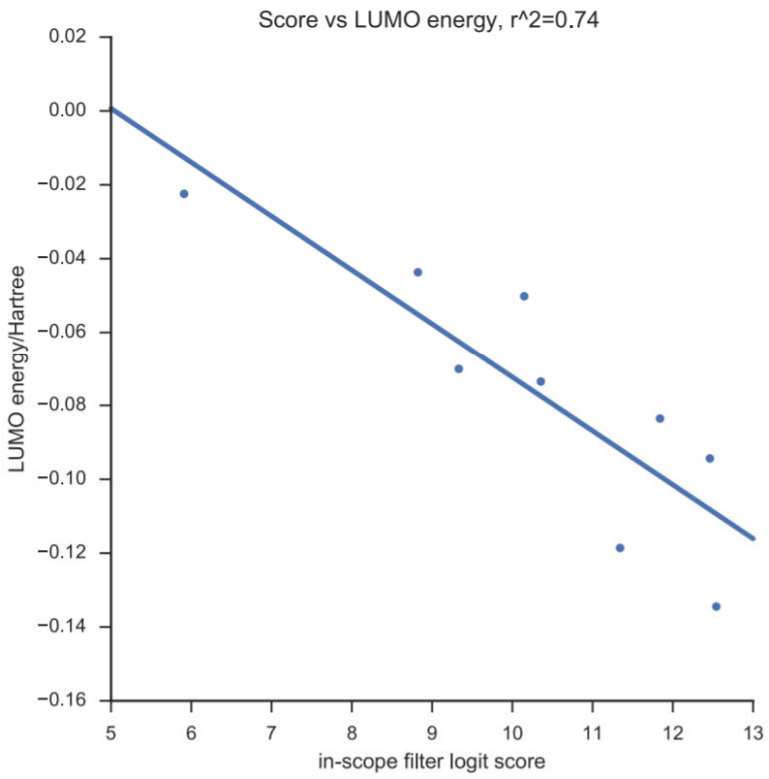
  For high-yielding reaction: A + B → C, hypothetical products D, E,... are not formed

  Generated 100 million negative reactions

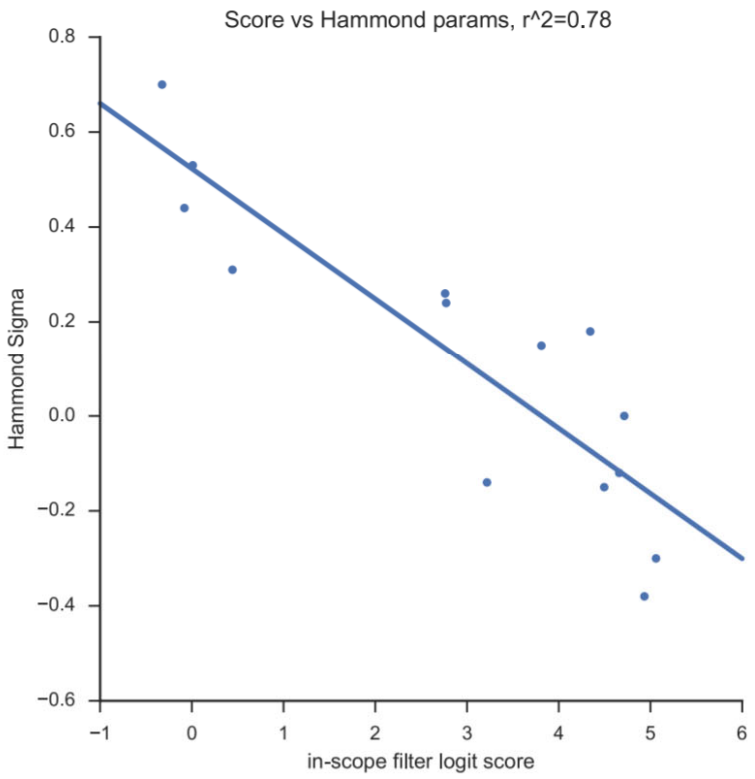
Nature, **2018**, *555*, 604

Filter Correlates with Electronic Properties

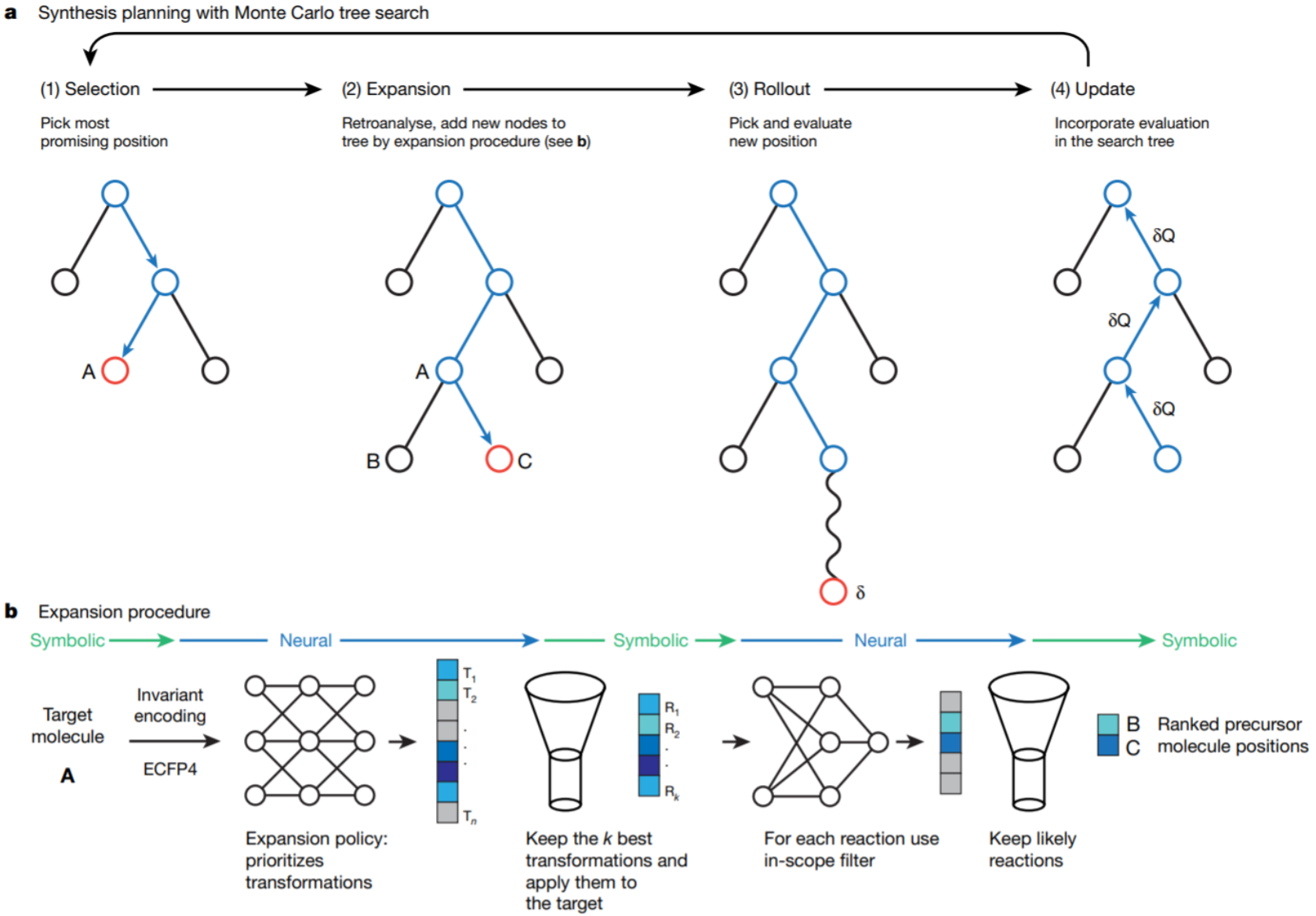
a) Diels-Alder reactions with Cyclopentadiene



b) para-Bromination of benzenes

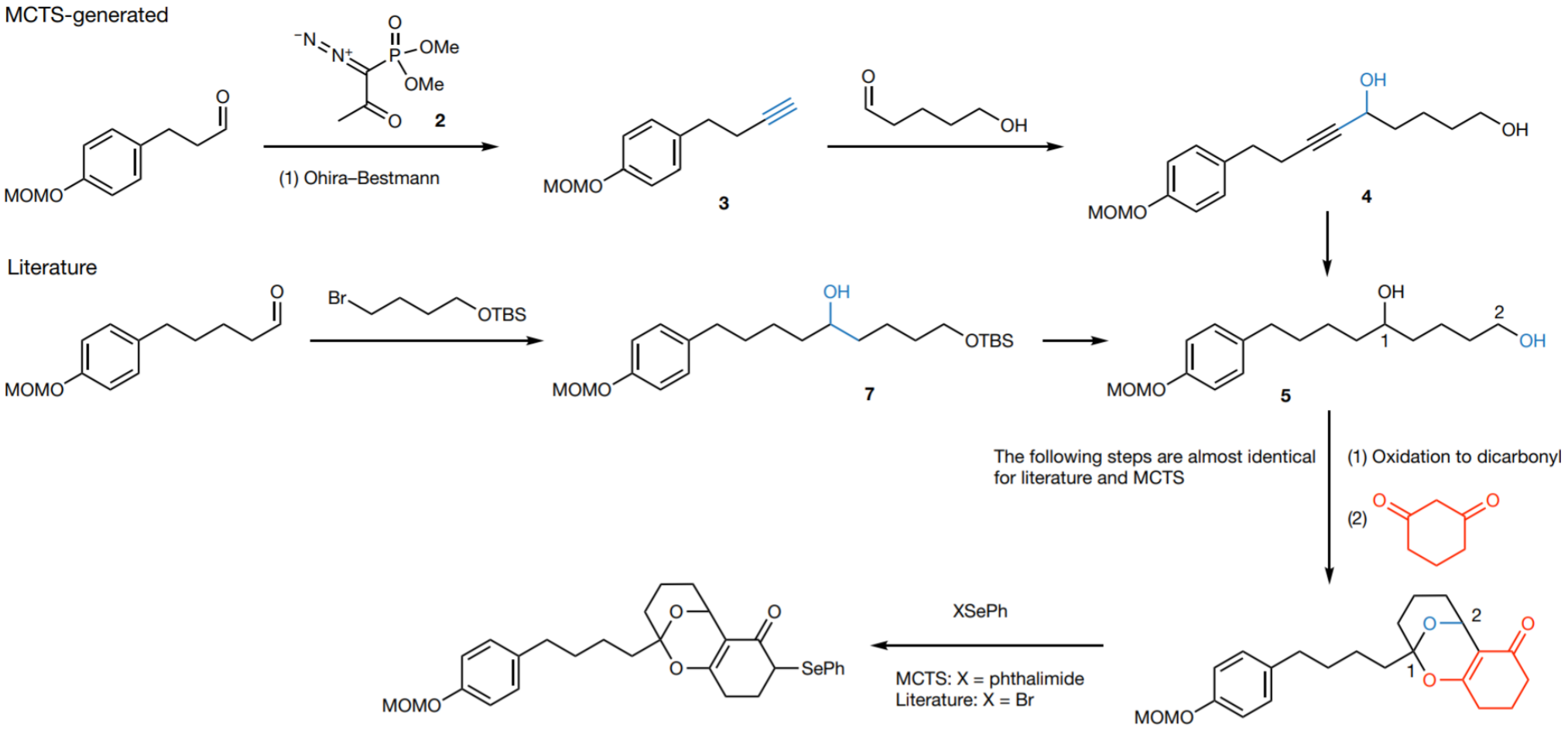


Integrating Neural Networks and MCTS

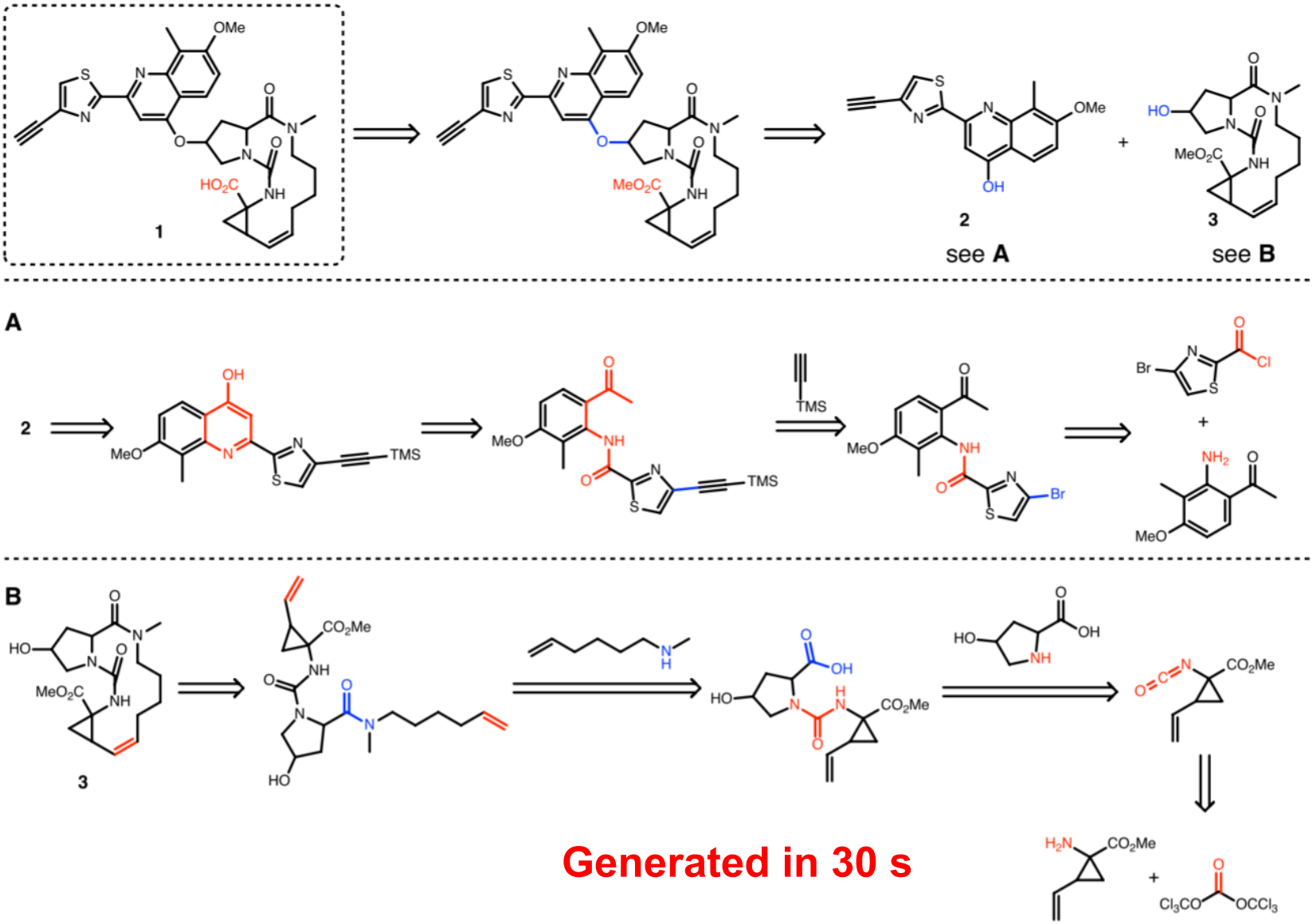


Performance of MCTS-3N

c Why did chemists prefer the literature over MCTS in task 1 of test a?



Example of a 10-step synthesis



Review of Deep Learning in Retrosynthesis

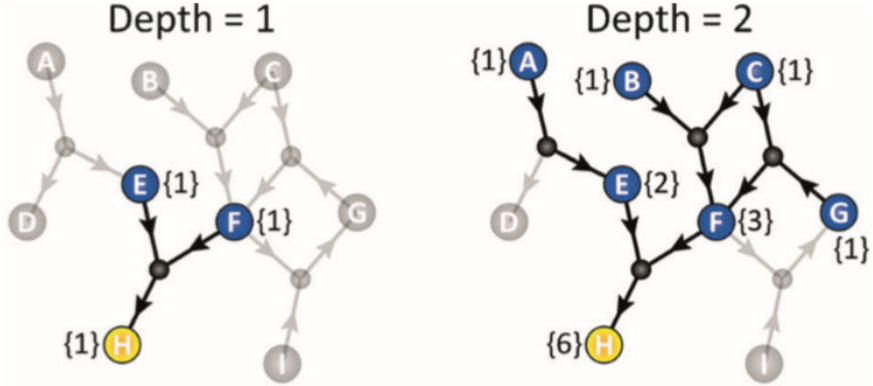
Year	2018
Author	Mark P. Waller (Shanghai University)
Database	Automated extracted from Reaxys (12.4 Million rxns)
Language gramma	SMILES/SMARTS
Interactive	No
Score	Neural Network
Retrosynthesis	MCTS-3N
Retrosynthesis depth	*infinite*
Predict new reaction	No
Perspective	Learn reaction rules from data; Few examples *in logic-centered* synthesis

Summary and Outlook

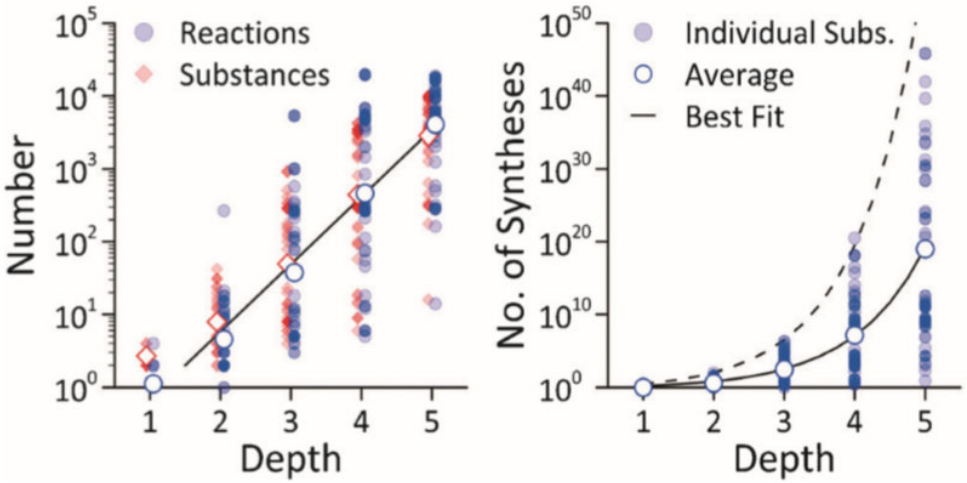
- Expert System:
  - Chemist-friendly
  - Precise but trouble-to-code rules
  - Poor scoring function
- Machine Leaning:
  - Preparation from big-data directly
  - Generation synthetic route rapidly
  - Lacking chemical meaning, black-box in algorithms
- Challenges in retrosynthesis of natural product:
  - Insufficient study than drug molecules
  - Scaffold complexity and diverse reactivity from small molecules
  - Require developing new methodology

Scoring Functions and Searches Methods

- Counting possible syntheses:



- Network searches in 51 different target:



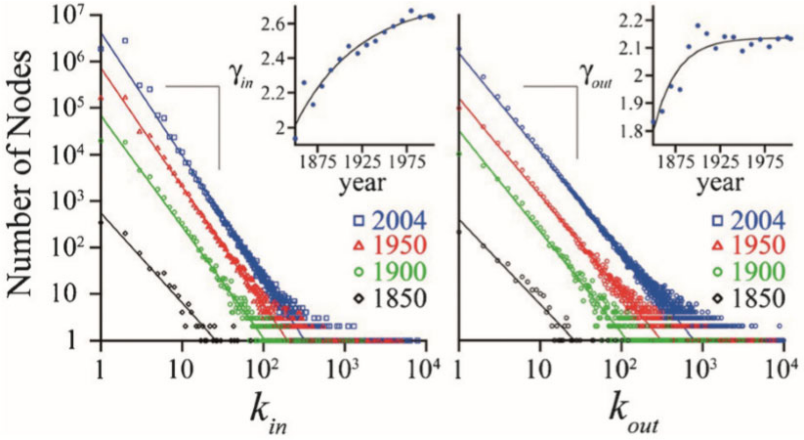
- Cost Function:

$$C_{\text{tot}} = C^{\text{o}}_{\text{rxn}} N_{\text{rxn}} + \sum_i C_{\text{sub}}(i)$$

labor, overhead, purification procedures

commercially available starting materials

- Popularity Function



$$P_{\text{tot}} = \sum_i 1/k(i)$$