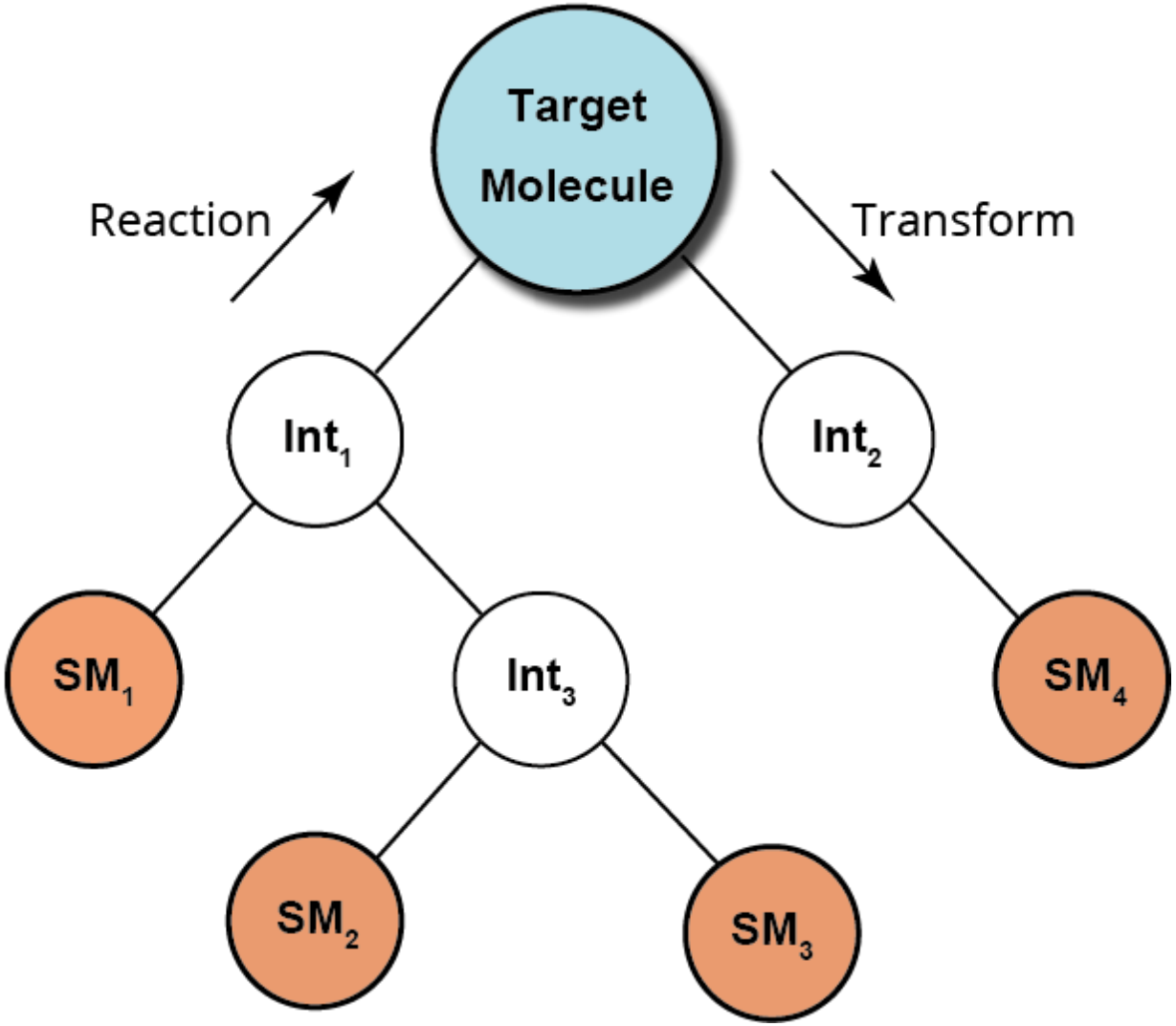
Computer-Assisted Retrosynthesis

2018/06/02

M1 Koki Sasamoto

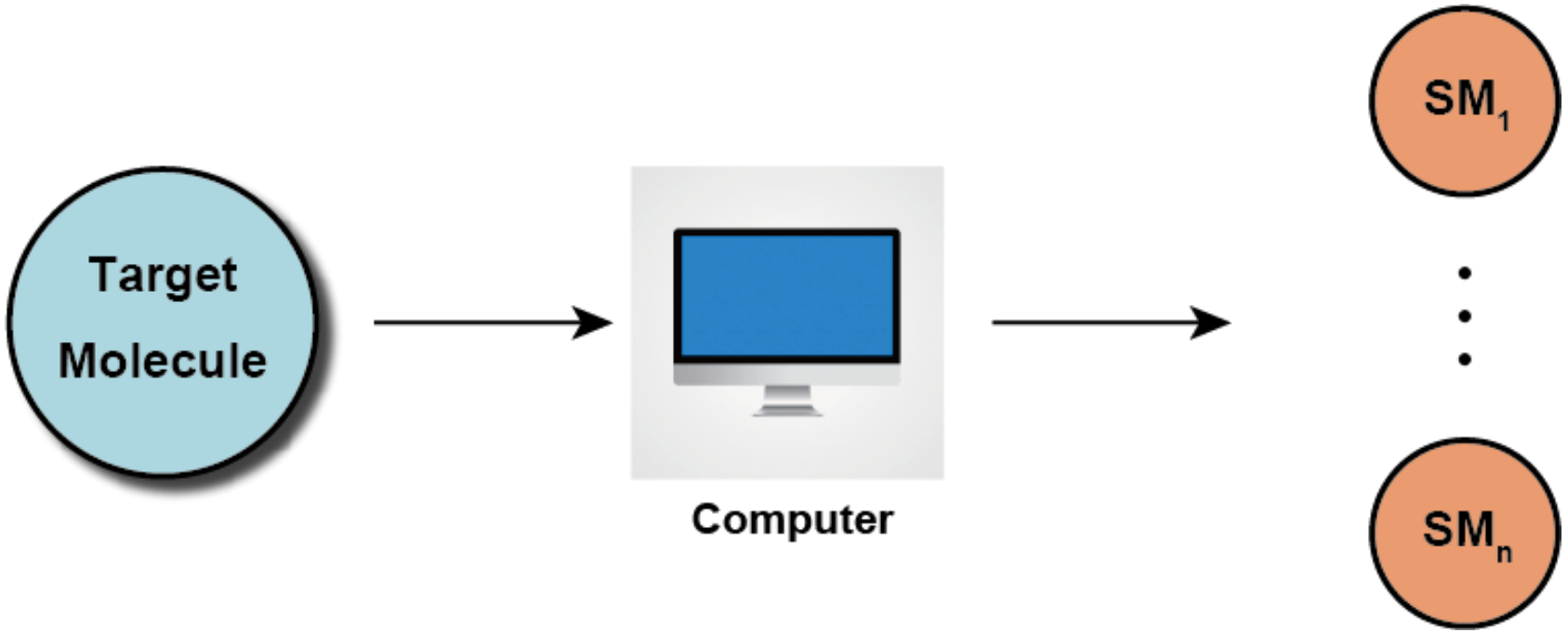
Introduction

- **Retrosynthesis**



Introduction

- **Computer-assisted synthesis planning (CASP)**

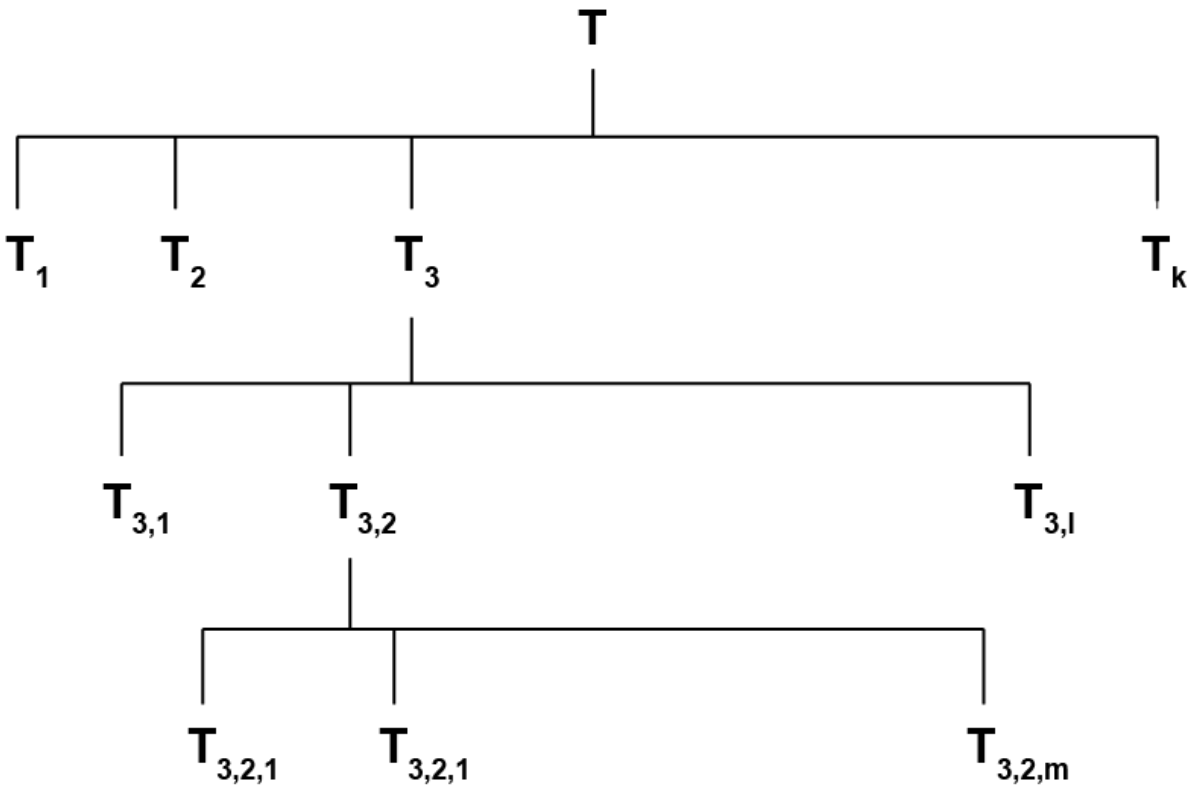


- ✓ It takes less time to devise synthetic routes.
- ✓ Proportion of successful synthesis rises.
- ✓ Scientists learn from the results of CASP.

Contents

LHASA

- **Interactive system using synthetic tree**

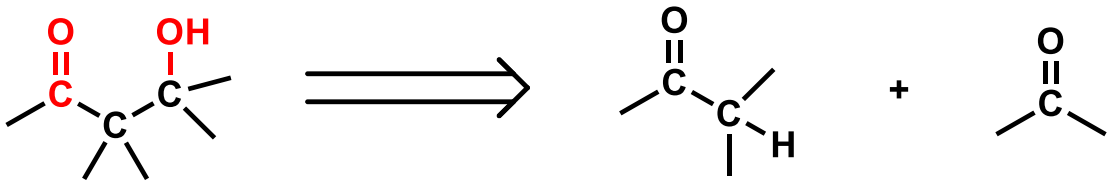


Synthetic Tree

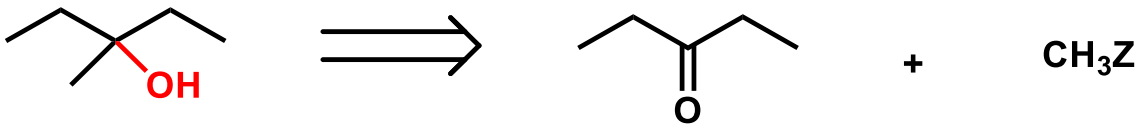
E. J. Corey and W. T. Wipke, *J. Am. Chem. Soc.*, **1972**, *94*, 431.

Transform Mechanism

- **Transform lists**
  - Two-group transform



  - One-group transform

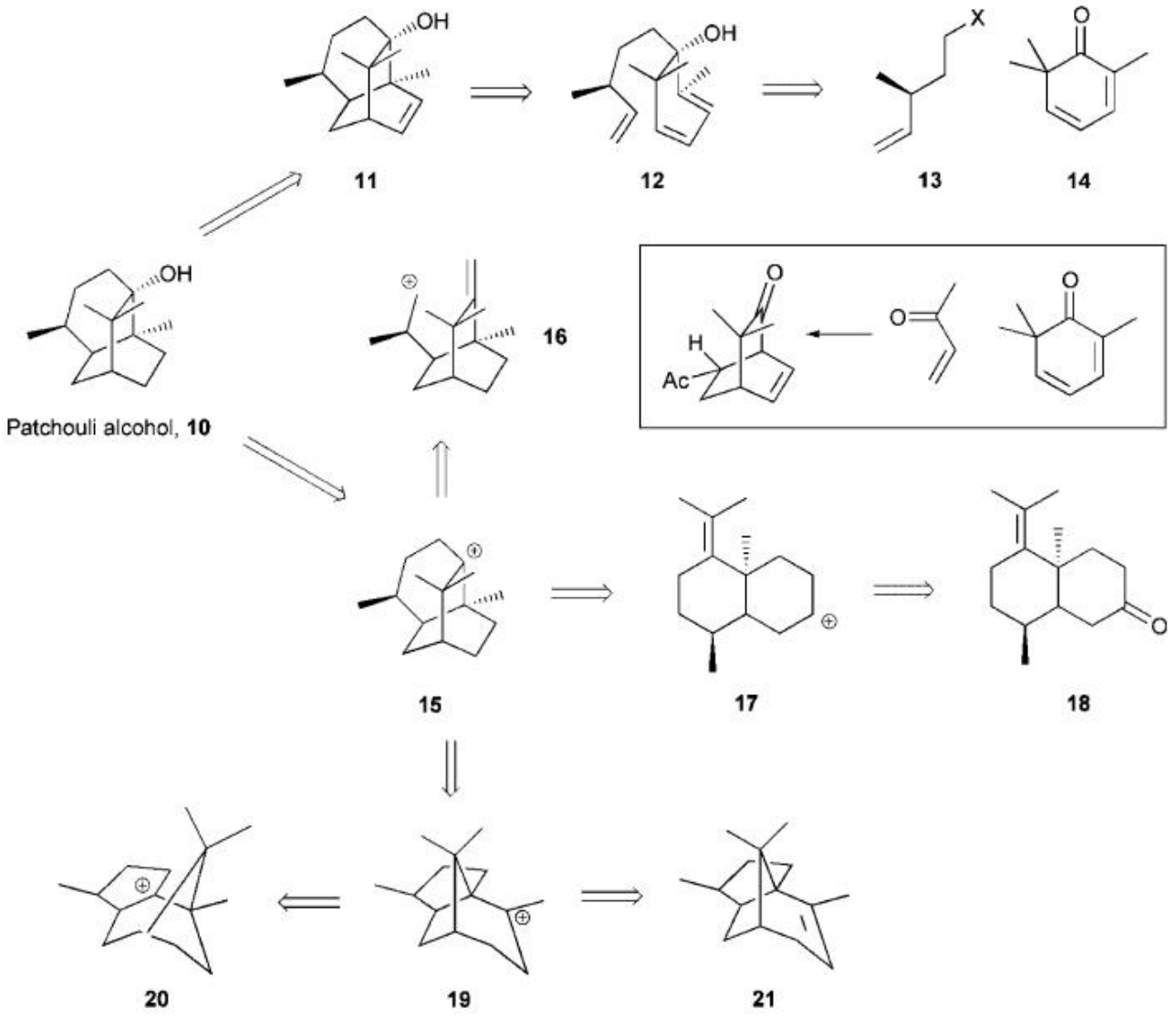


  - Functional Group Interchange (FGI) … etc.

- **Each transformation has data table.**
  - Which bond is cleaved
  - Rating depend on difficulties … etc.

E. J. Corey and W. T. Wipke, *J. Am. Chem. Soc.*, **1972**, *94*, 431.

Retrosynthesis Example



E. J. Corey and W. T. Wipke, *J. Am. Chem. Soc.*, **1972**, *94*, 431.

Other CASP Programs

- **Failure of CASP**

  SECS, SYNCHEM, SYNGEN, IGOR, WODCA, etc…

  ✖ These provided incompatible synthetic routes.

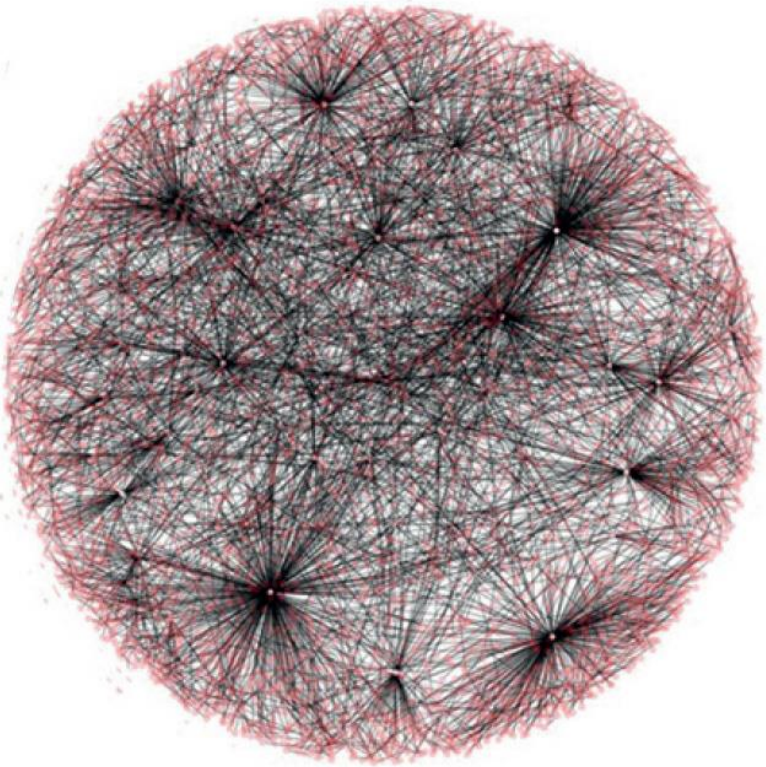
- **Lack of computing capacity**

  only having simplified rule set

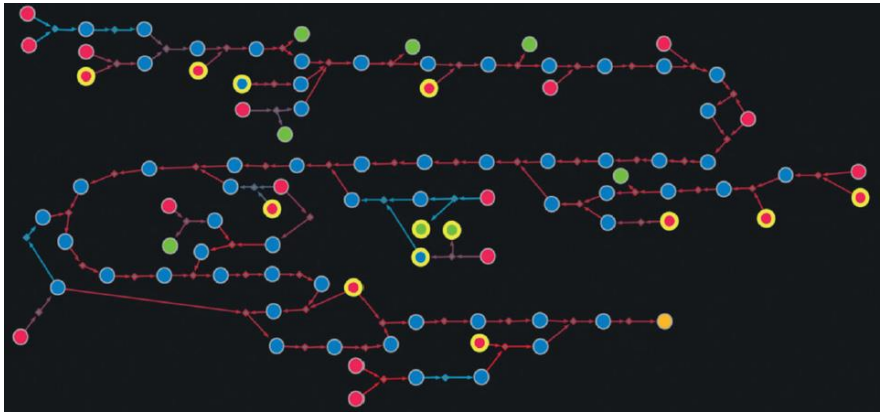
➡ Improved machine power solved this problem.

Chematica

- **Contains 10 millions reaction data**



NOC
(Network of Organic Chemistry)

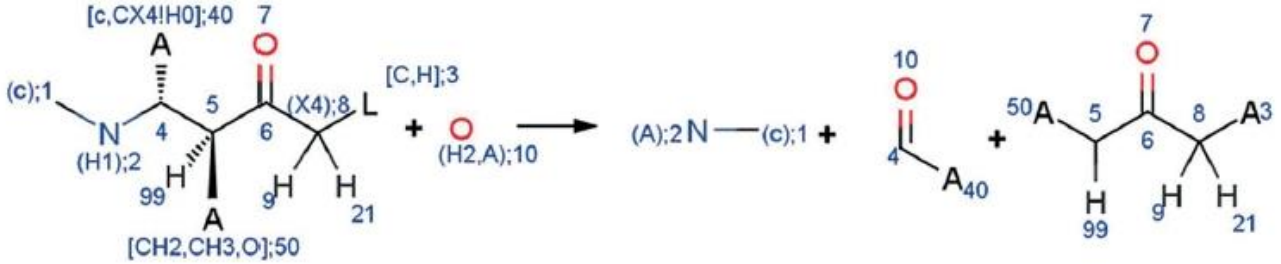


The lowest cost synthetic pathway of taxol (within 50steps)

B. A. Grzybowski *et al. Angew. Chem. Int. Ed*. **2016**, *55*, 5904.

Syntaurus

- **Algorithm for retrosynthesis**

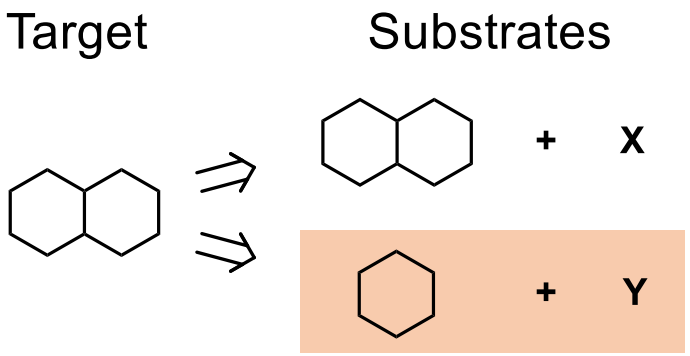


rxn_id: 8382,
name: "Proline-catalyzed Mannich Reaction",
reaction_SMARTS:[c:1][NH:2][C@H:4]([c,CX4!H0:40])[C@:5]([#1:99])([CH2,CH3,O:50])[C:6](=[O:7])[CX4:8]([#1:9])([#1:21])[#6,#1:3].[OH2:10]>>[c:1][N:2].[*:40][C:4]=[O:10].[*:50][C:5]([#1:99])[C:6](=[O:7])[C:8]([#1:9])([#1:21])[*:3]"
products:["[c][NH][C@H]([c,CX4!H0])[C@]([#1])([CH2,CH3,O])[C](=[O])[CX4]([#1])([#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"

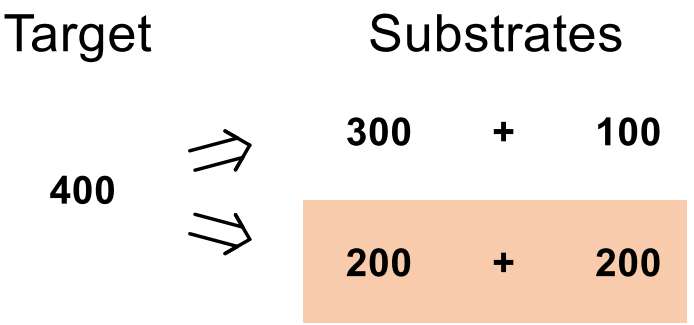
B. A. Grzybowski *et al. Angew. Chem. Int. Ed*. **2016**, *55*, 5904.

Scoring Functions

- **Chemical Scoring Function (CSF)**
  - Number of rings



  - Mass



- **Reaction Scoring Function (RSF)**
  - Necessity of protection
  - Yield etc…

B. A. Grzybowski *et al. Angew. Chem. Int. Ed*. **2016**, *55*, 5904.

Problems of Expert System

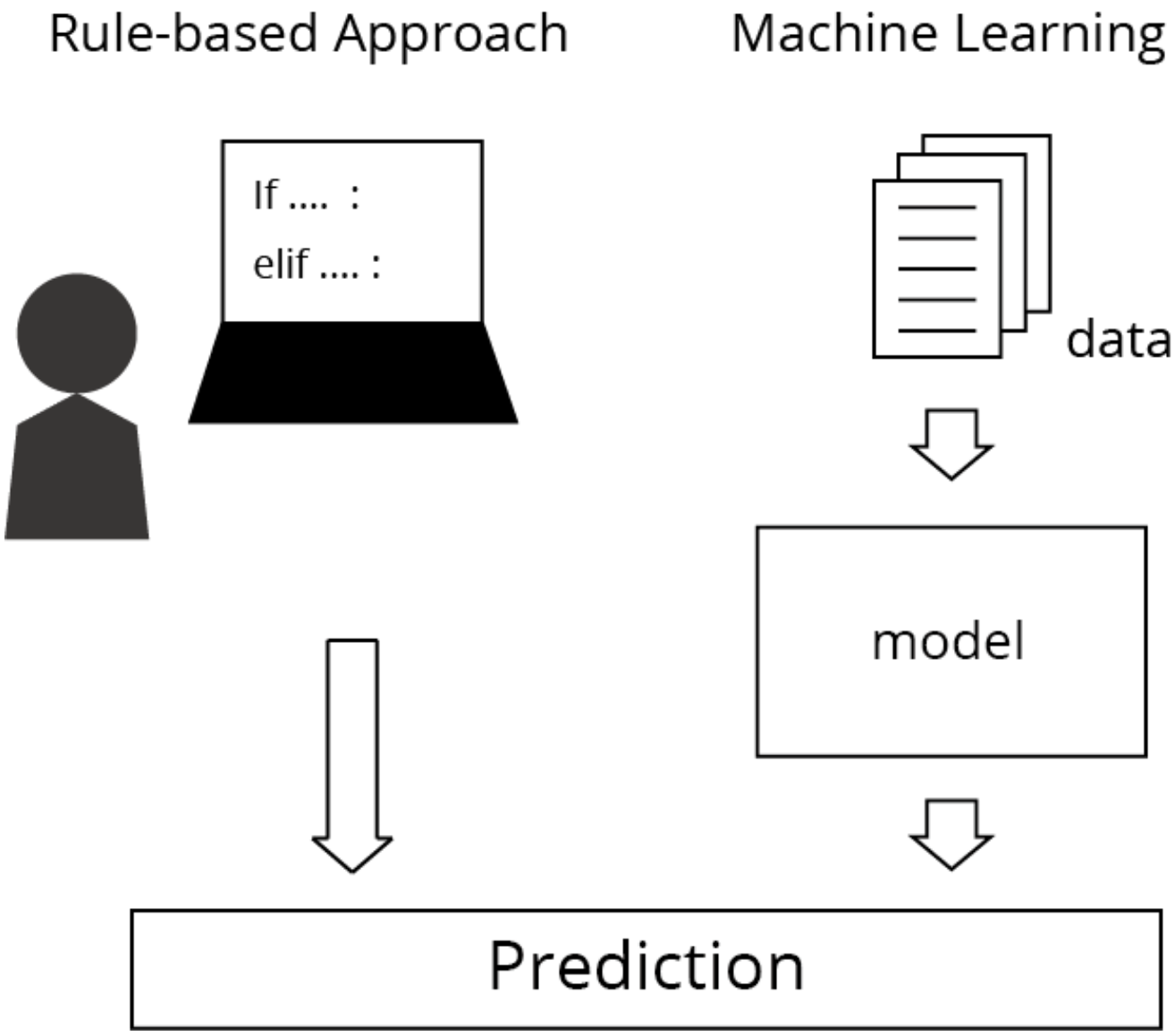
- **Forward Prediction**
  - Dependence of reaction templates

    ...Trouble of template creation
    ...Ignoring the context of molecules

  - Application of unknown reactions

- **Retrosynthesis**
  - Design of scoring functions

Machine Learning



Machine Learning

- **Supervised learning**

pattern

answer

⇒ *"Tiger"*

- **Unsupervised learning**

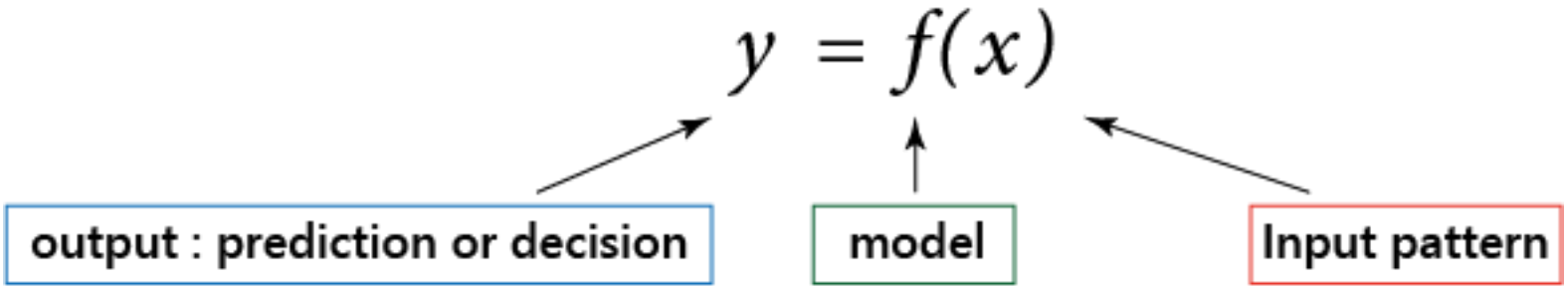
  Using unlabeled data

- **Reinforcement leaning**

  Maximizing rewards

Machine Learning

- **Supervised learning**



- **Classification**

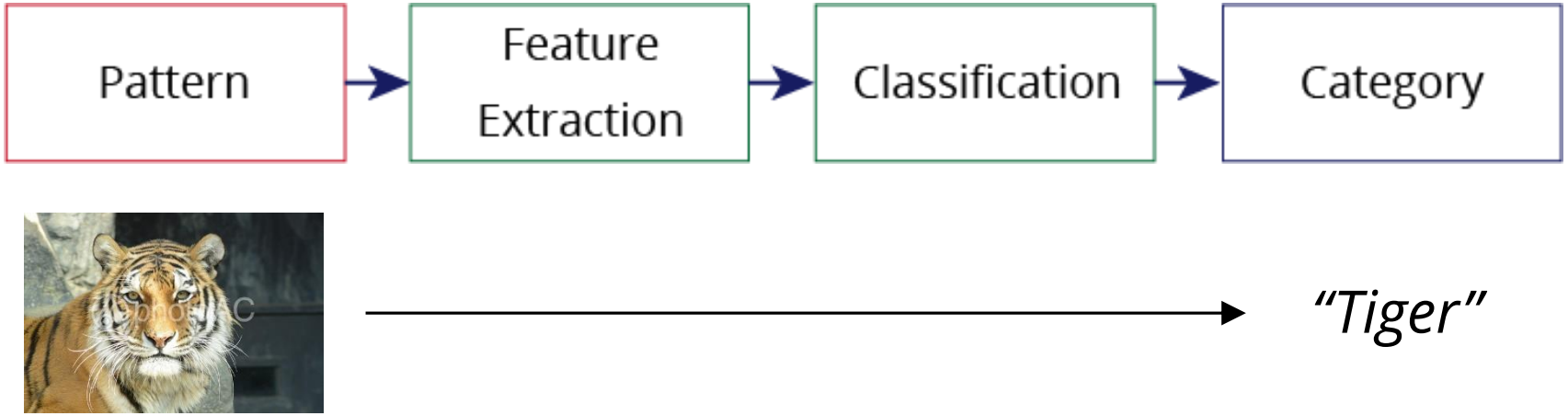
  output : discrete category
  example : estimation of animal type

- **Regression**

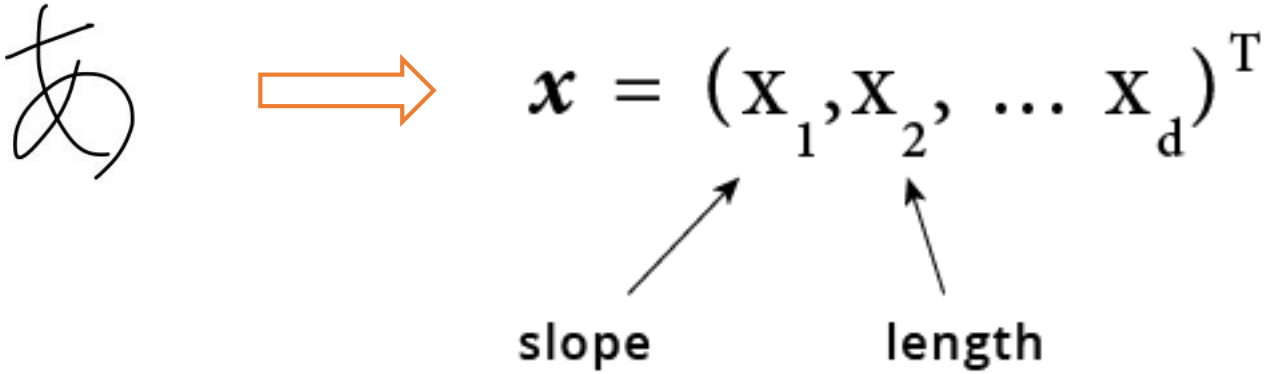
  output : continuous variable
  example : consumption of the entire economy

Machine Learning

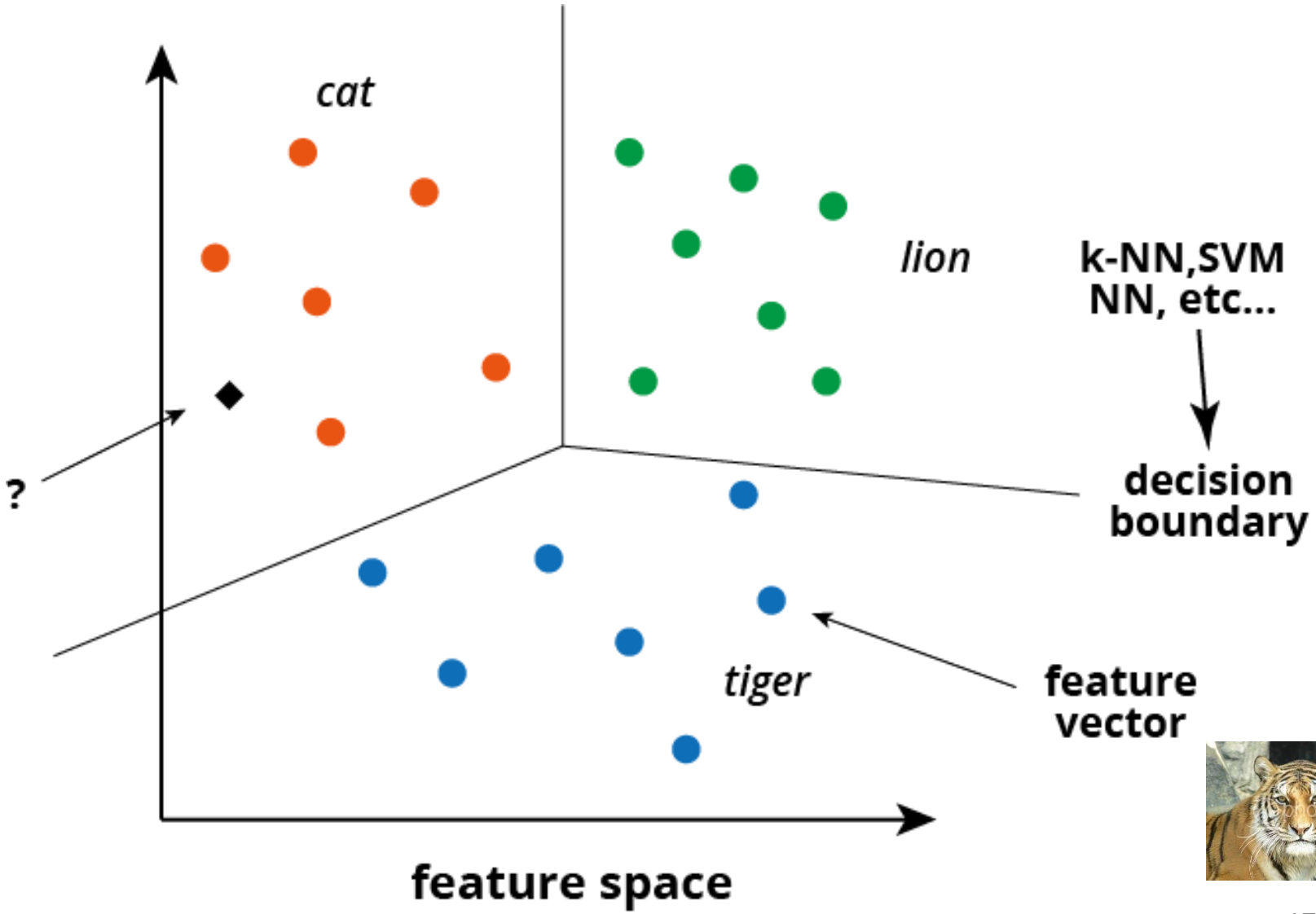
- **Classification**



- **Feature extraction**

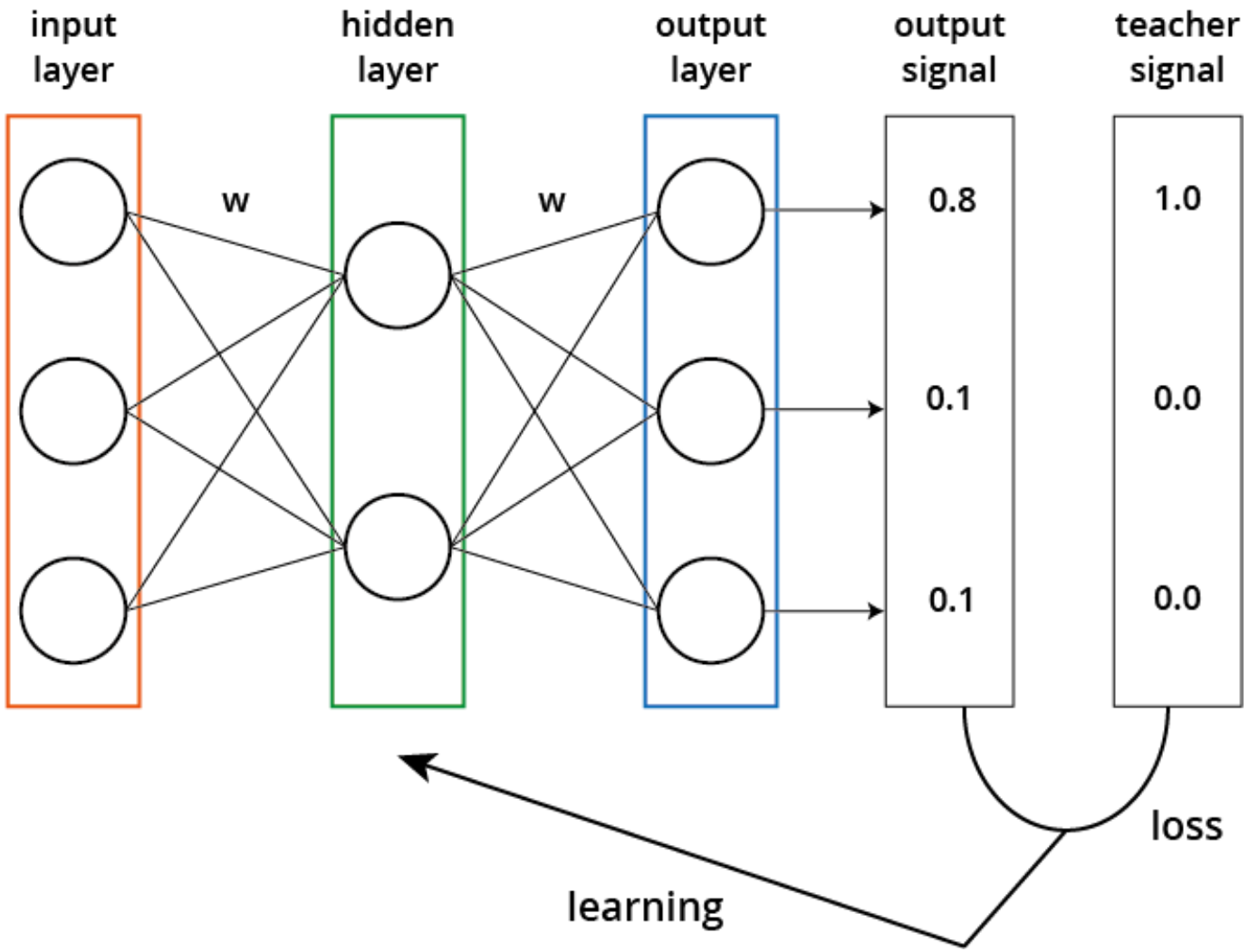


Feature Space



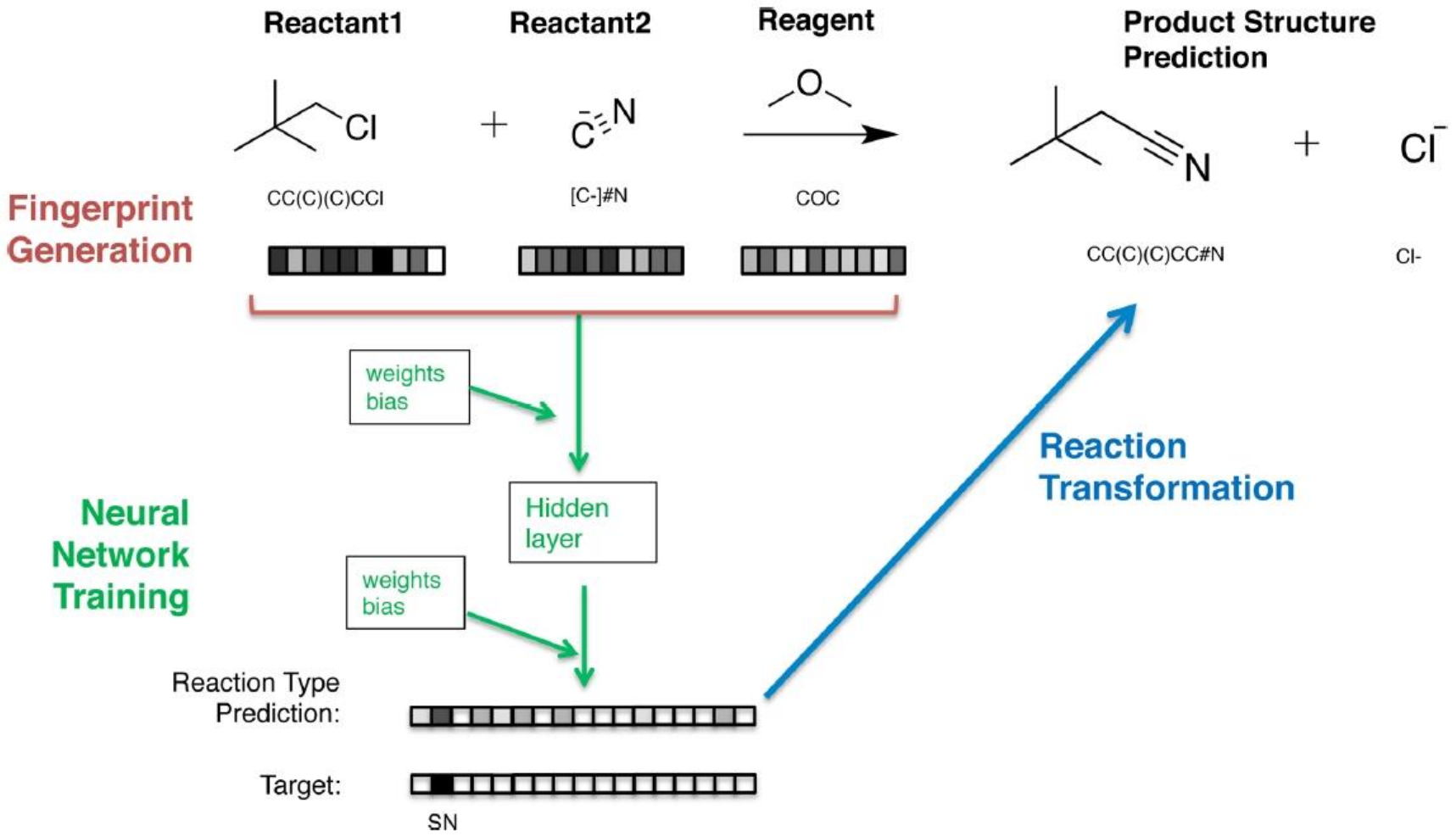
Neural Network

- Learn the best parameter automatically



Reaction Type Prediction

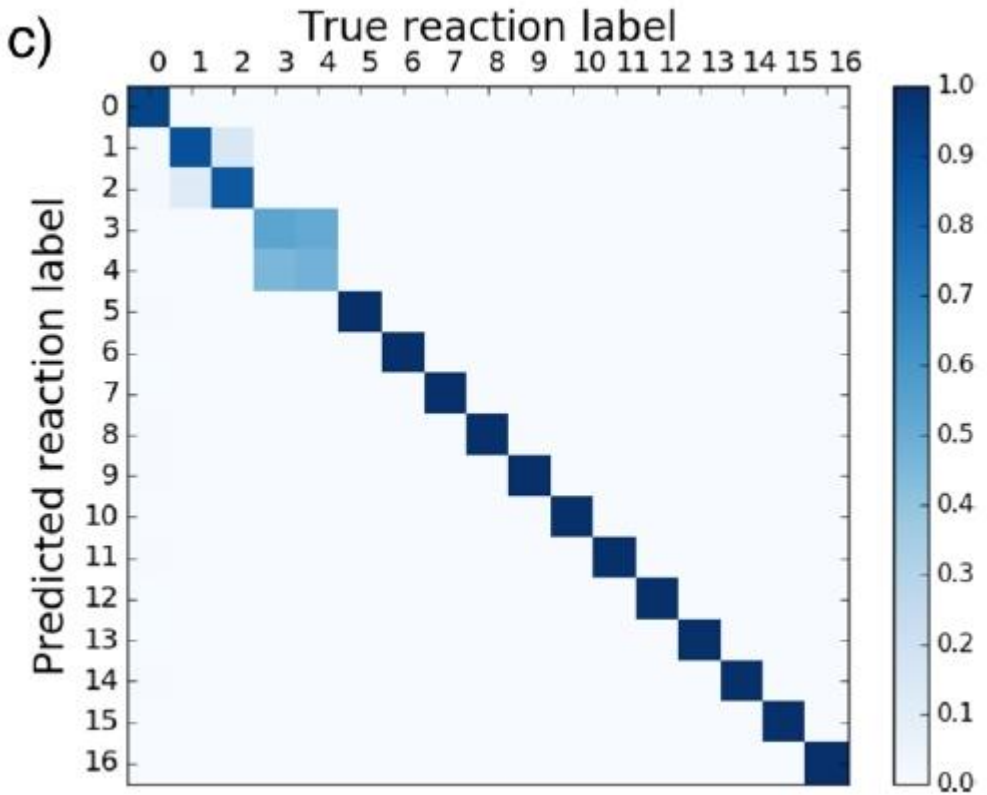
- **Predict 17 reaction types from reactants and reagent**



A. Aspulu-Guzik *et al*., *ACS Cent. Sci*., **2016**, *2*, 725.

Reaction Type Prediction

- **Predicted probability of each reaction type**



0. Null Reaction
1. Nucleophilic substitution
2. Elimination
3. Nucleophilic Substitution with Methyl Shift
4. Elimination with methyl shift
5. Hydrohalogenation (Markovnikov)
6. Hydrohalogenation (Anti-Markovnikov)
7. Hydration (Markovnikov)
8. Hydration (Anti-Markovnikov)
9. Alkoxymercuration-demercuration
10. Hydrogenation
11. Halogenation
12. Halohydrin formation
13. Epoxidation
14. Hydroxylation
15. Ozonolysis
16. Polymerization

Teacher signal

[0,0,0,0.5,0.5,0.0.....] (reaction type 3 or 4)

[0,1,0,0,0,.....] (others)

A. Aspulu-Guzik *et al*., *ACS Cent. Sci*., **2016**, *2*, 725.

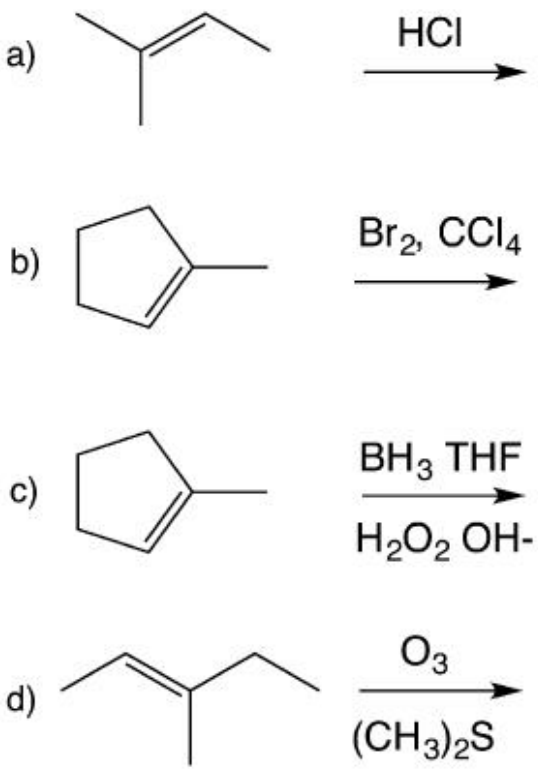
Reaction Type Prediction

- **Attempts to solve textbook problems**
  (Wade, *Organic Chemistry*, 6th ed.)

Results

8-47a	1.000	8-47i	0.998
8-47b	0.791	8-47j	0.001
8-47c	0.748	8-47l	1.000
8-47d	1.000	8-47m	1.000
8-47e	1.000	8-47n	0.855
8-47f	0.001	8-47o	0.999
8-47g	0.073	8-47p	1.000
8-47h	0.649		

Example



A. Aspulu-Guzik *et al*., *ACS Cent. Sci*., **2016**, *2*, 725.

Extended Reaction Templates

- **Improvement of Neural Network**

  ...dropout, highway network, ELU(activation function)

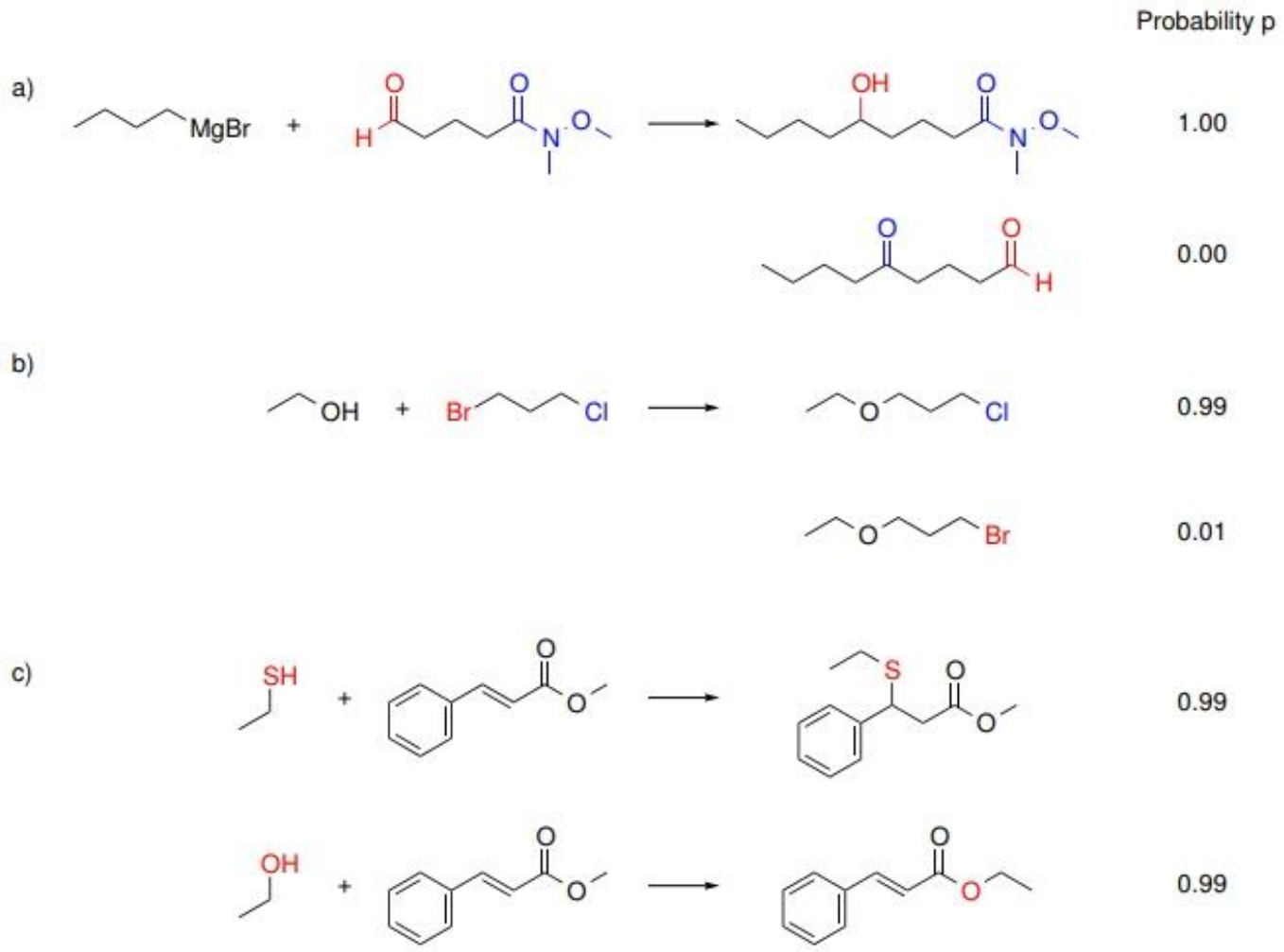
Table 2. Results for the study on 8720 automatically extracted rules.

Task/model	Acc	Top **10**-Acc.	MRR	W. Prec.
Reaction prediction				
random	0.00	0.00	0.00	0.00
expert system	0.02	0.18	0.02	0.06
logistic regression	0.41	0.65	0.49	0.31
highway network	0.78	0.98	0.86	0.77
FC512 ELU	0.77	0.97	0.85	0.76

M. H. S. Segler and M.P. Waller, *Chem. Eur. J.* **2017**, *23*, 5966.

Prediction Results

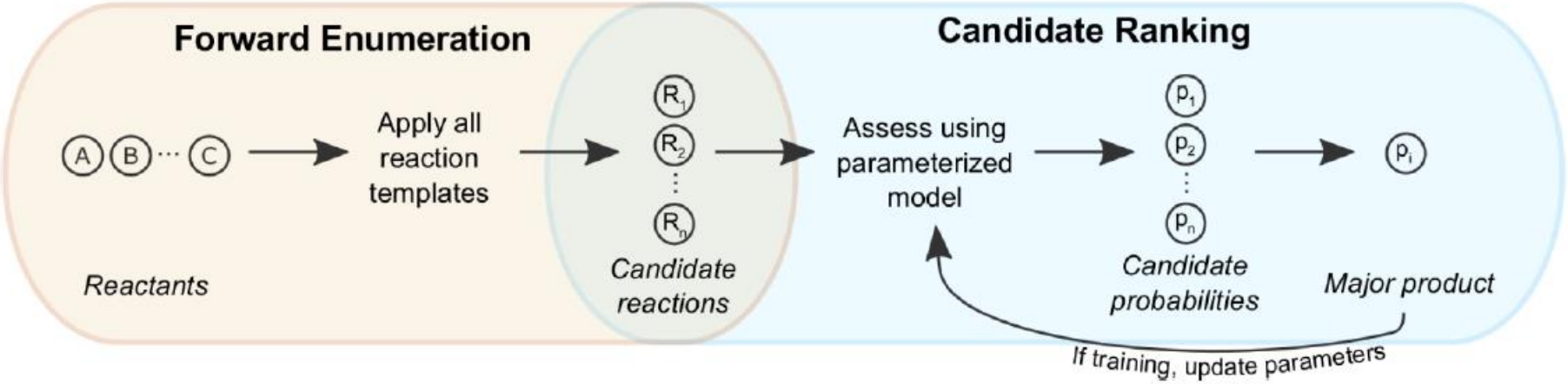
- **Neural Network learned molecular context?**



M. H. S. Segler and M.P. Waller, *Chem. Eur. J.* **2017**, *23*, 5966.

Candidate Generation and Ranking

- **Reaction type prediction by two frameworks**



- **Forward Enumeration**

  1689 reaction templates

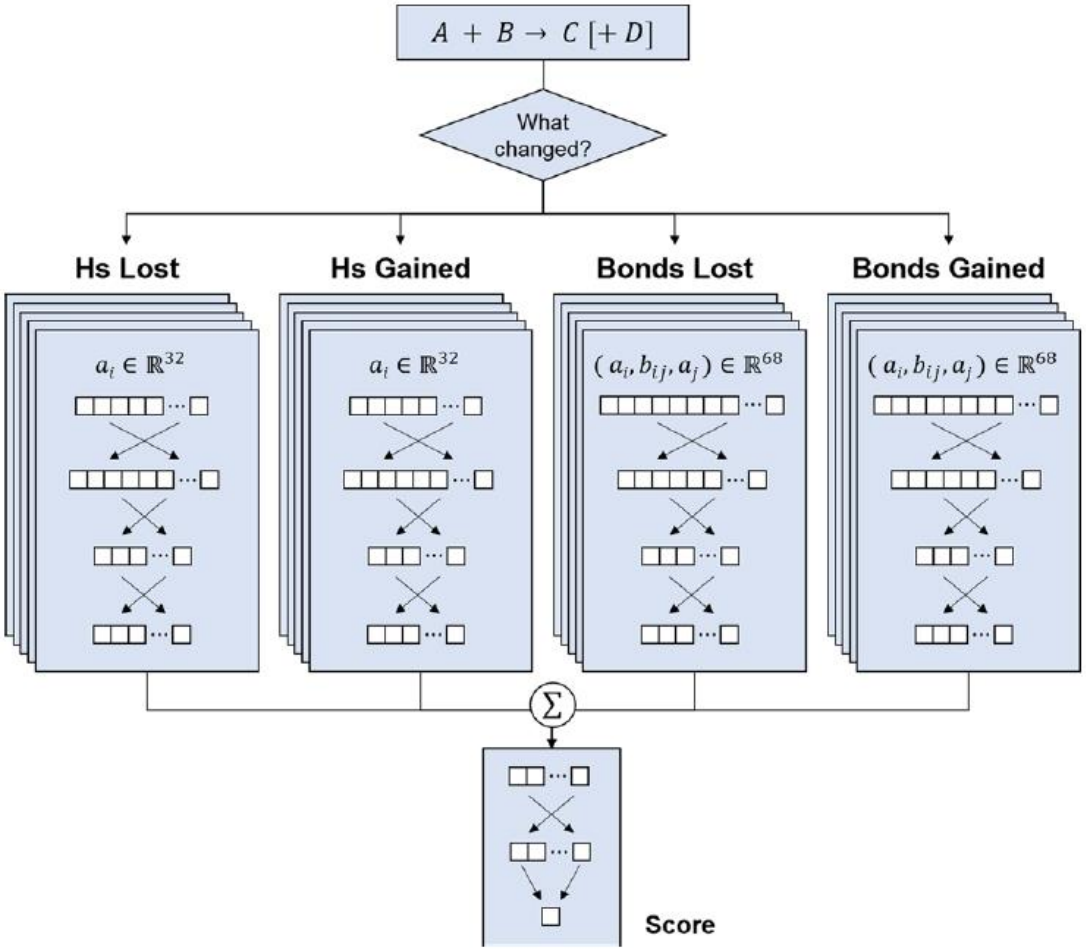
- **Candidate Ranking**

  focus only on changed atoms / bonds

C. W. Corey *et al.*, *ACS Cent. Sci.*, **2017**, *3*, 434.

Candidate Ranking

- **Focus on changed atoms / bonds**



C. W. Corey *et al.*, *ACS Cent. Sci.*, **2017**, *3*, 434.

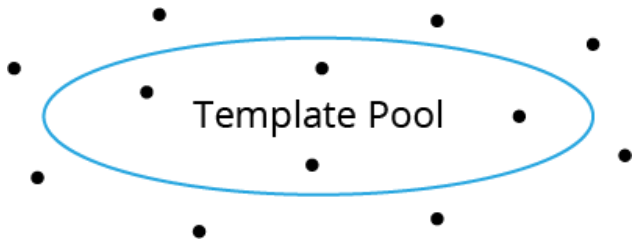
Reaction Prediction

- **Results**

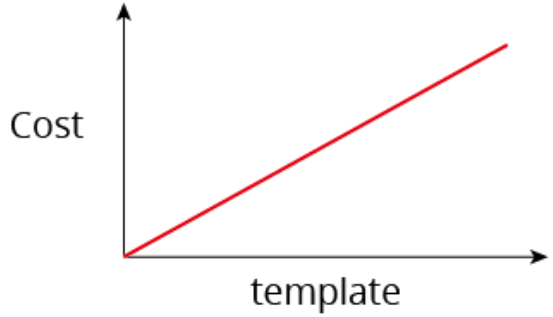
model	loss	acc. (%)	top-3 (%)	top-5 (%)	top-10 (%)
random guess	5.46	0.8	2.3	3.8	7.6
baseline	3.28	33.3	48.2	55.8	65.9
edit-based	1.34	68.5	84.8	89.4	93.6
hybrid	1.21	71.8	86.7	90.8	94.6

- **Problems of template-based model**

coverage



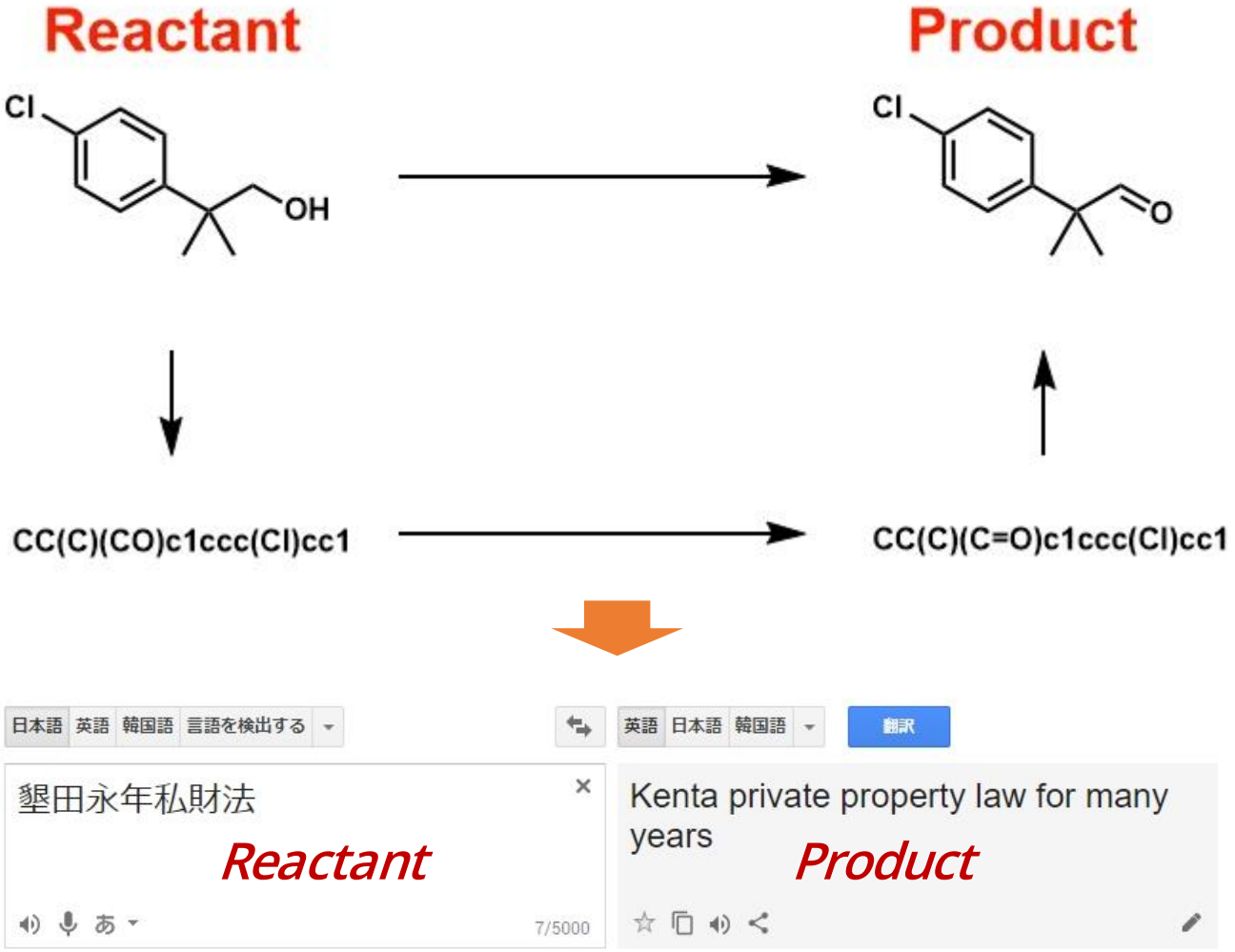
scalability



C. W. Corey *et al.*, *ACS Cent. Sci.*, **2017**, *3*, 434.

Sequence to Sequence (seq2seq)

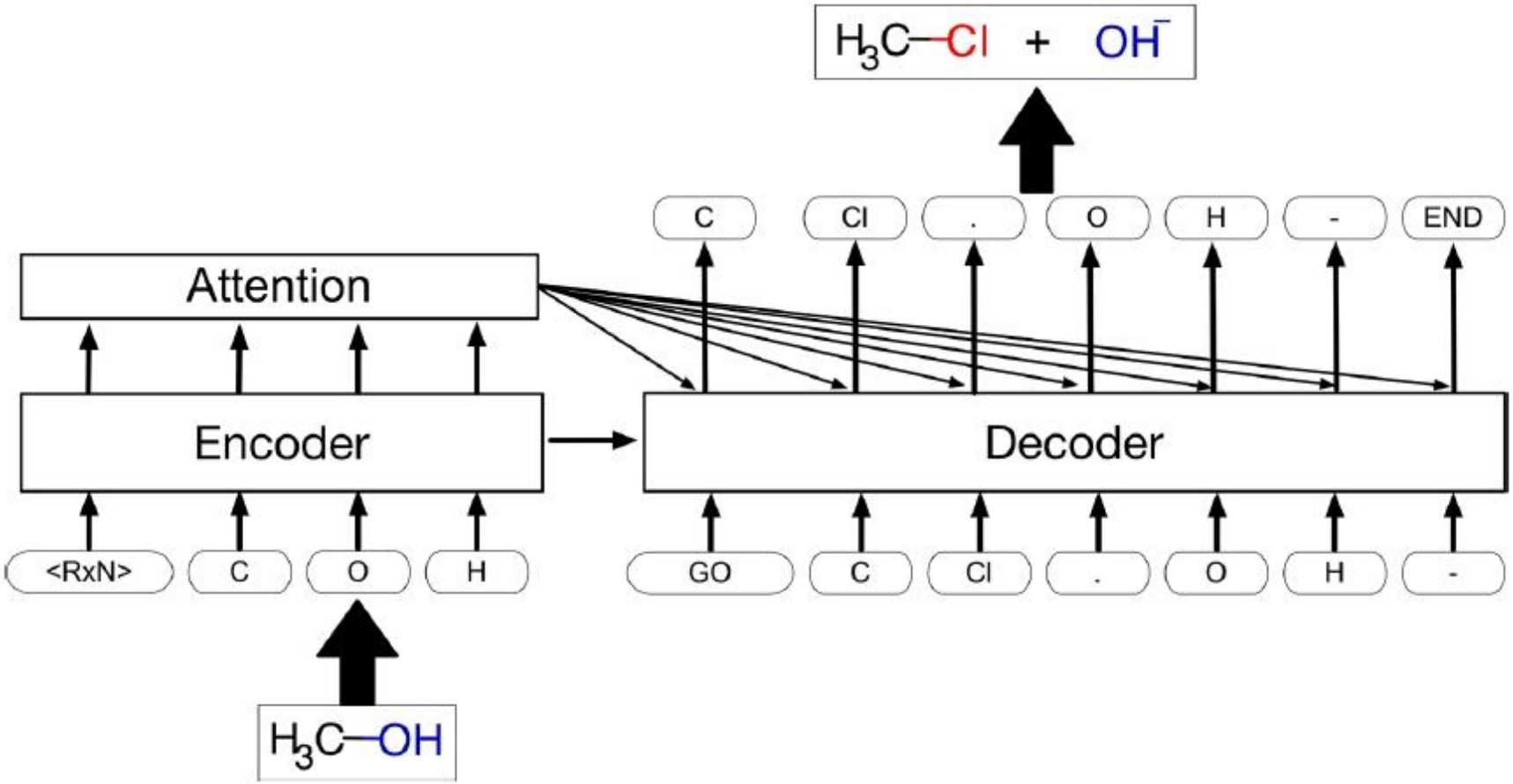
- **Is the chemical reaction similar to translation?**



P. Schwaller and T. Gaudin, arXiv:1711.04810v2, **2017**

Sequence to Sequence (seq2seq)

- **Reaction templates are not necessary.**



P. Schwaller and T. Gaudin, arXiv:1711.04810v2, **2017**

Sequence to Sequence (seq2seq)

- **Datasets**

Training

Jin's USPTO training set ... 395496

Test

Jin's USPTO test set ... 38648
Lowe's test set ... 50258

- **Results**

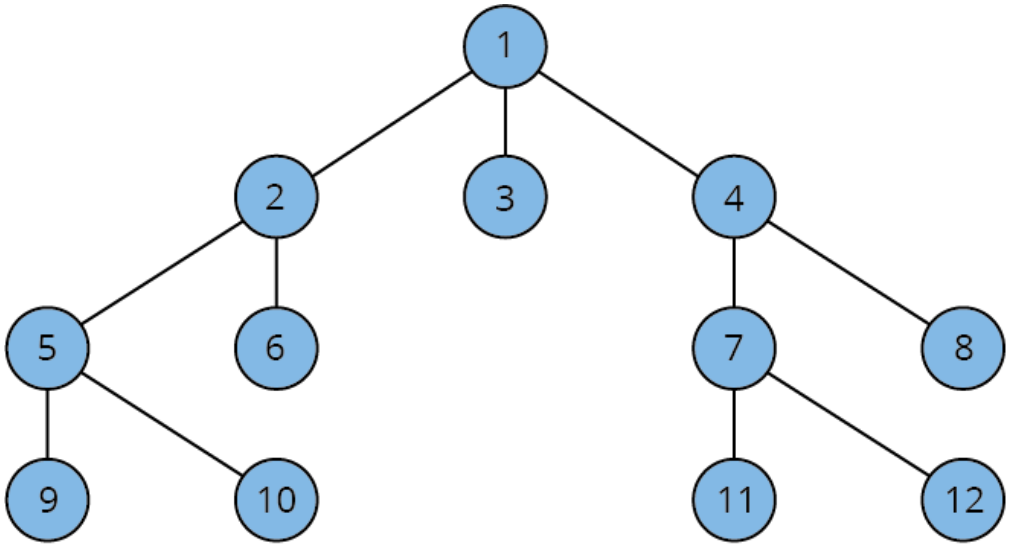
Dataset	Size	BLEU [36]	ROUGE [37]	Accuracies in [%] top-1	top-2	top-3
Jin's USPTO test set [17]	38,648	95.9	96.0	**83.2**	87.7	89.2
Lowe's test set [26]	50,258	90.3	90.9	**65.4**	71.8	74.1

P. Schwaller and T. Gaudin, arXiv:1711.04810v2, **2017**

Difficulties in Retrosynthesis

- **Very huge search space**
  - > 10000 reactions
  - 10^{30} ~ 10^{50} possible pathways

  ➡ Necessity of efficient search method

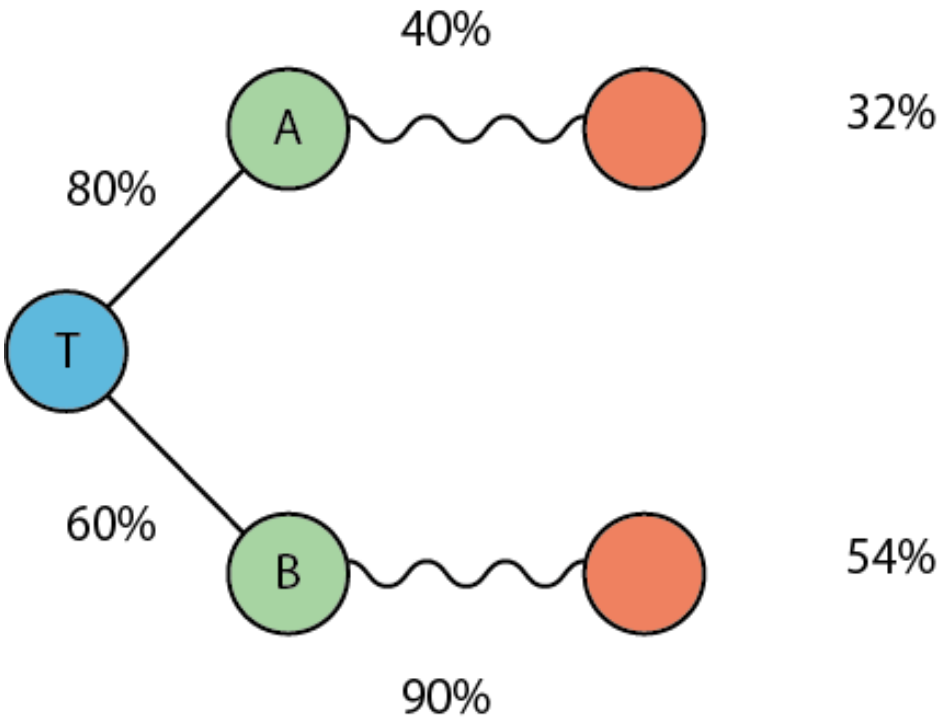


Breadth First Search

B. A. Grzybowski *et al. Angew. Chem. Int. Ed*. **2016**, *55*, 5904.

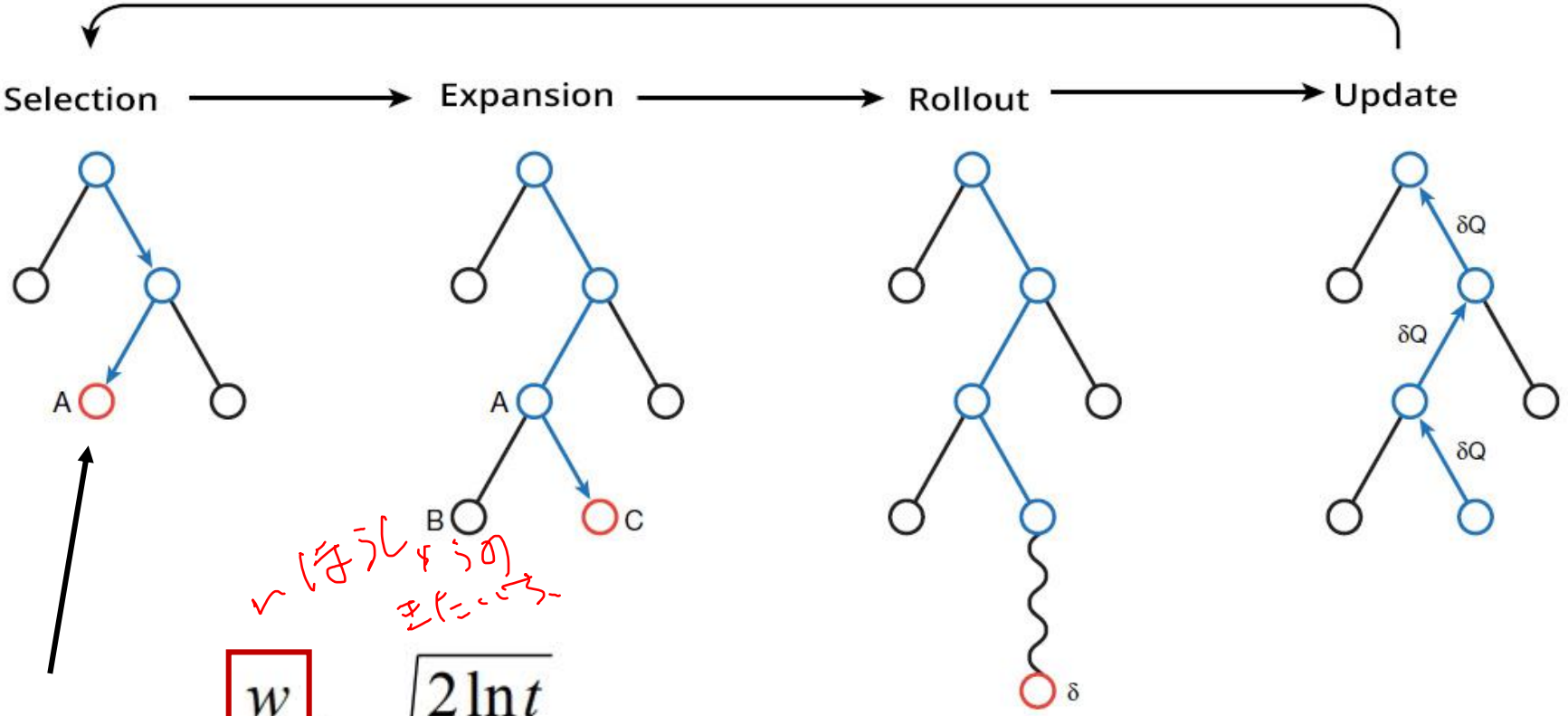
Difficulties in Retrosynthesis

- **Scoring Function**
  - Heuristic dependence
  - Necessity to expand synthetic tree to the end



Monte-Carlo Tree Search (MCTS)

- **Reinforcement learning to find the best route**

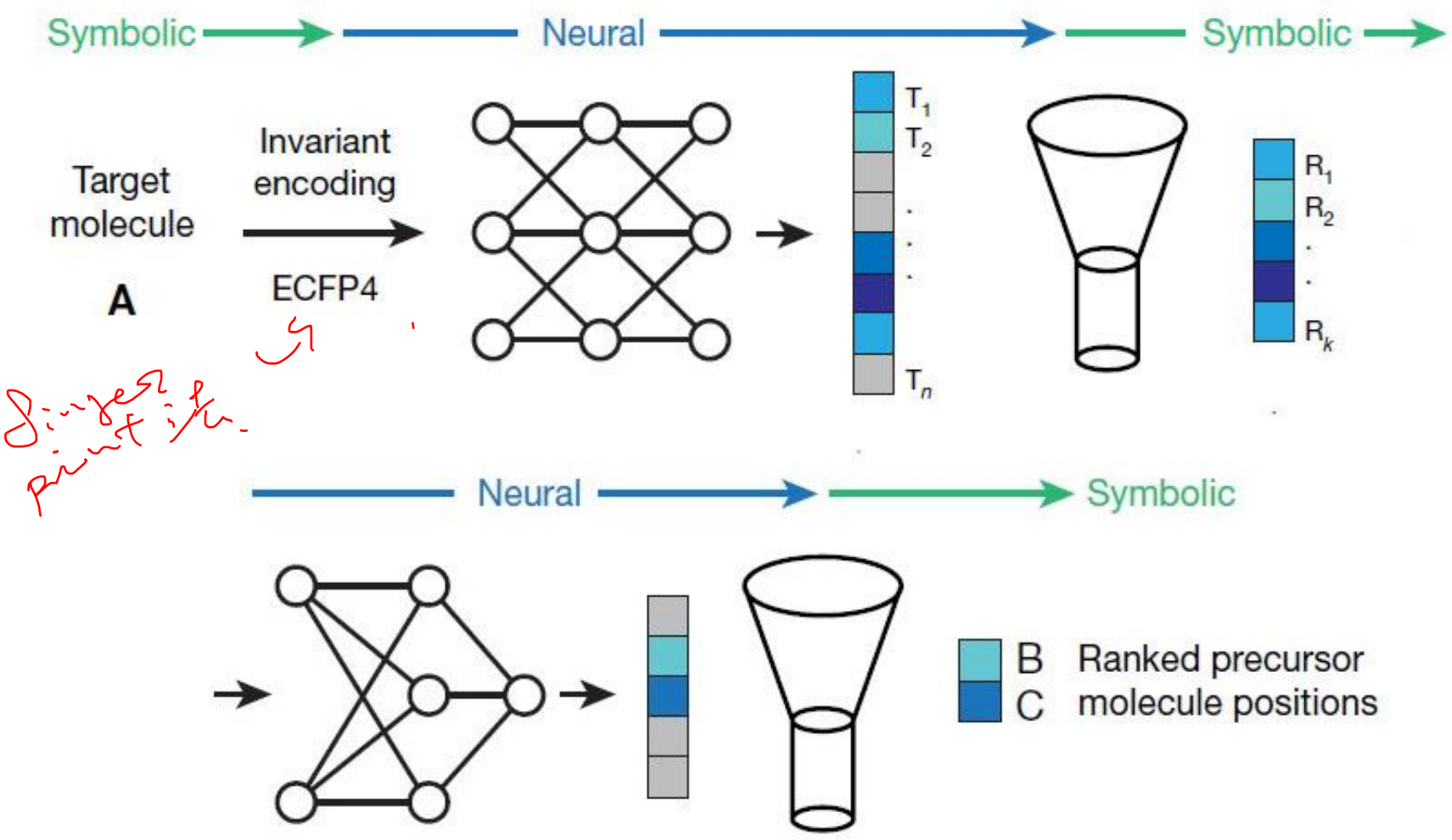


$$\text{UCB1} = \boxed{\frac{w}{s}} + c\sqrt{\frac{2\ln t}{s}}$$

(In fact, this parameter is more complicated…)

M. H. S. Segler and M.P Waller, *Nature*, **2018**, *555*, 604.

Expansion Procedure



M. H. S. Segler and M.P Waller, *Nature*, **2018**, *555*, 604.

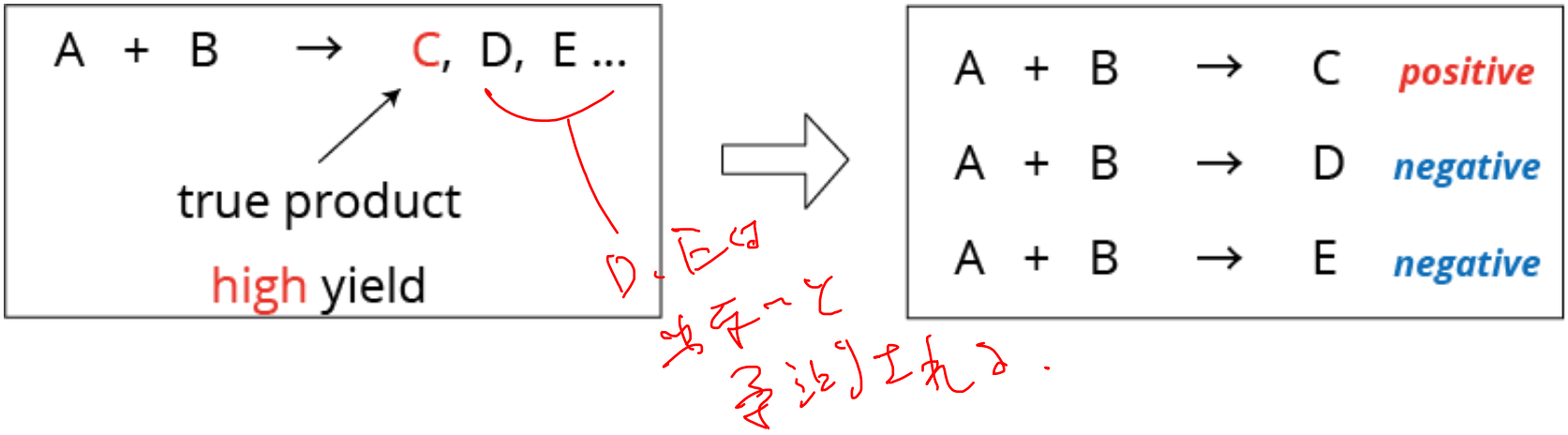
Datasets

Training datasets

12.4 million single-step reactions
- rollout rules
- expansion rules

Data Augmentation

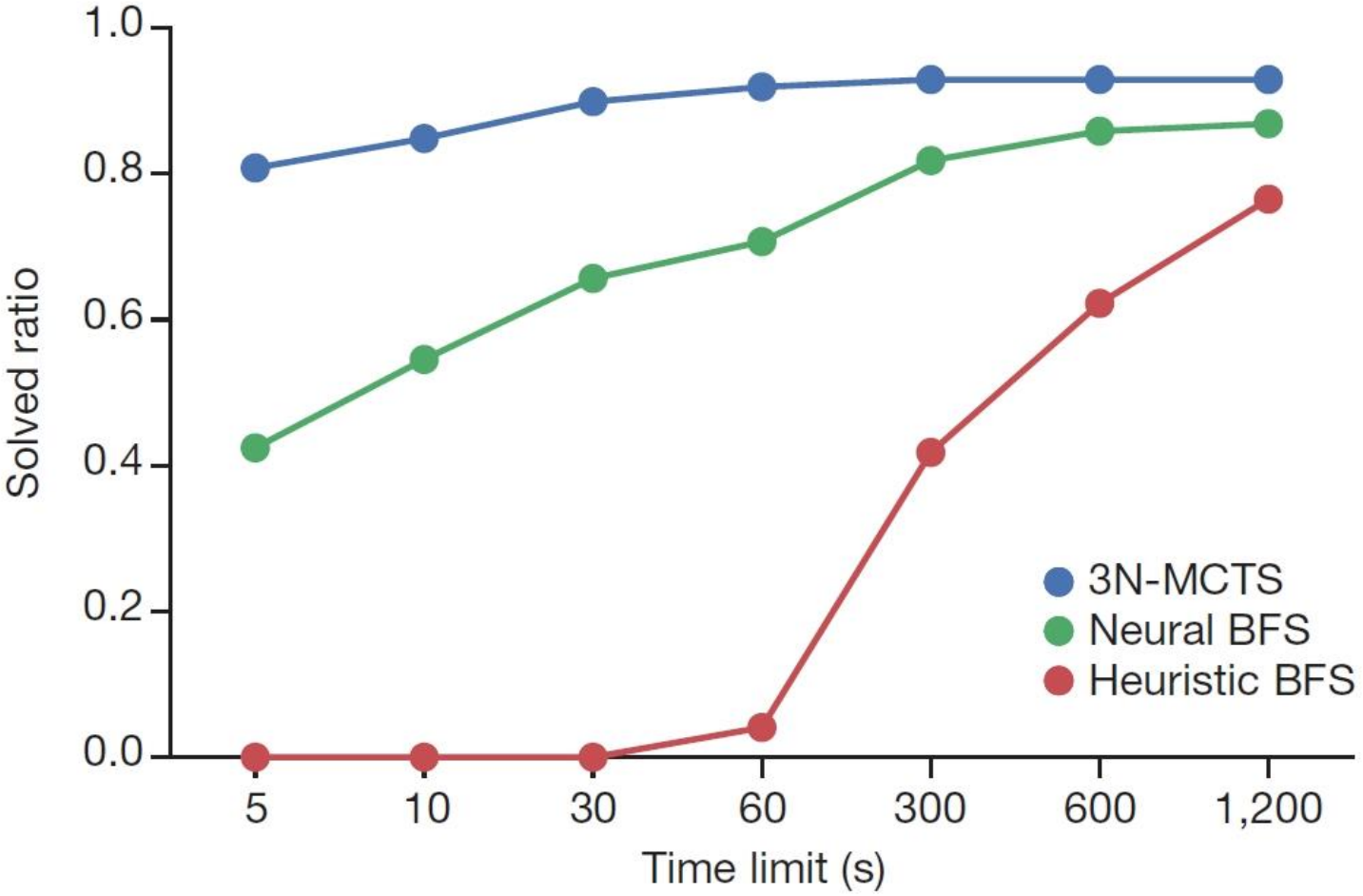
Generate 100 million negative reactions



M. H. S. Segler and M.P Waller, *Nature*, **2018**, *555*, 604.

Results – MCTS vs BFS

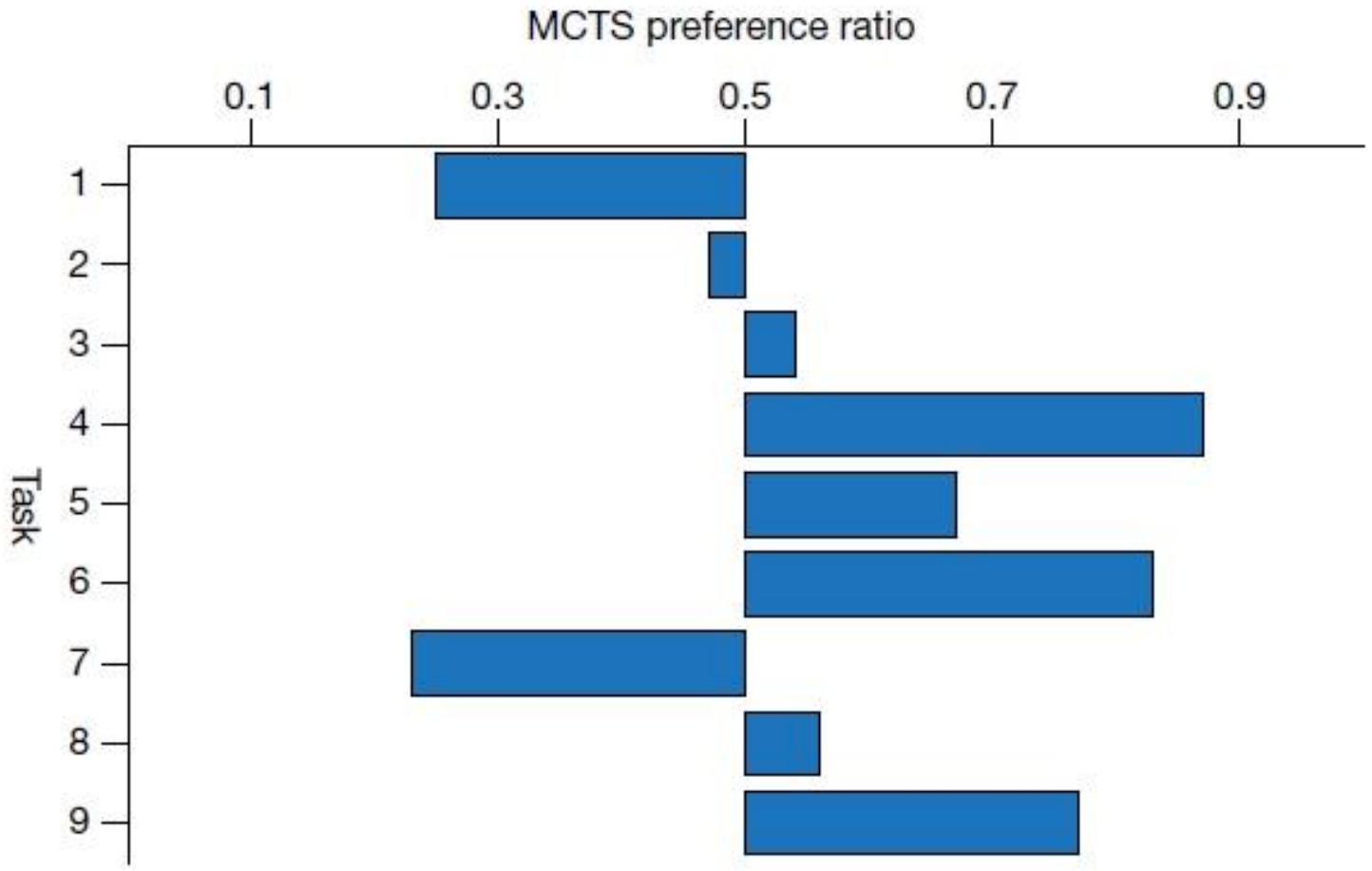
- **MCTS was faster than BFS (Breadth First Search).**



M. H. S. Segler and M.P Waller, *Nature*, **2018**, *555*, 604.

Results – MCTS vs Human

- **MCTS routes were more preferred.**



M. H. S. Segler and M.P Waller, *Nature*, **2018**, *555*, 604.

Summary

- **Expert system's problems**
  - Troublesome preparation of reaction rules
  - Application to unknown reactions
  - Lack of scoring function
- **Machine Learning**
  - The above problems can be solved.
  - Route design could be done at a level approaching humans.

Future

- **Current issues**
  - The best model was still unknown.

    (Finger Print, seq2seq, MCTS?)

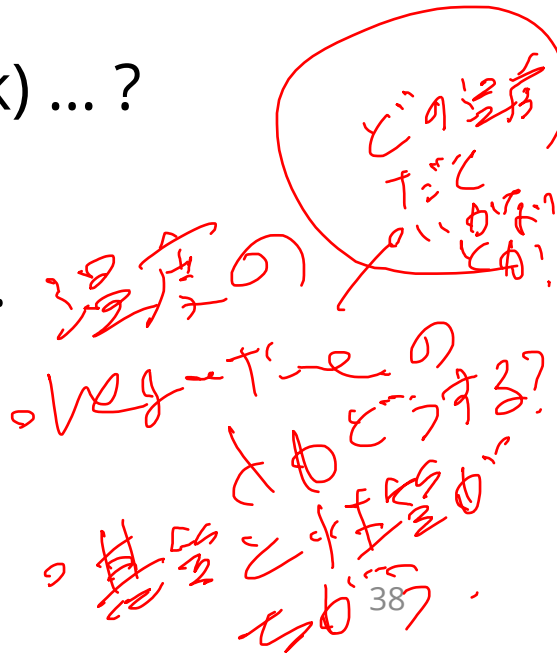
    ➡ Using images for compound cognition,

    Graph Representation,

    GAN (Generative Adversarial Network) … ?

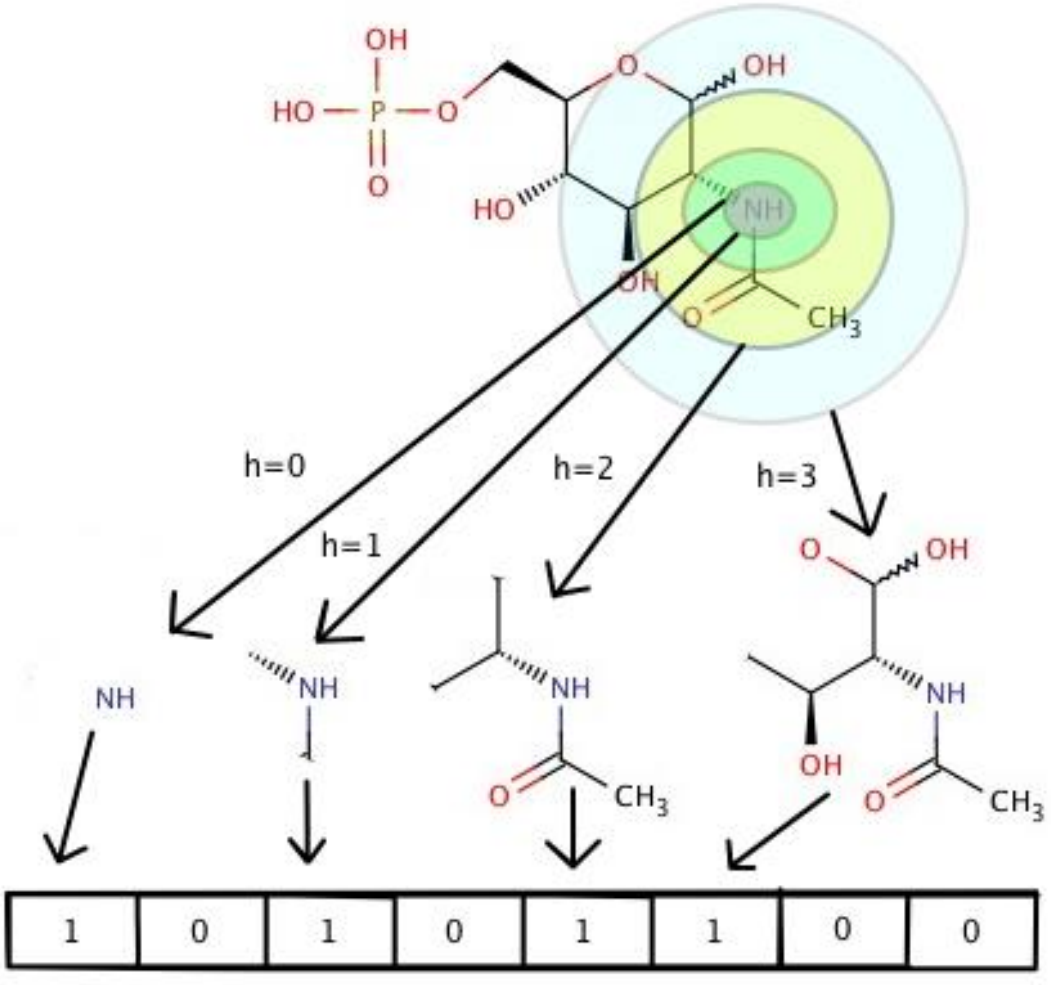
  - Reaction conditions are not considered.

    ➡ New reaction descriptor is required.



Appendix

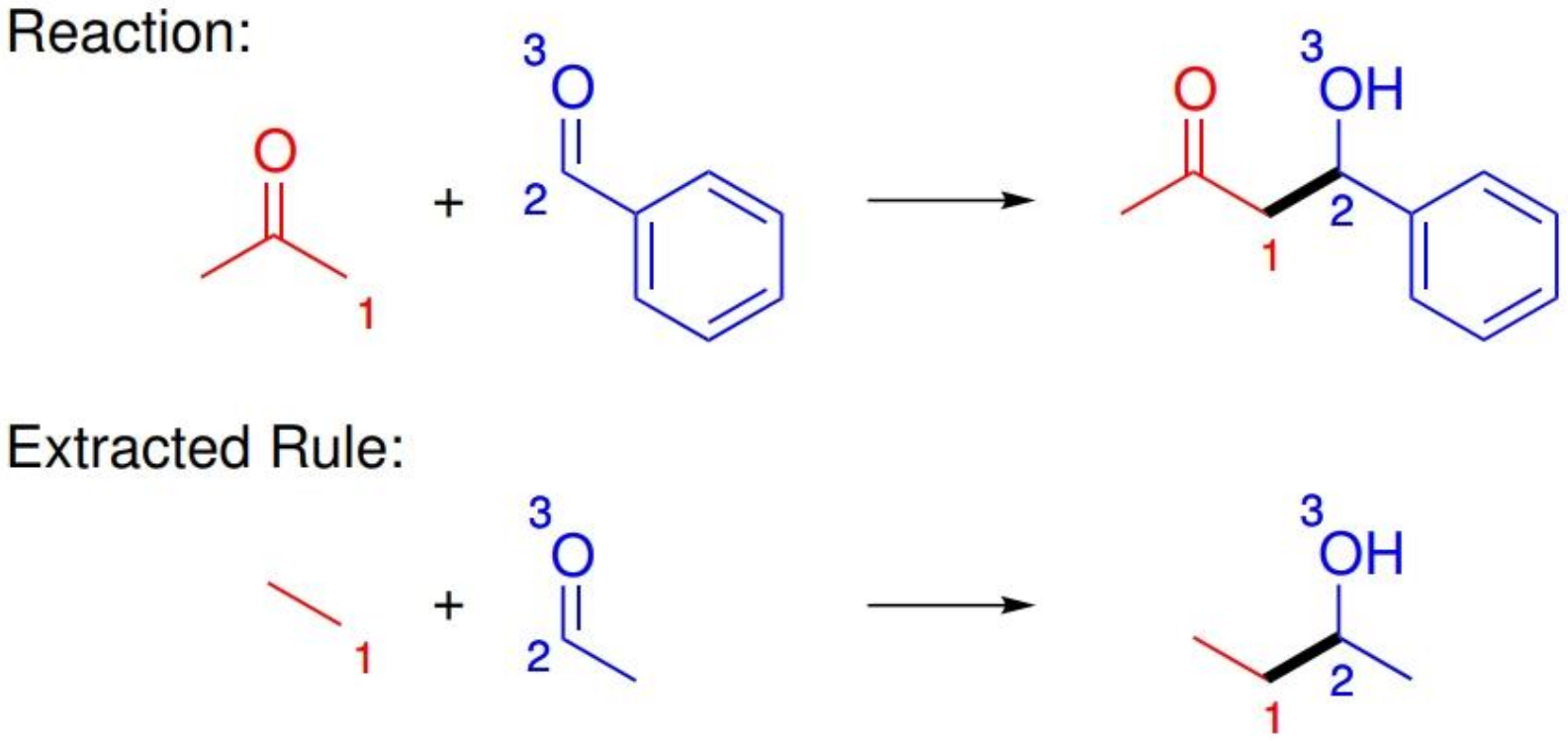
- **ECFP4**



https://chembioinfo.com/2011/10/30/revisiting-molecular-hashed-fingerprints/

Appendix

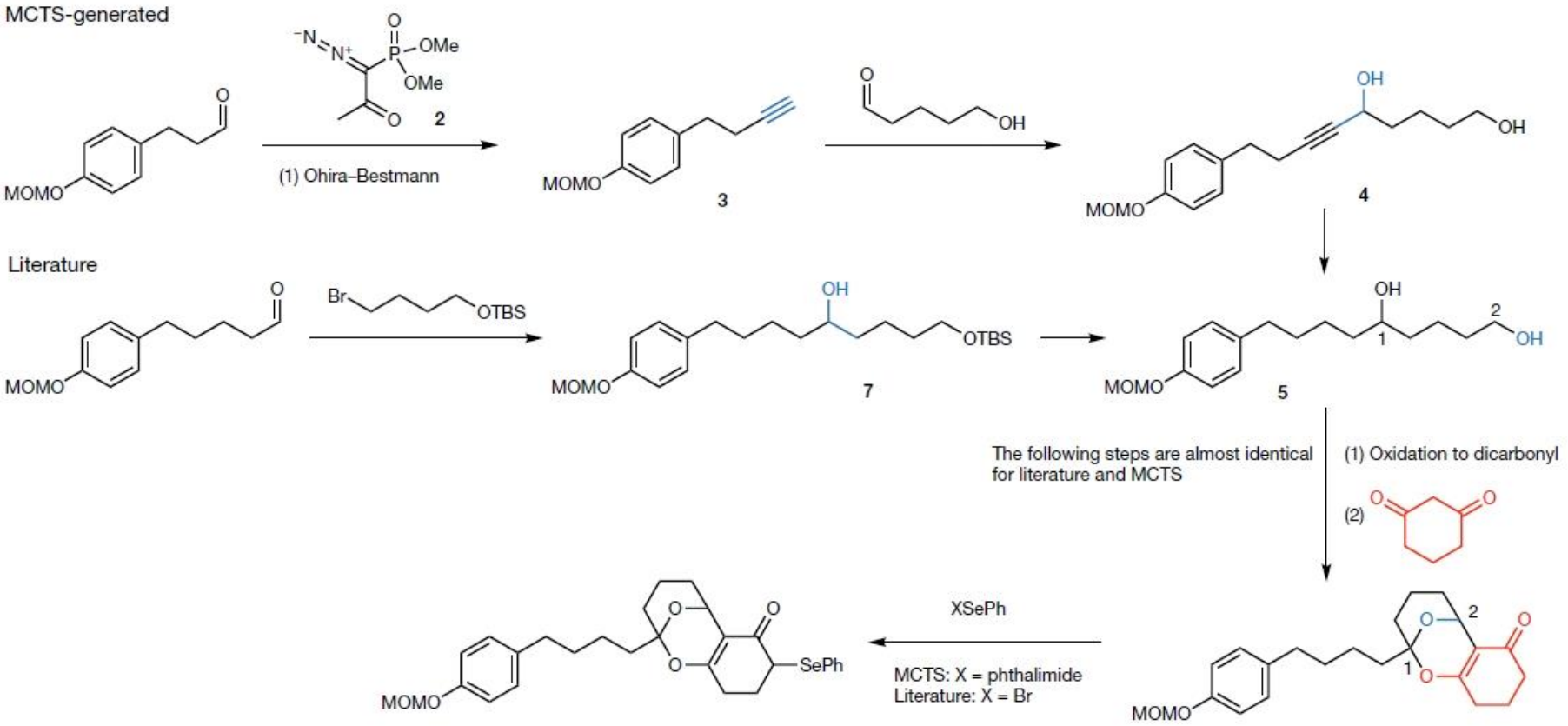
- **Rule extraction**



M. H. S. Segler and M.P. Waller, *Chem. Eur. J.* **2017**, *23*, 5966.

Appendix

- **In this example, chemists preferred literature routes.**



M. H. S. Segler and M.P Waller, *Nature*, **2018**, *555*, 604.