Introduction to Reaction ML

Marwin Segler

marwinsegler@microsoft.com

Twitter: @marwinsegler

Microsoft

Background Reading for this talk

Reviews

Strieth-Kalthoff, Sandfort, Segler, Glorius, *Chem. Soc. Rev.*, **2020,** 49, 6154-6168 **link**

Coley, Eyke, Jensen, *Angew. Chem. Int. Ed.* **2020**, 51, 22858 **link**

Johanson et al. (AZ), *Drug Discovery Today: Technologies,* **2019**, *32–33,* 65, **link**

Key Papers - Reaction ML

Segler, Waller, *Chem. Eur. J.*, **2017**, 23, 5966 - Reaction & Retrosynthesis Prediction

Coley et al ACS Cent Sci. **2017,** *5*, 434- Reaction Prediction

Segler, Preuss, Waller, *Nature*, **2018**, *555*, 604 - ML-Driven Multi-Step Retrosynthesis

Coley et al. *Science* **2019,** *365,* 6453- ML-based Retrosynthesis, Reaction & Condition Prediction + Robot Implementation

Schwaller et al ACS Cent. Sci **2019**, *11*, 3316 - Molecular Transformer

Segler, M. P. Waller, *Chem. Eur. J.,* **2017**, 23, 6118 - Reaction Knowledge Graphs

Molecular Design

Bradshaw, Paige, Kusner, Segler, Hernandez-Lobato, *NeurIPS* **2020 -** Reaction-Driven Generative Models

Segler, Kogej, Tyrchan, Waller, *ACS Cent. Sci.,* **2017**, 4, 120 **-** SMILES RNN

ANGEWANDTE
CHEMIE

91. Jahrgang 1979
Heft 2
Seite 99–184

New Applications of computers in Chemistry

Neue Anwendungsgebiete für Computer in der Chemie

Von Ivar Ugi, Johannes Bauer, Josef Brandt, Josef Friedrich, Johann Gasteiger, Clemens Jochum und Wolfgang Schubert[*]

"In chemistry the use of computers has been customary for a long time. **Nevertheless, only a modest part of the inherent capabilities of modern computers is utilized for the solving of chemical problems.** Numerical problems are solved, such as the ones encountered in quantum chemistry, and in the collection and evaluation of experimental data, or large sets of data are subjected to storage and retrieval operations.

The challenge to solve chemical problems by algorithms which simulate human intelligence in the sense of decision processes and deductive thought was felt at a rather early stage. It led to studies in a direction which is now associated with the term '**artificial intelligence**'."

Analytica Chimica Acta, 248 (1991) 1–30
Elsevier Science Publishers B V , Amsterdam

Review

Neural networks: A new method for solving chemical problems or just a passing phase?

J. Zupan *,[1] and J. Gasteiger

Organisch-chemisches Institut, Technische Universität München, D-8046 Garching (Germany)

(Received 3rd January 1991)

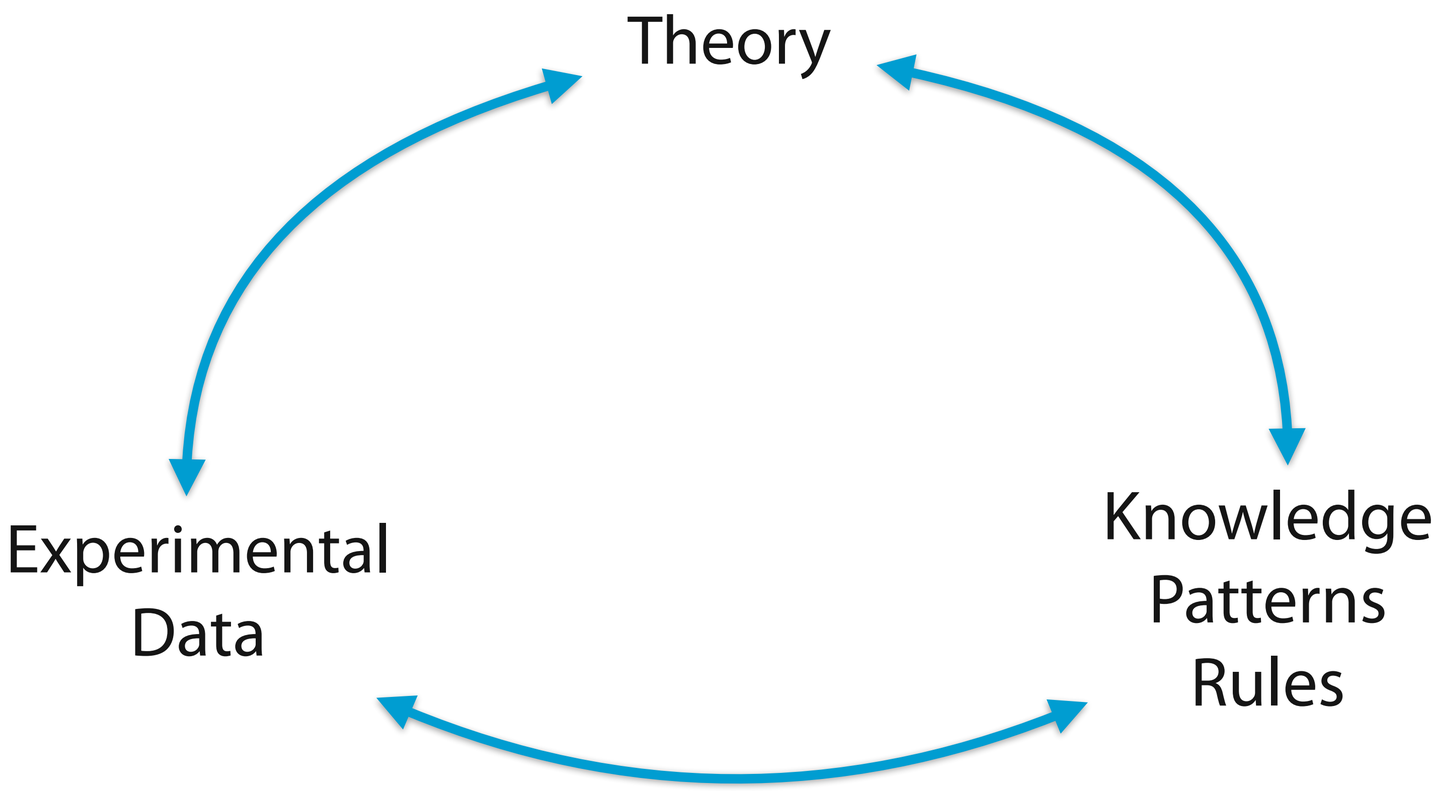
Two Paradigms for Program Solving with Computers

Programming: Encode all instructions to solve problem

Machine Learning: Use *example data* + *simple algorithm* to *derive* problem solving instructions

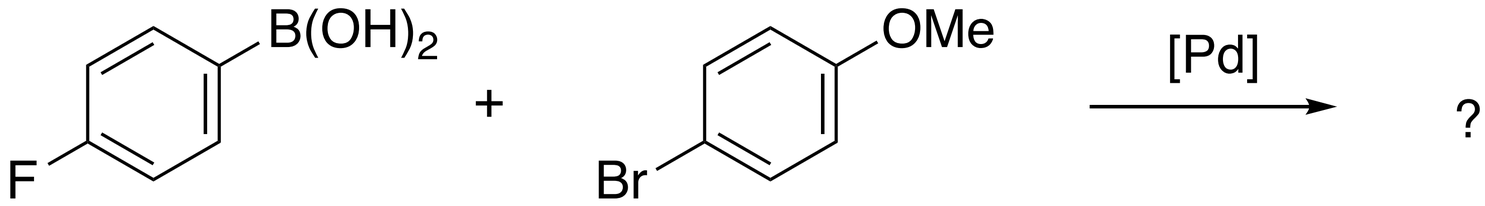
Chemical Reactions

The Trinity of Organic Chemistry
(We need all three!)



Strieth-Kalthoff, Sandfort, Segler, Glorius, *Chem. Soc. Rev.*, **2020**, *49*, 6154-6168

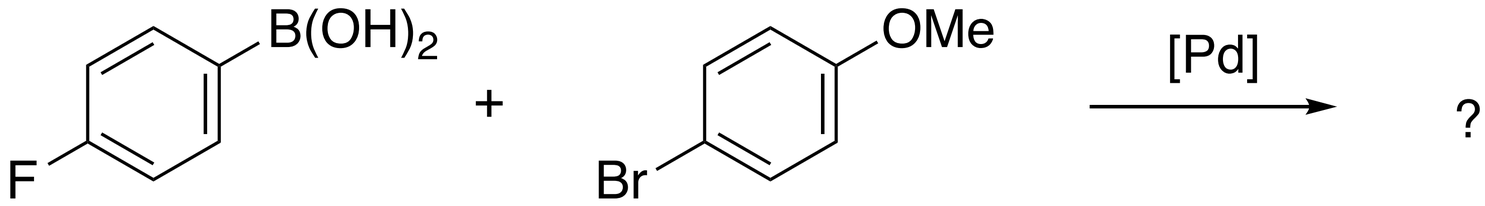
Pattern recognition in Organic Chemistry goes a long way.



Do you need ab-initio theory to
predict the likely outcome of this reaction?

Strieth-Kalthoff, Sandfort, Segler, Glorius, *Chem. Soc. Rev.*, **2020**, *49*, 6154-6168

Pattern recognition in Organic Chemistry goes a long way. But Not All…



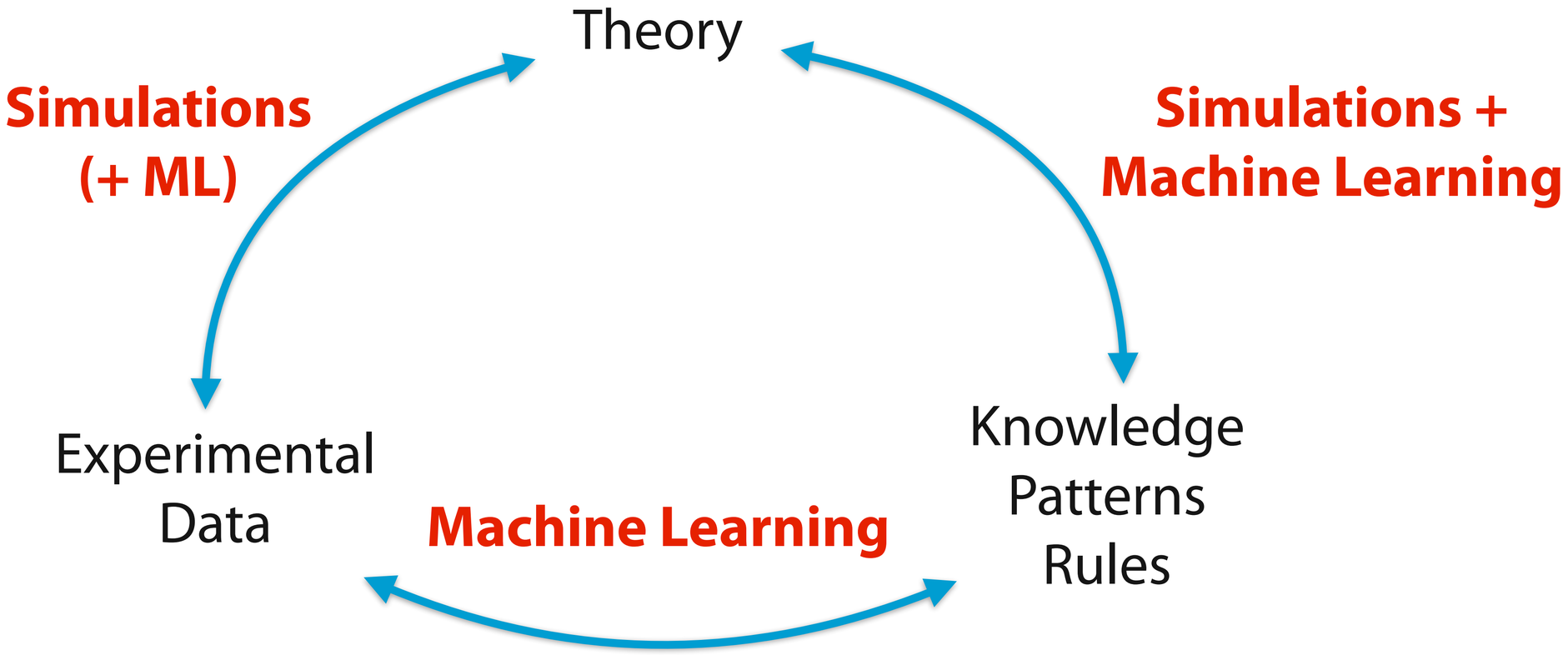
Do you need ab-initio theory to
predict the likely outcome of this reaction?

Will matching function groups give you an exact picture of the potential energy surface and transition states?

Strieth-Kalthoff, Sandfort, Segler, Glorius, *Chem. Soc. Rev.*, **2020**, *49*, 6154-6168

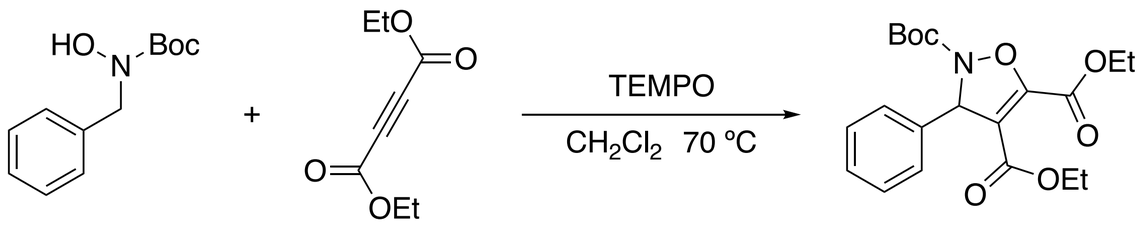
The Trinity of Organic Chemistry

We need all three!



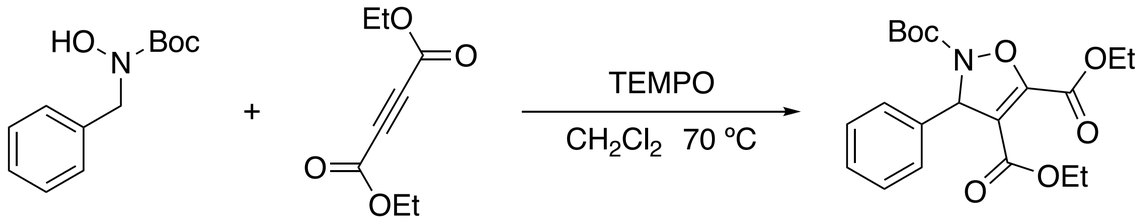
Strieth-Kalthoff, Sandfort, Segler, Glorius, *Chem. Soc. Rev.*, **2020**, *49*, 6154-6168

Chemical Reactions



Gini et al, *Chem. Eur. J.* **2015**, 21, 12053 – 12060

Experimental Procedures describe how to reproduce reactions



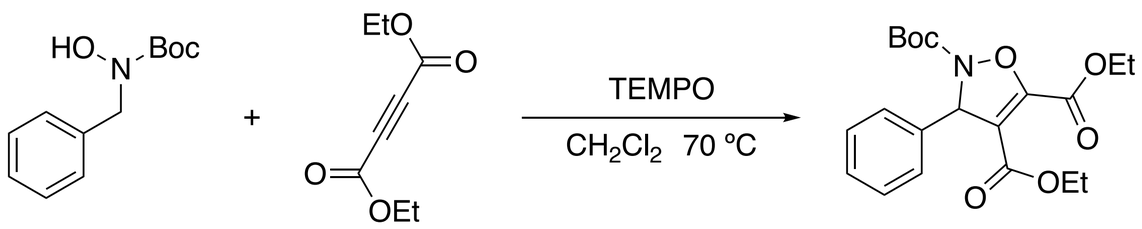
Yield: 87%

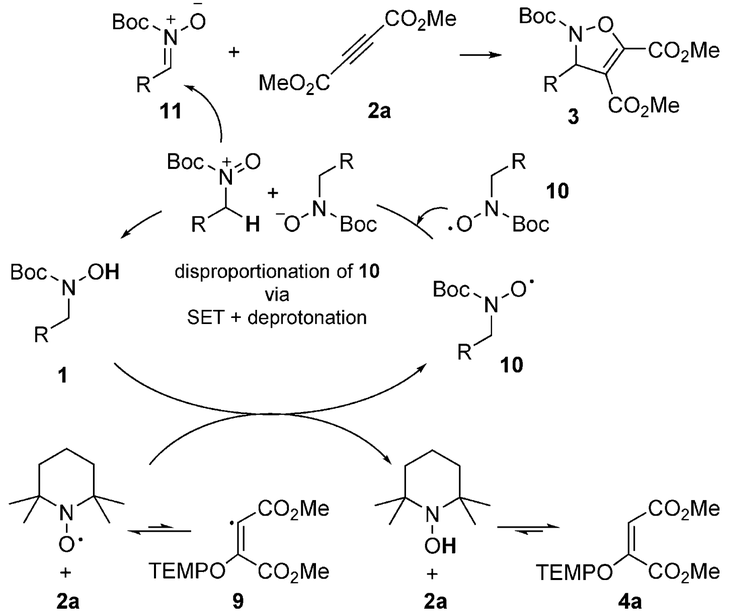
General procedure:
In a screw-cap Schlenk tube, the corresponding N-protected N-benzylhydroxyamine derivative 1 (1.00equiv) was dissolved in dry CH_2Cl_2 (2.00 mL). Dimethyl acetylenedicarbox- ylate (2 a) (4.00 equiv) and TEMPO (2.00 equiv) were added and the reaction mixture was stirred at 70 8C for 24 h. The solvent was re- moved under reduced pressure and the obtained crude product was purified by flash column chromatography on silica gel eluting with pentane/ AcOEt to give the corresponding N-protected isoxazoline 3.

2-tert-Butyl 4,5-dimethyl 3-phenylisoxazole-2,4,5-tricarboxylate (3a): According to the general procedure, N-Boc N-benzyl hydrox- ylamine (1 a) (0.25 mmol, 63.3 mg, 1.00 equiv), dry CH_2Cl_2 (2 mL), 2 a (122 mL, 1.00 mmol, 4.00 equiv), and TEMPO (78.8 mg, 0.50 mmol, 2.00 equiv) were reacted. The crude product was puri- fied by flash column chromatography on silica gel eluting with pentane/ AcOEt ((%AcOEt): 1 (30) ; 5 (50) ; 15 (300) 25 % (200 mL)) to give 3 a as a viscous oil (0.218 mmol, 79.3 mg, 87 %).
(300 MHz, $CDCl_3$): d = 7.34–7.20 (m, 5 H), 6.06 (s, 1 H), 3.87 (s, 3 H), 3.57 (s, 3H), 1.39ppm (s, 9H); ^{13}CNMR (75MHz, $CDCl_3$): d=161.4, 158.1, 155.2, 149.9, 138.7, 128.8, 127.4, 111.0, 84.3, 68.9 53.6, 52.2, 28.1 ppm; MS-ESI: m/z: calcd for $[C_{18}H_{21}NO_7Na]^+$: 386.1210; found: 386.1207.

Gini et al, *Chem. Eur. J.* **2015**, 21, 12053 – 12060

The Reaction Mechanism is the sequence of elementary steps from reactants to products



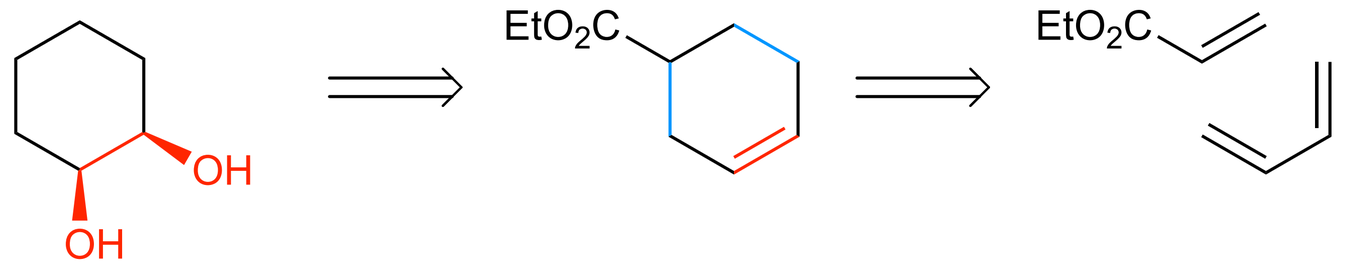


Scheme 5. Mechanistic proposal.

Gini et al, *Chem. Eur. J.* **2015**, 21, 12053 – 12060

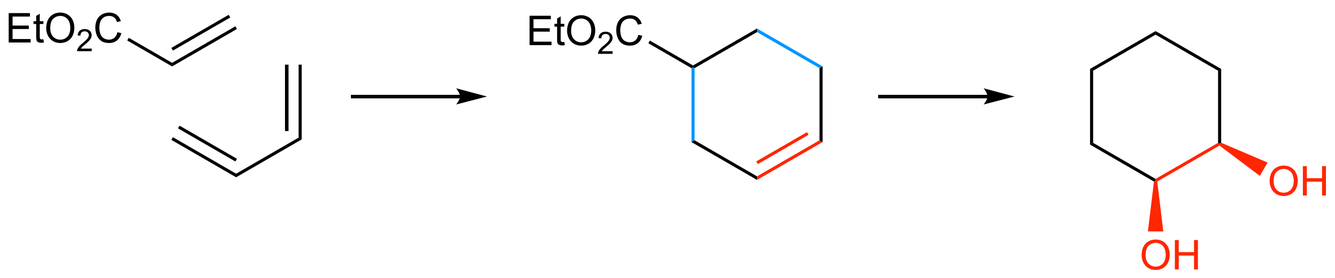
Retrosynthetic Analysis in Synthesis Planning

a) Retrosynthesis (backward)



Target Intermediate Starting materials

b) Synthetic route (forward)

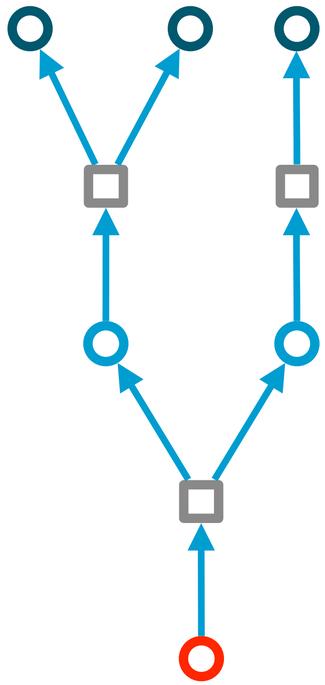


Starting materials Intermediate Target

Retrosynthesis vs Forward Synthesis: Synthesis Trees

Retrosynthesis

Backward



Building Blocks/
Starting materials

Reactions

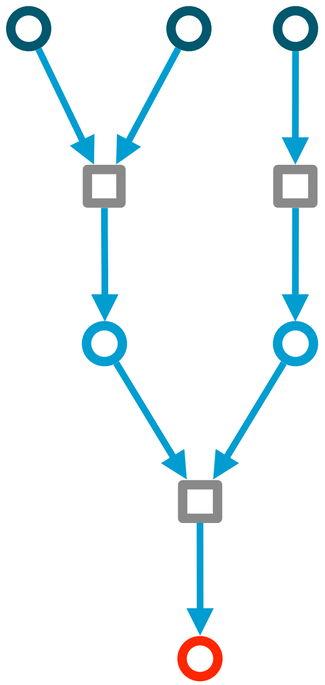
Intermediates

Reaction

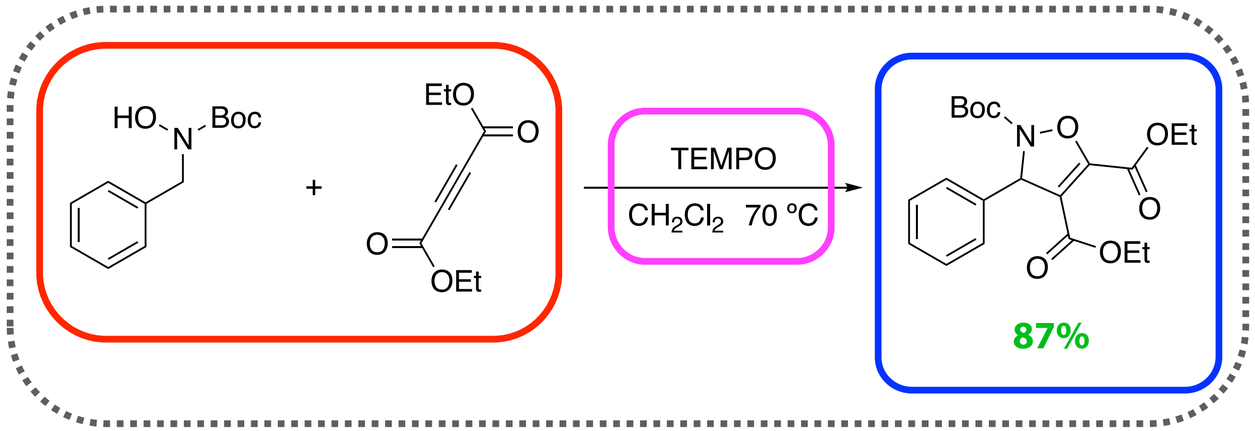
Desired Target Product

Actual Synthesis

Forward



Different Questions in Reaction Modelling



- What is the (major) product?
- What are the conditions?
- How can I make this product?
- What will be the yield, e.r., d.r.?
- Will this reaction run at all?
- What is the class of this reaction?
- Can I help to understand the mechanism?
- What is the procedure?

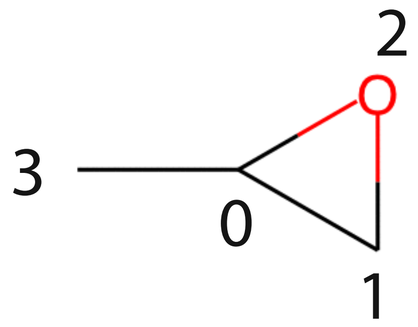
General procedure:
In a screw-cap Schlenk tube, the corresponding N-protected N-benzylhydroxyamine derivative 1 (1.00equiv) was dissolved in dry CH_2Cl_2 (2.00 mL). Dimethyl acetylenedicarbox- ylate (2 a) (4.00 equiv) and TEMPO (2.00 equiv) were added and the reaction mixture was stirred at 70 8C for 24 h. The solvent was re- moved under reduced pressure and the obtained crude product was purified by flash column chromatography on silica gel eluting with pentane/ AcOEt to give the corresponding N-protected isoxazoline 3.

2-tert-Butyl 4,5-dimethyl 3-phenylisoxazole-2,4,5-tricarboxylate (3a): According to the general procedure, N-Boc N-benzyl hydrox- ylamine (1 a) (0.25 mmol, 63.3 mg, 1.00 equiv), dry CH_2Cl_2 (2 mL), 2 a (122 mL, 1.00 mmol, 4.00 equiv), and TEMPO (78.8 mg, 0.50 mmol, 2.00 equiv) were reacted. The crude product was puri- fied by flash column chromatography on silica gel eluting with pentane/ AcOEt ((%AcOEt): 1 (30) ; 5 (50) ; 15 (300) 25 % (200 mL)) to give 3 a as a viscous oil (0.218 mmol, 79.3 mg, 87 %).
(300 MHz, $CDCl_3$): d = 7.34–7.20 (m, 5 H), 6.06 (s, 1 H), 3.87 (s, 3 H), 3.57 (s, 3H), 1.39ppm (s, 9H); ^{13}CNMR (75MHz, $CDCl_3$): d=161.4, 158.1, 155.2, 149.9, 138.7, 128.8, 127.4, 111.0, 84.3, 68.9 53.6, 52.2, 28.1 ppm; MS-ESI: m/z: calcd for $[C_{18}H_{21}NO_7Na]^+$: 386.1210; found: 386.1207.

How to represent molecules with computers

Molecular Graph

atoms & bonds as objects



SMILES: C1OC1C

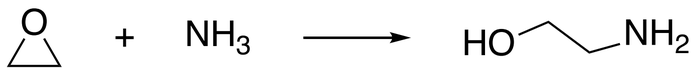
Adjacency Matrix
(does not always specify bond order!)

```
atom:    0  1  2  3        atom
array([[0, 1, 1, 0],      0
       [1, 0, 1, 0],      1
       [1, 1, 0, 1],      2
       [0, 0, 1, 0]])     3
```

Caveat: There is no inherent order in atoms/bonds: permutation invariance

Representing Reactions

reactants and products are sets of molecules



Reaction SMILES: `reactants>reagents>products`

`C1OC1.N>>OCCN`

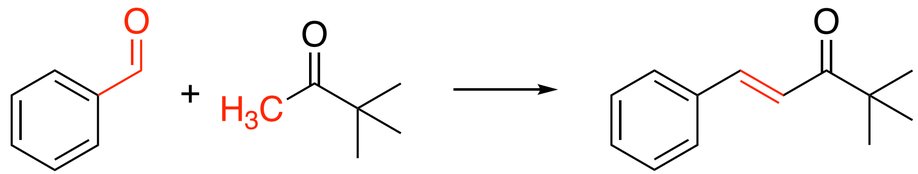
All reactants in one matrix

```
         O  C  C  N
array([[0, 1, 1, 0],    O
       [1, 0, 1, 0],    C
       [1, 1, 0, 0],    C
       [0, 0, 0, 0]])   N
```

Product matrix

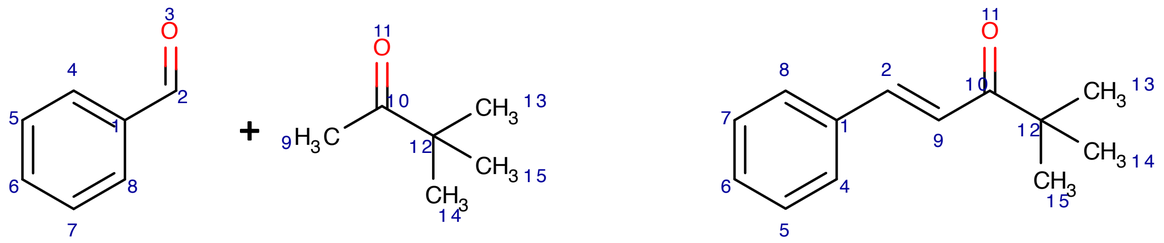
```
         O  C  C  N
array([[0, 1, 0, 0],    O
       [1, 0, 1, 0],    C
       [0, 1, 0, 1],    C
       [0, 0, 1, 0]])   N
```

Reaction Center



the set of atoms and bonds that get changed overall in the course of the reaction
not necessarily related to the mechanism

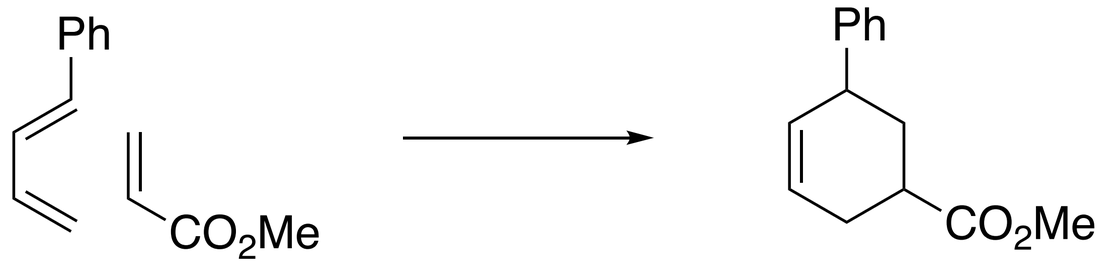
Reaction Mapping



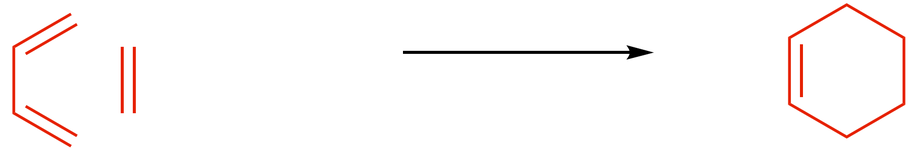
Automatic Reaction Mapping assignment is still not perfect
Manual Assignment significant work

Reaction Rules & Reaction Templates

a) Known reaction



b) Extraction of a general rule



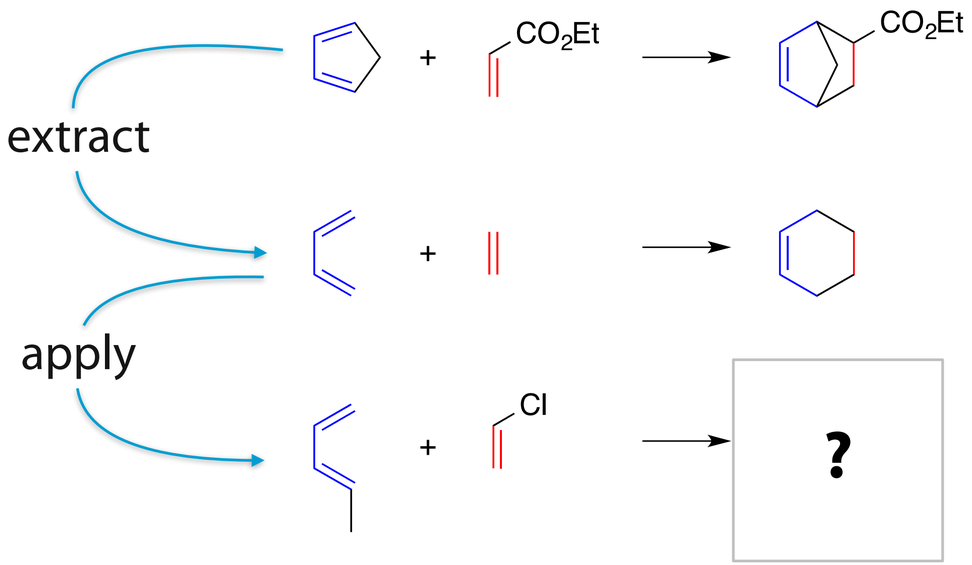
manual or automatic

Template: Usually refers to the reaction center + environment only
Rule: All templates are rules, but rules also contain additional information

Templates are a composition of graph edits (Ugi)

Reaction Rules & Reaction Templates

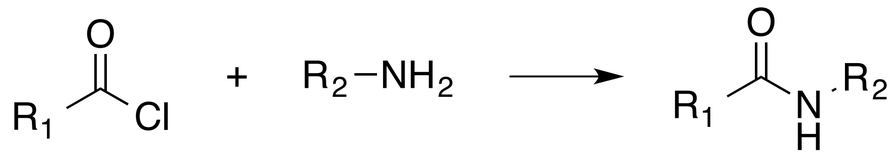
Work in both directions (forward/retro)



Rule application algorithm

1) Match left side of rule in starting graph
2) Cut out match
3) Glue in right side of rule
4) Return target graph

Reaction Rules & Reaction Templates: Advantages




```
[#6:1]-[#6:2]([Cl:3])=[O:4].[#6:6]-[#7;h2:5]>>[#6:6]-[#7;h1:5]-[#6:2](-[#6:1])=[O:4]
```

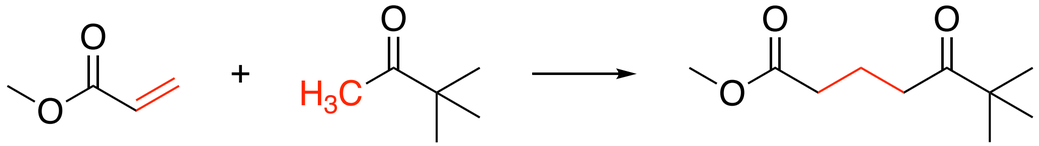
- Deeply rooted in chemists' language
- Perfect with perfect rule base, decent with good rule base
- no copying errors

Kayala, M., Baldi, P.; *J. Chem. Inf. Model.* **2012**, *52*, 2526–2540

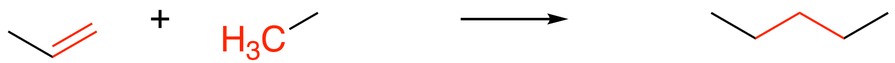
B.A. Grzybowski *et al. Angew. Chem. Int. Ed.* **2016**, *55*, 5904-5937

Activating Groups need to be captured

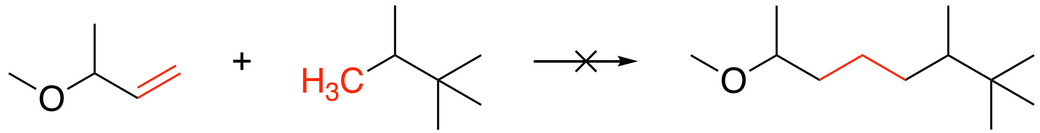
observed reaction



oversimple template



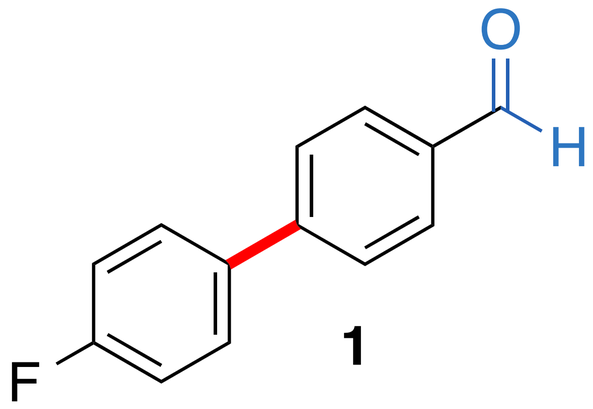
this reaction will fail



Tolerated Functional Groups need to be captured

Problem: Which is the correct rule to apply?

(1) Query



1

⟹ **?**

(2a) Rule A
Suzuki

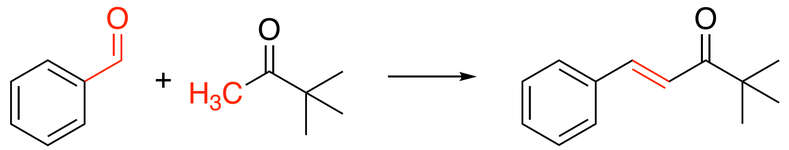
R^1-R^2 ⟹ $R^1-B(OH)_2 + Br-R^2$

(2b) Rule B
Kumada

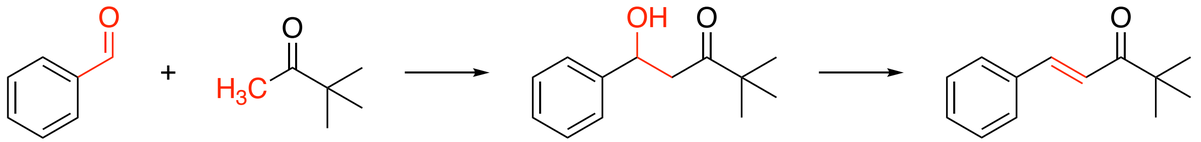
R^1-R^2 ⟹ $R^1-MgBr + Br-R^2$

Templates do not always capture intermediates

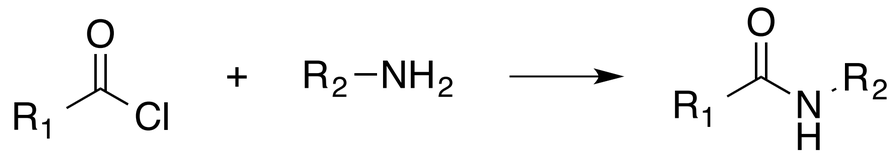
observed reaction



What actually happens



Reaction Rules & Reaction Templates




```
[#6:1]-[#6:2]([Cl:3])=[O:4].[#6:6]-[#7;h2:5]>>[#6:6]-[#7;h1:5]-[#6:2](-[#6:1])=[O:4]
```

- Deeply rooted in chemists' language
- Perfect with perfect rule base, decent with good rule base
- no copying errors
- rules have to be created (manually, automatically extracted)
- reactivity conflicts, selectivity have to be captured
- many reaction mechanisms and scope not well understood
- purification, solubility, stability not taken into account
- no inherent ranking mechanism

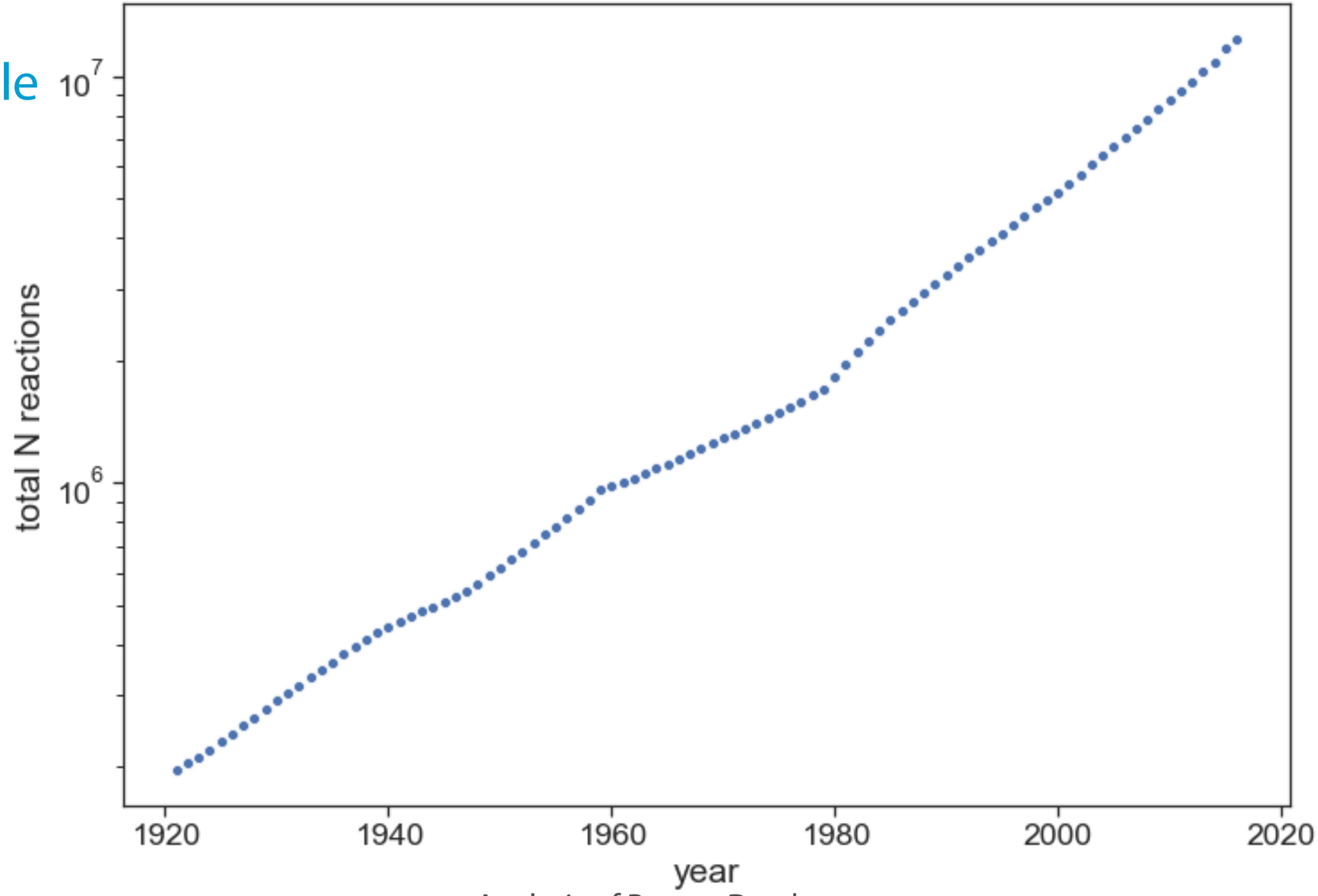
Kayala, M., Baldi, P.; *J. Chem. Inf. Model.* **2012**, *52*, 2526–2540

B.A. Grzybowski *et al. Angew. Chem. Int. Ed.* **2016**, *55*, 5904-5937

Why are data driven approaches for reaction modelling appealing?

Growth of Chemical Data is unbroken!

log scale

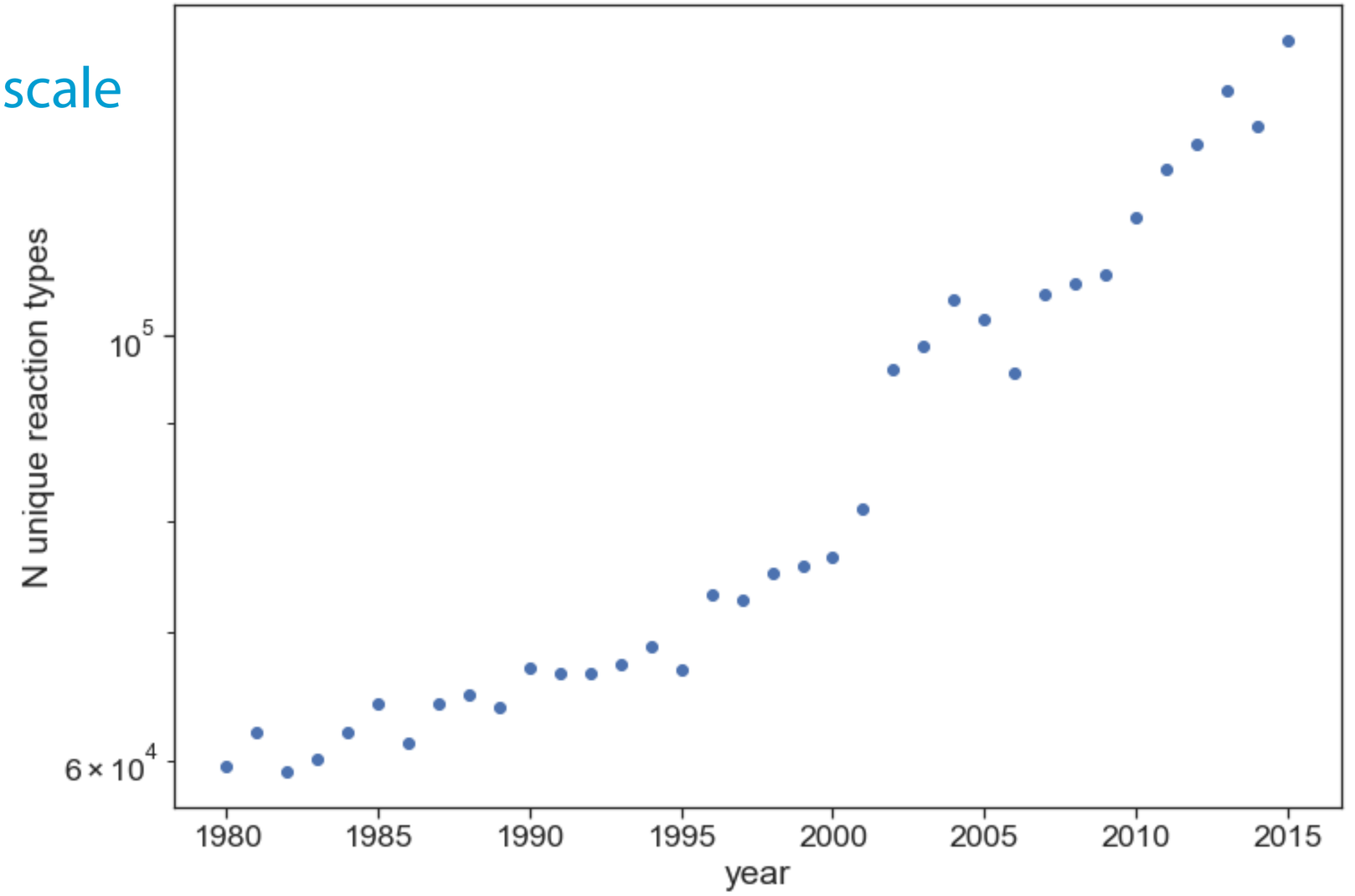


Analysis of Reaxys Database

Chemists' creativity does not slow down!

Number of **unique** reaction types / year

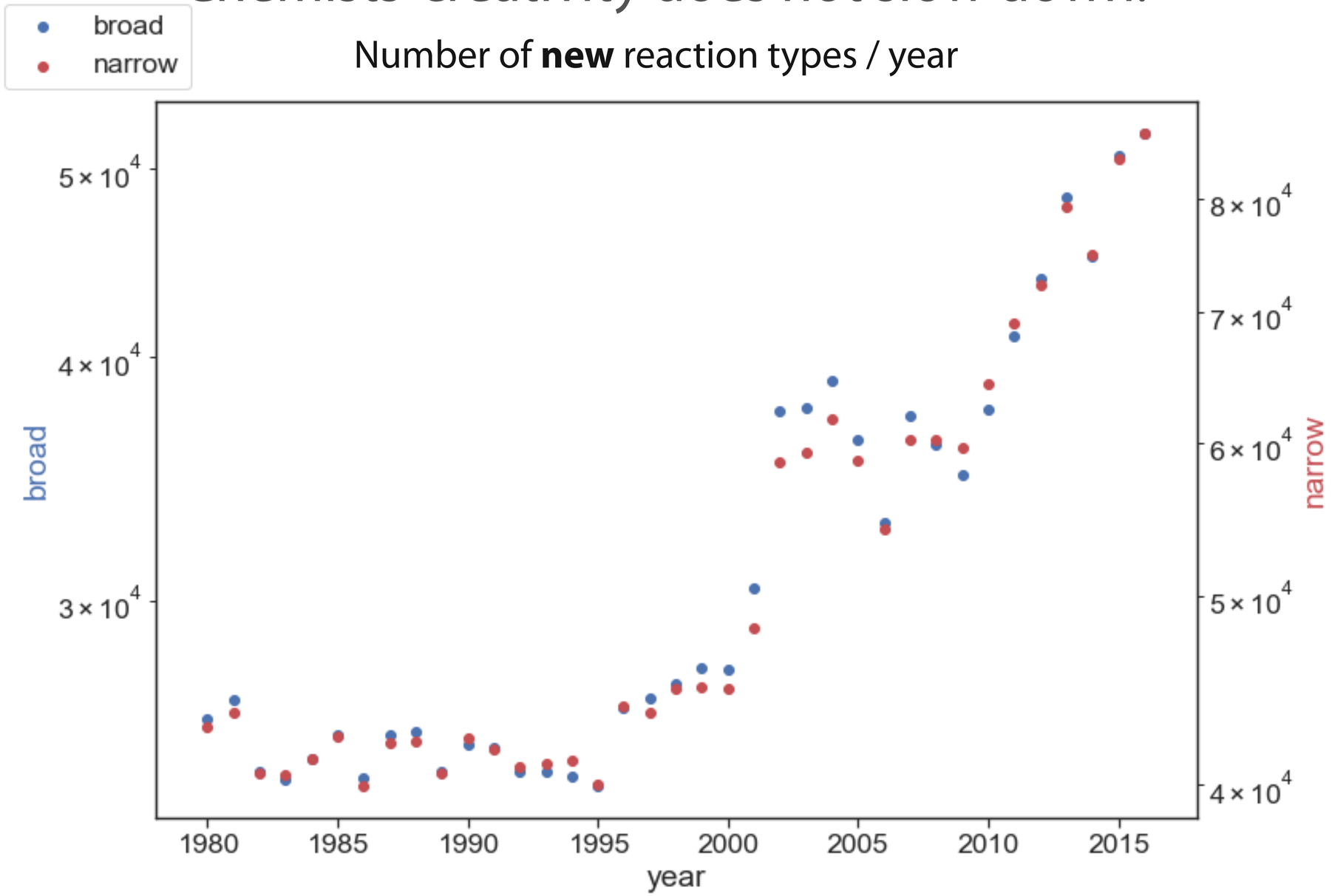
log scale



Via extracted reaction rules/templates, Analysis of Reaxys Database

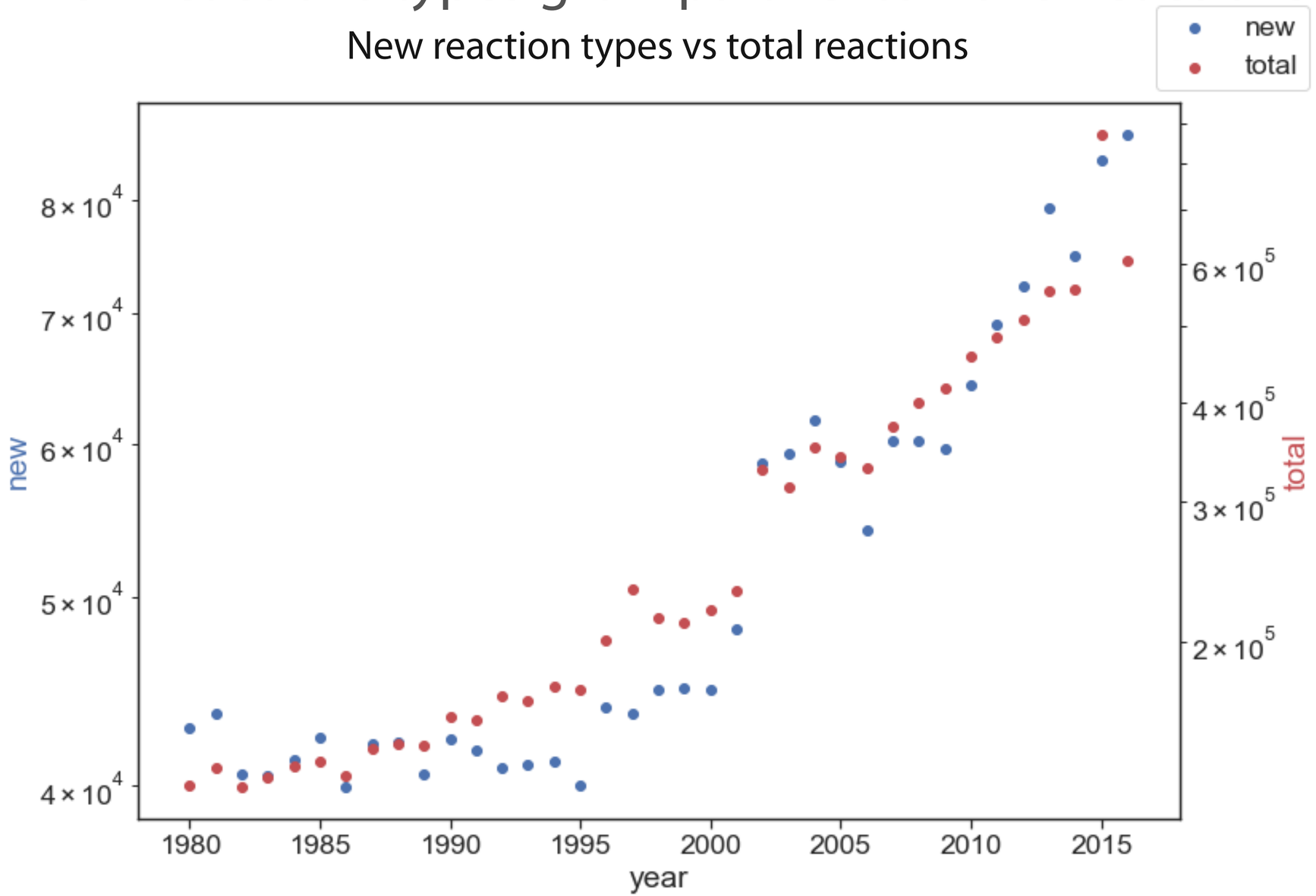
Chemists' creativity does not slow down!

log scale

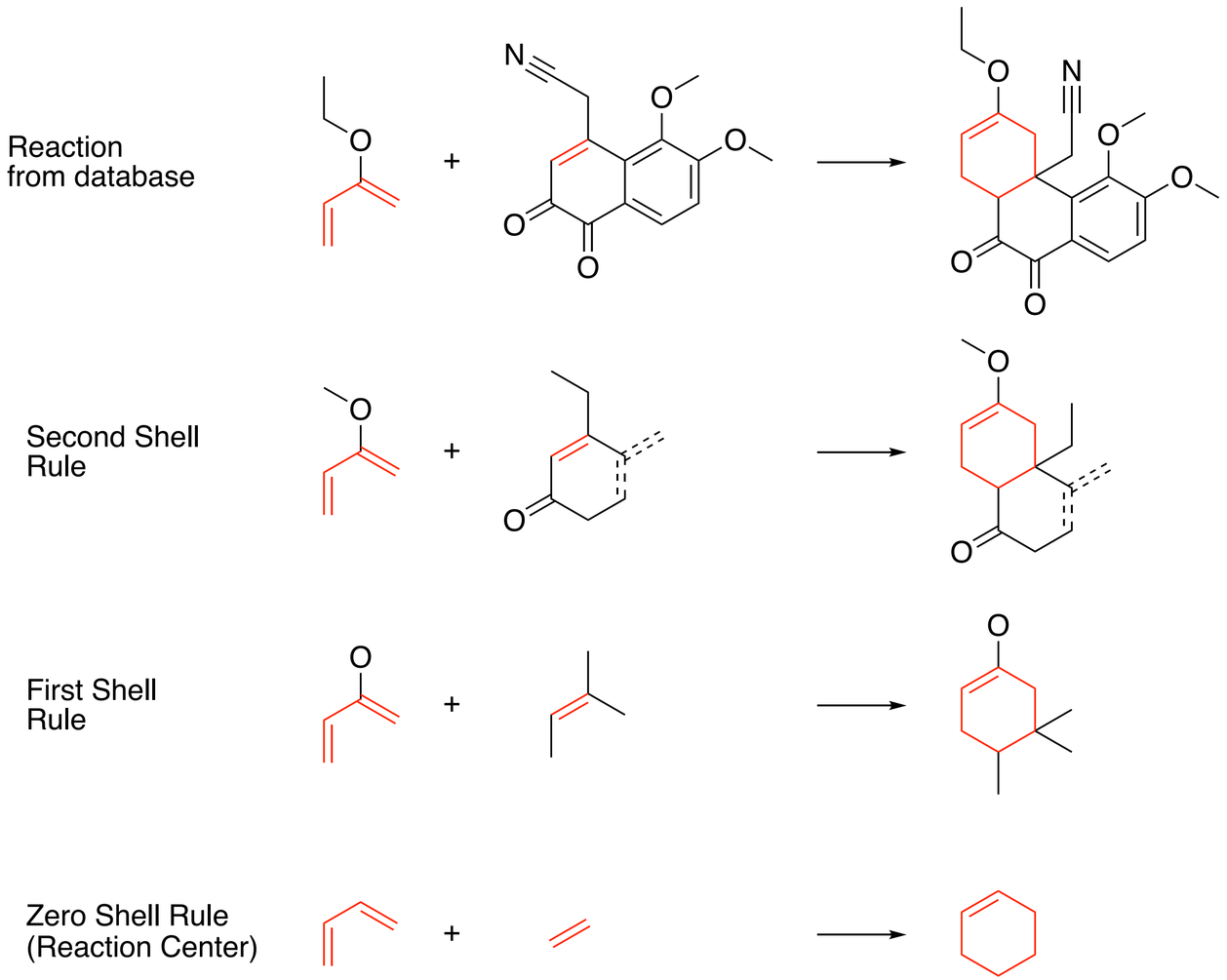


Via extracted reaction rules/templates, Analysis of Reaxys Database

New reactions types grow parallel to novel reactions



Automatic Template Extraction via Algorithms



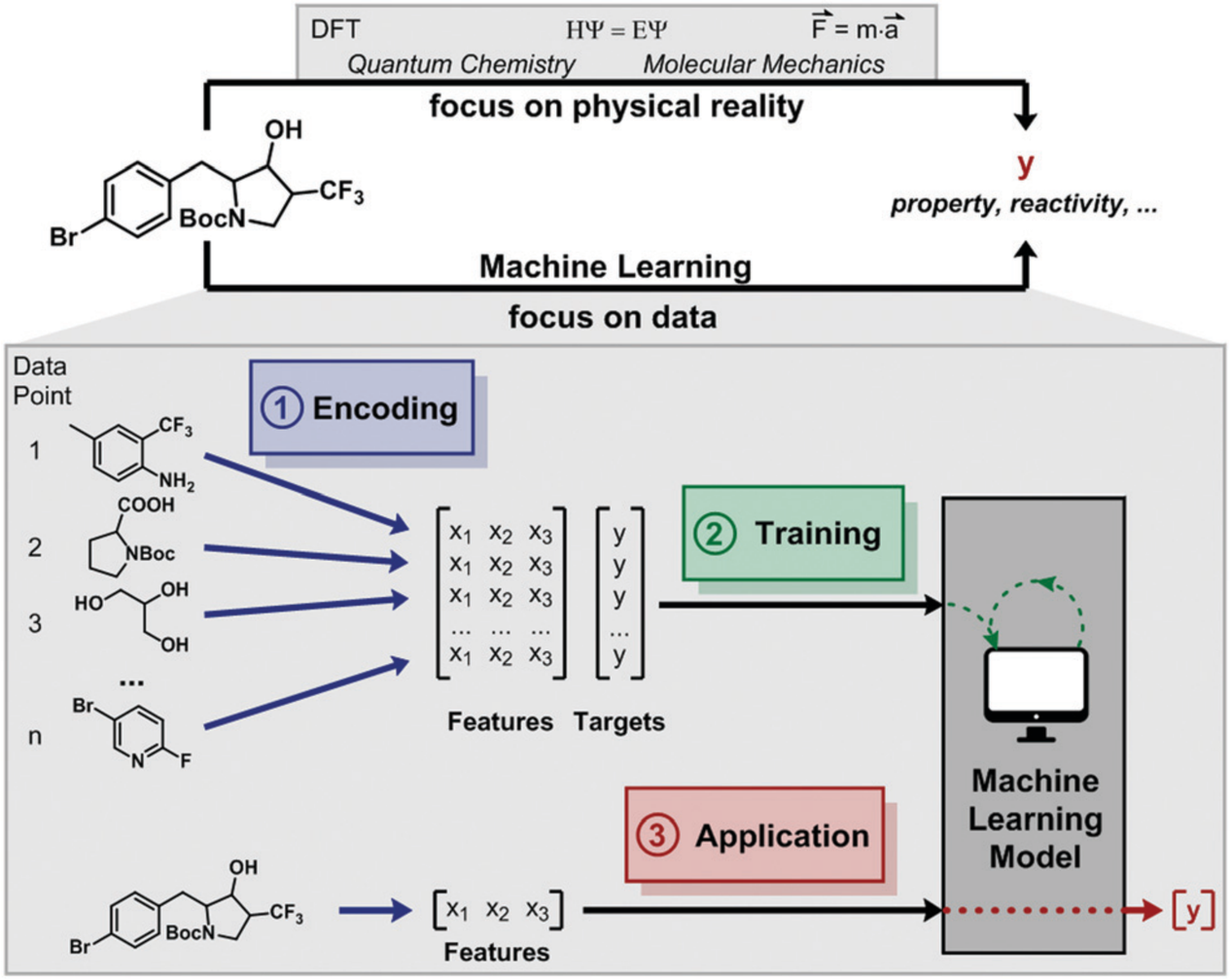
Law et al. *JCIM* **2009**, 593–602
Christ, Zentgraf, Kriegl *JCIM* **2012** 1745
Saller et al. *Org. Process Res. Dev.* **2015**, 357–368

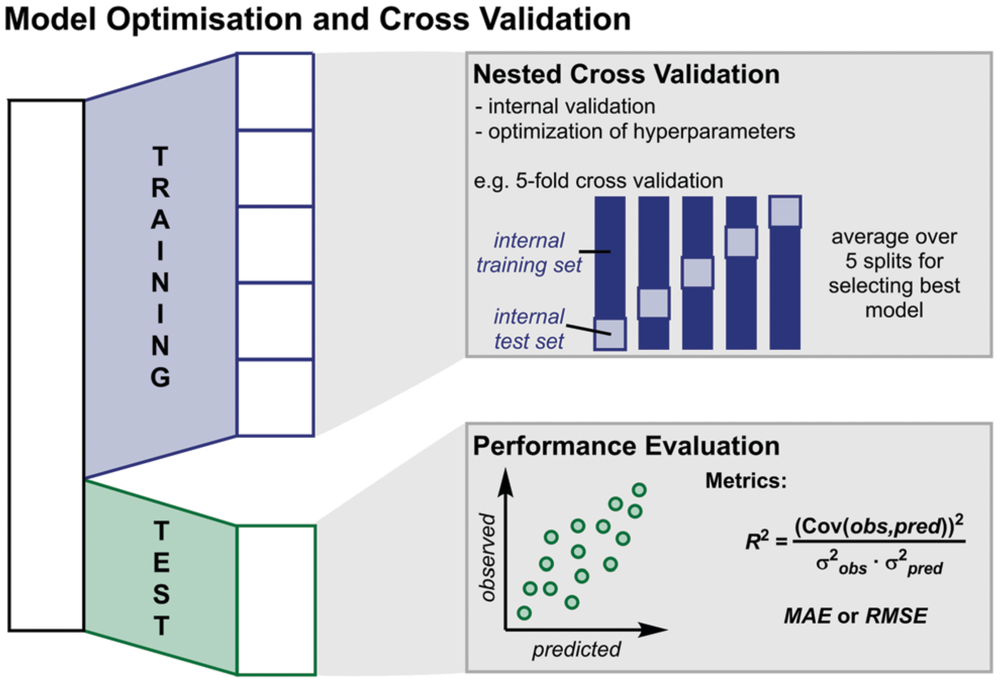
Manual coding vs Automatic Template Extraction

Method	Manual Coding	Automatic Extraction
Human Effort	Very high (decades)	Very Little
Requirements	A large team of organic chemists (expensive)	Reaction Database (depends)
Scalability to new reactions	Low, need to be encoded anew	15 million reactions over night (laptop)
Updating Rulebase	Complex, need to revisit old rules	Simple (see above)
Error Sources	Expertise of chemist, many reactions are not well enough understood	Current extraction algorithms often do not capture activating groups and scope well, lack of negative data

Law et al. *JCIM* **2009**, 593–602
Christ, Zentgraf, Kriegl *JCIM* **2012** 1745
Saller et al. *Org. Process Res. Dev.* **2015**, 357–368

Supervised Machine Learning





Molecular Representations for Machine Learning: Featurization

$$f : \mathcal{M} \rightarrow \mathbb{R}^D$$

the set of all possible molecules (graphs or 3D)

D-dimensional real vectors

- fingerprints (often sparse)
- physicochemical/topological descriptor vectors
- Graph Neural Networks

Reaction Representations for Machine Learning

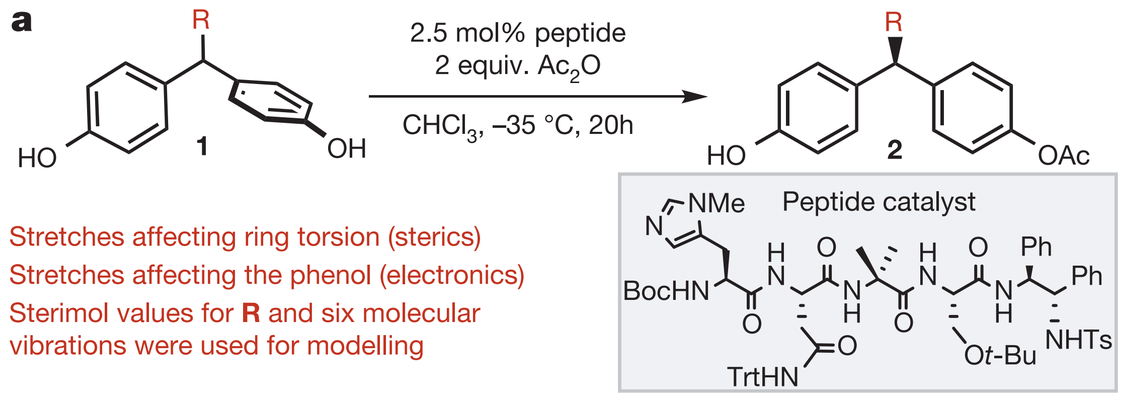
- reaction difference fingerprints (sum of products - sum of reactants)
- Reaction Graph Neural Networks based on reaction center
- Seq2Seq Descriptors based on Reaction SMILES

Supervised Machine Learning: Classification and Regression

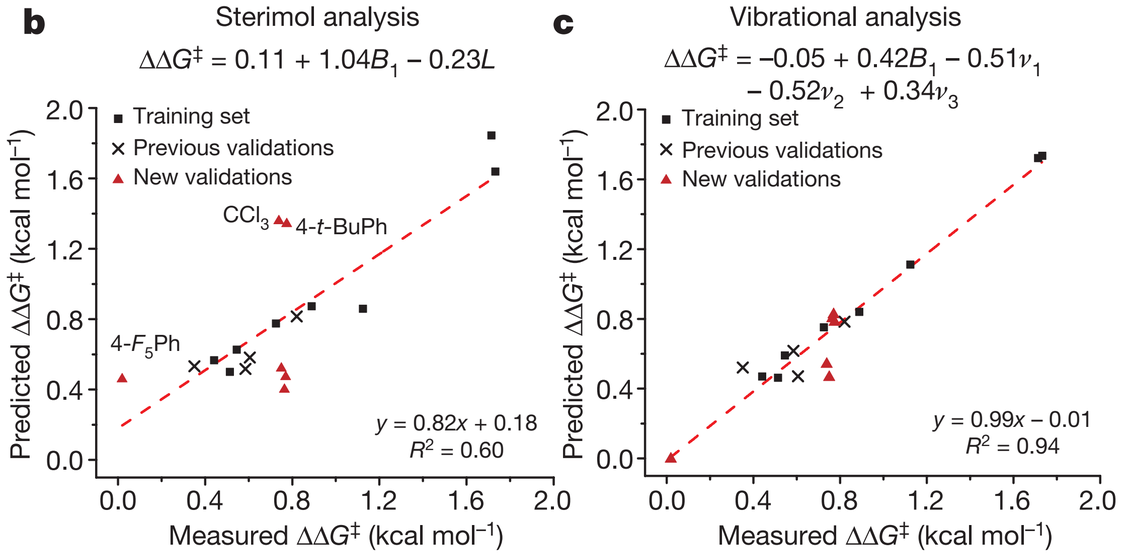
$$y = f_\theta(\phi(m))$$

	Regression	**Binary Classification**	**Multi-Class Classification**
$y =$	real number	probability [0,1]	probability vector $[0,1]^c$
	%Yield, e.r.	reactive/unreactive	Reaction Class [DielsAlder: 0.2, Suzuki: 0.7, Aldol: 0.1]

(Mechanistic) QSRR Modeling



- Stretches affecting ring torsion (sterics)
- Stretches affecting the phenol (electronics)
- Sterimol values for **R** and six molecular vibrations were used for modelling

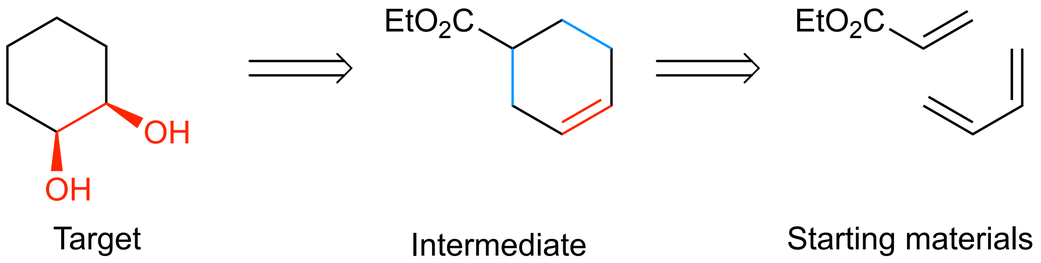


Milo, Sigman, Doyle, Gensch, Gasteiger, Glorius, ….

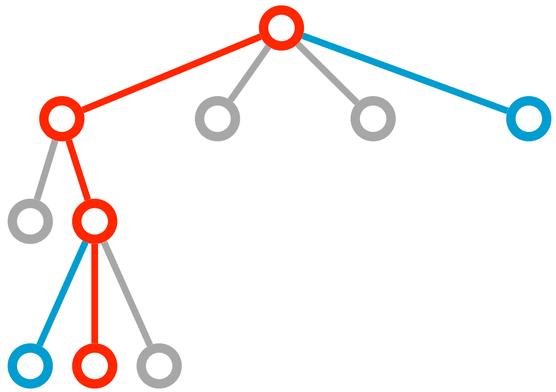
Ingredients of Computer-Aided Synthesis Planning Algorithms

- Module to propose feasible retrosynthetic disconnections (now ML)
- efficient search algorithm
- stop criteria (building blocks) and ranking

Target or Intermediate Molecule → [redacted] → Ranked List of 1-step Precursor Sets



Target Intermediate Starting materials



Brief History of Computer-Aided Synthesis Planning

1963
Vladuts (USSR)
Proposes Computer-Aided Synthesis Planning

1967
Corey (USA)
Formalises Retrosynthesis & First Expert System Implementation

1970ies
Ugi (D)
Formal Logic

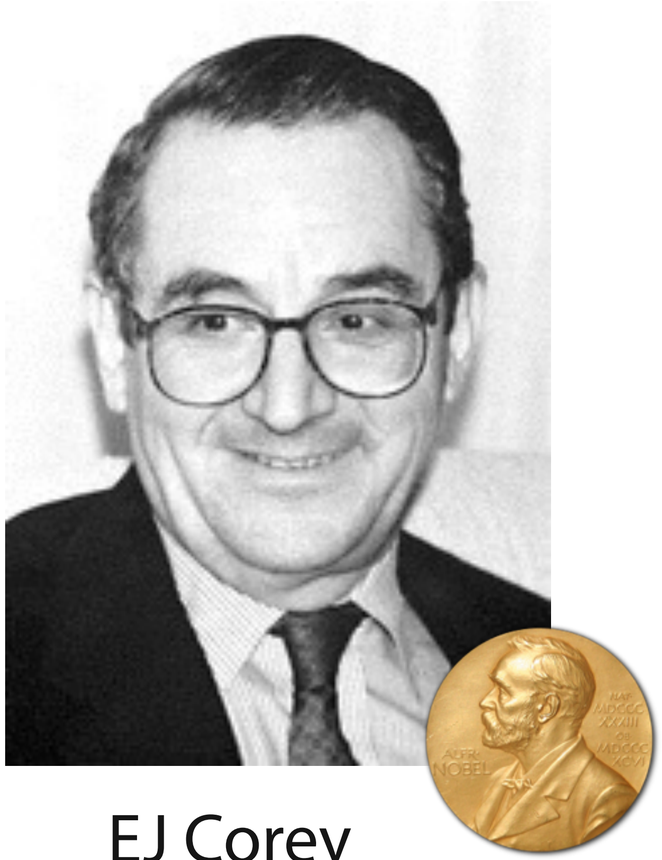
1970-1990
Expert Systems

1990-2010
Automatic Rule Extraction

1996
Gasteiger & Ihlenfeldt
Current Approaches: Dead End?

2016
Segler, Coley, Schwaller, …
Machine Learning

2016
Synthia (**Expert System**, Grzybowski)



EJ Corey

Nobel prize 1990

Vleduts (1963), Corey (1968)

Write down all of chemical knowledge in logic form

Great for humans!

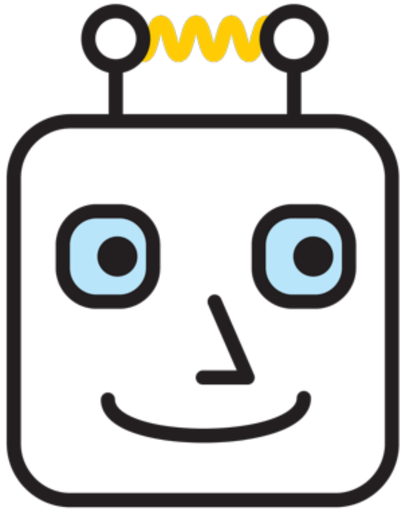
Not so much for machines?

"The synthetic chemist is more than a logician and strategist; […] These added elements provide the touch of artistry which can hardly be included in a cataloguing of the basic principles of synthesis, but they are very real and extremely important." (Corey)

Vléduts, G. *Inform. Storage Retrieval* 1, 117–146 (1963).

Corey, *The Logic of Chemical Synthesis*

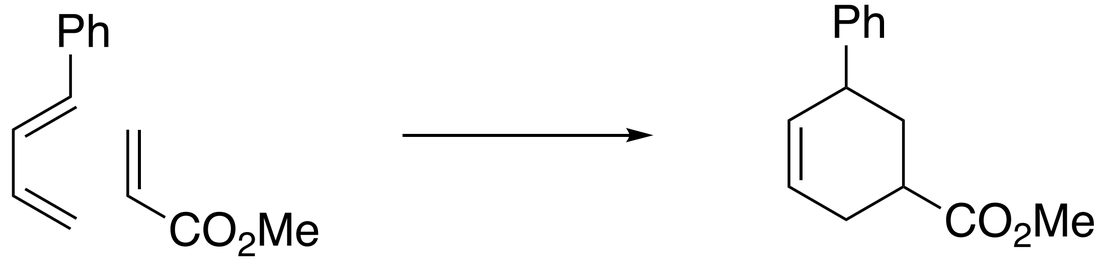
Idea: Machine Learning & Reinforcement Learning for Search



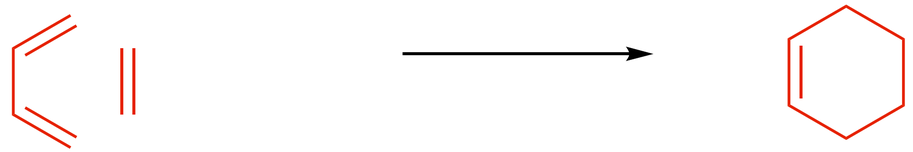
- Module to propose feasible retrosynthetic disconnections
  - Learn to predict disconnections
  - Learn to predict reactions
- modern efficient search
- ML provides a rigorous metrics framework!

Assigning Molecules to Rules gives Labels for ML

a) Known reaction

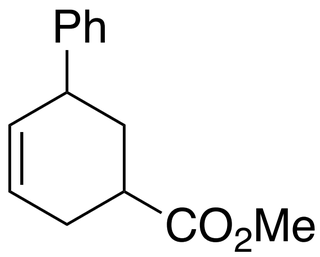


b) Extraction of a general rule



#987128 (Diels Alder)

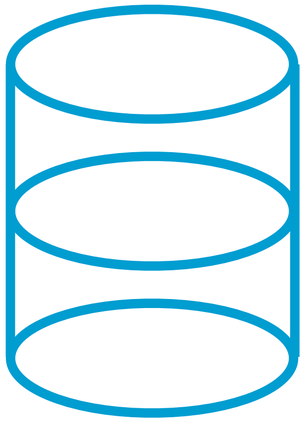
This product X can be made with this rule y: P(Y|X)



Label: #987128 (Diels Alder)

Rule assignment gives us labeled dataset for classification

Data of our entire discipline!



- 11 M reactions

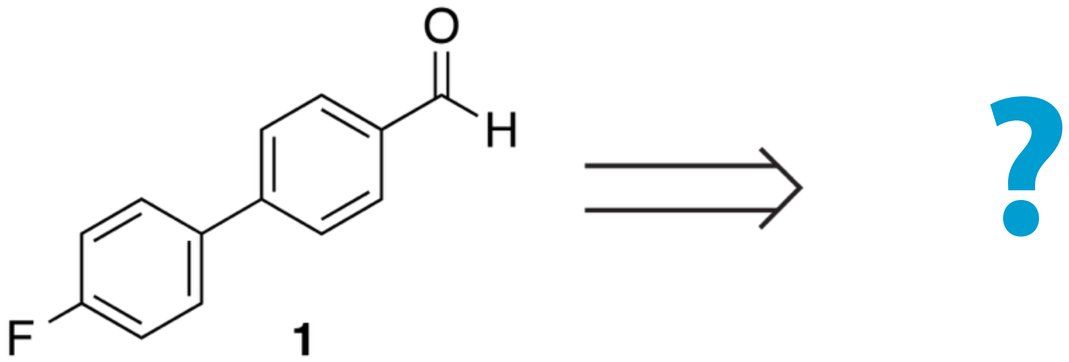
Successful Reactions contain implicit knowledge!

Data: Reaxys

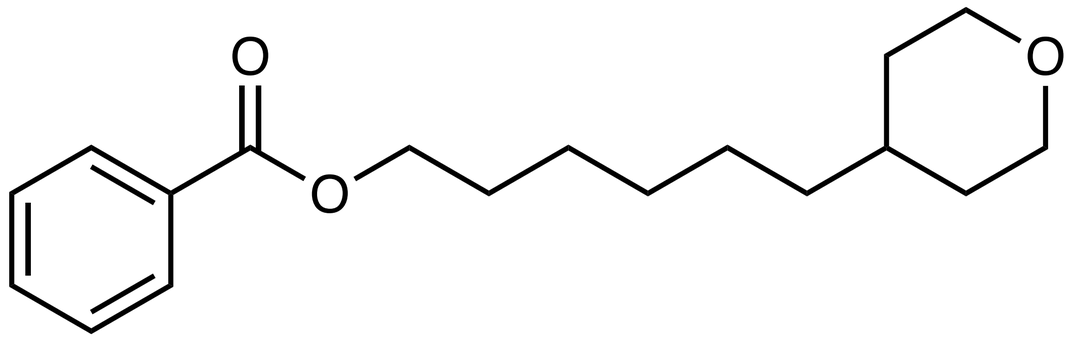
Challenges

- learn the rules
- ***predict likely disconnections***
- filter out infeasible reactions
- efficient search

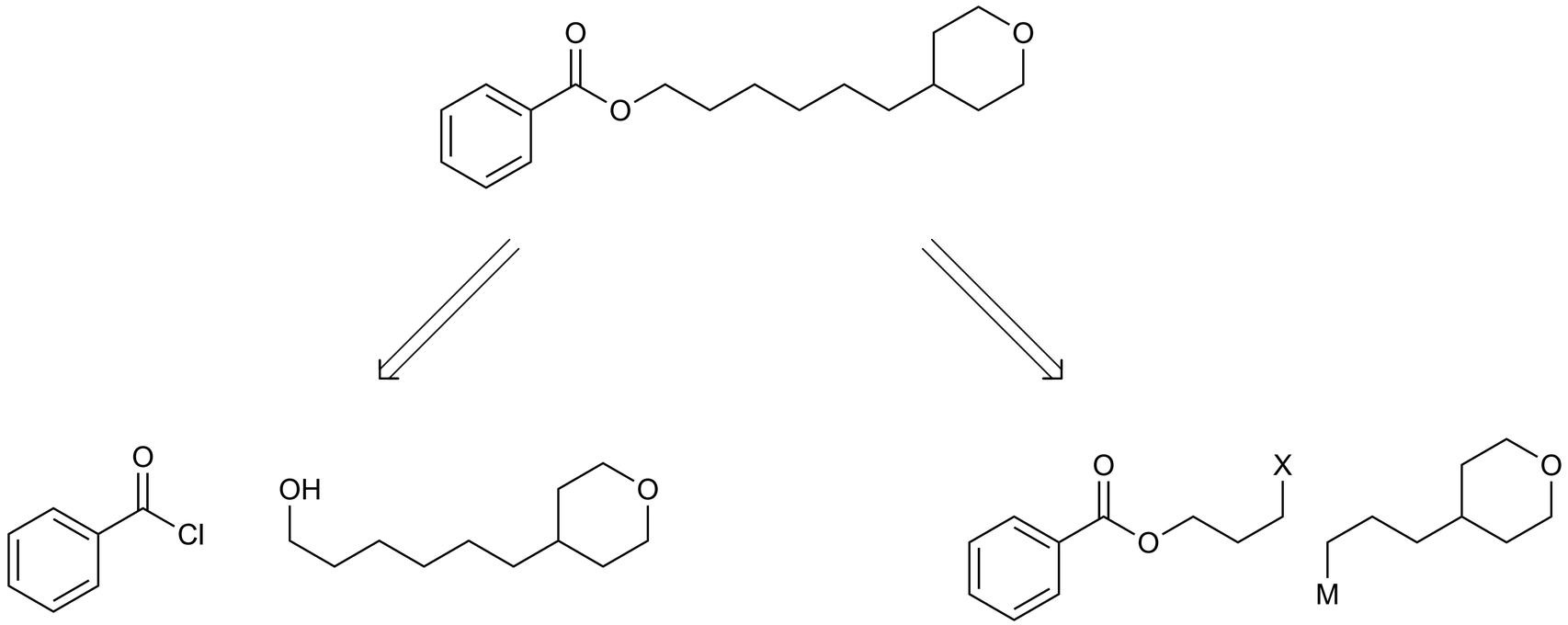
Retrosynthetic disconnection prediction: Multi-class Classification



How to make this molecule?

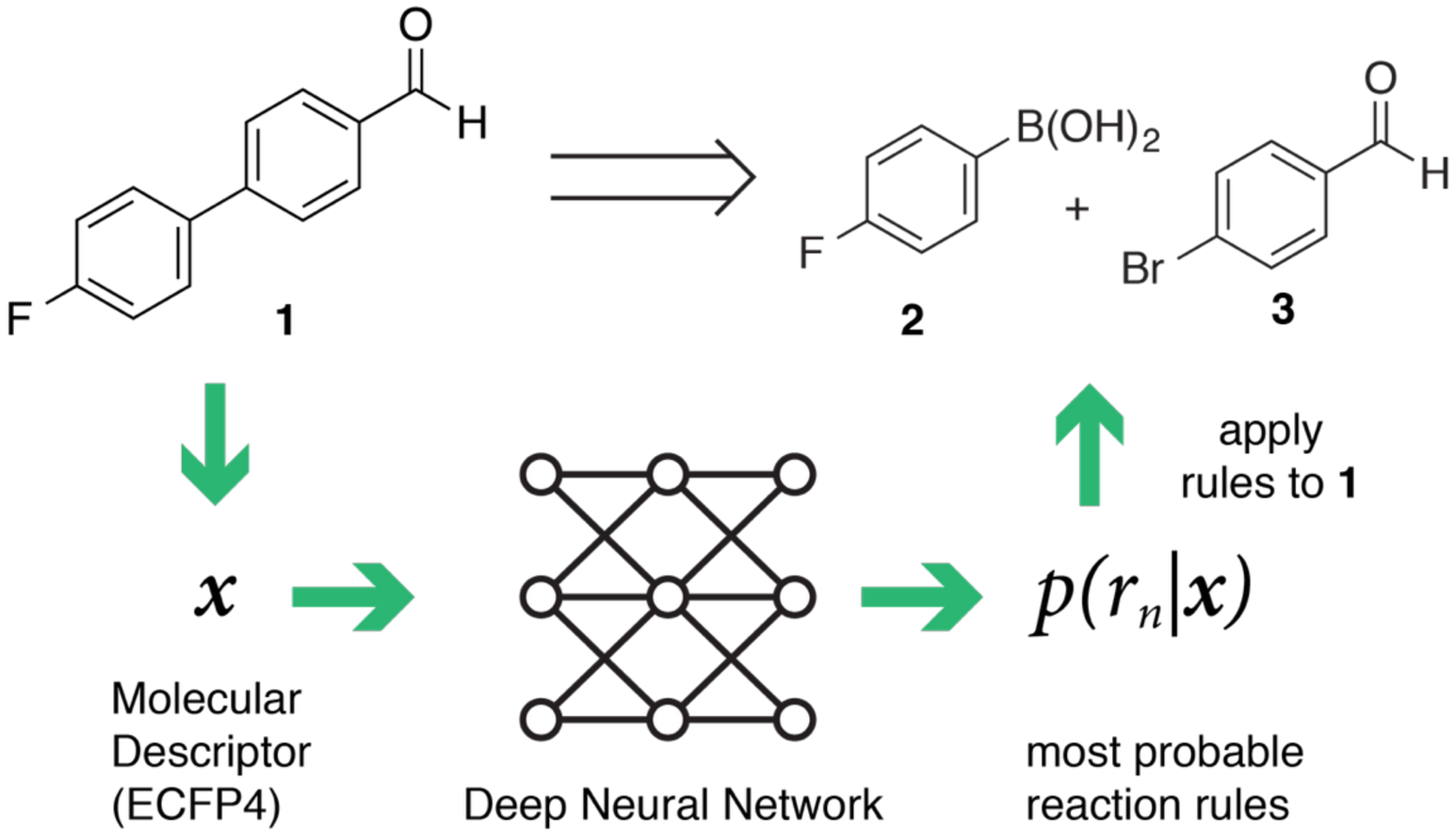


Pattern Recognition



Retrosynthetic disconnection prediction: Multi-class Classification

Mimics Chemical Intuition & Allows to learn tolerated molecular context!

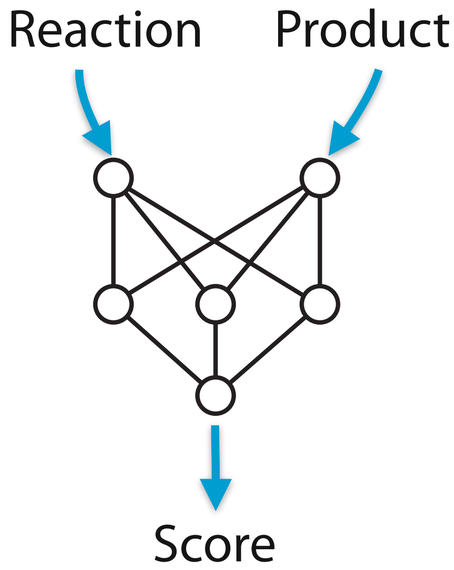


Segler, Waller, *Chem. Eur. J.* **2017,** *DOI: 10.1002/chem.201605499*
Deep Highway Networks (Schmidhuber), ELU nonlinearity (Clevert, Unterthiner, Hochreiter)

Challenges

- learn the rules
- focus on most promising routes first
- ***filter out infeasible reactions***
- efficient search

Reaction Prediction: In-scope Filter



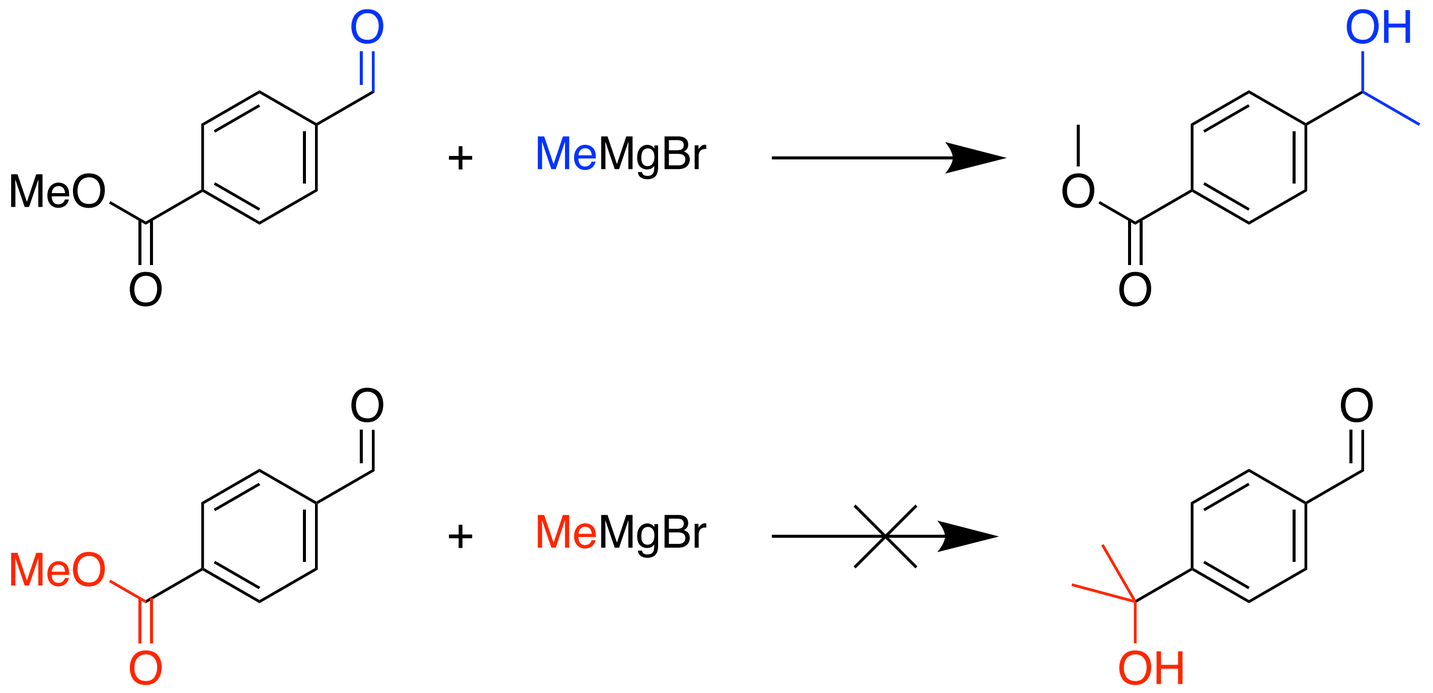
Binary Classification using real positive and mined negative data [1, 2]

[1] Segler, Waller, *Chem. Eur. J.* **2017,** 6118 [2] Coley, Jensen, ACS Cent. Sci. **2017**, 434

Reaction Prediction

No failed reaction in literature?

Make your own![1,2]

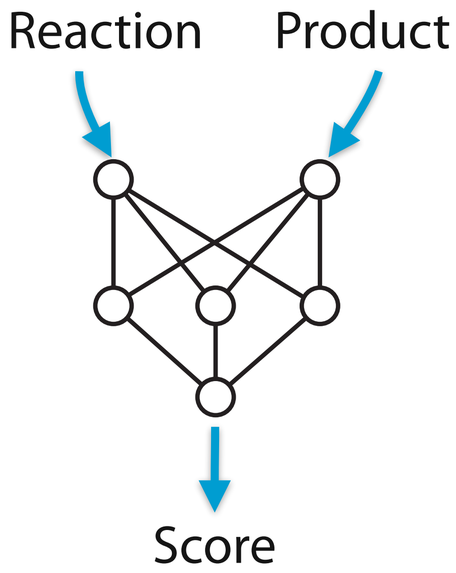


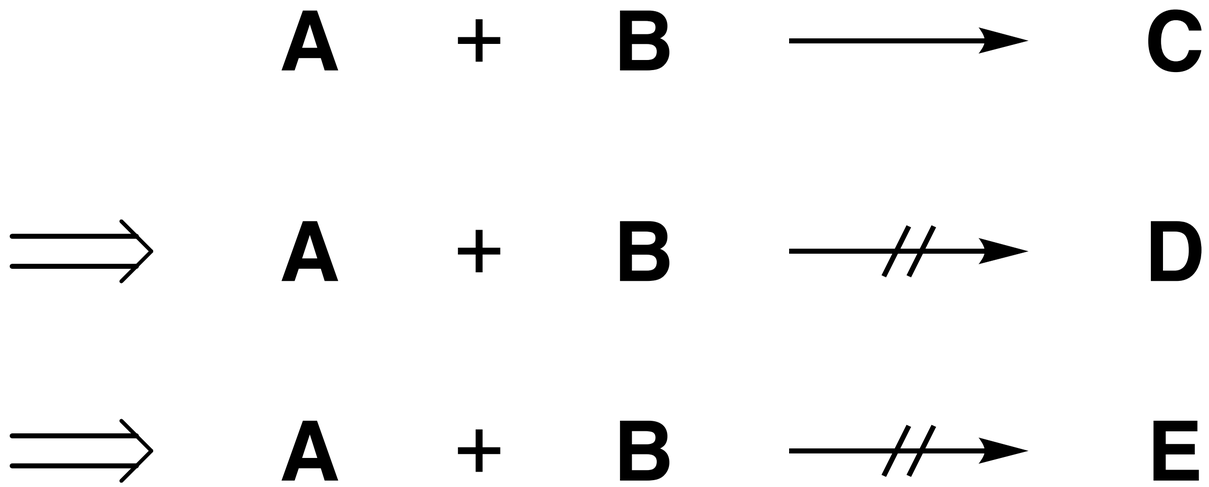
[1] Segler, Waller, *Chem. Eur. J.* **2017,** 6118 [2] Coley, Jensen, ACS Cent. Sci. **2017**, 434

Reaction Prediction: In-scope Filter

No failed reaction in literature?

Make your own!





Neural Network: ROC AUC 0.986

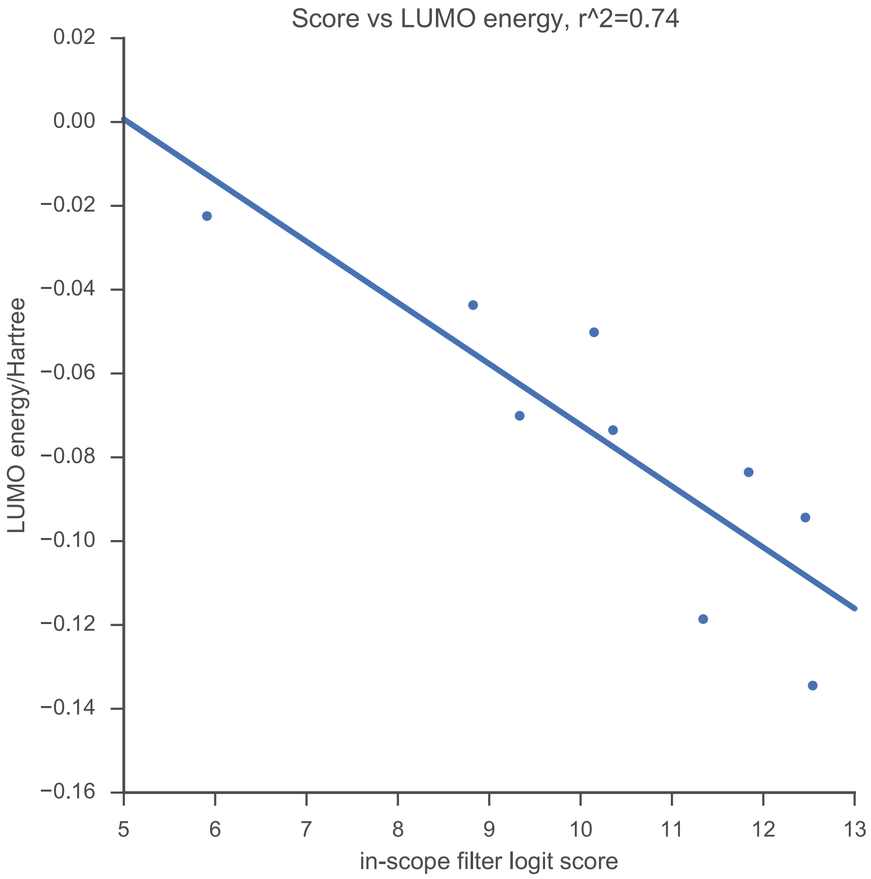
False positive rate: 1.5%

[1] Segler, Waller, *Chem. Eur. J.* **2017,** 6118 [2] Coley, Jensen, ACS Cent. Sci. **2017**, 434

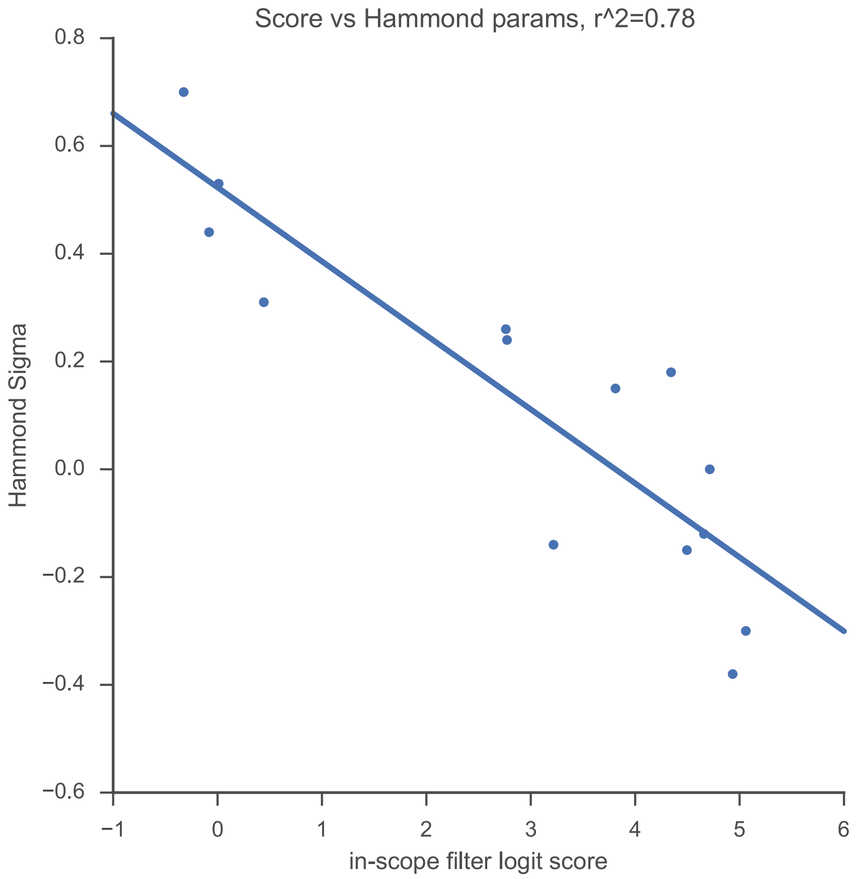
In-scope Filter

f: (ProductFP, ReactionFP) -> [0,1]

a) Diels-Alder reactions with Cyclopentadiene



b) para-Bromination of benzenes



Output correlates with LUMO energies and Hammett parameters!

Challenges

- learn the model
- focus on most promising actions first
- filter out infeasible reactions
- ***efficient search***

Heuristic Best First Search

Idea: Define strong heuristic function to score nodes

For example: Split up molecules in equally sized parts, simplify molecule, cleave strategic bonds first…

Problems:

- Chemists disagree about good solutions, intuition is not addressed
- Synthesis only solved at the end
- Molecular complexity needs to be tactically increased (Protecting groups!)

Monte Carlo Tree Search (MCTS): Idea

- Approximate values online by random MC simulation (Agent picks transforms randomly until end of synthesis)
- Use these approximated values to build search tree

=> Not dependent on strong heuristic!
=> Can deal with very high branching factors
=> can be guided by predicted value or probability of disconnection

R. Coulom, "Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search," in *Proc. 5th Int. Conf. Comput. and Games*, **2006**, pp. 72–83.
L. Kocsis and C. Szepesvari, "Bandit based Monte-Carlo Planning," in Euro. Conf. Mach. Learn. Berlin, Germany: Springer, **2006**, pp. 282–293.

Quantitative analysis on 500 random molecules

Method	Scoring	solved/%	time per molecule/s
BFS	Heuristics [1]		
BFS	Neural Net		
MCTS	Neural Net		

trained on data < 2015, molecules first reported >= 2015

[1] B.A. Grzybowski *et al. Angew. Chem. Int. Ed.* **2016**, *55*, 5904-5937

Quantitative analysis on 500 random molecules

Method	Scoring	solved/%	time per molecule/s
BFS	Heuristics [1]	56	422
BFS	Neural Net	84	39
MCTS	Neural Net	95	13

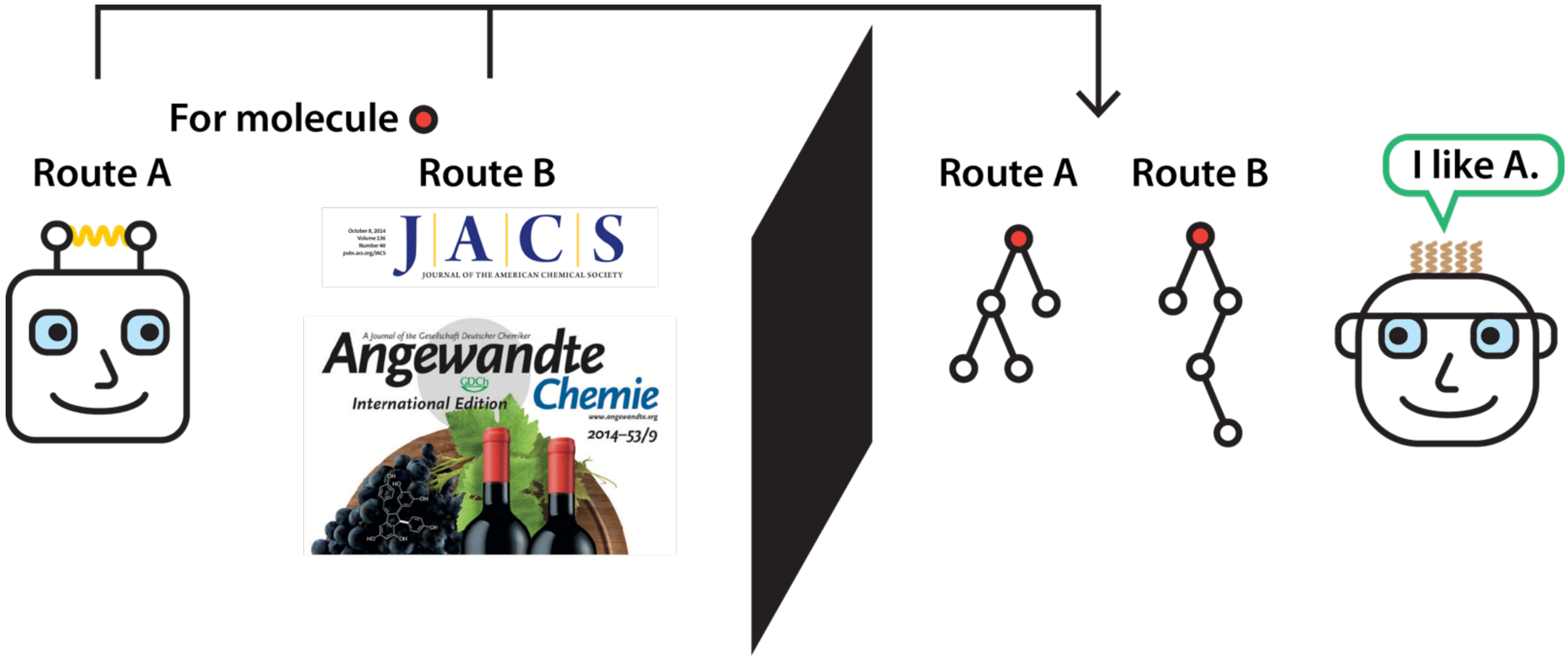
trained on data < 2015, molecules first reported >= 2015

[1] B.A. Grzybowski *et al. Angew. Chem. Int. Ed.* **2016**, *55*, 5904-5937

How to test the quality of a retrosynthesis system?

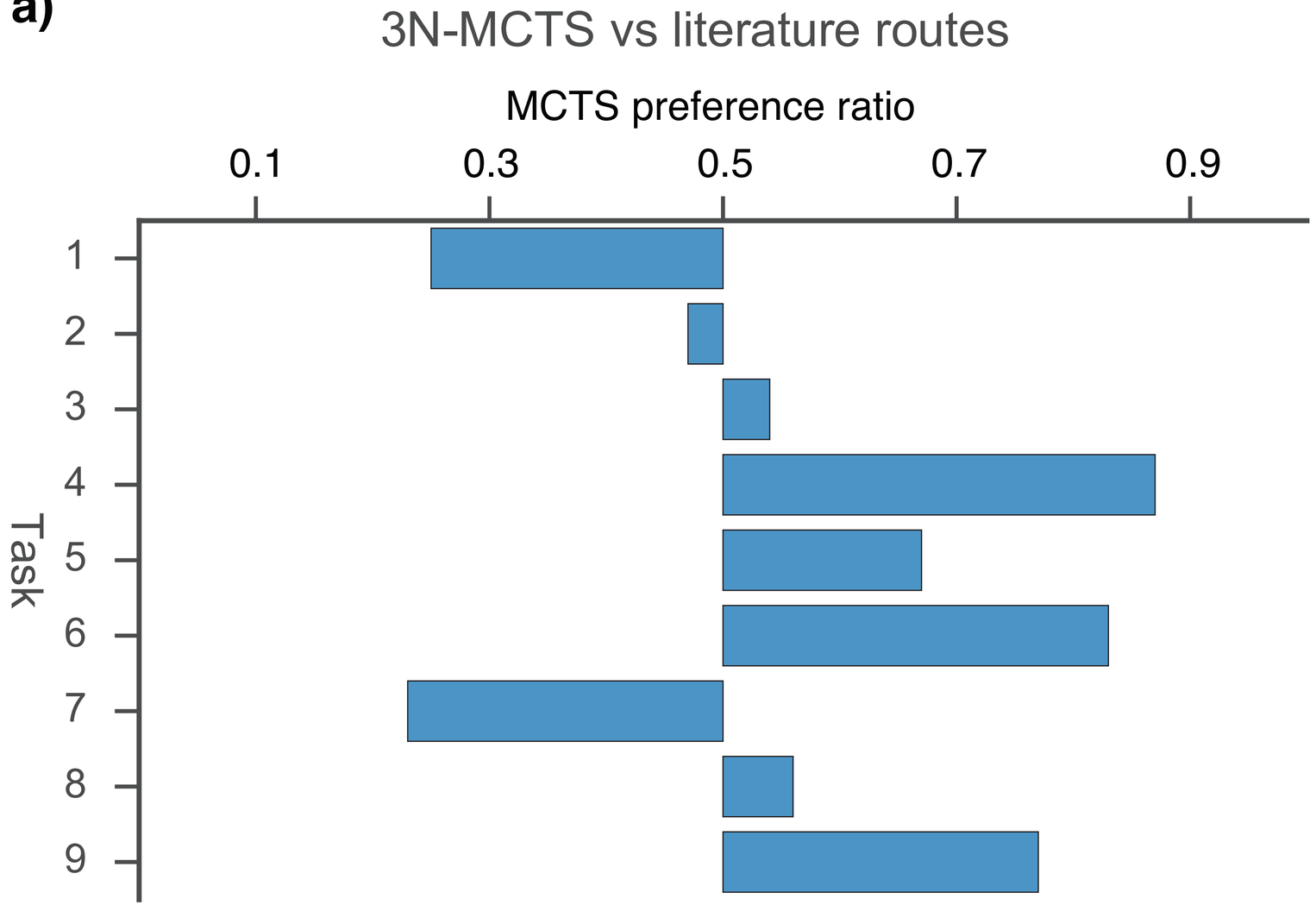
Null Hypothesis: Experts won’t like Computer’s solutions

Qualitative Analysis: Chemical Turing Test



- Double Blind
- 45 PhD students, postdocs,++ from Shanghai (CN) and Münster (DE)

a)



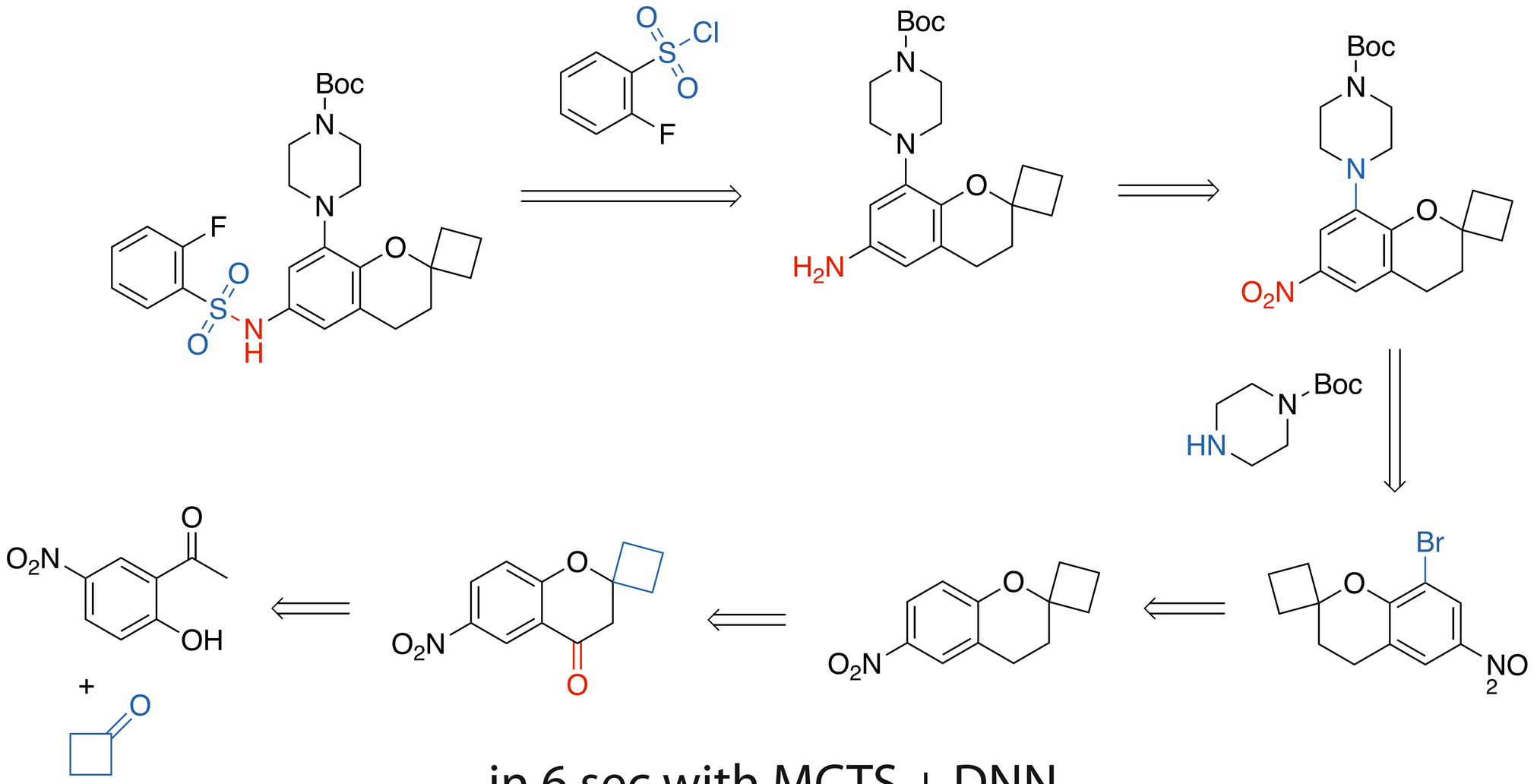
57:43

insignificant!

=> Expert & Computer routes cannot be distinguished!

Segler, Preuss, Waller, *Nature*, **2018**, (555), 604–610

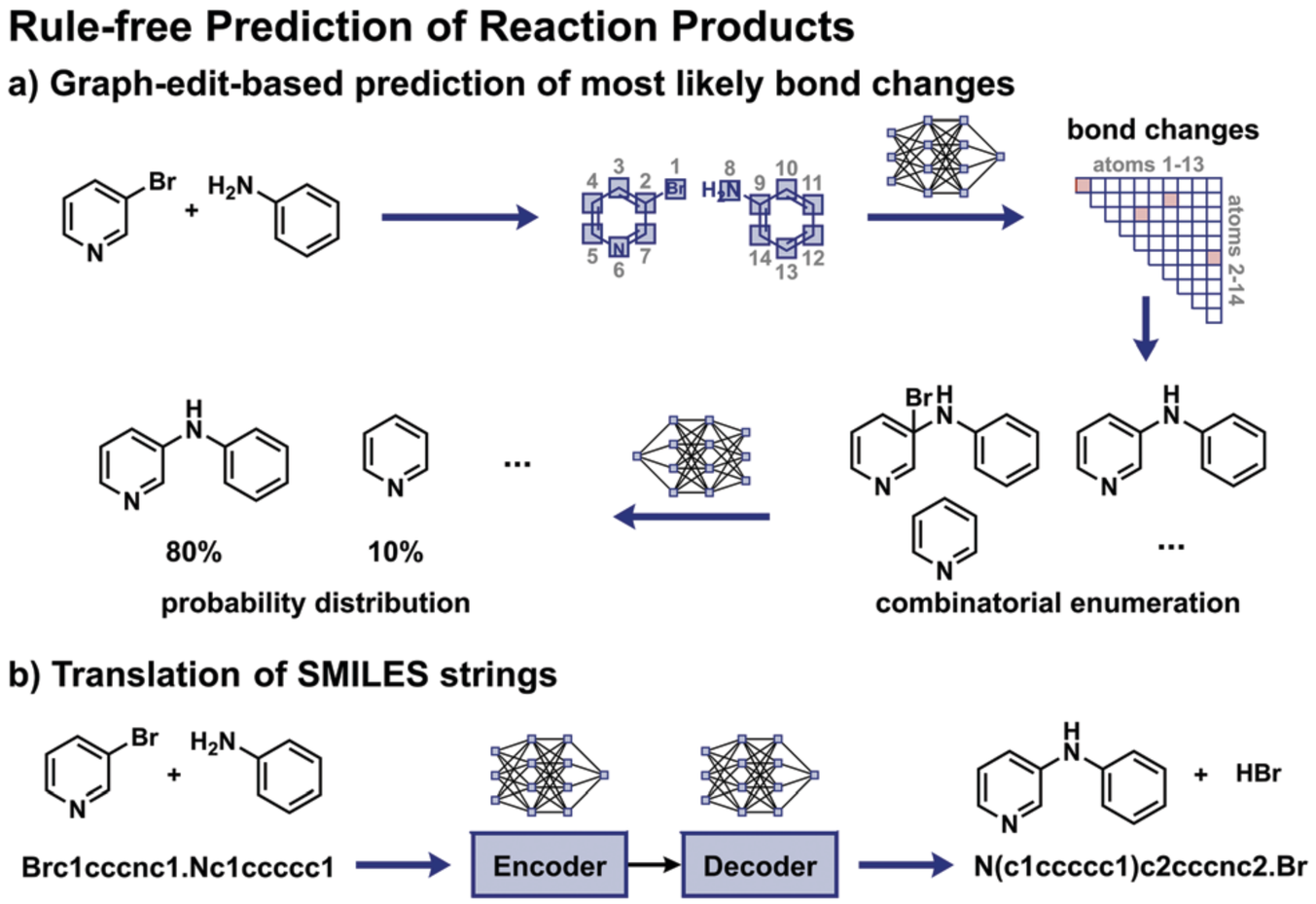
Example



in 6 sec with MCTS + DNN

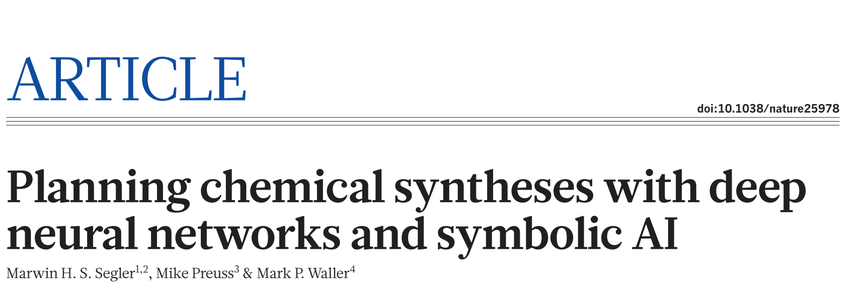
Segler, Preuss, Waller, *Nature*, **2018**, (555), 604–610

Alternative Approaches for Reaction and Retrosynthesis Prediction



Coley et al. *Chem. Sci*, **2019;** Schwaller et al. *Chem. Sci.* **2020;**

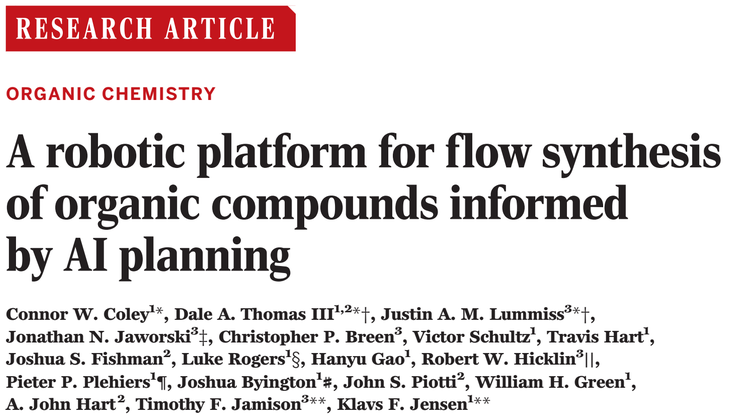
Machine Learning now core part of Computer Aided Synthesis Planning


ARTICLE

doi:10.1038/nature25978

Planning chemical syntheses with deep neural networks and symbolic AI

Marwin H. S. Segler[1,2], Mike Preuss[3] & Mark P. Waller[4]


RESEARCH ARTICLE

ORGANIC CHEMISTRY

A robotic platform for flow synthesis of organic compounds informed by AI planning

Connor W. Coley[1]*, Dale A. Thomas III[1,2]*†, Justin A. M. Lummiss[3]*†, Jonathan N. Jaworski[3]‡, Christopher P. Breen[3], Victor Schultz[1], Travis Hart[1], Joshua S. Fishman[2], Luke Rogers[1]§, Hanyu Gao[1], Robert W. Hicklin[3]||, Pieter P. Plehiers[1]¶, Joshua Byington[1]#, John S. Piotti[2], William H. Green[1], A. John Hart[2], Timothy F. Jamison[3]**, Klavs F. Jensen[1]**


Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy†

Philippe Schwaller, *[a] Riccardo Petraglia,[a] Valerio Zullo,[b] Vishnu H. Nair,[a] Rico Andreas Haeuselmann,[a] Riccardo Pisoni,[a] Costas Bekas,[a] Anna Iuliano [b] and Teodoro Laino[a]

Segler et al. *Nature*, **2018**; Coley et al. *Science*, **2019;** Schwaller et al. *Chem. Sci.* **2020;** Genheden, Thakkar et al. *J.Cheminf.* **2020;** Grzybowski et al. *Angew. Chem.* **2016**;

open source (e.g. AiZynthfinder, ASKCOS),

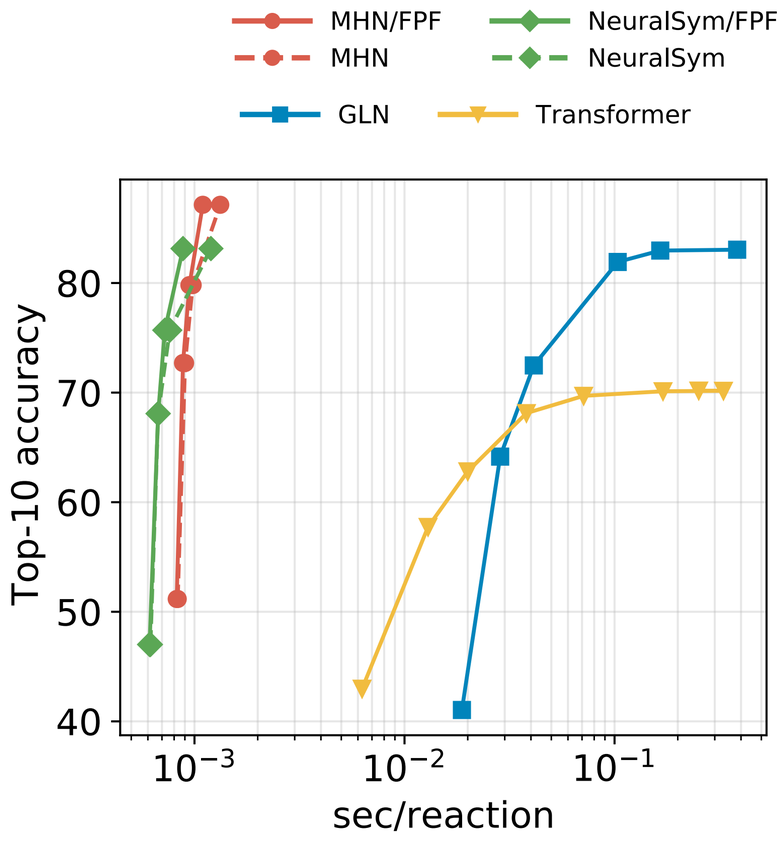
commercial tools (Reaxys, CAS, IBM, MoleculeOne, Iktos, …)

Comparison of Different Approaches for Reaction and Retrosynthesis Prediction

Method	**Purely Rule-based**	**ML + Graph Manipulation**	**Seq2Seq**
Classification	Symbolic	Neural-Symbolic	Neural
Uses Machine Learning	No	Yes	Yes
Molecule Repr.	Graphs	Graphs	SMILES
Reaction Repr.	Rules	Rules, Graph-Manipulation at different granularity	Implicit within neural network
Works by	Applying Rules	Predicting with parts of graph to manipulate with ML, then apply rule or edits	Generate target molecule from scratch with ML
Bottleneck	Need to Specify Exact Rules	Need to Specify Rough Rules, Data hungry	Very data hungry
Ease of getting started	-	0	+
Error Sources And Types	Rule Base, Chemical Errors (-)	Rule/Edit Base, Chemical Errors (0), Data	Copy Errors, Chemical Errors (+), Data

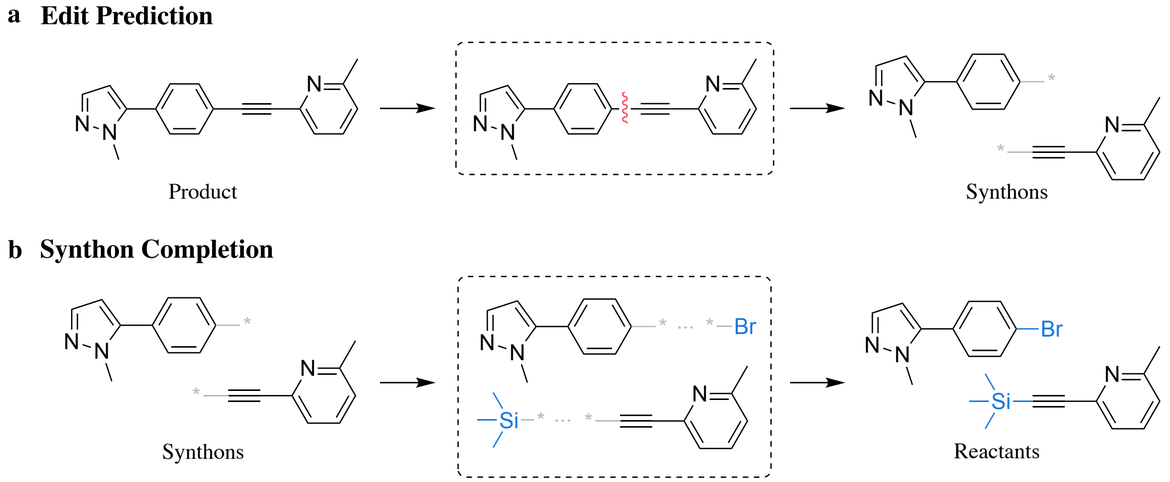
Recent Directions in ML for Reactions/Retrosynthesis

Disconnection Prediction with Modern Hopfield Networks

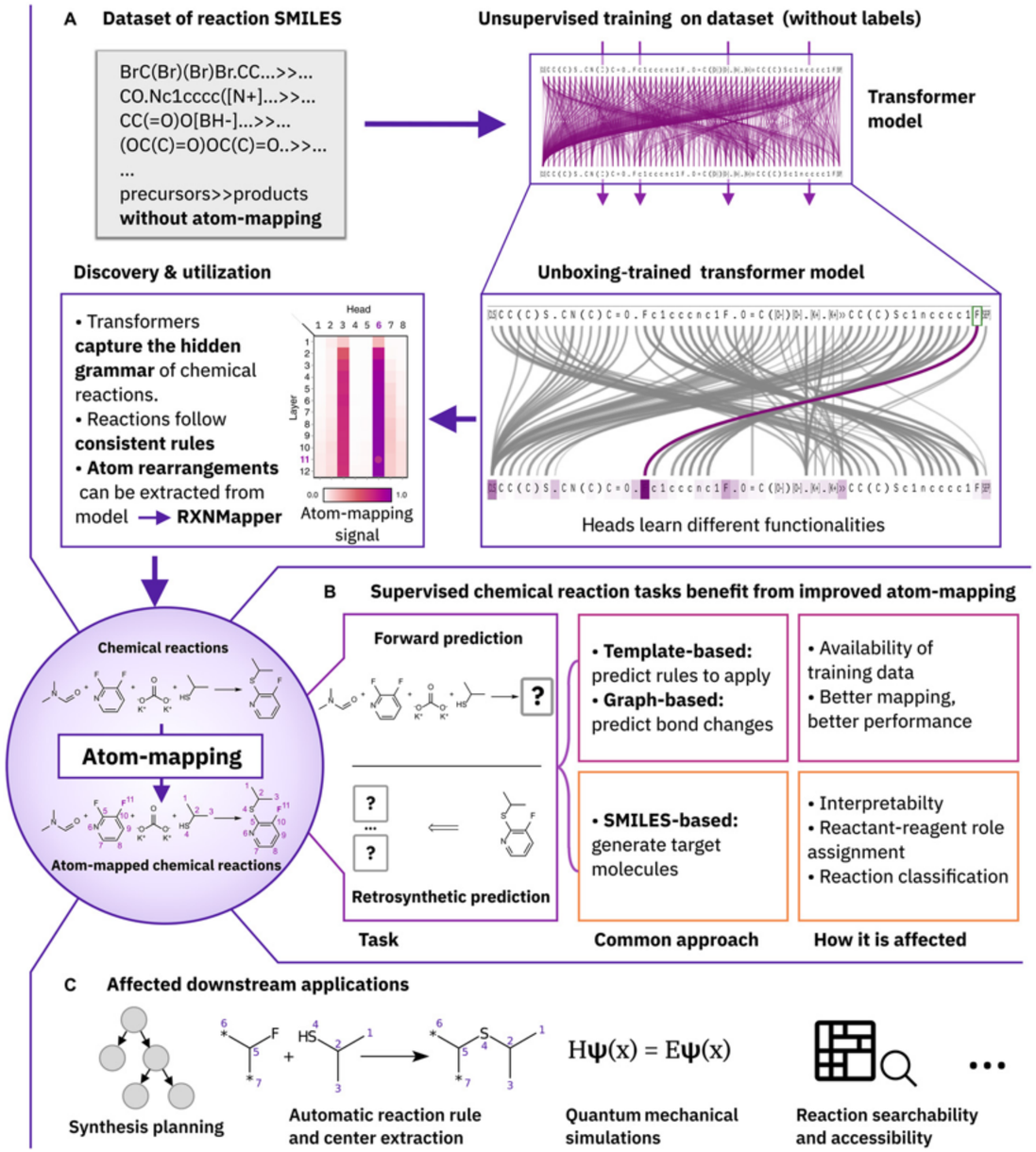


Seidl, Renz et al, MS, arXiv:2104.03279 **2021**

Learning Graph Models for Retrosynthesis Prediction
Somnath, Coley, et al arXiv:2006.07038 **2021**



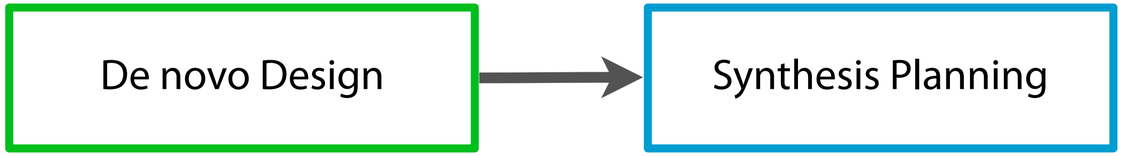
Recent Directions in ML for Reactions/Retrosynthesis



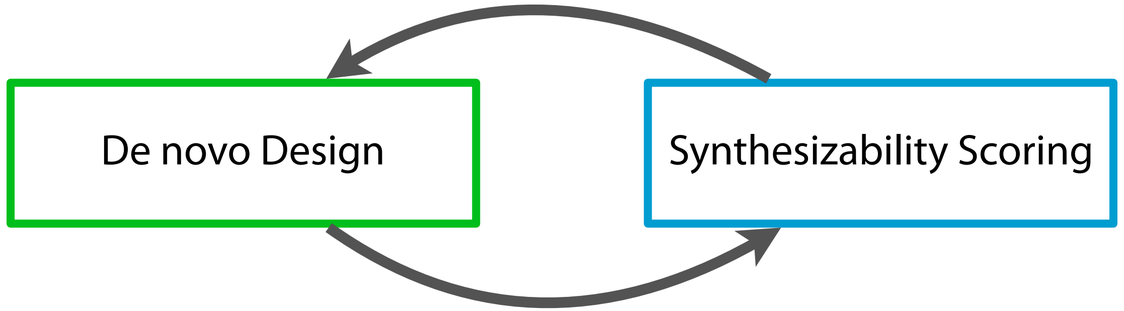
Schwaller et al, Extraction of organic chemistry grammar from unsupervised learning of chemical reactions, Sciences Adv. **2021**

How to integrate Synthesis Planning with De Novo Design?

1)



2)



3)

Gao, Coley *JCIM* **2020 ,** Boda et al JCAMD **2007**, Vinkers et al *J. Med. Chem.* **2003**; Segler, Preuss, Waller, *ICLR Workshop* **2017**

Synthesizability Scoring

Boda/Gasteiger => Fragments
SAScore - Ertl, Schuffenhauer => Fragments
SCScore - Coley et al => ML, Heuristic for Synthesis Planning

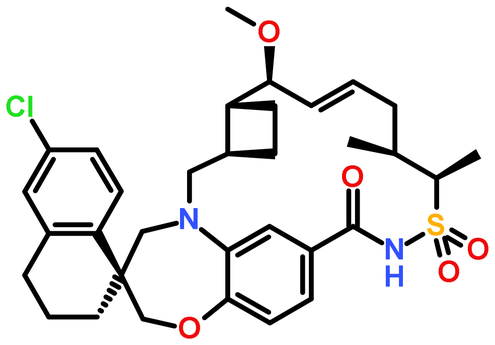
Boda; Seidel, Gasteiger, *J. Comput.-Aided. Mol. Des.* **2007** 10.1007/s10822-006-9099-2
Ertl, Schuffenhauer, *J. Cheminf.* **2009** *10.1186/1758-2946-1-8*
Coley, Rogers, Green, Jensen, *JCIM* **2018** 10.1021/acs.jcim.7b00622

Synthesizability: Not a well-defined concept

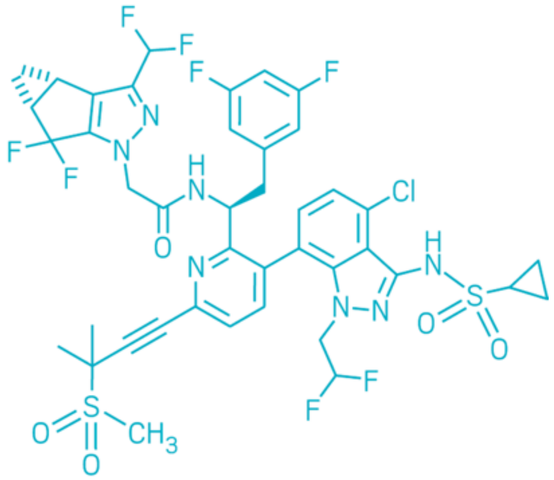
Not fully defined by structure

Context dependent — hit expansion vs. late lead opt vs. scale-up

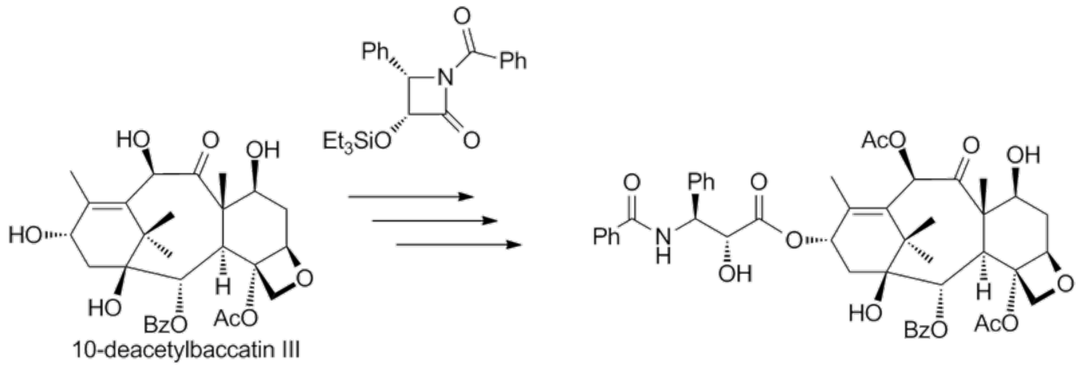
Starting-material dependent — Availability reduces complexity



AMG-176



Gilead's GS-CA1

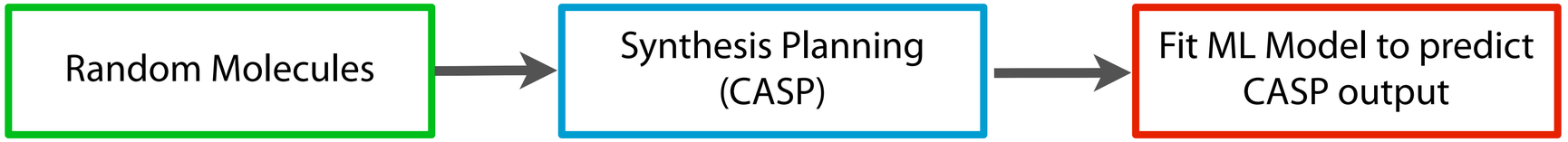


10-deacetylbaccatin III

Image Credits: Ingo Hartung, C&EN Mag, Wikipedia

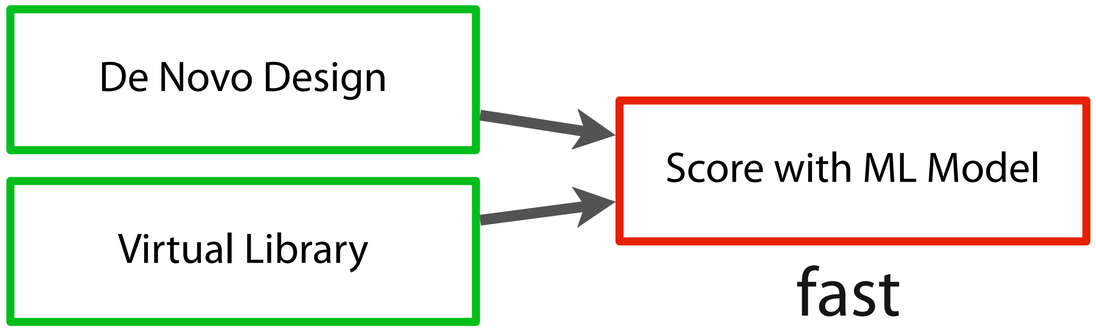
Learning to approximate a full synthesis planner

1)

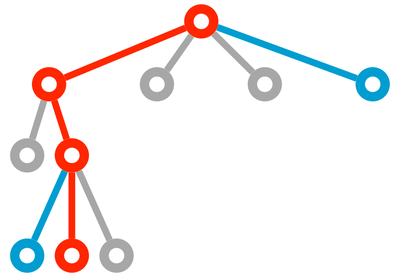


slow

2)



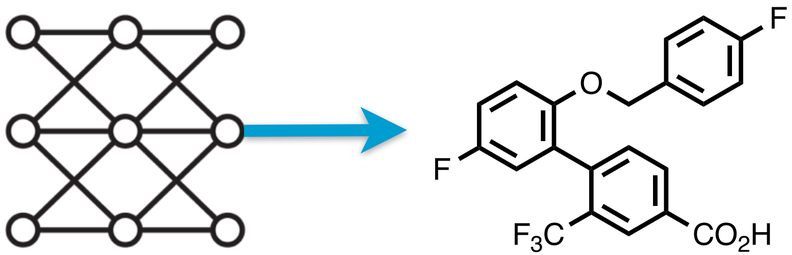
fast



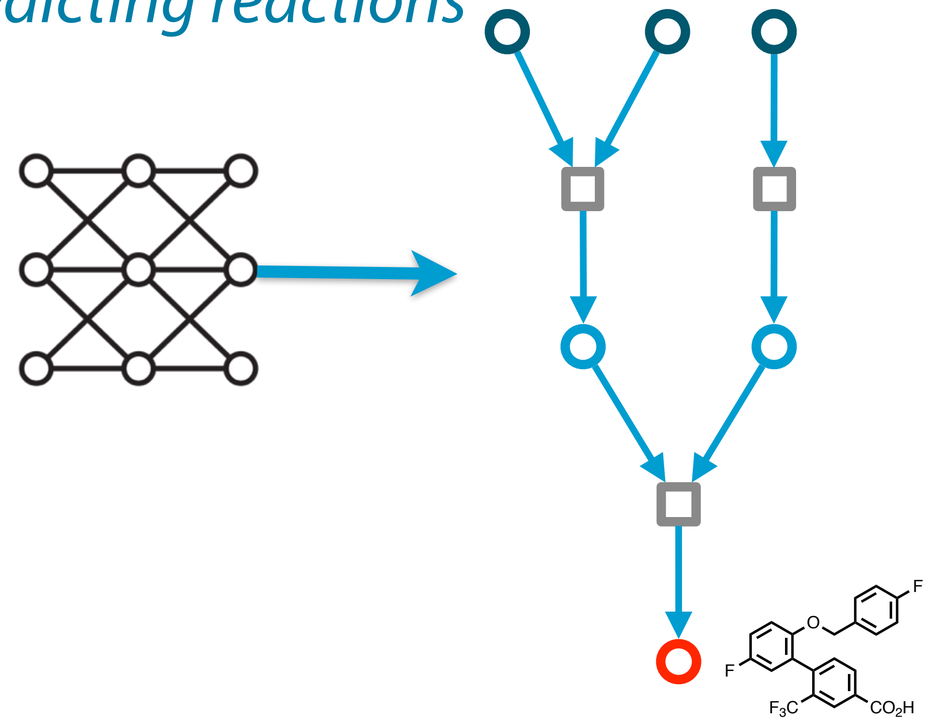
Liu, Korablyov, Jastrzębski, Włodarczyk-Pruszyński, Bengio, Segler, *arXiv:2011.13042*, **2020**
Parallel work: Thakkar et al. *Chem. Sci.* **2021**

Generative Models for Synthesis Trees?

Generative models for Molecules
build atom by atom



Alternative:
build synthesis tree by picking reactants and predicting reactions

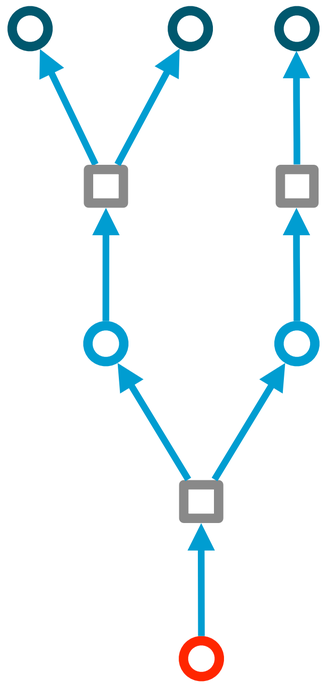


Non-Neural: Vinkers et al **-** SYNOPSIS, ***J. Med. Chem.*** **2003**; Hartenfeller, Schneider, *WIRES,* **2011;**
Neural: Bradshaw et al. *NeurIPS*, **2019,** Gottipatti *ICML* **2020**, Horwood, Noutahi, *ACS Omega,* **2020**

Retrosynthesis vs Forward Synthesis

Retrosynthesis

Backward chaining



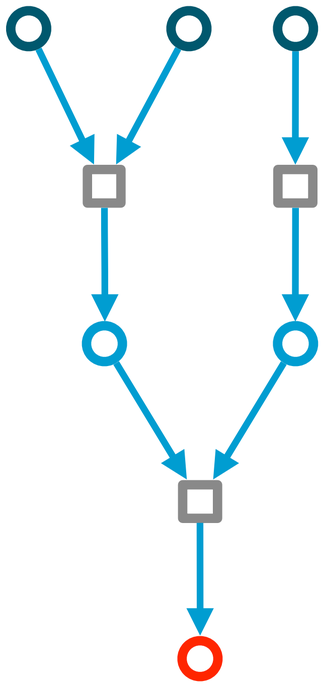
Building Blocks

Intermediates

Product

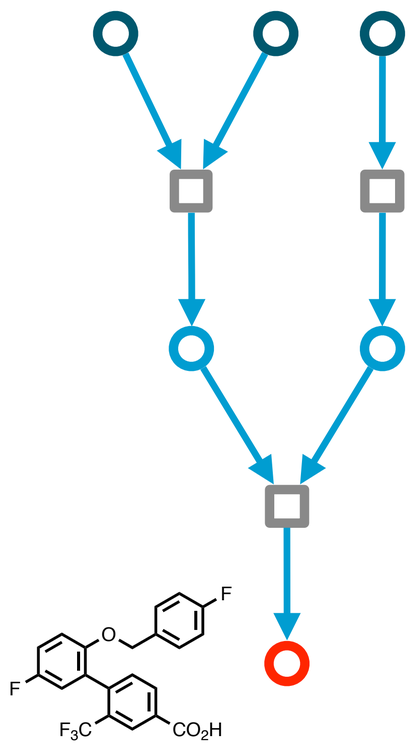
Actual Synthesis

Forward chaining



Generative Models for Synthesis Trees?

DAG (Directed Acyclic Graph) of Graphs



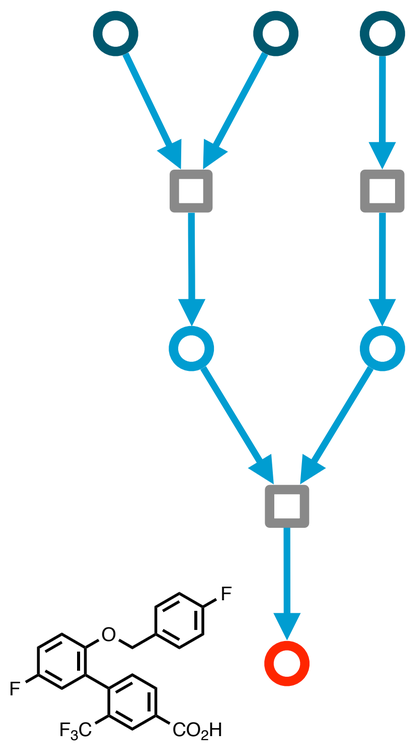
Building Blocks

Intermediates

Product

Generative Models for Synthesis Trees?

DAG (Directed Acyclic Graph) **o**f **G**raphs



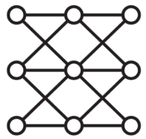
Building Blocks

Intermediates

Product

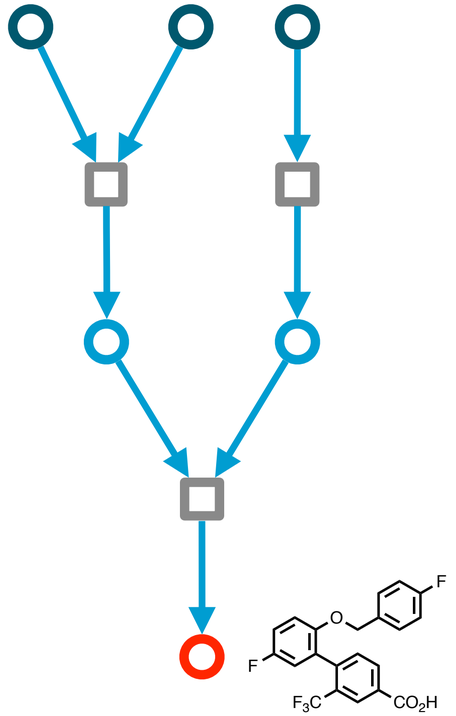
DoG Algorithm

Provide Building Library & Reaction Predictor (MT; Schwaller et al. 2019)

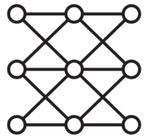


Model chooses steps:

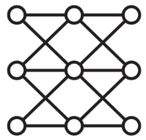
1) Pick Reactants
2) Pick Intermediates
3) Predict Reaction
4) Stop



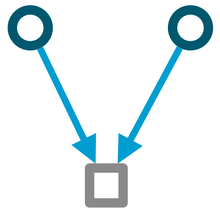
DoG Algorithm



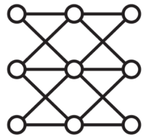
DoG Algorithm



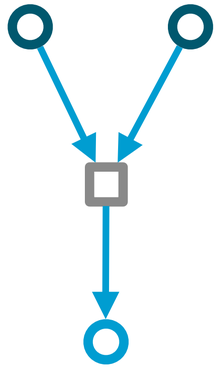
Pick reactants



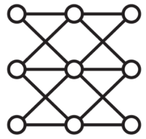
DoG Algorithm



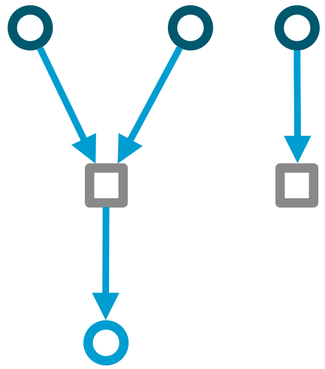
Predict reaction



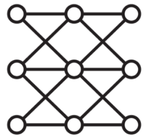
DoG Algorithm



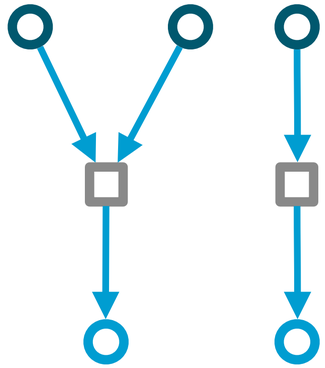
Pick reactants



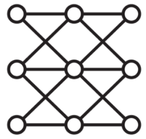
DoG Algorithm



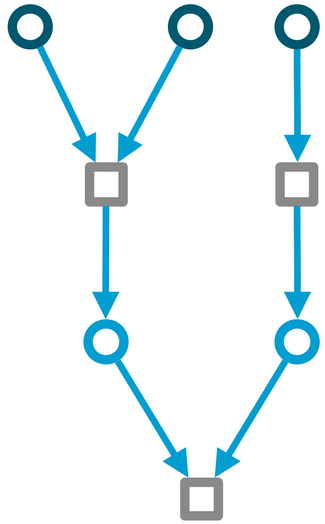
Predict reaction



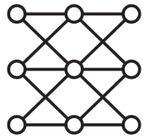
DoG Algorithm



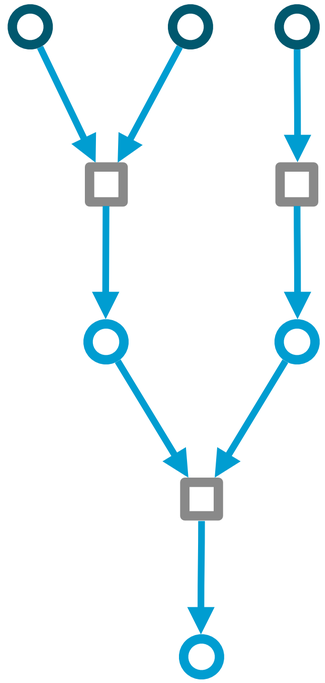
Pick intermediates



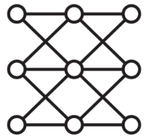
DoG Algorithm



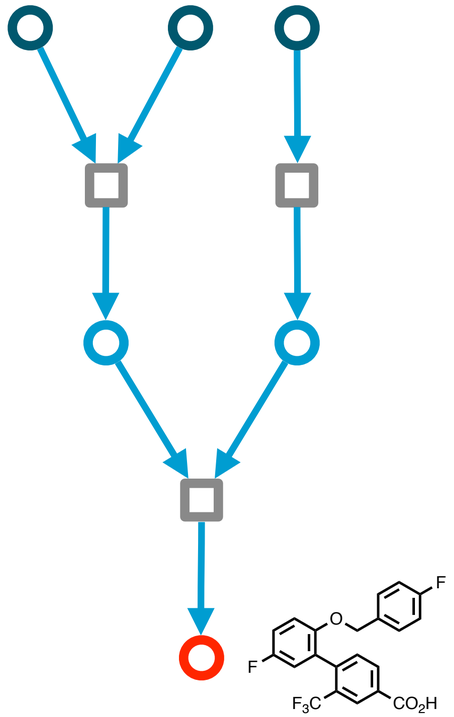
Predict reaction



DoG Algorithm



Stop



Optimisation Experiments

DoG-Generator + Cross-Entropy Method
Guacamol Optimisation Benchmarks

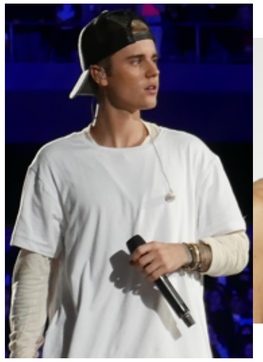
Guacamol: Brown et al. *JCIM* **2019;**

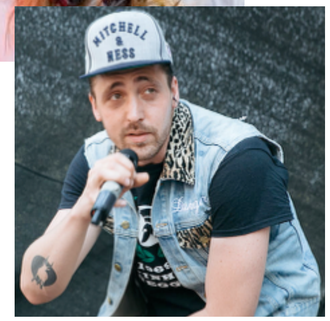
Maximum Scores vs Quality Tradeoff

Guacamol: Brown et al. *JCIM* **2019;**

It’s not just about leaderboard performance…

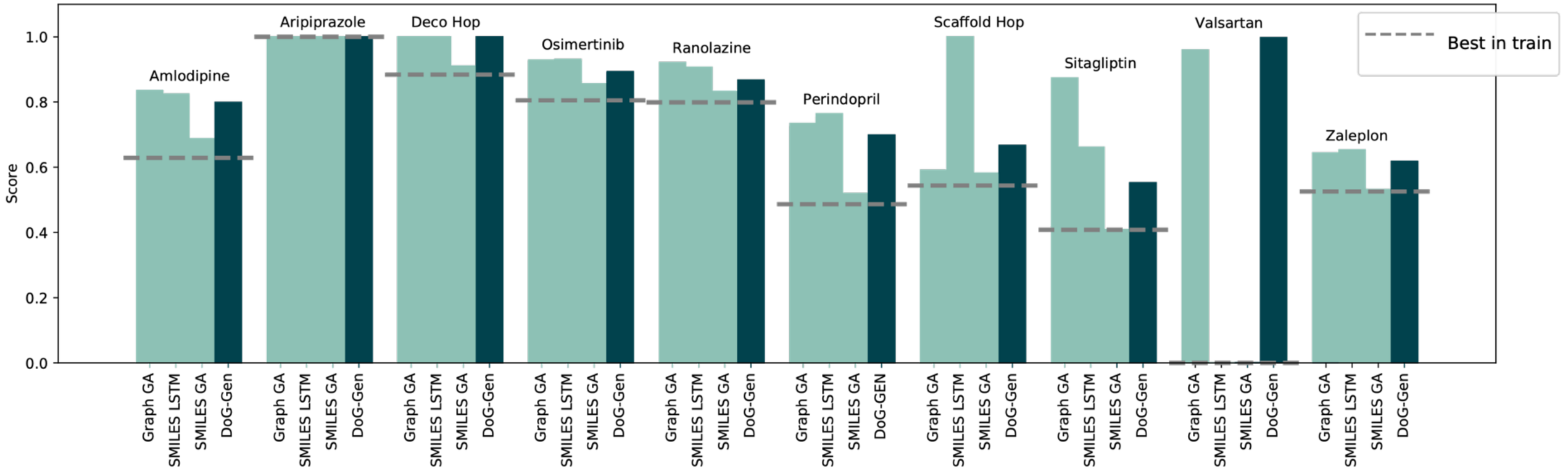
Score





Quality

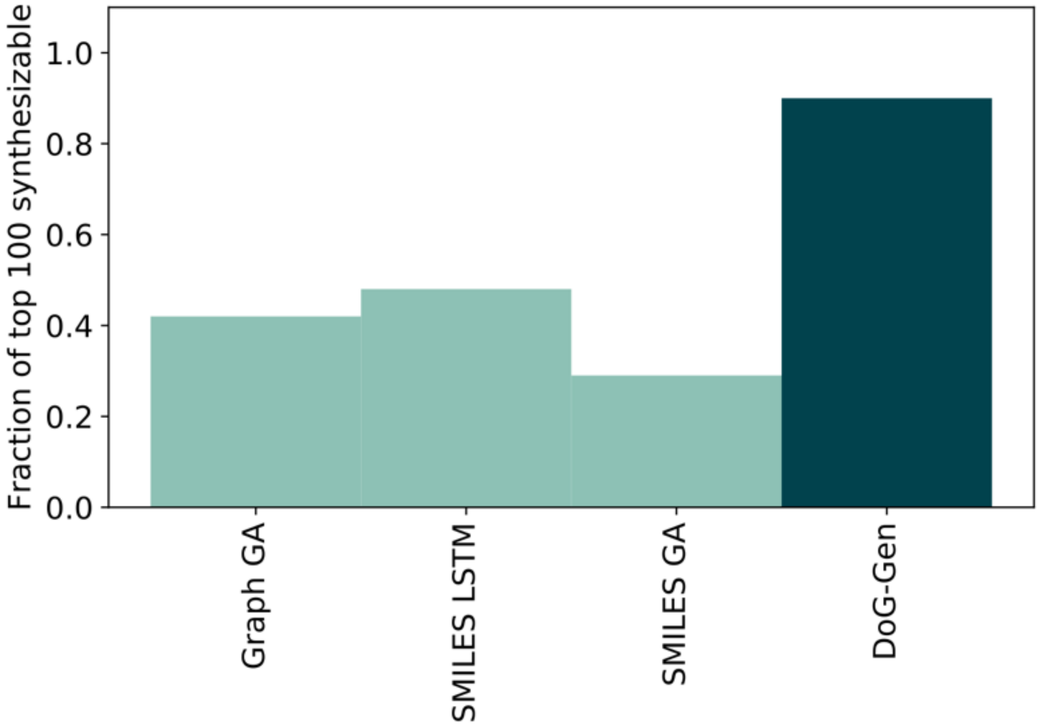
Performance on Guacamol Optimisation Tasks



Guacamol: Brown et al. *JCIM* **2019;**

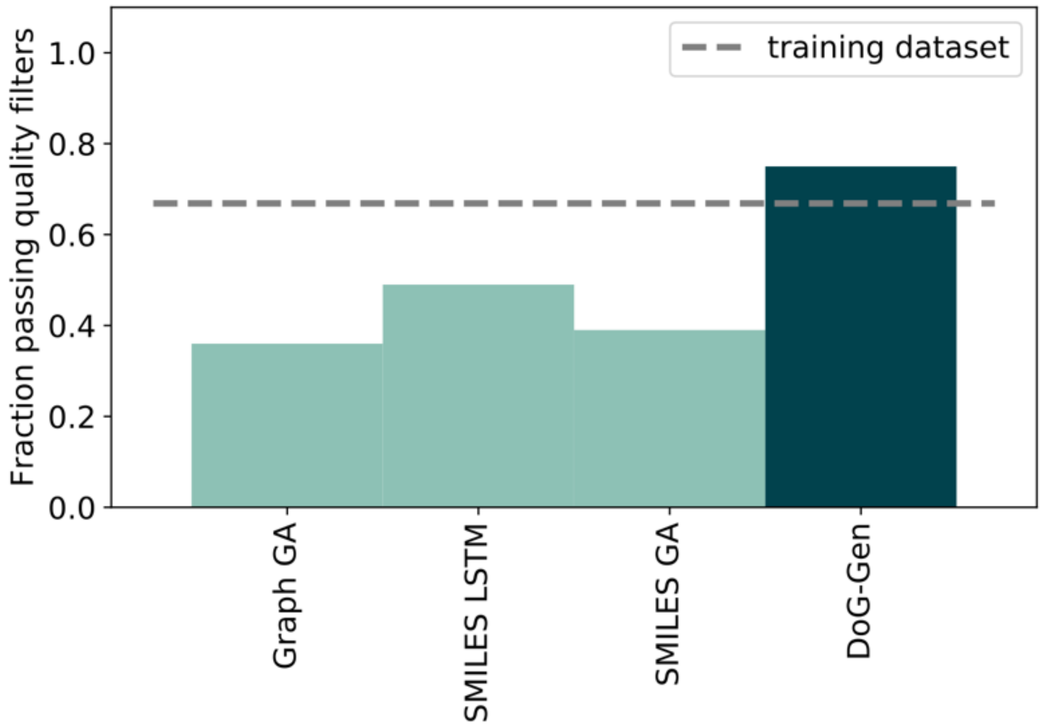
Performance on Guacamol Optimisation Tasks

Synthesizable against CASP oracle [1]



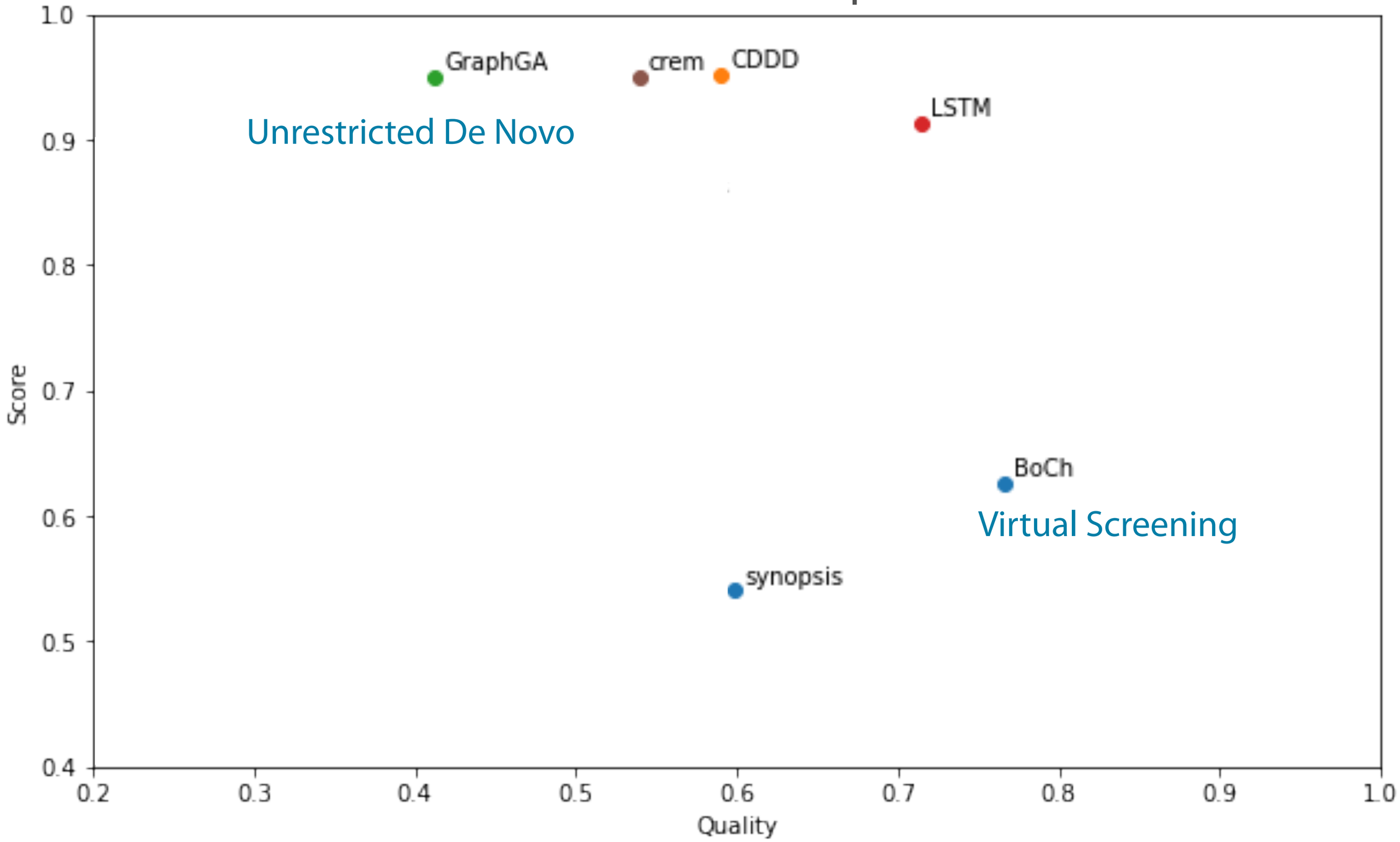
higher better

Quality Score [2]



[1] Gao, Coley, *JCIM,* **2020**; Segler et al. *ICLR Workshop,* **2017**
[2] *Guacamol*: Brown et al. *JCIM* **2019;**

Performance on Guacamol Optimisation Tasks



Performance on Guacamol Optimisation Tasks