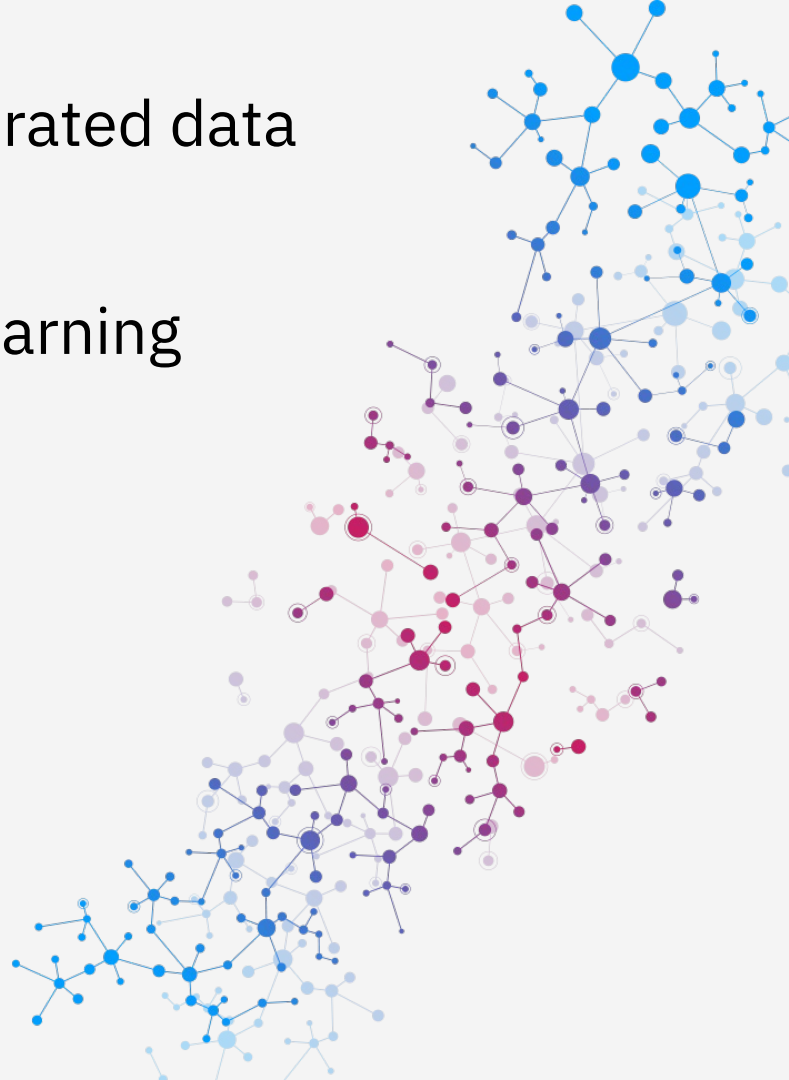
Combining high-quality, humanly curated data with language models:

The dawn of on-demand machine learning models for digital chemistry

Alain Vaucher, *IBM Research*
Fiona Shortt de Hernandez, *Thieme Group*
Sascha Hausberg, *Thieme Group*
Teodoro Laino, *IBM Research*

ACS Fall 2022
24 August 2022



Outline

1. Introduction:
   AI for chemical reactivity & data
2. Study (IBM Research & Thieme):
   combining curated data & data-driven models
   - Data & model training
   - Examples, challenges, expert feedback
3. Where do we go from here?

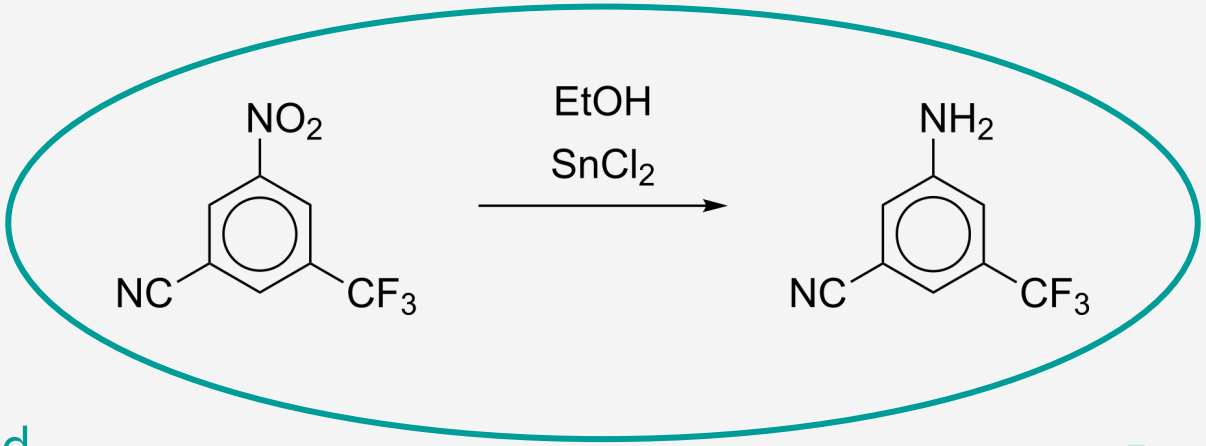
Outline

1. Introduction:
   AI for chemical reactivity & data
2. Study (IBM Research & Thieme):
   combining curated data & data-driven models
   - Data & model training
   - Examples, challenges, expert feedback
3. Where do we go from here?

AI and chemical reactivity

Experimental conditions?

Product?



Related reactions?

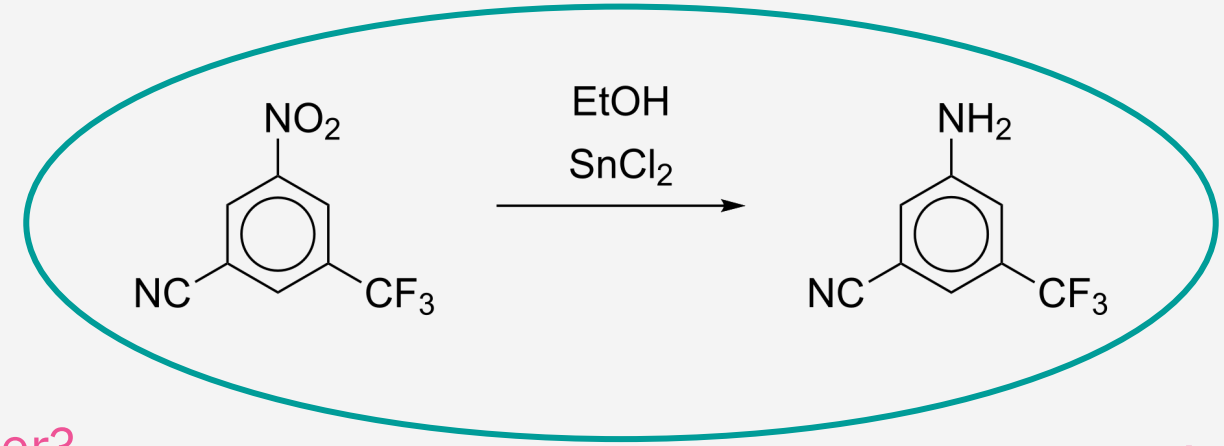
Retrosynthesis?

Yield?

AI and chemical reactivity

Accuracy

State-of-the-art

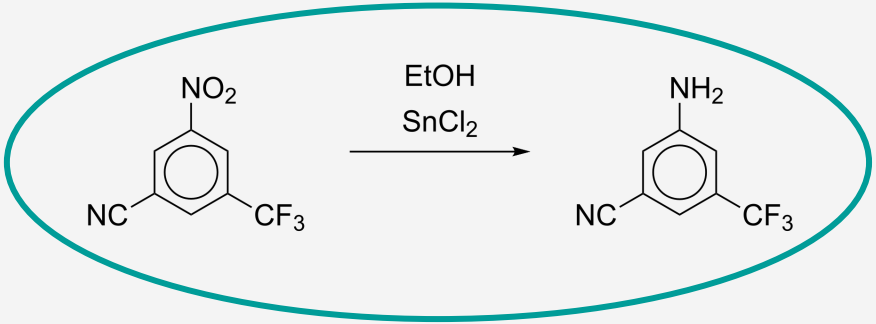


End user?

Applicability?

Usefulness?

AI and chemical reactivity



Applicability?

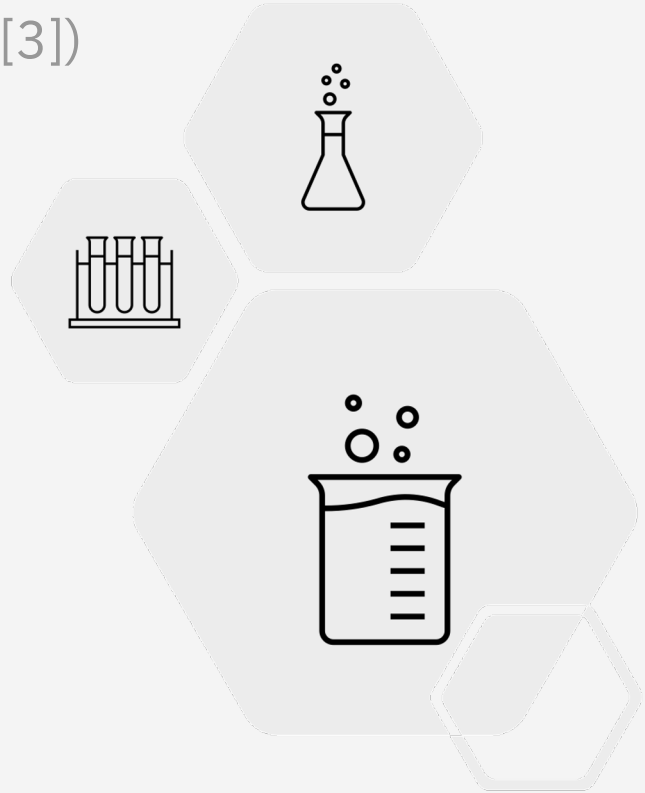
Training data must cover adequate chemistry

End user? Usefulness?

What models are needed? How are they used?

AI and chemical reactivity – data sources

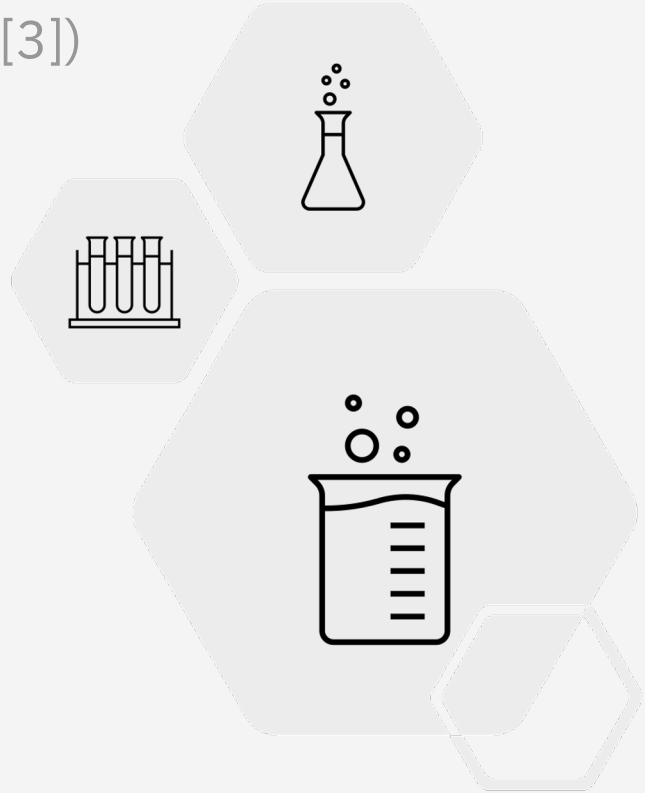
- Patents (USPTO [1], Pistachio [2], SciWalker [3])
- Scientific publications
- Proprietary reactions (industry)
- Publishers (Reaxys, CASFinder, Thieme)
- Others (Open Reaction Database [4])
- Etc.



[1] https://dx.doi.org/10.6084/m9.figshare.5104873.v1
[2] https://www.nextmovesoftware.com/pistachio.html
[3] https://www.sciwalker.com
[4] https://open-reaction-database.org

AI and chemical reactivity – data sources

- Patents (USPTO [1], Pistachio [2], SciWalker [3])
- Scientific publications
- Proprietary reactions (industry)
- Publishers (Reaxys, CASFinder, Thieme)
- Others (Open Reaction Database [4])
- Etc.



[1] https://dx.doi.org/10.6084/m9.figshare.5104873.v1
[2] https://www.nextmovesoftware.com/pistachio.html
[3] https://www.sciwalker.com
[4] https://open-reaction-database.org

AI and chemical reactivity – going beyond patents

- Patent data:
  - Good coverage of mid-size organic compounds
  - Good coverage of common organic reactions
  - essential for development of data-driven models

- Limitations:
  - Reporting errors
  - Incorrect extraction
  - Reproducibility concerns
  - Limited coverage of reaction space

AI and chemical reactivity – going beyond patents

- If patent data is not enough:
  - Fine-tuning models for specific application
  - Push for more general models

- Open questions:
  - What data sources?
  - Data safety?

- Study: reactivity models trained on curated data

AI and chemical reactivity – IBM RXN & Thieme SOS

Combine IBM RXN models with curated data from Thieme

- How useful are the predictions for chemists?
- How valuable is the curated data for ML applications?

Outline

1. Introduction:
   AI for chemical reactivity & data
2. Study (IBM Research & Thieme):
   combining curated data & data-driven models
   - Data & model training
   - Examples, challenges, expert feedback
3. Where do we go from here?

Outline

1. Introduction:
   AI for chemical reactivity & data
2. Study (IBM Research & Thieme):
   combining curated data & data-driven models
   - Data & model training
   - Examples, challenges, expert feedback
3. Where do we go from here?

Curated data (Thieme)

Science of Synthesis

~450,000 reactions

Collections & volumes covering diverse topics in chemical synthesis

Synfacts

~16,000 reactions (2017-2018)

Highlights in chemical synthesis

Focus on total synthesis

Data

Science of Synthesis Reference Library
Stereoselective Synthesis Volume 1, section 1.13.1.2

1-Methyl-3-(octan-2-yl)imidazolidin-2-one (2, R¹ = Me); Typical Procedure:[21]

A suspension of 1-methylimidazolidin-2-one (20 mg, 0.20 mmol), oct-1-ene (1.3 g, 12 mmol), $(AuCl)_2[(S)\text{-}\mathbf{1}]$ (8.0 mg, 5.0 μmol), and AgOTf (2.6 mg, 10 μmol) in *m*-xylene (0.5 mL) was stirred at 100 °C for 48 h. The crude mixture was filtered through a plug of silica gel, concentrated, and chromatographed (hexanes/EtOAc 5:1 to 1:1) to give a colorless oil; yield: 37 mg (86%); 76% ee.

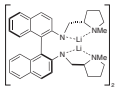
1.13.1.2 Cyclization of Aminoalkenes

1.13.1.2.1 Using Chiral Alkali Metal Based Catalysts

The first reports on base-catalyzed additions of amines to alkenes date back 60 years. In particular, multiple catalyst systems utilizing alkali metals have been reported.[9,10,13] However, application of chiral alkali metal complexes in the asymmetric hydroamination of nonactivated aminoalkenes has drawn little attention to date.[22–26] Attempts to perform asymmetric hydroamination utilizing chiral alkaline earth metal complexes have been thwarted by facile Schlenk equilibria of the metal species in solution.[24,25]

The proline-derived dimeric diamidobinaphthyl dilithium salt (*S*,*S*,*S*)-**3**, which is prepared via deprotonation of the corresponding tetraamine with butyllithium, catalyzes asymmetric intramolecular hydroamination reactions of aminopentenes **4** at or below ambient temperatures to form pyrrolidines **5** with enantioselectivities of up to 74% ee (Scheme 3).[22,26] The enantioselectivities may be improved to up to 85% ee by lowering the reaction temperature to −10 °C.[26] The unique reactivity of (*S*,*S*,*S*)-**3** is believed to derive from the close proximity of the two lithium centers chelated by the proline substituents, because more simple lithium amides require significantly higher reaction temperatures and give inferior selectivities.

Scheme 3 Lithium-Catalyzed Asymmetric Hydroamination/Cyclization of Aminoalkenes[22,26]



(*S*,*S*,*S*)-**3**

4 **5**

R¹	R²	Catalyst (mol%)	Temp (°C)	Time (h)	Yield (%)	ee (%)	Ref
Me	Me	2.5	22	45	93	67	[22]
$(CH_2)_2$		5	20	2	82	74	[22]
$(CH_2)_2$		2	−10	22	84	85	[26]

for references see p 727

Synfacts 2021; 17(12): 1314
DOI: 10.1055/s-0041-1737090

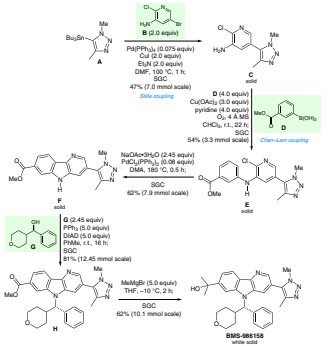
Category
Synthesis of Natural Products and Potential Drugs

Key words
BMS-986158
Stille coupling
Chan–Lam coupling
copper catalysis
palladium-catalyzed C–H activation
Mitsunobu reaction

A. V. GAVAI* ET AL. (BRISTOL MYERS SQUIBB COMPANY, PRINCETON, USA)

Discovery and Preclinical Pharmacology of an Oral Bromodomain and Extra-Terminal (BET) Inhibitor Using Scaffold-Hopping and Structure-Guided Drug Design
J. Med. Chem. **2021**, *64*, 14247–14265, DOI: 10.1021/acs.jmedchem.1c00625.

Synthesis of BMS-986158



Significance: BMS-986158 is an inhibitor of the bromodomain and extra-terminal (BET) family of adaptor proteins that are involved in the transcriptional regulation of key oncogenes. It has entered phase 1/2a clinical trials in patients with advanced cancers and hematologic indications including myelofibrosis.

Comment: Key steps in the small-scale discovery synthesis of the 5*H*-pyrido[3,2-*b*]indole core of BMS-986158 are (1) the copper-catalyzed oxidative coupling of the chloropyridine **C** with the boronic acid **D** (Chan–Lam coupling) and (2) the palladium-catalyzed C–H activation reaction **E** → **F**.

SYNFACTS Contributors: Philip Kocienski
Synfacts 2021, 17(12), 1314 Published online: 17.11.2021
DOI: 10.1055/s-0041-1737090; **Reg-No.:** R07021SF

Data - summary

Science of Synthesis

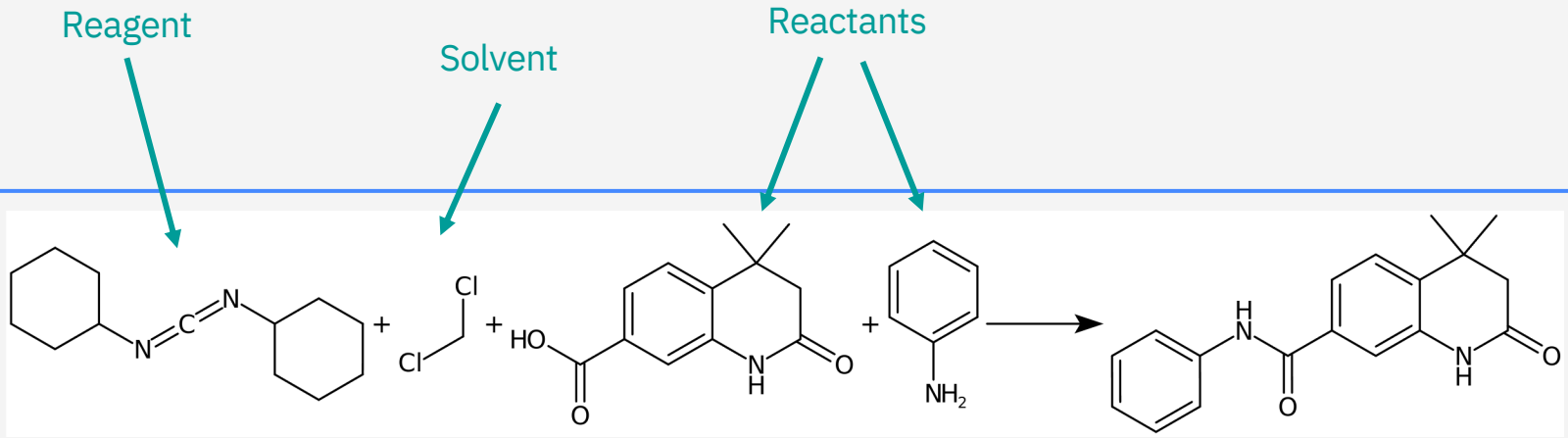
76,011 documents

Synfacts (2017-2018)

2,597 documents

PDF, XML and RDF files

Target reaction format



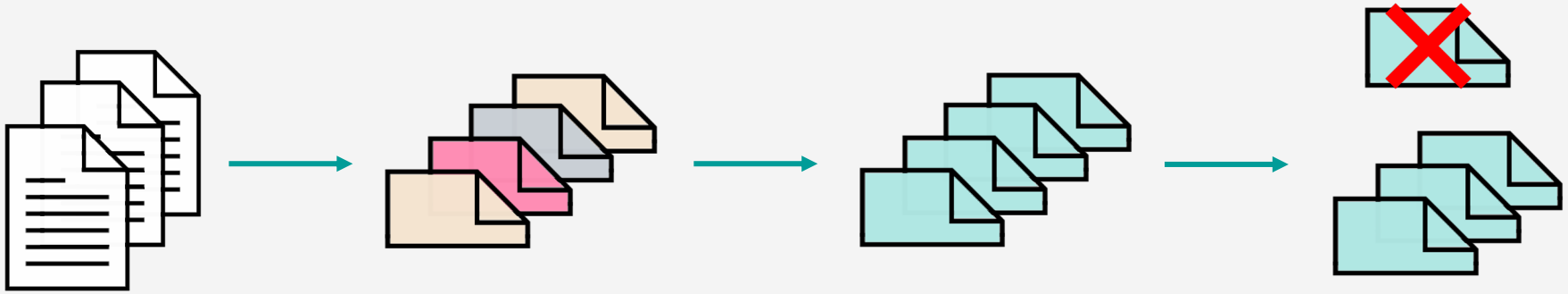
Reaction SMILES:

```
C(=NC1CCCCC1)=NC1CCCCC1.ClCCl.CC1(C)CC(=O)Nc2cc(C(=O)O)
ccc21.Nc1ccccc1>>CC1(C)CC(=O)Nc2cc(C(=O)Nc3ccccc3)ccc21
```

Data Processing

Steps to produce training data:

1. Conversion to reaction SMILES
2. Standardization
3. Sanity checks and filters



Processed Data

Science of Synthesis

318,383 reactions

Synfacts (2017-2018)

13,818 reactions

Processed Data

Science of Synthesis

318,383 reactions

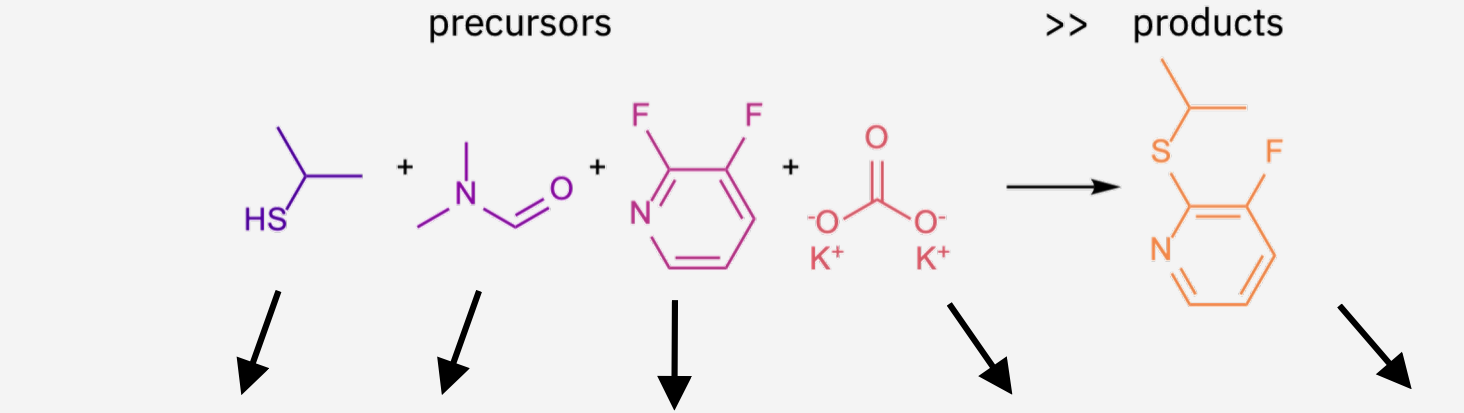
Synfacts (2017-2018)

13,818 reactions

Pistachio

~2M reactions

Atoms as *letters*, molecules as *words*



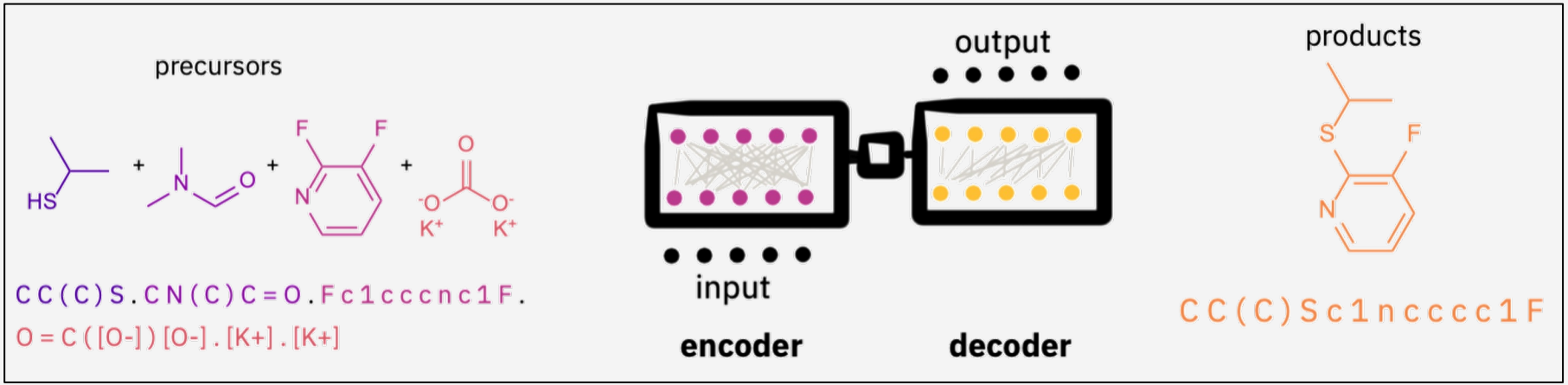
SMILES: CC(C)S.CN(C)C=O.Fc1cccnc1F.O=C([O-])[O-].[K+].[K+]>>CC(C)Sc1ncccc1F

Split -> “sequences of atoms”

Tokens: C C (C) S . C N (C) C = O . F c 1 c c c n c 1 F . O = C ([O-]) [O-] . [K+] . [K+] >> C C (C) S c 1 n c c c c 1 F

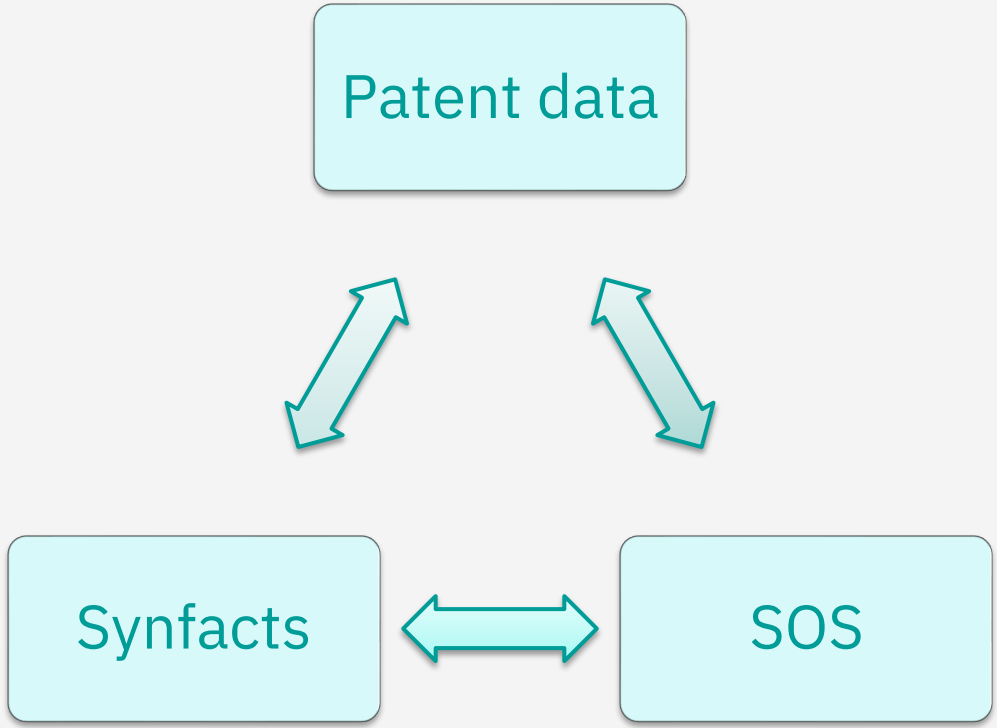
→ *Borrow methods developed for human languages*

Training the models



Forward reaction prediction and retrosynthesis

Training the models



3 sources included for training (weighted differently)

Model accuracy (top-1)

Forward reaction prediction

Patent-only: 22.9%

Patent + Thieme: **69.5%**

Retrosynthesis

Patent-only: 1.1%

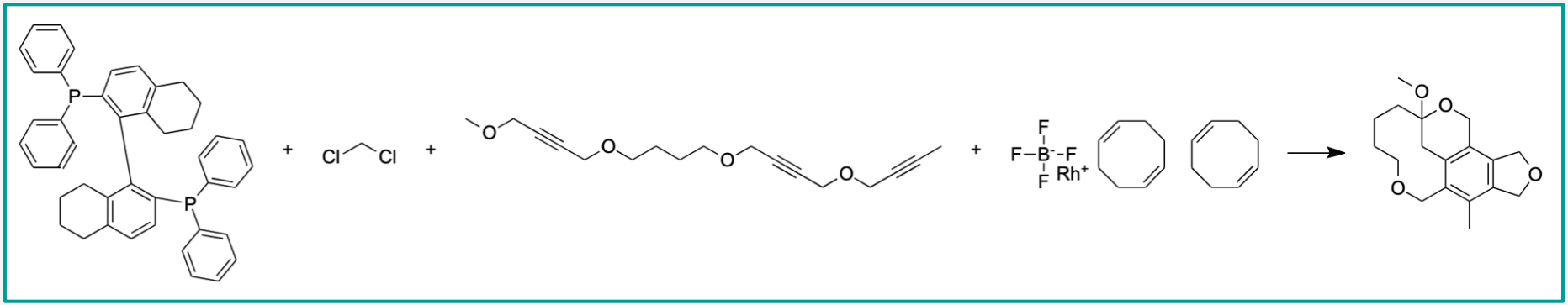
Patent + Thieme: **17.8%**

Outline

1. Introduction:
   AI for chemical reactivity & data
2. Study (IBM Research & Thieme):
   combining curated data & data-driven models
   - Data & model training
   - Examples, challenges, expert feedback
3. Where do we go from here?

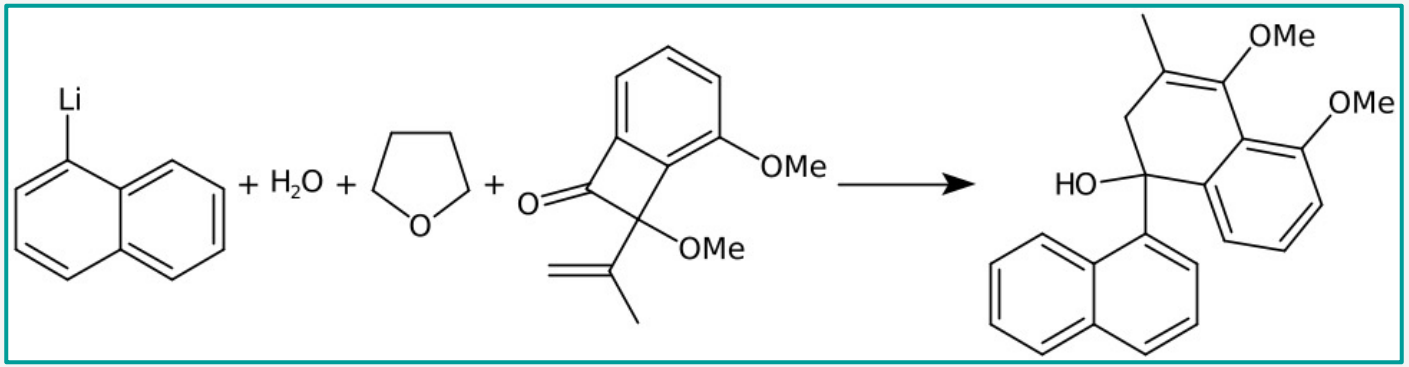
Forward model: inspection of a few predictions

Correct product predicted



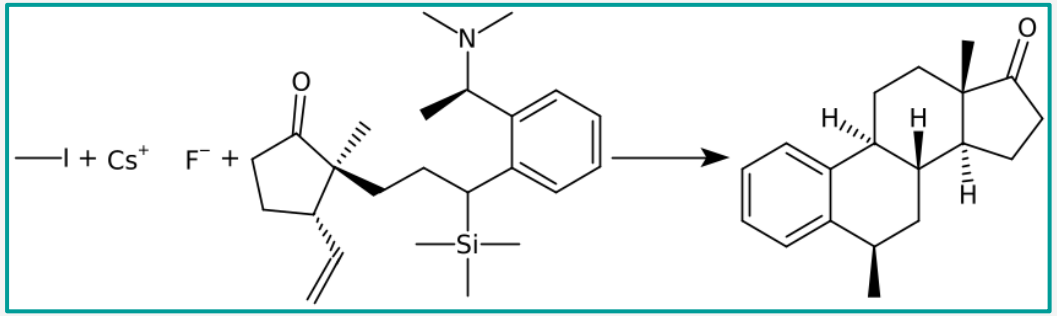
Forward model: inspection of a few predictions

Correct product predicted

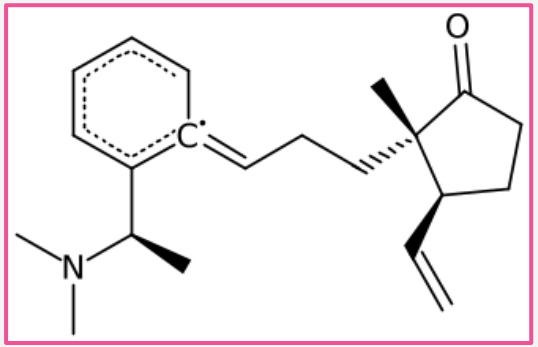


Forward model: inspection of a few predictions

Thieme reaction

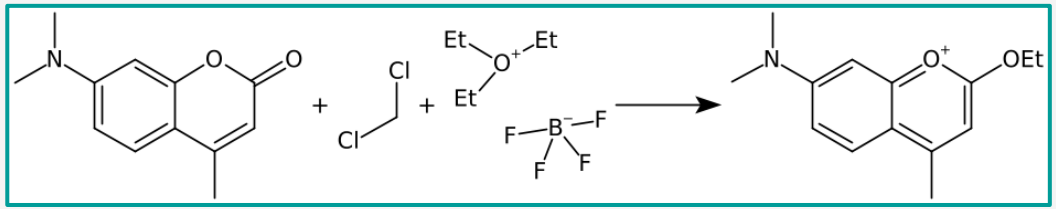


Predicted product

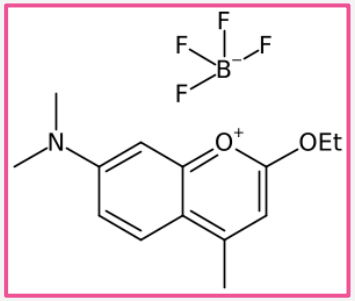


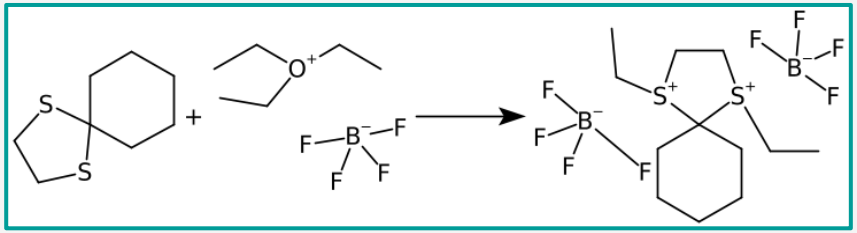
Invalid SMILES: 1.0% of predictions

Forward model: inspection of a few predictions

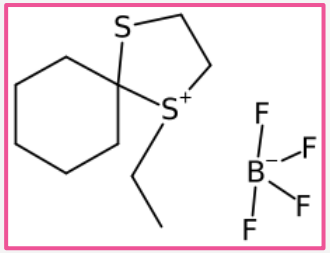


Predicted product

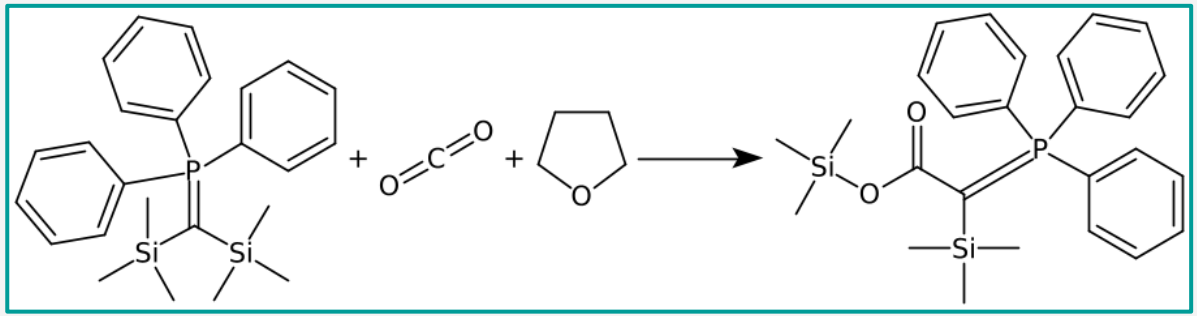




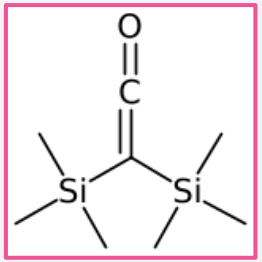
Predicted product



Forward model: inspection of a few predictions

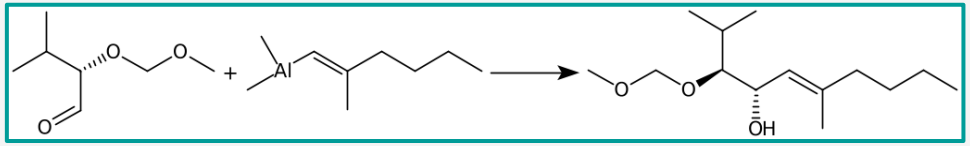


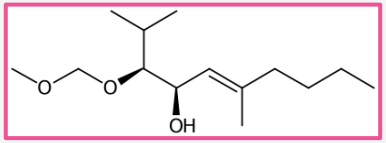
Predicted product



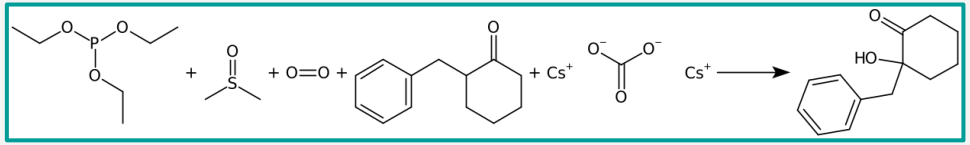
Forward model: inspection of a few predictions

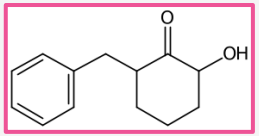
27%: identical molecular formula



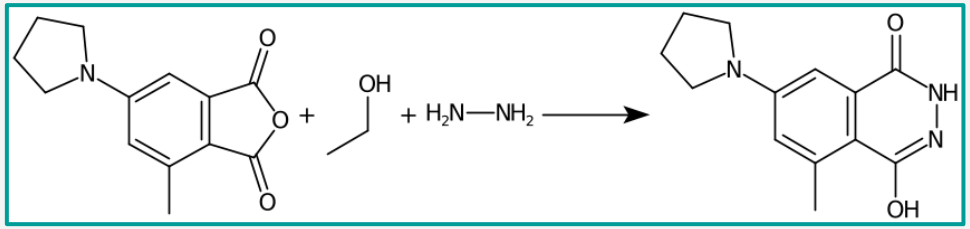


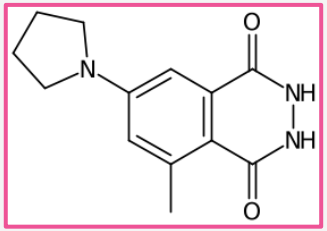
stereochemistry





regiochemistry





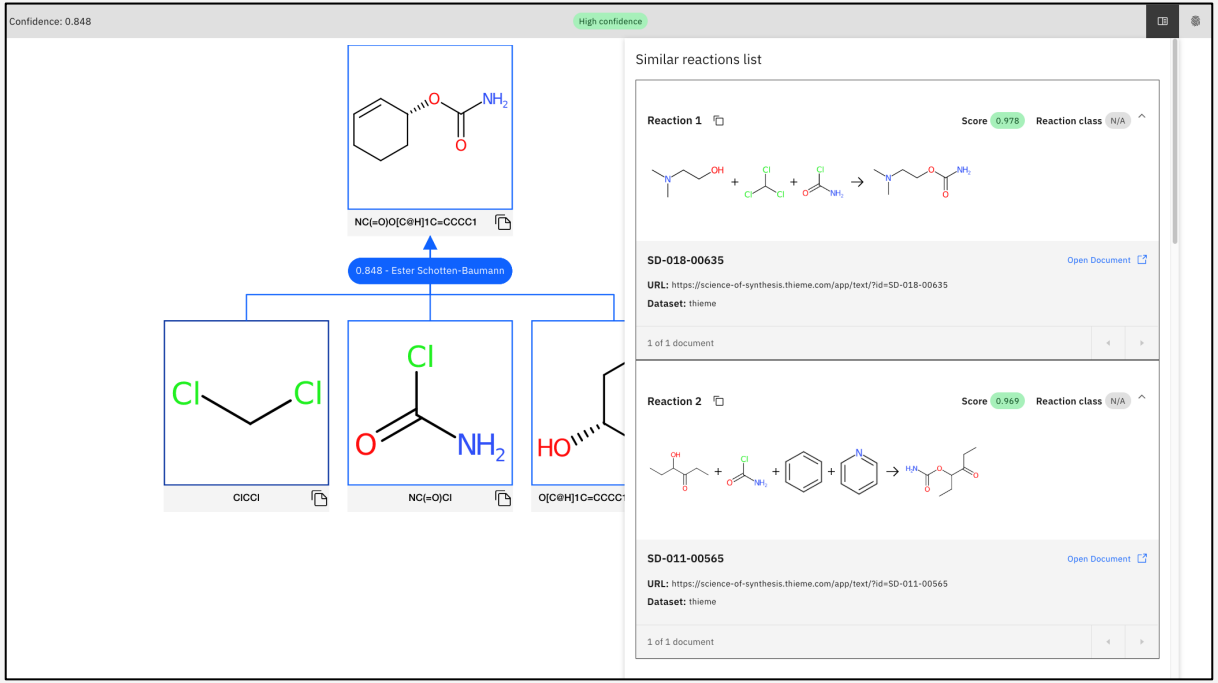
tautomers

Expert feedback

- What we asked:
  - Does curated data help?
  - What would make the models even more useful?
- Feedback:
  - Curated data helps! Considerable difference to patent-only.
  - There are still some errors
  - Comments on diversity, usability (see next slides)

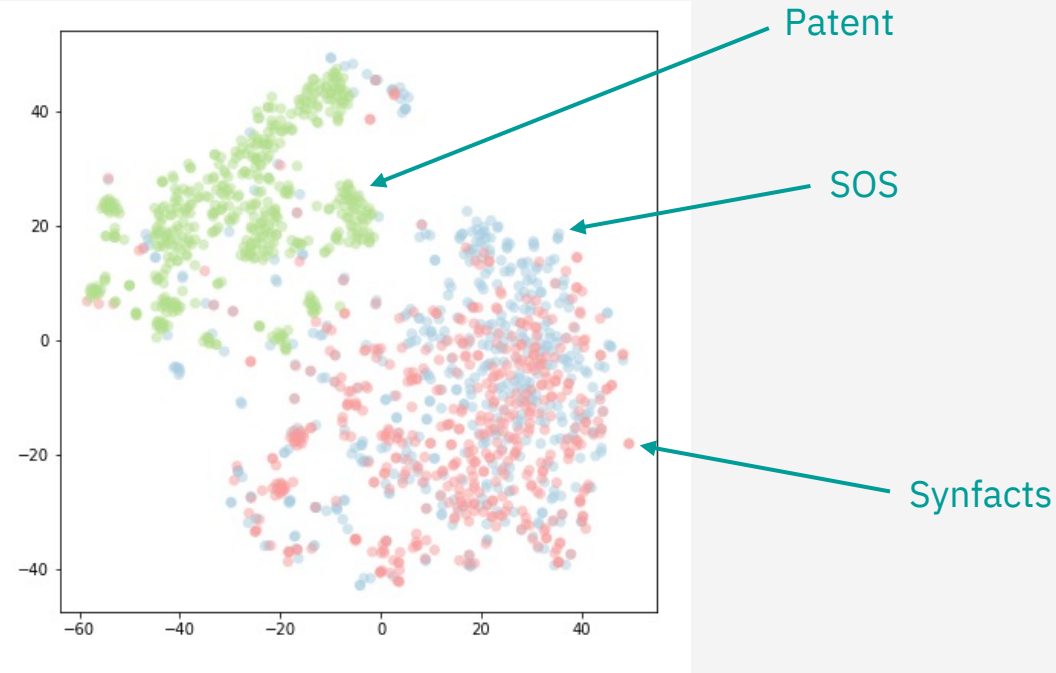
Related reactions

“It would be nice to see literature references for similar/related reactions.”



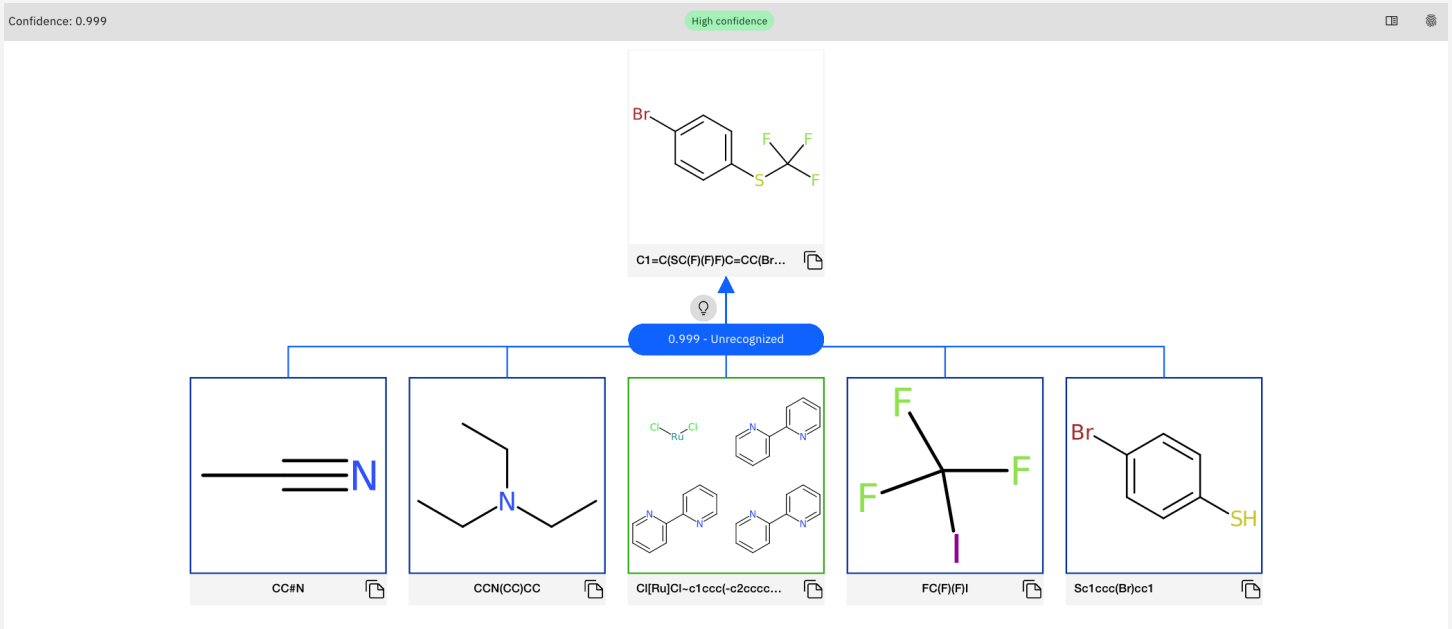
Diversity of predictions

"The retrosynthetic sequences lack variety."

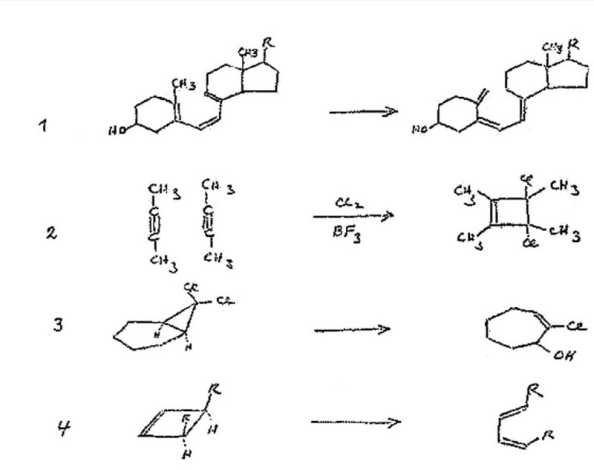


Photochemical and thermal reactions

"Photochemistry and unimolecular reactions do not work."

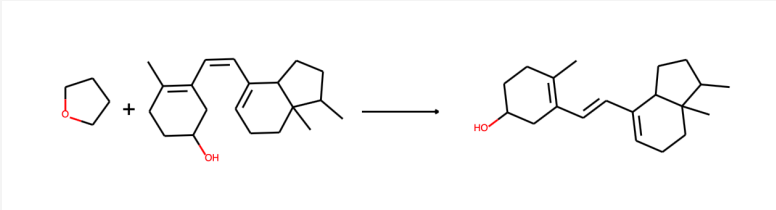


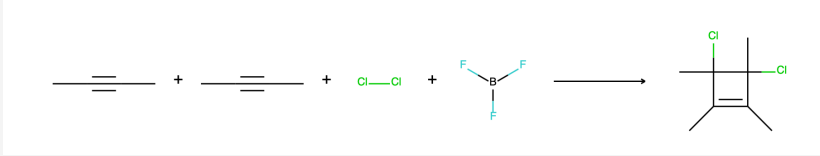
Photochemical and thermal reactions

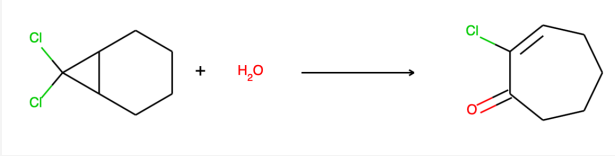


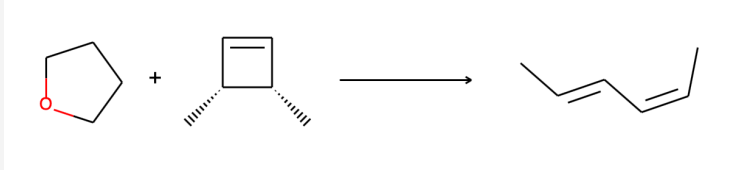
3. Woodward's Four Mysterious Reactions

Tantillo, Seeman, Chem. Eur. J. 2021, 27, 7000 – 7016

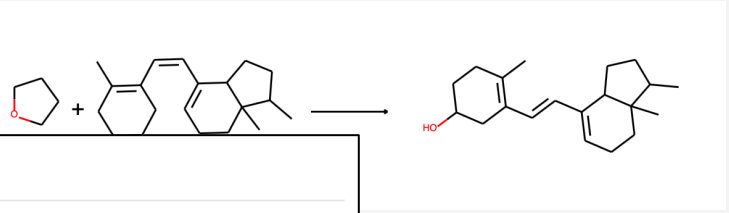


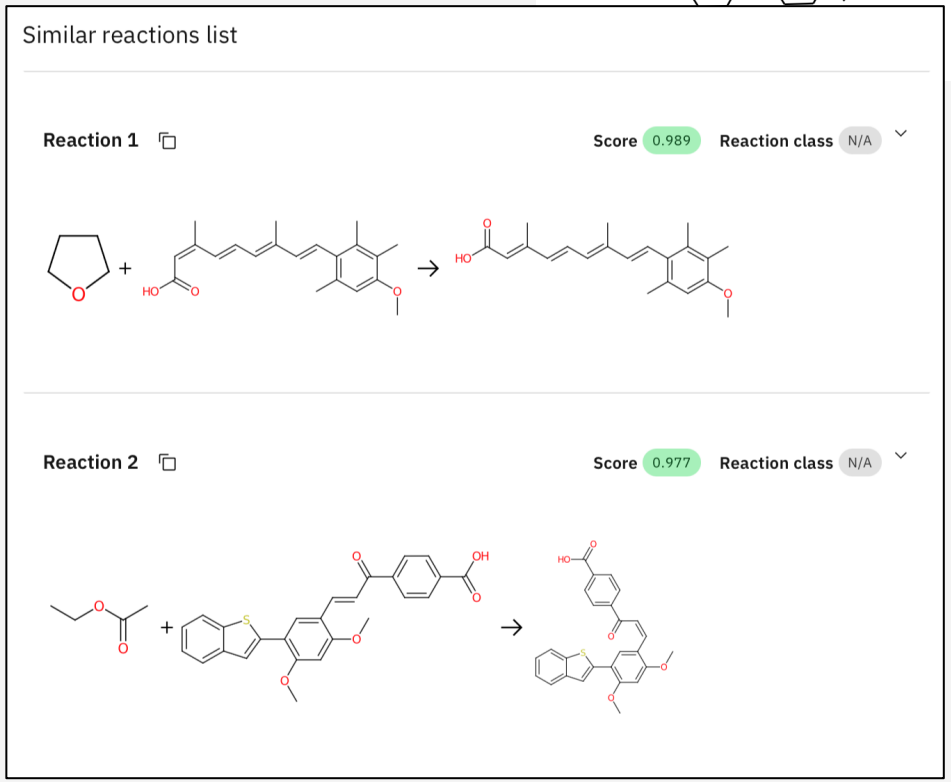






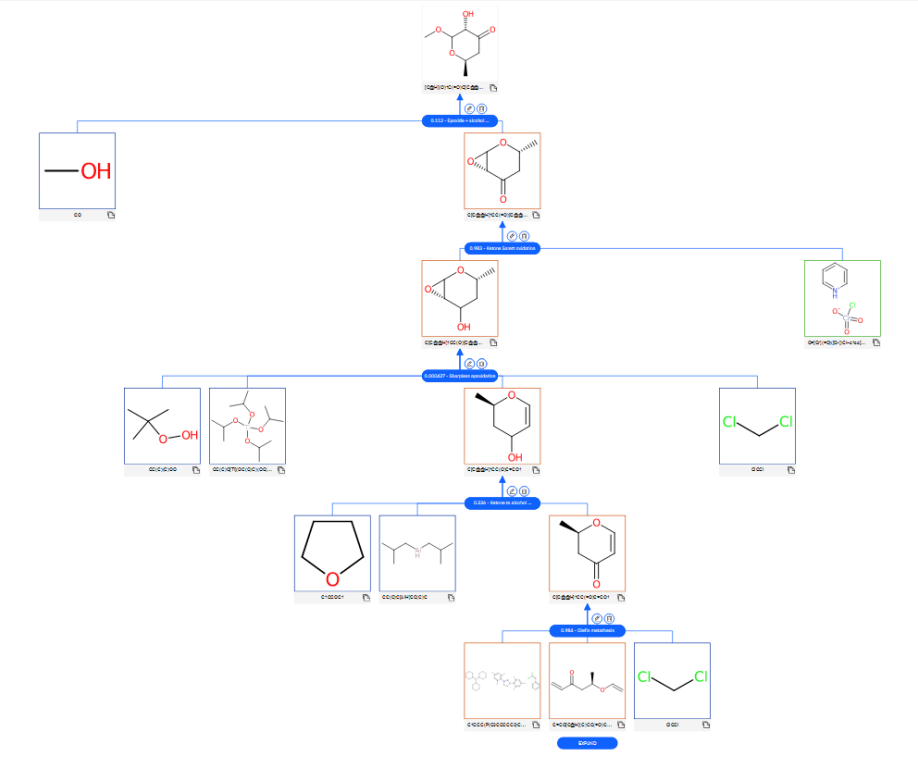
Photochemical and thermal reactions





Closest reactions come from patents and are questionable

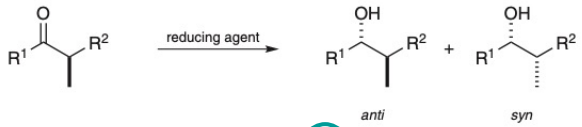
To keep in mind: still some work to do!



- “The **configuration** of the proposed epoxide is wrong!”
- “The computer model **does not seem to know/understand** OH-directed epoxidations.”
- “The model **does not consider a cycloaddition** (Hetero Diels-Alder) reaction.”

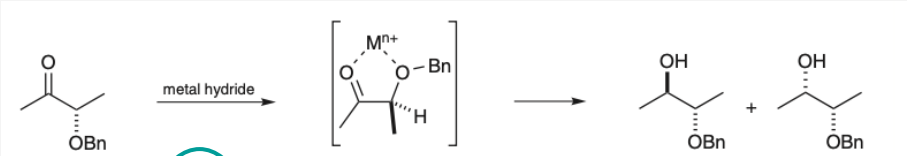
Challenges: Multiple products

Scheme 59 Diastereoselective Reduction of α-Substituted Acyclic Ketones with Borohydrides and Disiamylborane[364,365]



R^1	R^2	Reducing Agent	Ratio (*anti*/*syn*)	Ref
Ph	Et	$NaBH_4$/MeOH	57:43	[364]
Ph	CH=CH_2	$NaBH_4$/MeOH	70:30	[364]
Ph	C≡CH	$NaBH_4$/MeOH	89:11	[364]
Me	CHMePh	LiB*s*-Bu_3H/THF	96:4	[365]
Me	CHMePh	Sia_2BH/THF	20:80	[365]

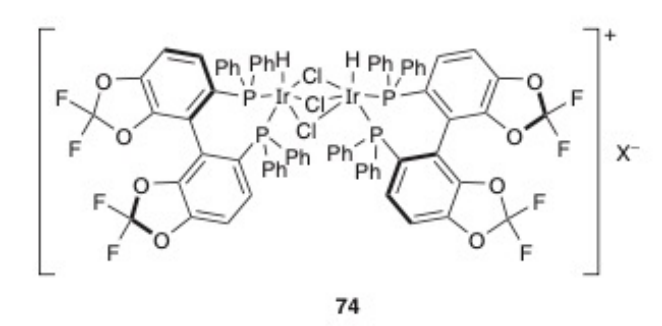
[a] Sia = siamyl (3-methylbutan-2-yl).



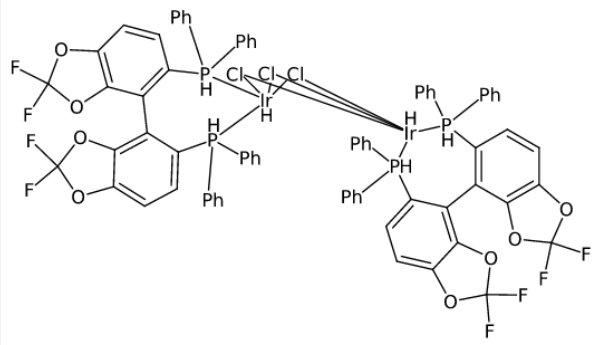
Metal Hydride	Ratio (*anti*/*syn*)	Ref
$Zn(BH_4)_2$	86:14	[385]
$LiBEt_3H$	96.5:3.5	[385]
$LiBBu_3H$	99:1	[385]

Challenges: Compound representation

PDF



Extracted structure



- Hydrogen atoms added to P
- Standardization failing (hypervalent chlorine)

Challenges

- Multiple products
- Compound representation
- Standardization, normalization
- Reagents with no SMILES
- Multiple steps drawn as one reaction
- Etc.

Note: Most of these challenges apply to all datasets!

Outline

1. Introduction:
   AI for chemical reactivity & data
2. Study (IBM Research & Thieme):
   combining curated data & data-driven models
   - Data & model training
   - Examples, challenges, expert feedback
3. Where do we go from here?

Where do we go from here?

- Curated data does have value for ML

- Handling of proprietary data for ML:
  - Data safety must be preserved
  - Train models on trusted servers? Federated learning?
  - Address concern of data leaking
- Possibility to select which data to train new models on

- Keep listening to and learning from chemists!

Thank you for your attention!

If you have any questions:

E-mail: ava@zurich.ibm.com

Twitter: @acvaucher

Acknowledgments:

IBM Research
Teodoro Laino
Philippe Schwaller

Thieme Group
Fiona Shortt de Hernandez
Sascha Hausberg
Klaus Köberlein

Experts (evaluation of models)
Prof. Dame Margaret Brimble (University of Auckland, New Zealand)
Prof. Alois Fürstner (MPI Mülheim, Germany)
Prof. Karl Gademann (University of Zurich, Switzerland)
Prof. Ang Li (Shanghai Institute of Organic Chemistry, China)
Prof. Cristina Nevado (University of Zurich, Switzerland)
Prof. Richmond Sarpong (University of California, Berkeley, USA)
Prof. Dirk Trauner (University of Pennsylvania, USA)
... and their research groups