Inferring missing molecules in incomplete chemical equations

Alain Vaucher, Philippe Schwaller, Zeineb Ayadi, Alessandra Toniato, Teodoro Laino

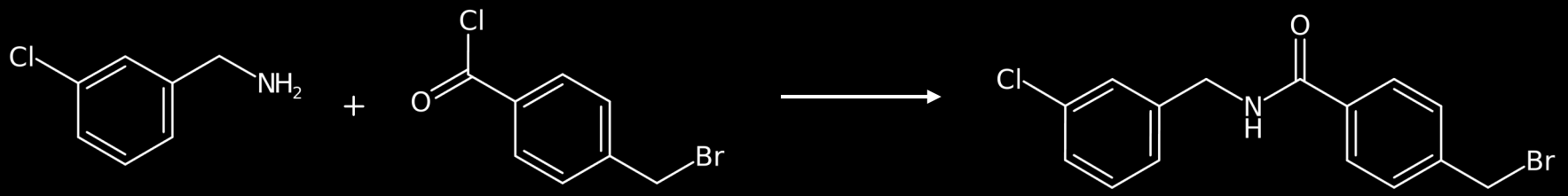
IBM Research Europe – Zurich, Switzerland

@acvaucher

ACS Fall 2021
August 22, 2021

Background – IBM RXN

Reaction prediction



Textual representation (SMILES)

NCc1cccc(Cl)c1 O=C(Cl)c1ccc(CBr)cc1 O=C(NCc1cccc(Cl)c1)c1ccc(CBr)cc1

“Sentence of atoms”

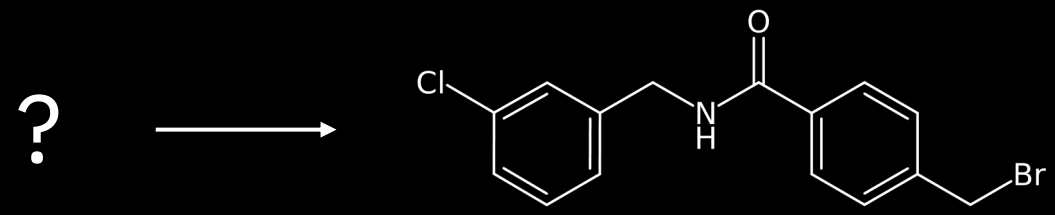
“Translation”

N C c 1 c c c c (Cl) c 1 . O = C (Cl) c 1 c c c (C Br) c c 1 → O = C (N C c 1 c c c c (Cl) c 1) c 1 c c c (C Br) c c 1

Molecular Transformer

Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C. & Lee, A. A., *ACS Cent. Sci.*, **2019**, *5*, 1572-1583.

Retrosynthetic analysis



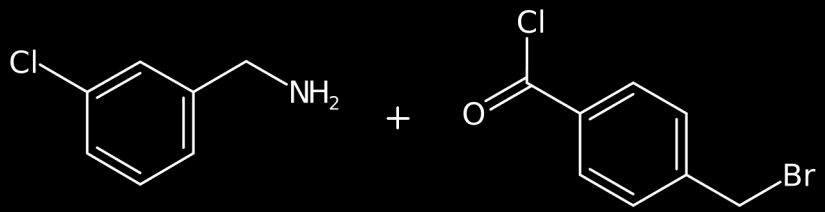
Similar approach, both sides switched

“Translation”

`O = C ( N C c 1 c c c c ( Cl ) c 1 ) c 1 c c c ( C Br ) c c 1` ⟹ `N C c 1 c c c c ( Cl ) c 1 . O = C ( Cl ) c 1 c c c ( C Br ) c c 1`

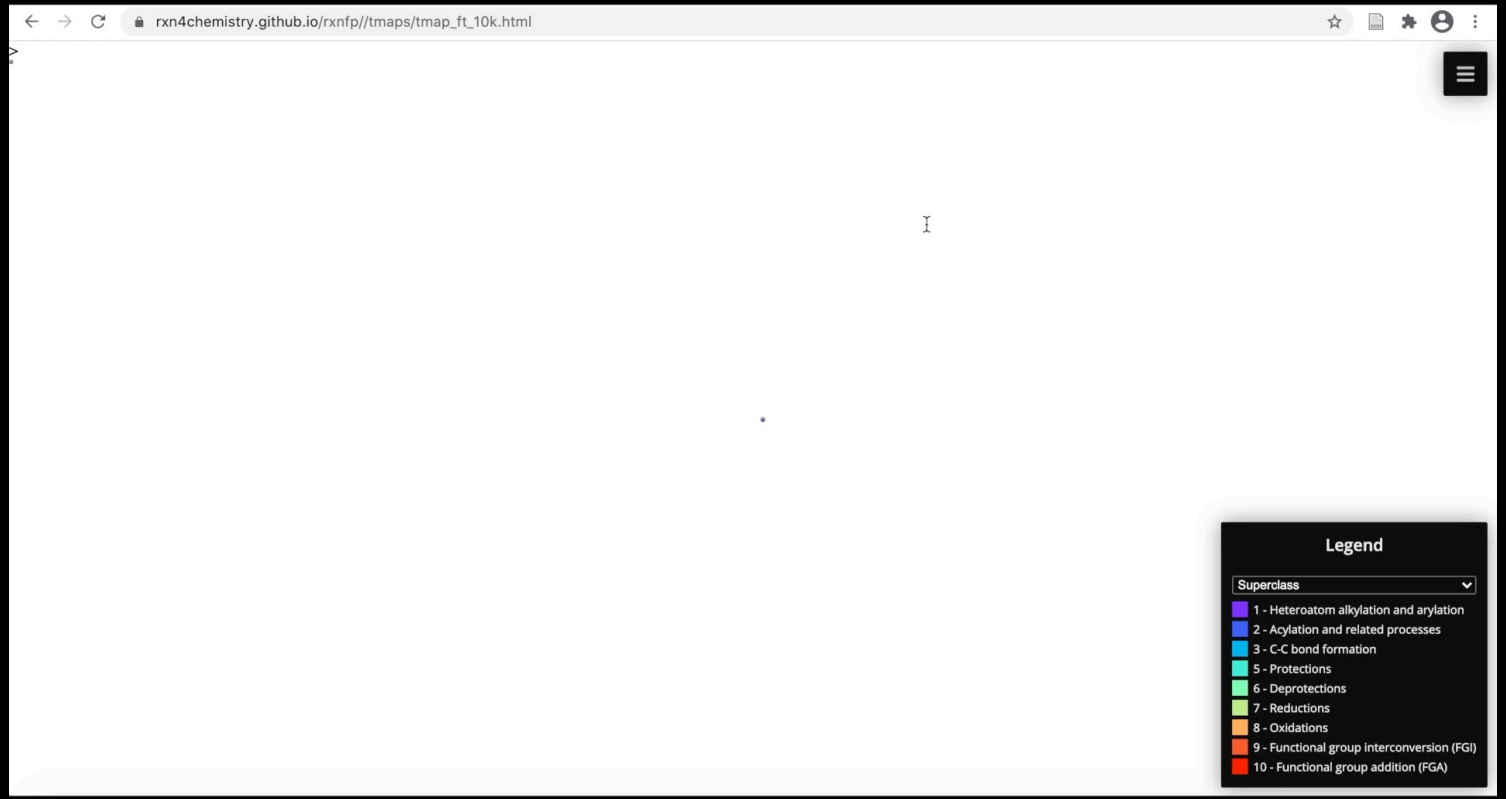
Transformer

One among many correct sets of precursors



Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A. & Laino, T., *Chem. Sci.*, **2020**, *11*, 3316-3325.

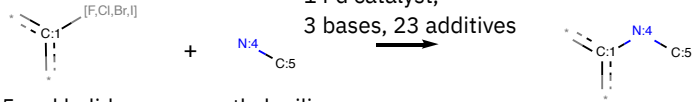
Classifying and mapping reactions



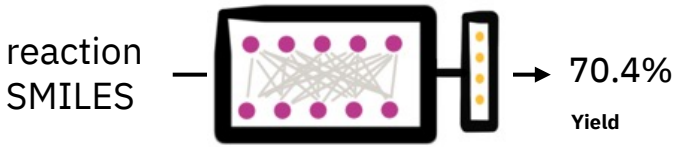
Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T. & Reymond, J.-L., *Nat. Mach. Intell.*, **2021**, *3*, 144-152.

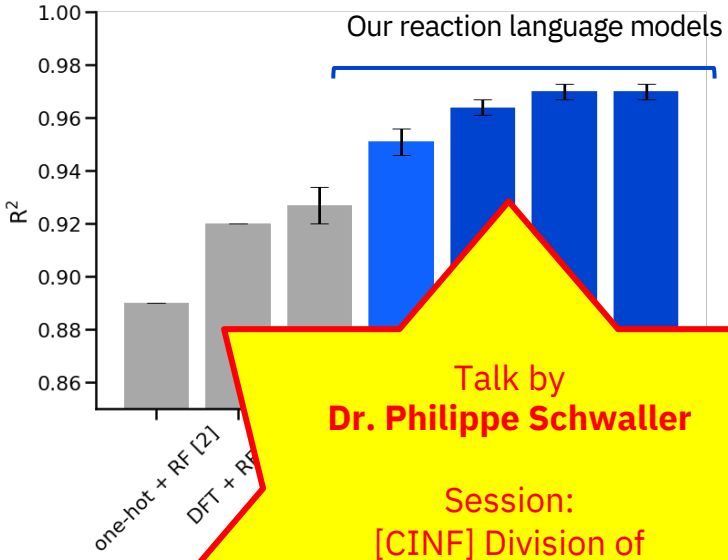
Prediction of chemical reaction yields

3955 **Buchwald-Hartwig reactions**



15 aryl halides methylaniline





Talk by
Dr. Philippe Schwaller

Session:
[CINF] Division of Chemical Information

Sunday, August 22, 5.15 pm (ET)

[1] Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).

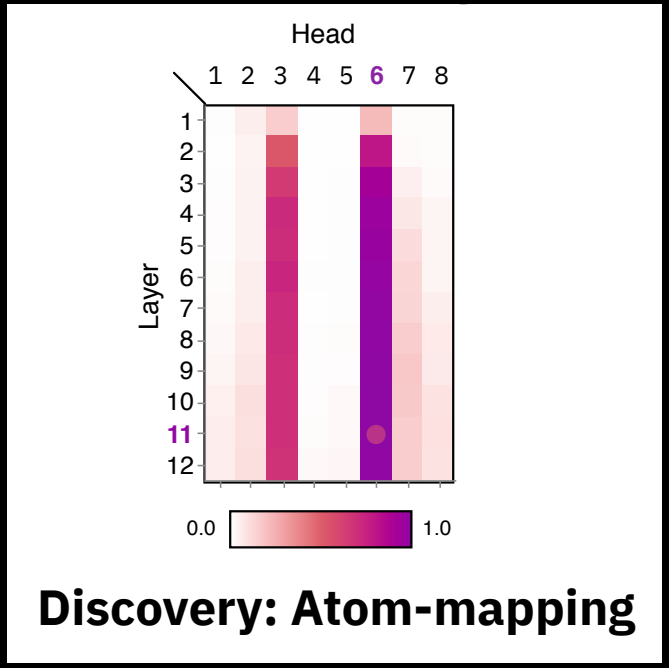
[2] Chuang, K. V. & Keiser, M. J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* **362** (2018).

[3] Sandfort, F., Strieth-Kalthoff platform for predicting ch

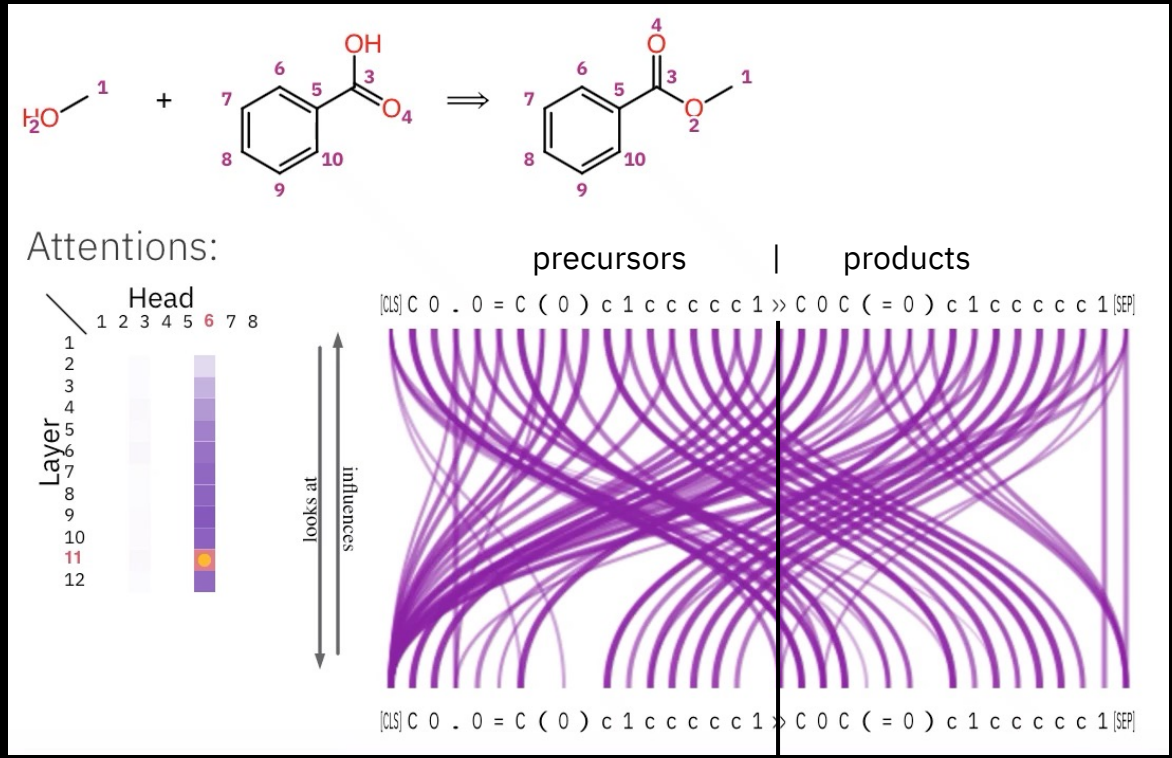
[4] Schwaller, P., Vaucher, A. C., La using deep learning. *ChemRxiv prepri*

Schwaller, P.; Vaucher, A. C.; Laino, T. & Reymond, J.-L., *Mach. Learn.: Sci. Technol.*, **2021**, *2*, 015016.

Atom mapping: RXNMapper

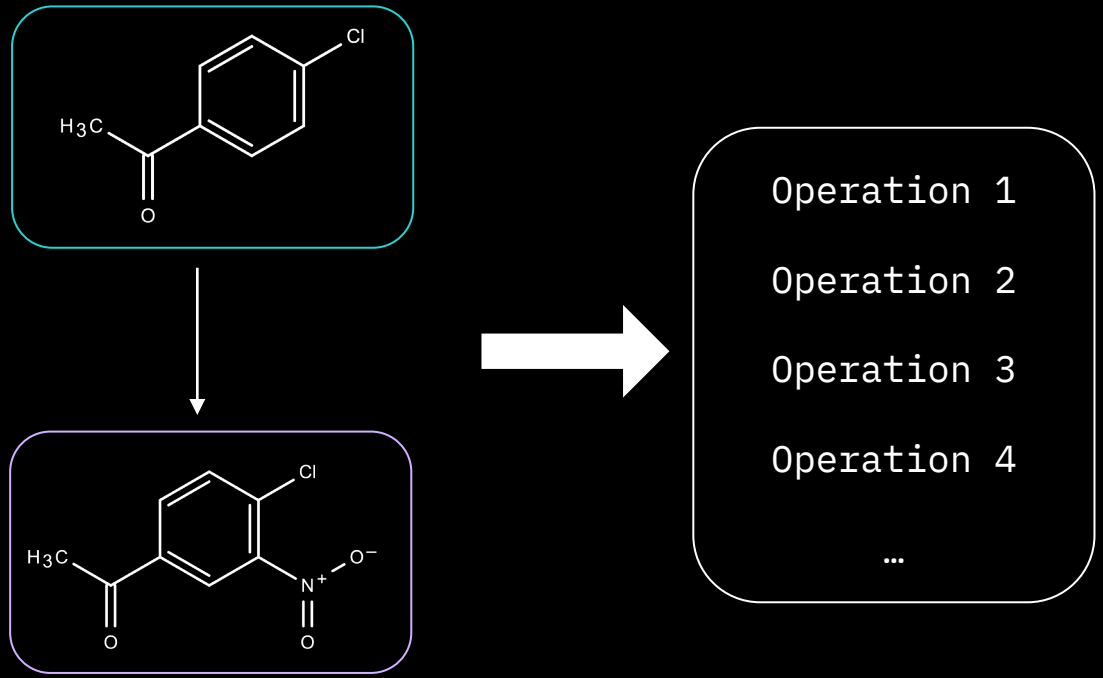


Discovery: Atom-mapping



Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H. & Laino, T., *Sci. Adv.*, **2021**, *7*, eabe4166.

Synthesis actions & synthesis automation



C1=CC(C(=O)C)=CC=C1Cl>>C1=CC(C(=O)C)=CC([N+]([O-])=O)=C1Cl

Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T., *Nat. Commun.*, **2021**, *12*, 2573.

Synthesis actions & synthesis automation

Talk by
Dr. Teodoro Laino

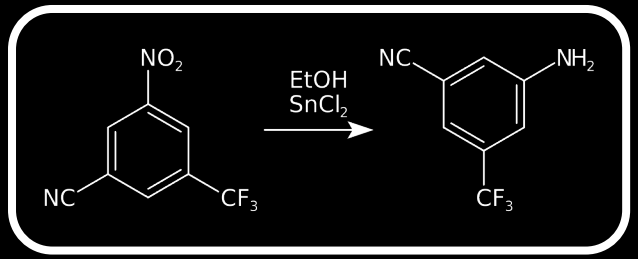
Session:
[ENFL] Division of Energy and Fuels

Monday, August 23, 10.30 am (ET)

Completion of partial chemical equations

Motivation: predictions on reactions

Forward reaction prediction,
retrosynthetic analysis, ...



Reaction class

Nitro to amino

Schwaller et al., Nat. Mach. Intell., 2021, 3, 144-152.

Reaction yield

87.1% yield

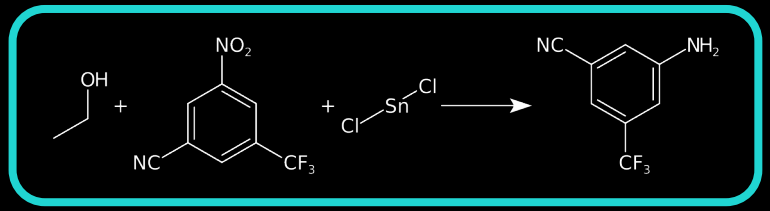
Schwaller et al., Mach. Learn.: Sci. Technol., 2021, 2, 015016.

Synthesis actions

1. Add 3-nitro-5-(trifluoromethyl)benzonitrile
2. Add SnCl2
3. Add ethanol
4. Reflux for 1 hour
5. Concentrate
6. Add ethyl acetate
7. Add NaHCO3
8. Filter
9. Collect organic layer
10. Wash with brine
11. Dry with MgSO4
12. Concentrate

Vaucher et al., Nat. Commun., 2021, 12, 2573.

Motivation: predictions on reactions

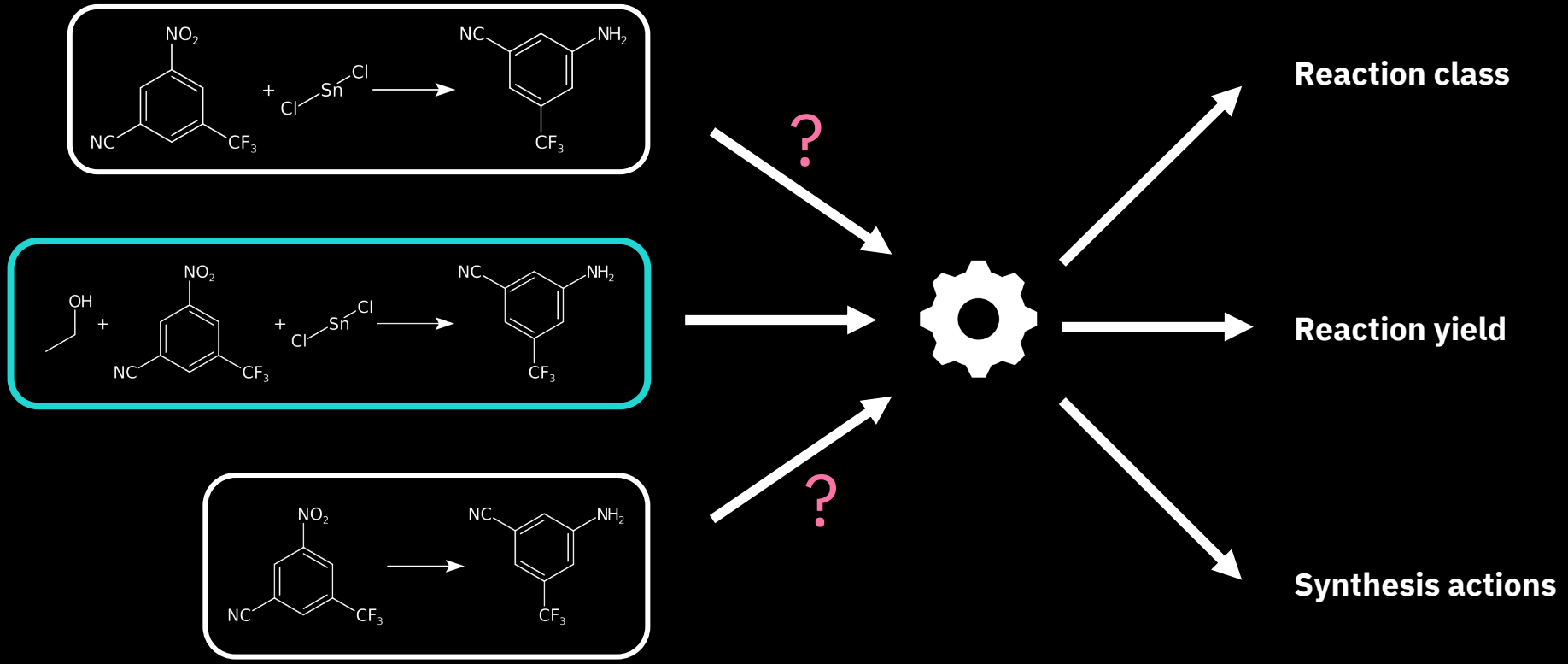


Reaction class

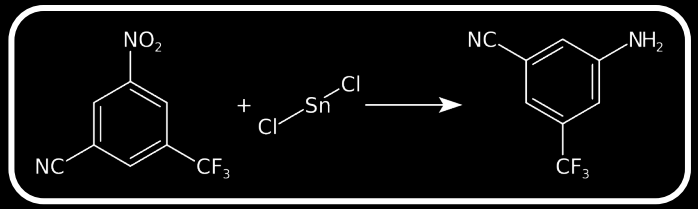
Reaction yield

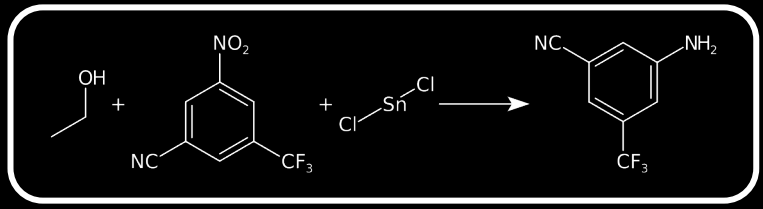
Synthesis actions

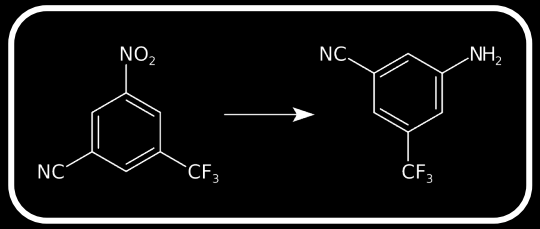
Motivation: predictions on reactions

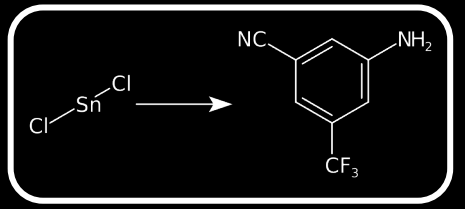


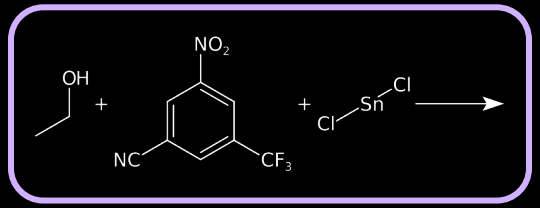
Completing partial chemical equations



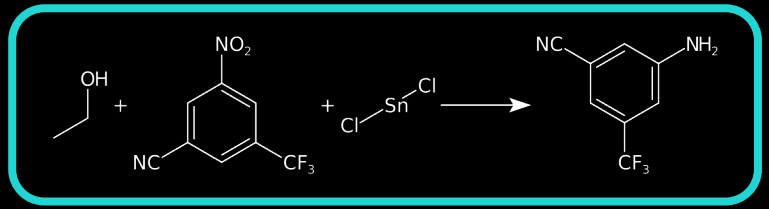


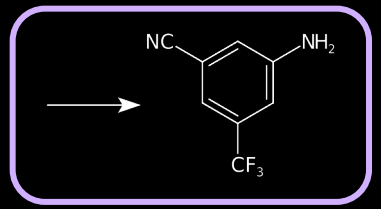






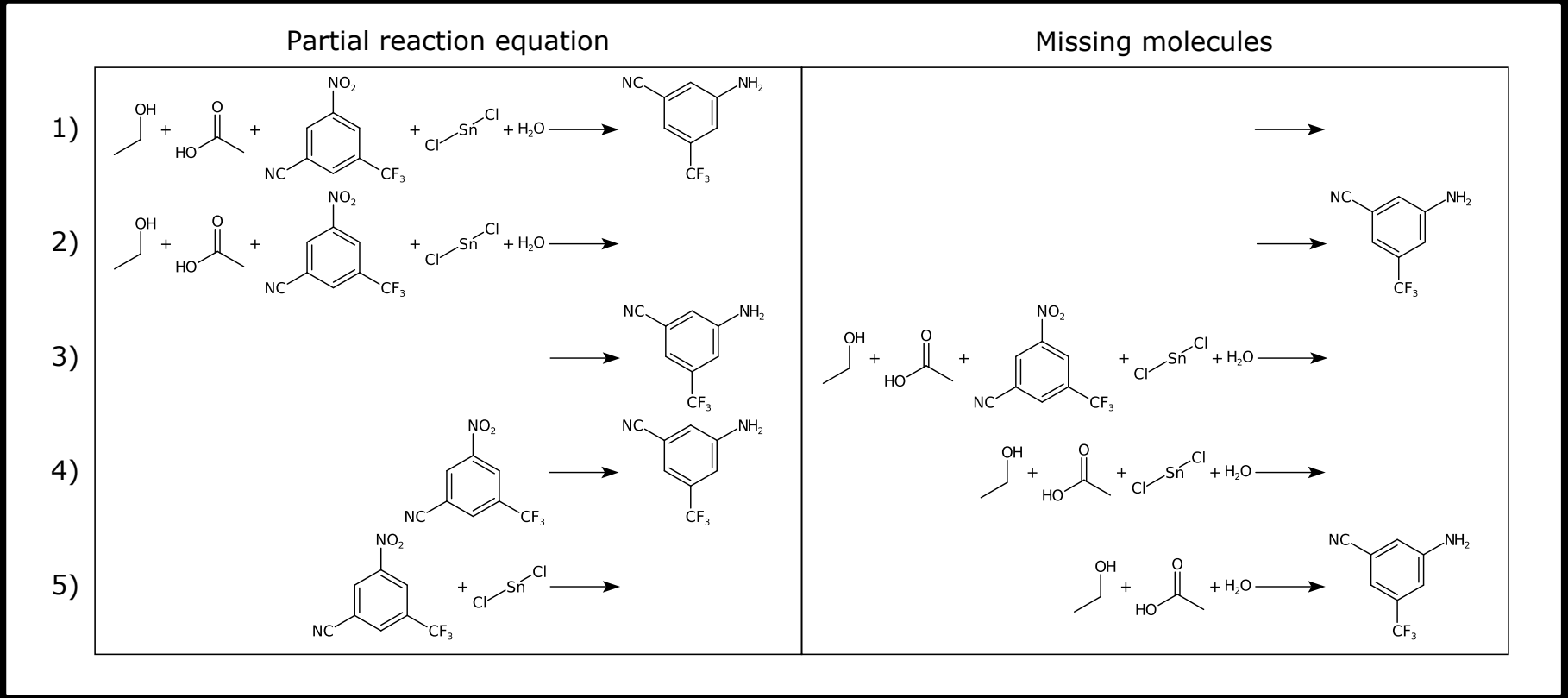
Forward reaction prediction





Single-step retrosynthesis

Completing partial reaction equations

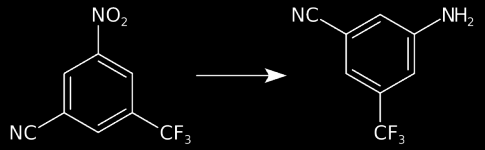


Machine learning model

Inspired by the **Molecular Transformer**

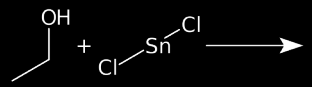
Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C. & Lee, A. A., *ACS Cent. Sci.*, **2019**, *5*, 1572-1583.

Input*



"Translation"

Output*



`N#Cc1cc([N+](=O)[O-])cc(C(F)(F)F)c1>>N#Cc1cc(N)cc(C(F)(F)F)c1`

`CCO.Cl[Sn]Cl>>`

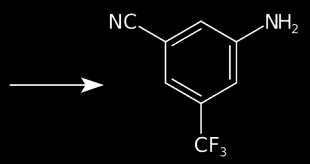
*Not shown here for readability: the Model uses tokenized SMILES strings: "`CCO.Cl[Sn]Cl>>`" → "`C C O . Cl [Sn] Cl >>`"

Machine learning model

Inspired by the **Molecular Transformer**

Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C. & Lee, A. A., *ACS Cent. Sci.*, **2019**, *5*, 1572-1583.

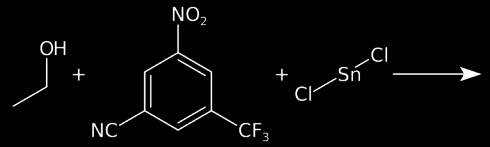
Input*



>>N#Cc1cc(N)cc(C(F)(F)F)c1

"Translation"

Output*

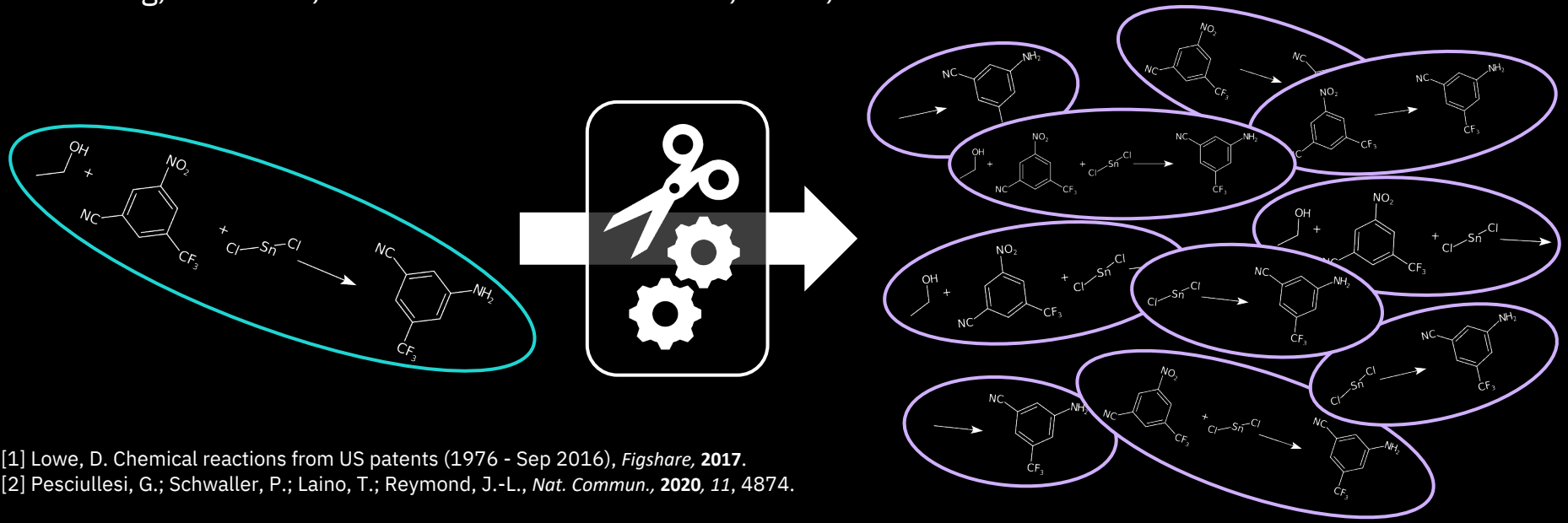


CCO.Cl[Sn]Cl.N#Cc1cc([N+](=O)[O-])cc(C(F)(F)F)c1>>

*Not shown here for readability: the Model uses tokenized SMILES strings: "CCO.Cl[Sn]Cl>>" → "C C O . Cl [Sn] Cl >>"

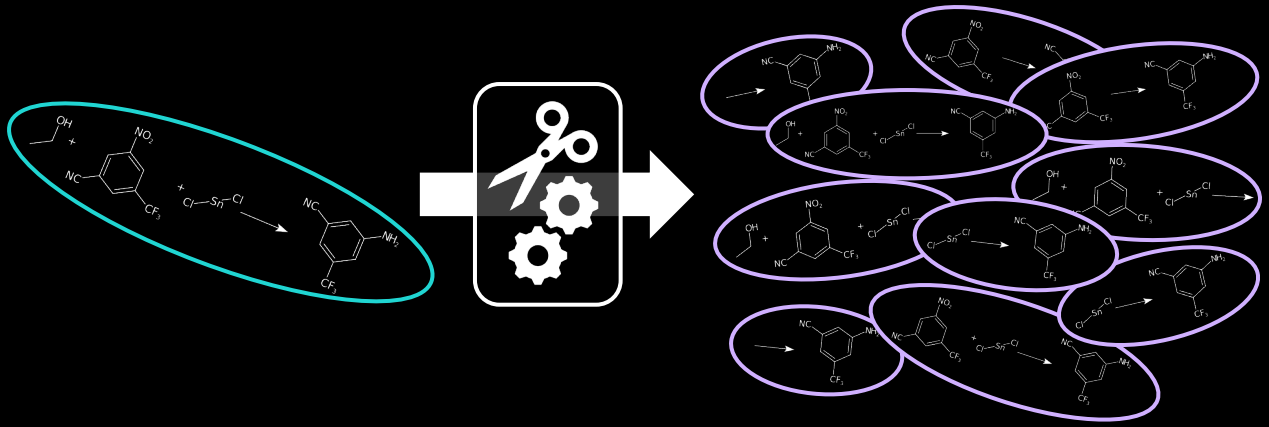
Data

- US patent reactions (from Ref. [1]), as post-processed by Ref. [2].
- **10** partial reaction SMILES per reaction.
- training, validation, and test sets of sizes **10.9 M**, **0.6 M**, and **0.6 M**.



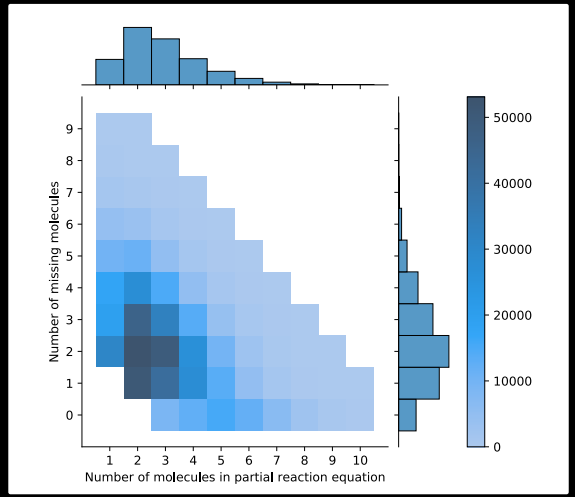
[1] Lowe, D. Chemical reactions from US patents (1976 - Sep 2016), *Figshare,* **2017**.
[2] Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L., *Nat. Commun.,* **2020**, *11*, 4874.

Data

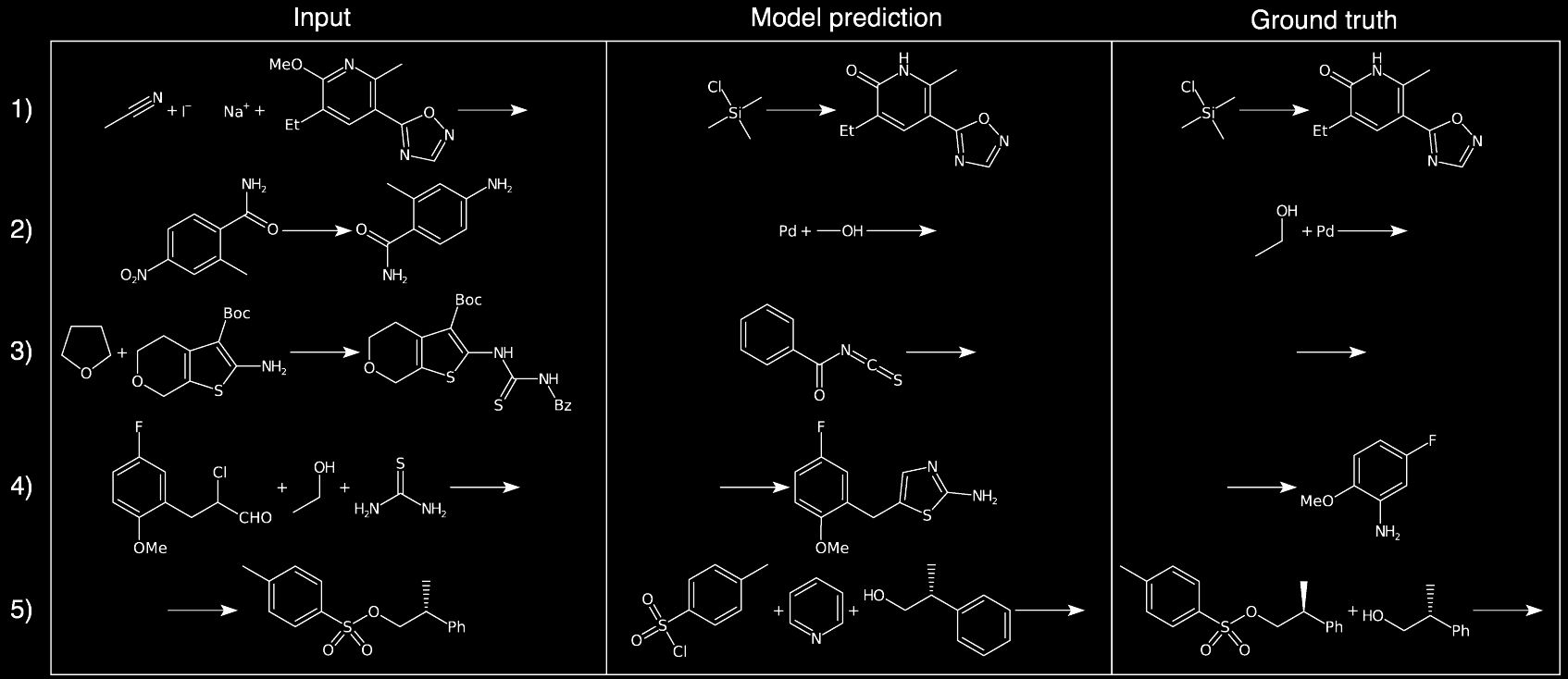


Challenges:

- Adequate balance between sub-tasks
- How many compounds to remove
- Ambiguity (many correct answers)
- Etc.

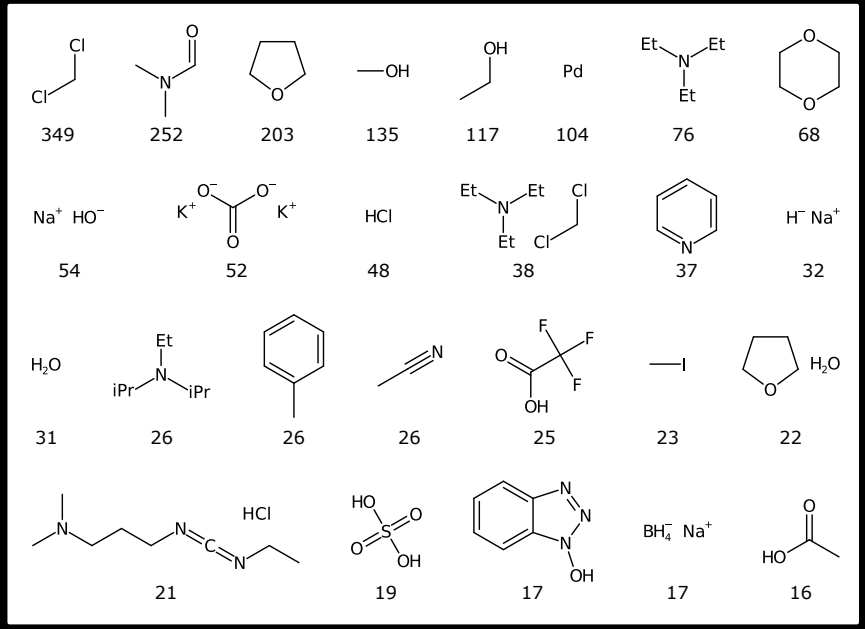


Initial results



Initial results

- Partial reaction completion:
  - Accuracy (on ground truth): **30.4%**
  - Validity of resulting reactions: **77.6%**
- Forward reaction prediction:
  - Accuracy: **68.1%**
  - Reference [1]: 77.6%
- Single-step retrosynthesis:
  - Round-trip accuracy: **81.5%**
  - Reference [2]: 81.2%
- Application on ground truth data:
  - 2.7k (out of 60.5k in the test set) reactions considered to be incomplete

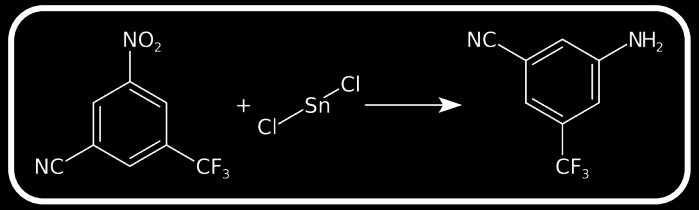


Common missing molecules in dataset

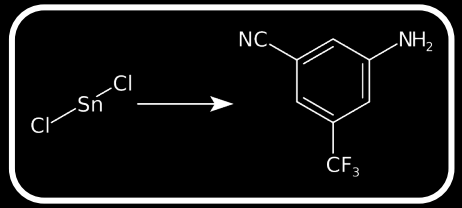
[1] Pesciullesi et al., *Nat. Commun.,* **2020**, *11*, 4874.
[2] Schwaller et al., T., *Chem. Sci.,* **2020**, *11*, 3316-3325.

Ongoing work

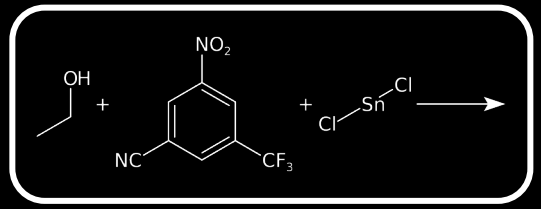
Sub-tasks



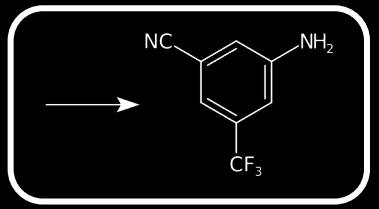
“Predict the solvent”



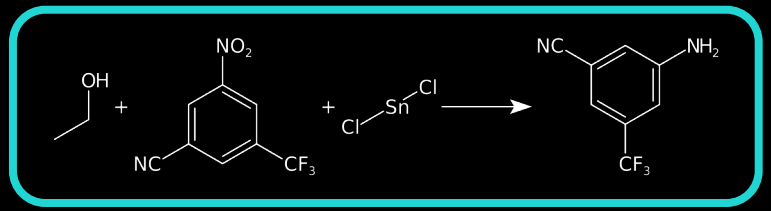
”Predict whatever is missing”



“Predict the product”



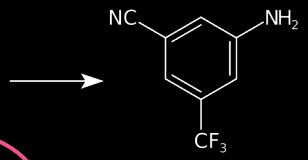
“Predict the precursors”



Still one single model!

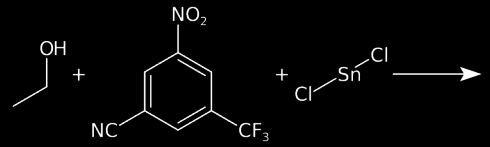
Sub-tasks with prefixes

Input*



“Translation”

Output*



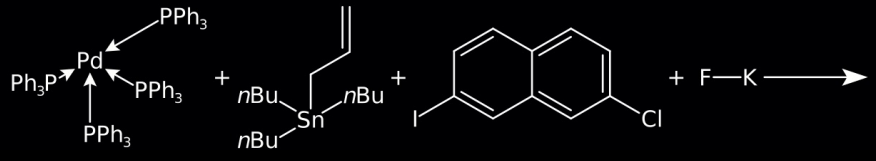
`[retro] >>N#Cc1cc(N)cc(C(F)(F)F)c1`

`CCO.Cl[Sn]Cl.N#Cc1cc([N+](=O)[O-])cc(C(F)(F)F)c1>>`

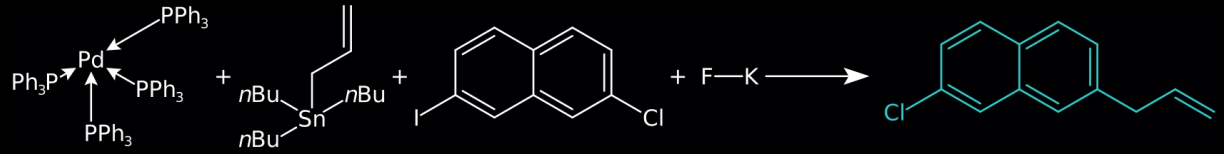
Prefix in input string,
specifies the sub-task

Sub-tasks with prefixes

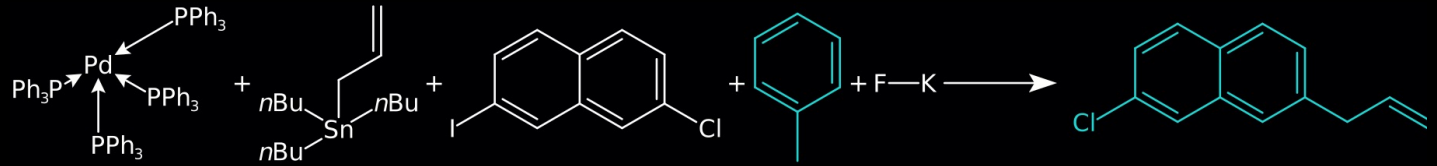
Incomplete reaction:



“forward” prefix:



“any” prefix:



Conclusions and open questions

- AI model for determining missing molecules
- One single model for multiple tasks:
  - Forward reaction prediction
  - Single-step retrosynthesis
  - Solvent/catalyst prediction
  - …
- Better (?) than models trained individually
- Compatibility of downstream AI models

Thank you for your attention!

Questions or comments

E-mail: ava@zurich.ibm.com

Twitter: @acvaucher

Preprint with initial results: ibm.biz/rxn-completion