Human-in-the-loop for a disconnection aware retrosynthesis

Andrea Byekwaso, **Alain Vaucher**, Philippe Schwaller, Alessandra Toniato, Teodoro Laino

@acvaucher

ACS Spring
March 23, 2022

IBM **Research**

IBM and the data science / chemistry ecosystem?

- Accelerated discovery – IBM RXN
- Two sides in our relationship with the ecosystem:
  - User: existing data/tools to develop new AI models
  - Provider: make technology available & usable via platform
- Interesting interplay between research and platform development!

- Example (today's talk): disconnection aware retrosynthesis

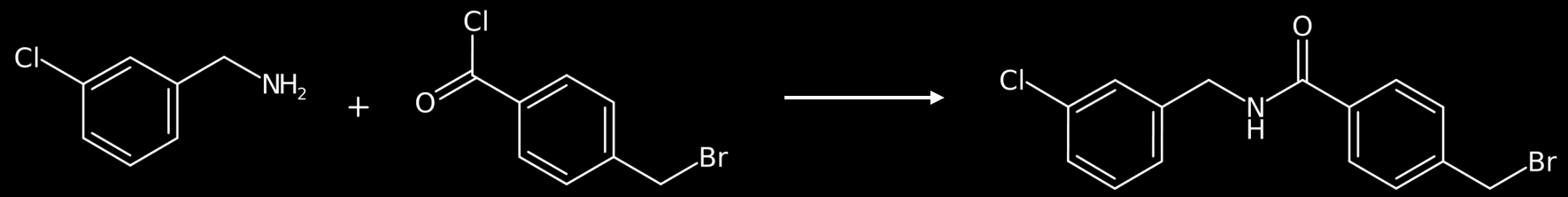
OUTLINE

1. IBM RXN research
2. IBM RXN platform
3. Data-driven retrosynthetic models in practice
4. Disconnection aware retrosynthesis

OUTLINE

1. IBM RXN research
2. IBM RXN platform
3. Data-driven retrosynthetic models in practice
4. Disconnection aware retrosynthesis

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”

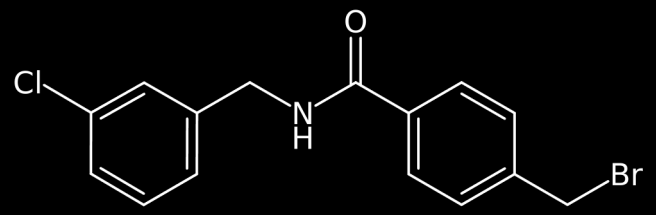
N C c 1 c c c c (Cl) c 1 . O = C (Cl) c 1 c c c (C Br) c c 1 → O = C (N C c 1 c c c c (Cl) c 1) c 1 c c c (C Br) c c 1

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

?

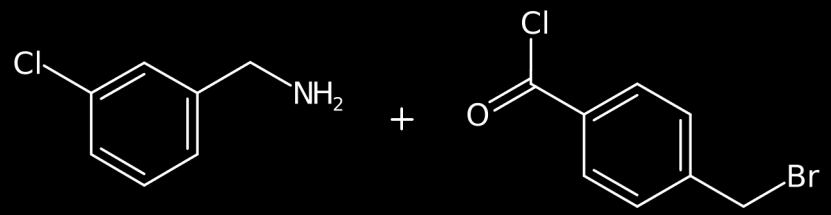


Similar approach, both sides switched

“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

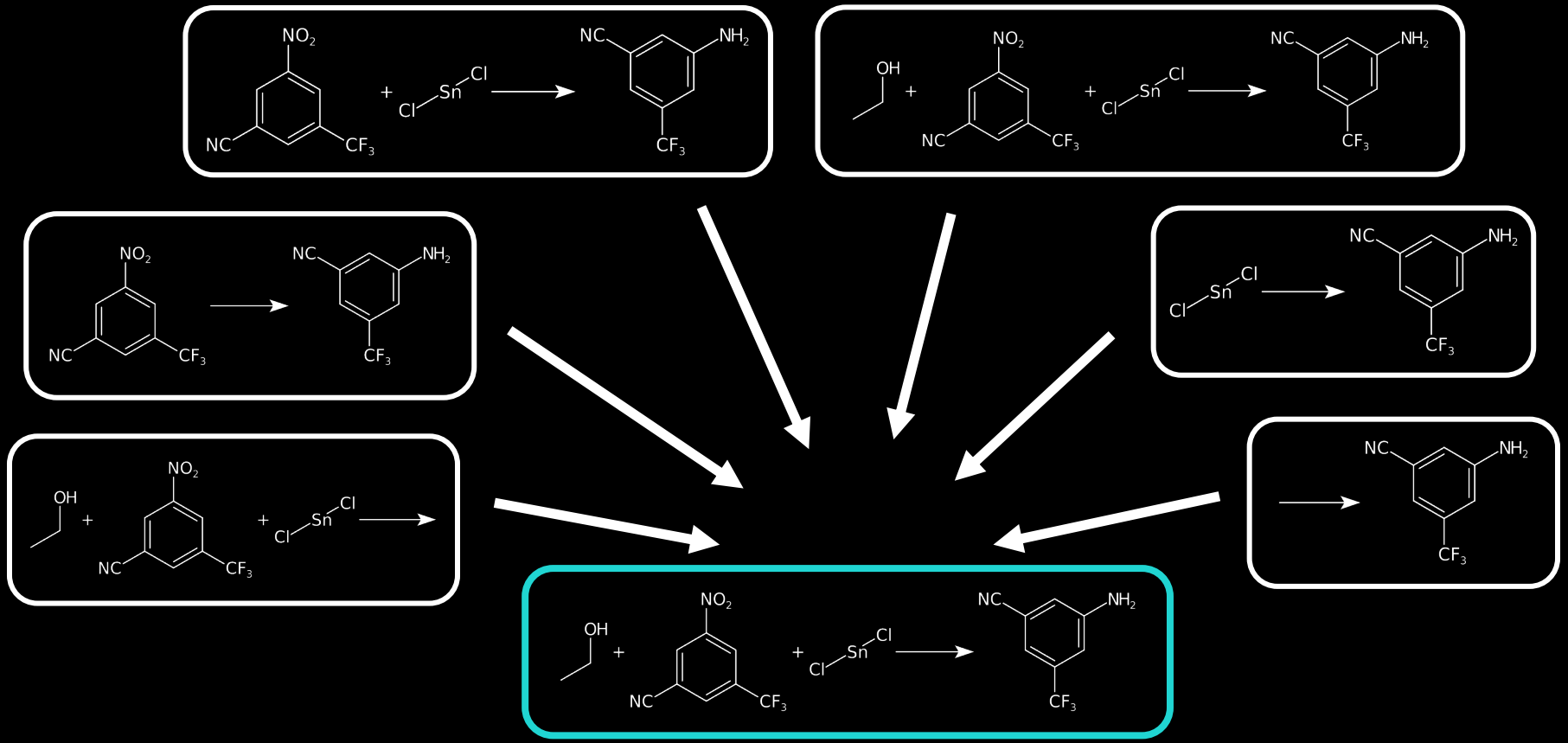
Transformer



One among many correct sets of precursors

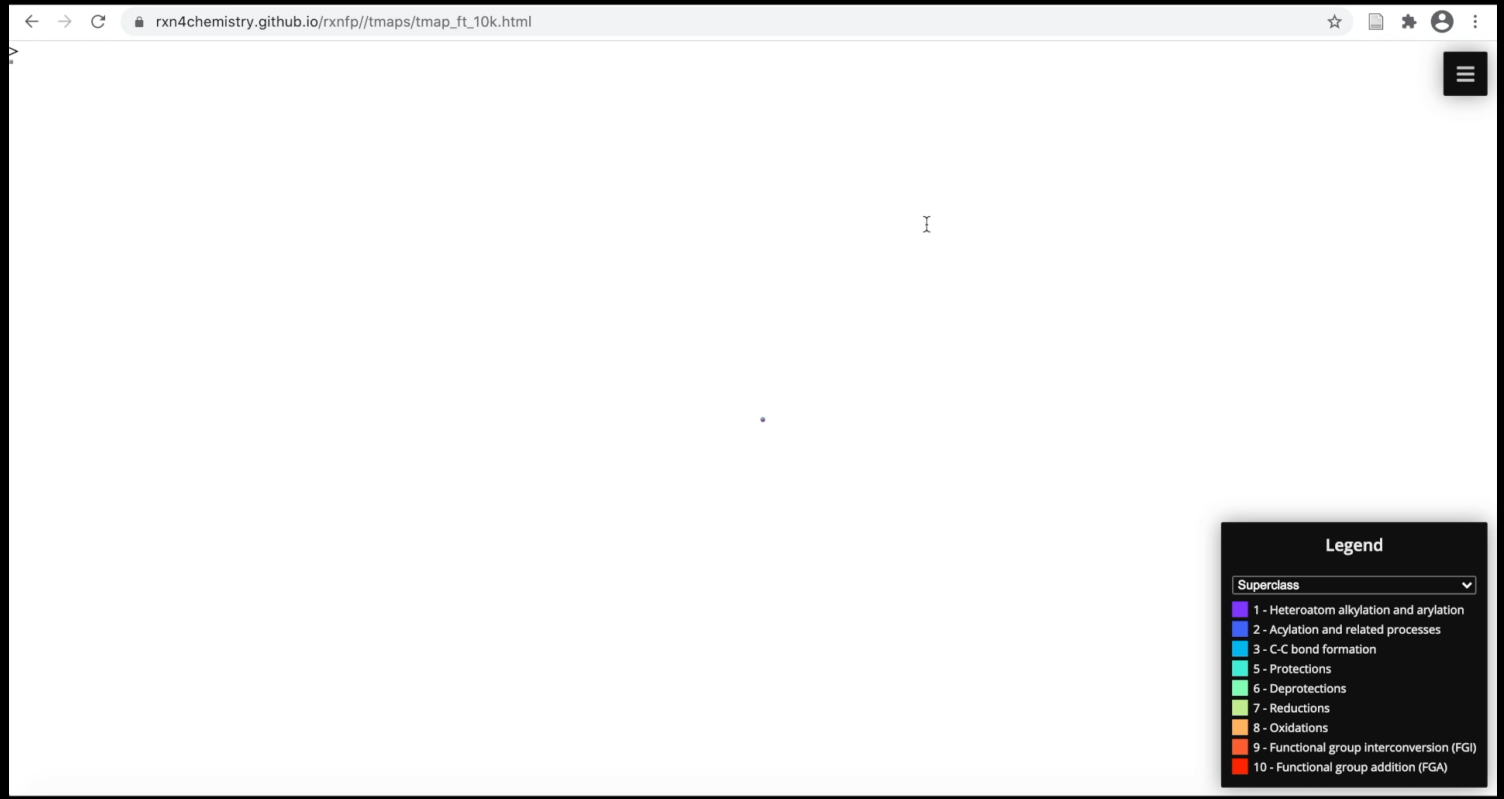
Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A. & Laino, T., *Chem. Sci.*, **2020**, *11*, 3316-3325.

Completing partial chemical equations



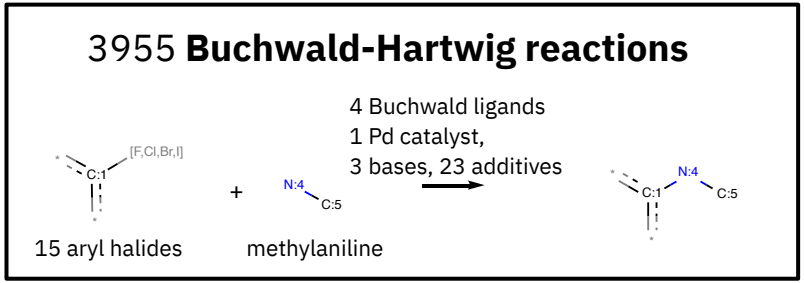
Vaucher, A. C.; Schwaller, P. & Laino, T., Chemrxiv.13273310, **2020**.

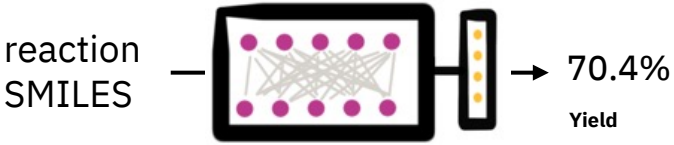
Classifying and mapping reactions

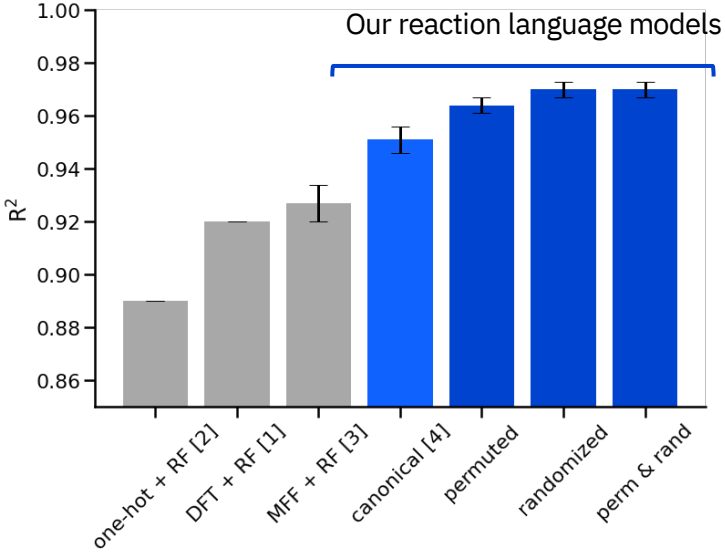


Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T. & Reymond, J.-L., *Nat. Mach. Intell.*, **2021**, *3*, 144-152.

Prediction of chemical reaction yields







[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).

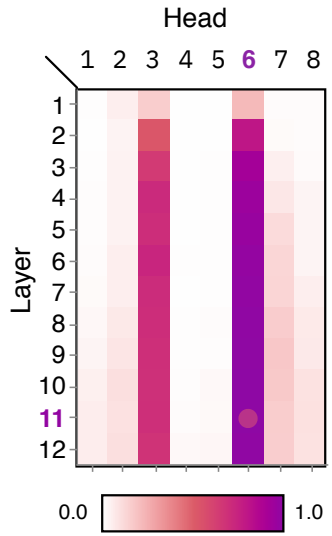
[2] Chuang, K. V. & Keiser, M. J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* **362** (2018).

[3] Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* (2020).

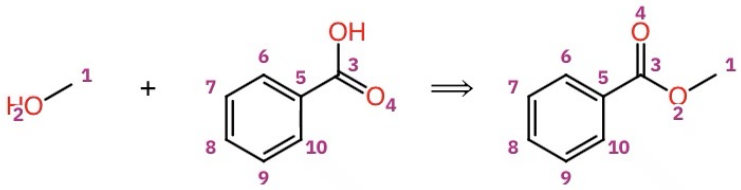
[4] Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *ChemRxiv preprint* doi:10.26434/chemrxiv.12758474 (2020).

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

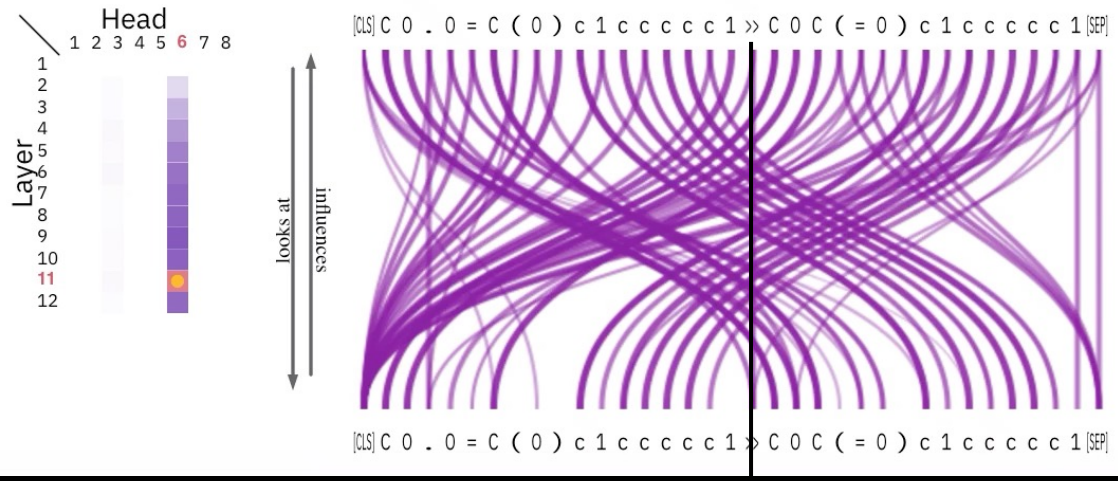
Atom mapping: RXNMapper



Discovery: Atom-mapping

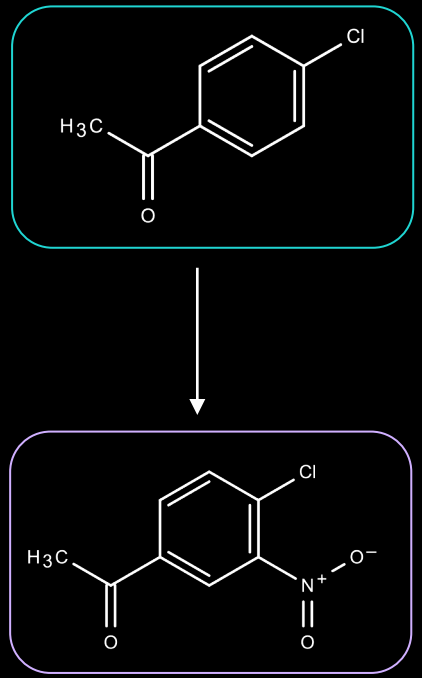


Attentions:



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

Synthesis actions & synthesis automation



Operation 1

Operation 2

Operation 3

Operation 4

...

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

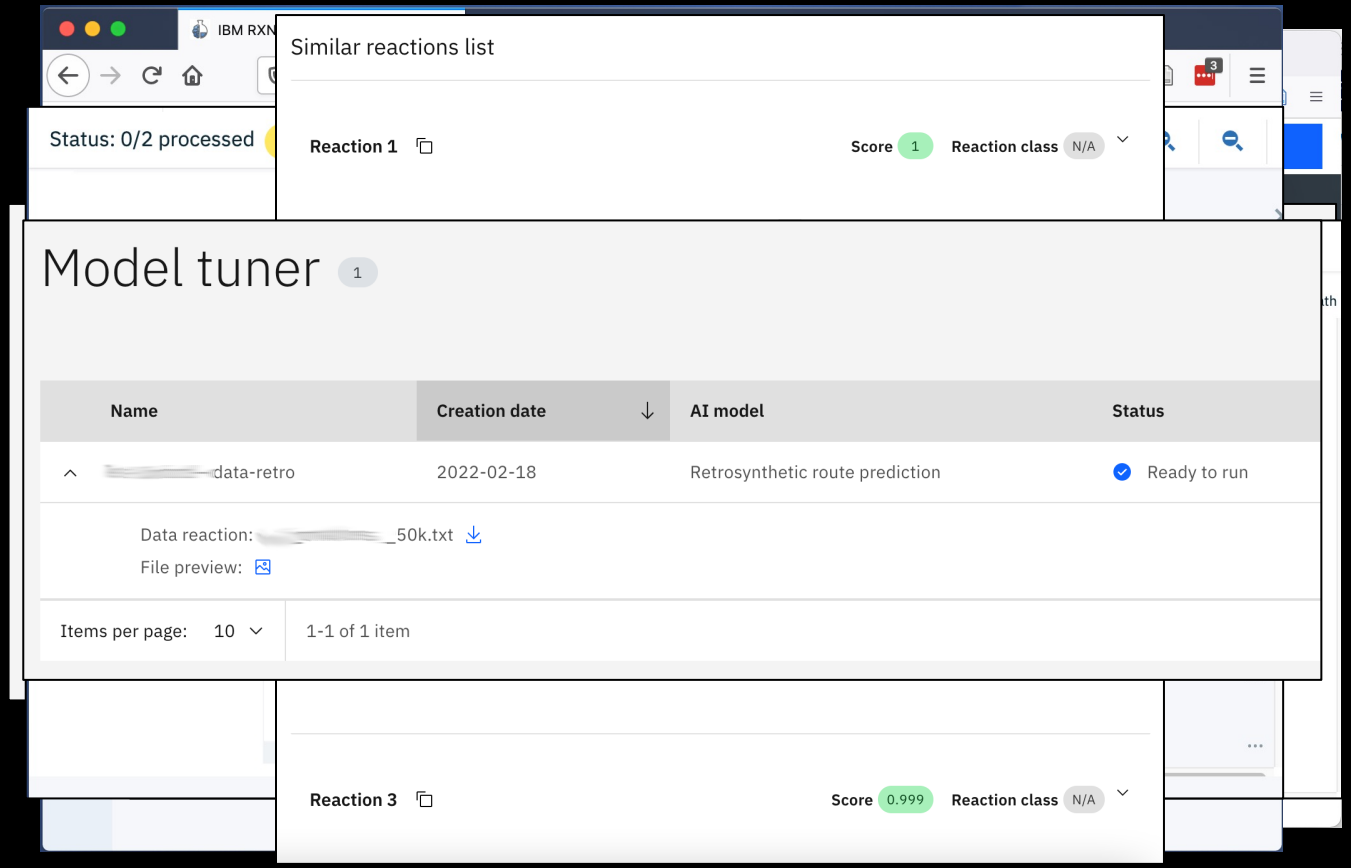
Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T., *Nat. Commun.*, **2021**, *12*, 2573.

Synthesis actions & synthesis automation

Background: 7/7

OUTLINE

1. IBM RXN research
2. IBM RXN platform
3. Data-driven retrosynthetic models in practice
4. Disconnection aware retrosynthesis



Freely available platform:
`rxn.res.ibm.com`

API wrapper on GitHub:
github.com/rxn4chemistry/rxn4chemistry

Launcher | rxn4chemistry_tour.ipynb

Markdown | Python 3 (ipykernel)

Running a reaction prediction is as simple as:

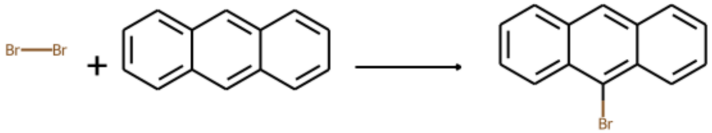
[5]:
```
predict_reaction_response = rxn4chemistry_wrapper.predict_reaction(
    'BrBr.c1ccc2cc3ccccc3cc2c1'
)
```

NOTE: we have set limitations on the number of calls per second and per minute in the public version of RXN for Chemistry. These limits can be tweaked or removed in on-premise deployments. Those limitations are currently set to 5 calls per minute, in most cases this is not a problematic limitation.

[6]:
```
predict_reaction_results = rxn4chemistry_wrapper.get_predict_reaction_results(
    predict_reaction_response['prediction_id']
)
```

[7]:
```
get_reaction_from_smiles(predict_reaction_results['response']['payload']['attempts'][0]['smiles'])
```

[7]:



It is possible to run reaction prediction in batches (not storing the information in any project) to use the service in a highthroughput fashion:

[8]:
```
predict_rection_batch_response = rxn4chemistry_wrapper.predict_reaction_batch(
```

OUTLINE

1. IBM RXN research
2. IBM RXN platform
3. Data-driven retrosynthetic models in practice
4. Disconnection aware retrosynthesis

Retrosynthesis: numerous data-driven models!

TABLE 1 Different retrosynthetic prediction approaches

Input (product)	Output	Single-step approach	Multi-step algorithm
Sequence-based			
SMILES	Reactants SMILES	Seq-2-seq LSTM	None: 56
SMILES	Largest reactant SMILES	Transformer	None: 83
SMILES	Reactants SMILES	Transformer	None: 83,95–99, MCTS: 112
SMILES	1: Synthons prediction, 2: Synthons completion	Transformer	None: 105
MACCS keys	Reactants MACCS keys	Transformer	None: 100
SMILES	Precursors SMILES	Transformer	Beam search: 28
Fingerprint-based			
Fingerprint	Reaction template	Similarity	None: 93
Fingerprint	Reaction template	Feed-forward NN	None: 78,106,107, MCTS: 23,29,109,110, RL: 114, A*: 115,116
Fingerprint	Reaction template	Modern Hopfield network	None: 91
Graph-based			
Molecular graph	Reaction template	Graph neural network	None: 52,101, SMC: 117
Molecular graph	1: Synthons prediction, 2: Synthons completion	Graph neural network	None: 102,103,118
Molecular graph	Sequence of graph-edits	Graph neural network	None: 30

Only a selection!

Schwaller, P. et al., Machine intelligence for chemical reaction space, *WIREs Comput. Mol. Sci.*, **2022**, e1604.

Retrosynthesis: recent models

PAPER • OPEN ACCESS

Chemformer: a pre-trained transformer for computational chemistry

Ross Irwin[1], Spyridon Dimitriadis[1,2], Jiazhen He[1] and Esben Jannik Bjerrum[3,1]

arXiv > physics >

Physics > Chemical P

[Submitted on 29 Jan 2022]

Retrofor
Retrosyn

Yue Wan, Be

AI-Driven Synthetic Route Design Incorporated with Retrosynthesis Knowledge

Shoichi Ishida, Kei Terayama, Ryosuke Kojima, Kiyosei Takasu, and Yasushi Okuno*

Share Add to Export

RIS

Research article | Open Access | Published: 15 March 2022

Improving the performance of models for one-step retrosynthesis through re-ranking

Min Htoo Lin, Zhengkai Tu & Connor W. Coley

Journal of Cheminformatics **14**, Article number: 15 (2022) | Cite this article

269 Accesses | **9** Altmetric | Metrics

Retrosynthesis: metrics?

PAPER • OPEN ACCESS

Chemformer: a pre-trained transformer for computational chemistry

Ross Irwin[1], Spyridon Dimitriadis[1,2], Jiazhen He[1] and Esben Jannik Bjerrum[3,1]

arXiv > physics >

Physics > Chemical P

[Submitted on 29 Jan 2022]

Retrofor
Retrosyn

Yue Wan, Be

AI-Driven Synthetic Route Design Incorporated with Retrosynthesis Knowledge

Shoichi Ishida, Kei T a, Ryosuke Kojima, Kiyoshi Takas , and Yasushi Okuno*

Share Add to Export

RIS

Metrics?

Research article | Open Access | Published: 15 March 2022

Improving the performance of models for one-step retrosynthesis through re-ranking

Min Htoo Lin, Zhengkai Tu & Connor W. Coley

Journal of Cheminformatics **14**, Article number: 15 (2022) | Cite this article

269 Accesses | **9** Altmetric | Metrics

Retrosynthesis: metrics?

“On direct synthesis and retrosynthesis prediction benchmark datasets we publish state-of-the-art results for top-1 accuracy.”

“We adopt the conventional top-k accuracy of the full reactants to evaluate the retrosynthesis performance.”

“Typically, these data-driven methods are evaluated in terms of top-*N* accuracy”

“We acknowledge that top-*N* accuracy alone does not paint a complete picture of a one-step model’s performance, as others have also argued.”

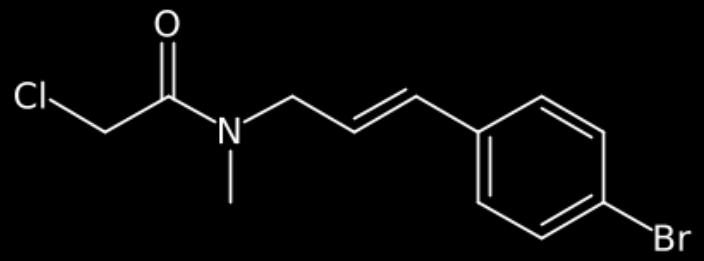
Retrosynthesis: metrics?

- top-N accuracy is not fully satisfactory – a necessary evil
- Multi-step: hard to assess as well!
  - Top-N
  - Turing test
  - Percentage of solved molecules

- What do the chemists need?

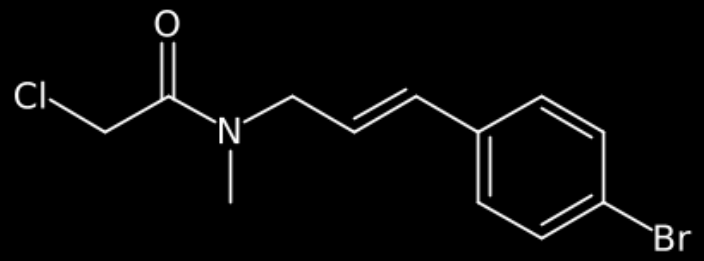
Our experience with IBM RXN

- DEMO: `CN(C/C=C/c1ccc(Br)cc1)C(=O)CCl`



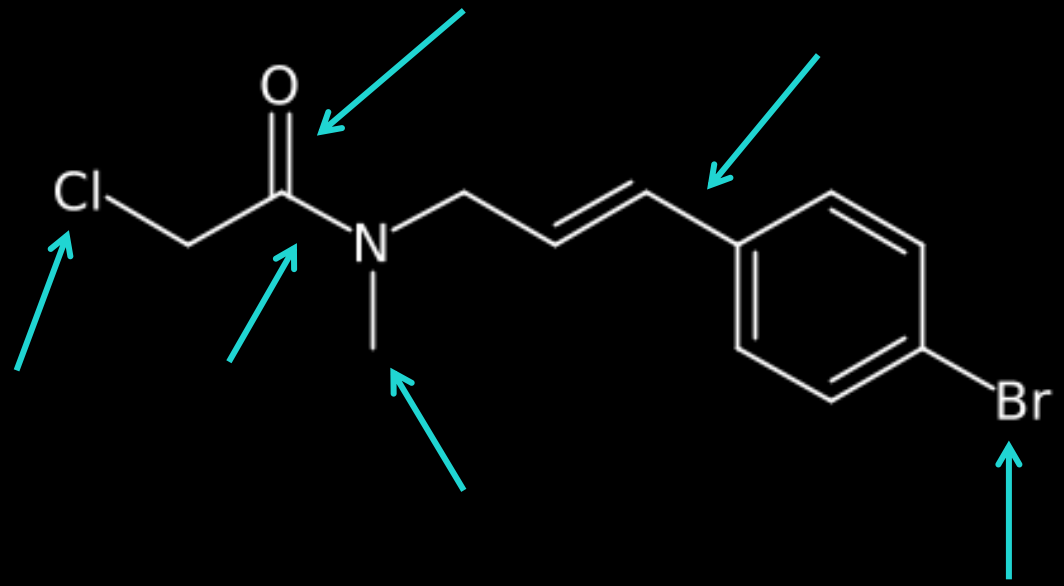
Our experience with IBM RXN

- Many chemists prefer “interactive” mode
- DEMO: `CN(C/C=C/c1ccc(Br)cc1)C(=O)CCl`



Even more interactive control?

- Let the chemists decide where to break the compound?



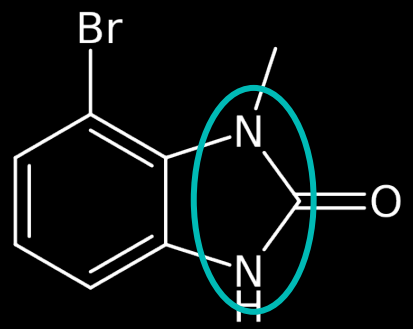
OUTLINE

1. IBM RXN research
2. IBM RXN platform
3. Data-driven retrosynthetic models in practice
4. Disconnection aware retrosynthesis

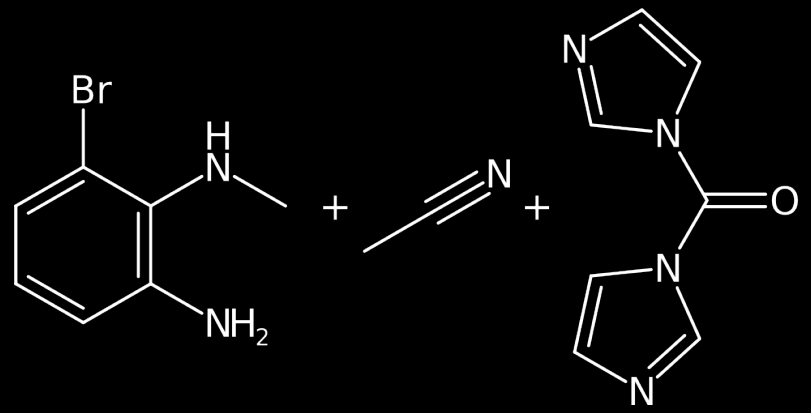
Disconnection-aware retrosynthesis

Goal:

Input (target compound)



Output (precursors)

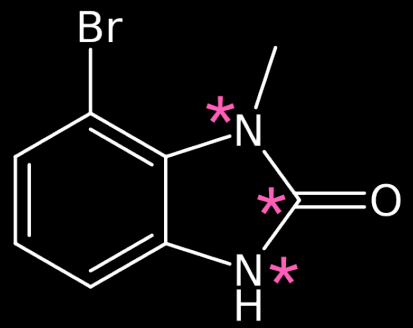


Machine learning model

Inspired by the **Molecular Transformer**

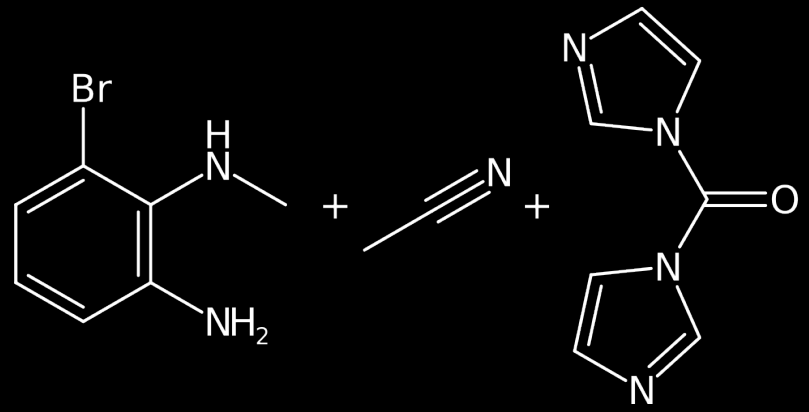
Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C. & Lee, A. A., *ACS Cent. Sci.*, **2019**, *5*, 1572-1583.

Input



“Translation”

Output



`Cn1c2c(Br)cccc2[nH]c1=O`

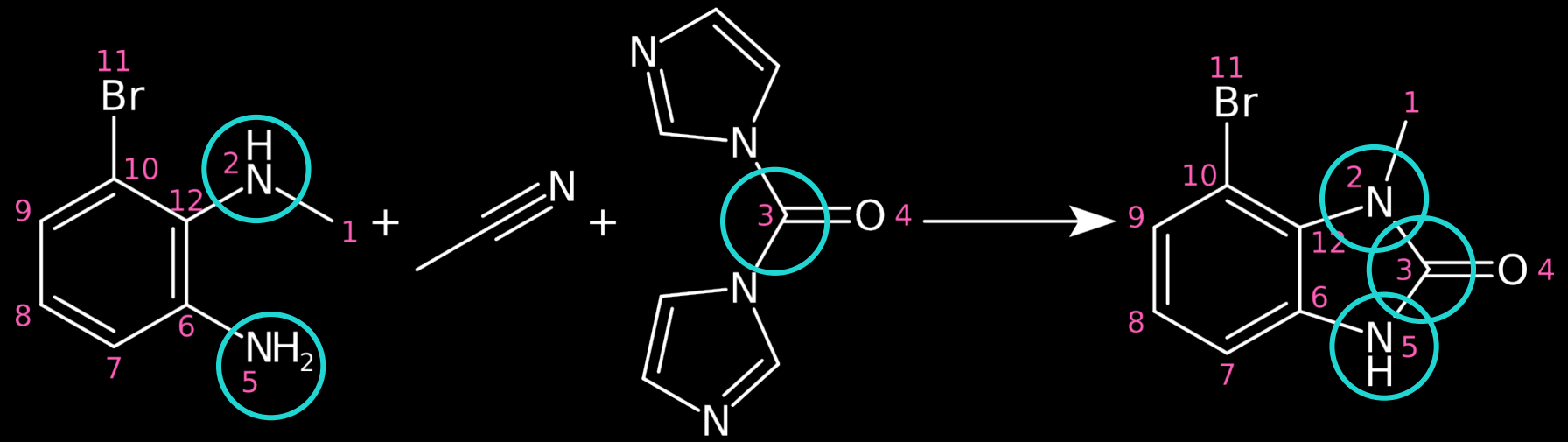
`C[n:1]1c2c(Br)cccc2[nH:1][c:1]1=O`

`CNc1c(N)cccc1Br.CC#N.O=C(n1ccnc1)n1ccnc1`

*Not shown here for readability: the Model uses tokenized SMILES strings: “`C[n:1]1c2c(Br)…`” → “`C [n:1] 1 c 2 c ( Br ) …`”

Dataset generation

- Start from atom-mapped reaction:

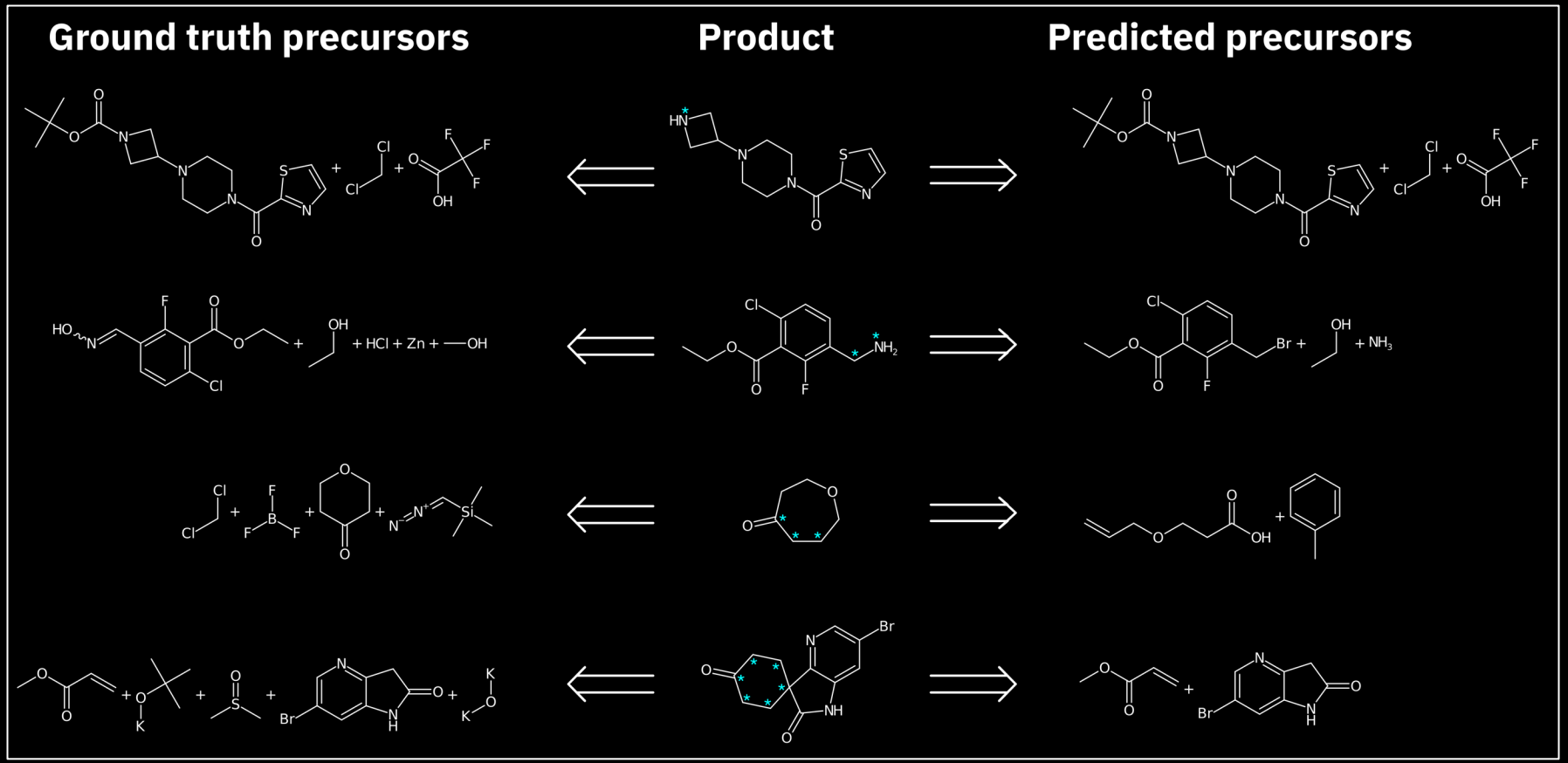


Data and model

- >2M patent reactions from Pistachio [1]
- Atom-mapped with RXNMapper [2]
- Training, validation, and test sets of sizes **2.27M**, **10.0k**, and **126k**.
- Transformer-based seq-2-seq model implemented with OpenNMT [3]

[1] Nextmove Software Pistachio, http://www.nextmovesoftware.com/pistachio (Accessed Sep 23, 2021).
[2] Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H. & Laino, T., *Sci. Adv.*, **2021**, *7*, eabe4166..
[3] Klein, G. et al, OpenNMT: Open-Source Toolkit for Neural Machine Translation. ACL 2017.

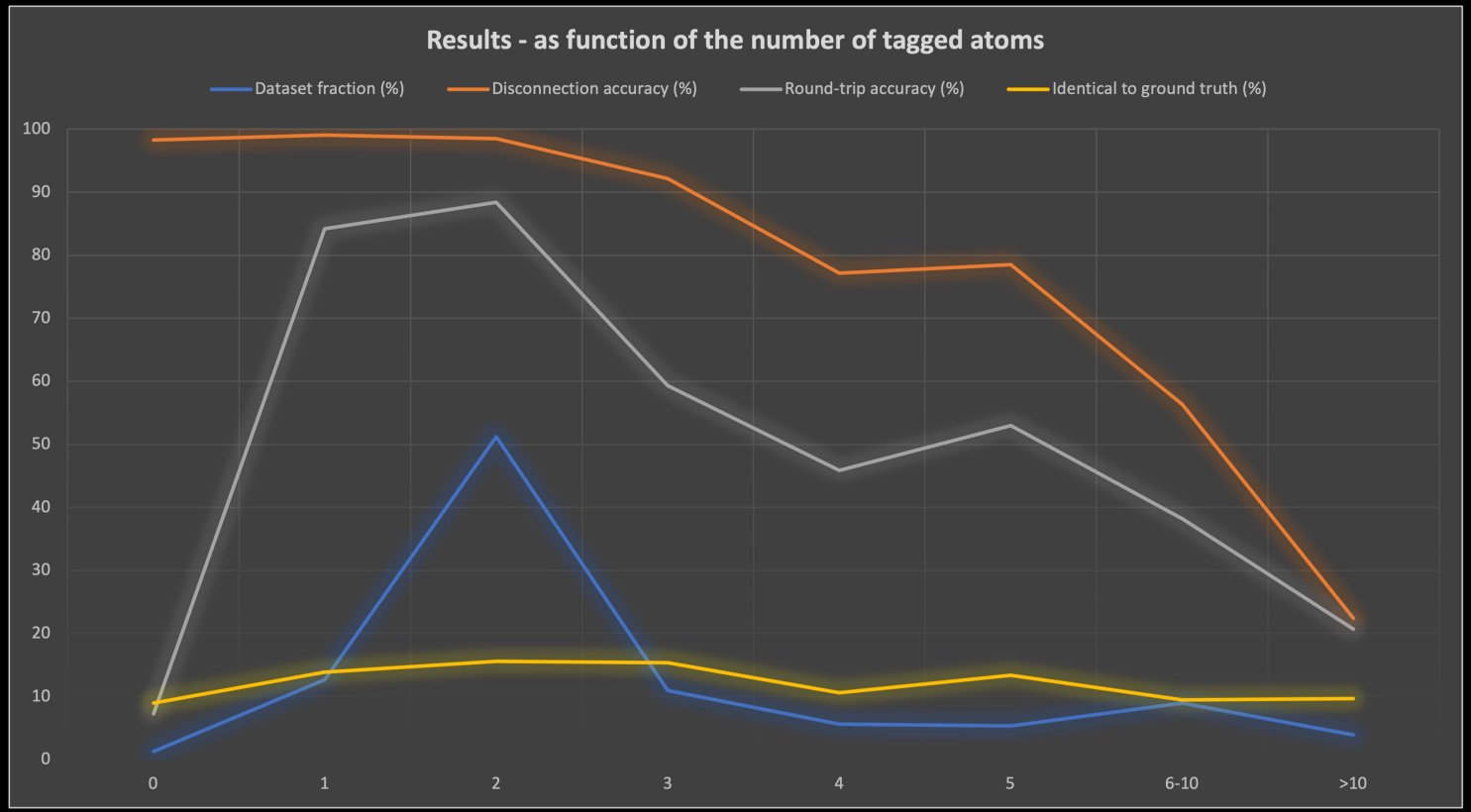
Results: examples



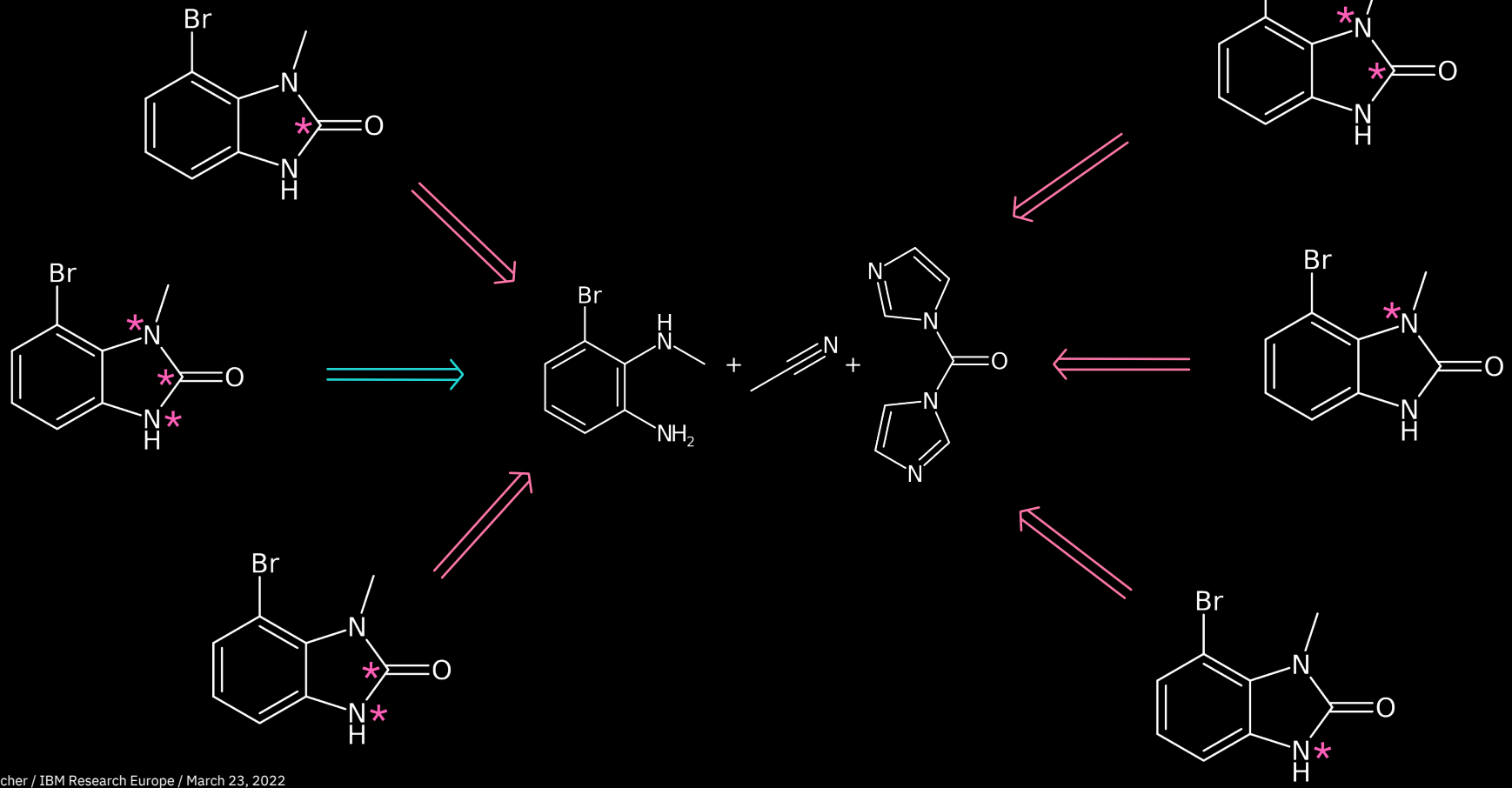
Results

Number of tagged atoms	Dataset fraction (%)	Disconnection accuracy (%)	Round-trip accuracy (%)	Identical to ground truth (%)
0	1.29	98.3	7.3	9.0
1	12.65	99.1	84.2	13.9
2	51.22	98.5	88.4	15.6
3	10.99	92.2	59.3	15.4
4	5.64	77.2	45.9	10.6
5	5.34	78.5	53.0	13.4
6-10	8.96	56.3	38.2	9.5
>10	3.91	22.4	20.7	9.7
Overall	**100.0**	**88.9**	**76.0**	**14.1**

Results


Results - as function of the number of tagged atoms
Dataset fraction (%)
Disconnection accuracy (%)
Round-trip accuracy (%)
Identical to ground truth (%)
100
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6-10
>10

Next steps



Thank you for your attention!

Questions or comments

E-mail: ava@zurich.ibm.com

Twitter: @acvaucher

Preprint with initial results: ibm.biz/disconnection-aware-retro