Molecular Transformer-aided Biocatalysed Synthesis Planning

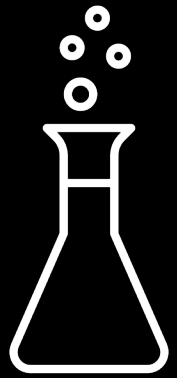
Alain Vaucher
15 August 2022

IBM **Research**

Introduction

Data

Chemistry



Artificial intelligence

(Bio)catalysis

Automation

Outline

1. AI models for chemistry
2. Synthesis automation – RoboRXN
3. AI for biocatalysis

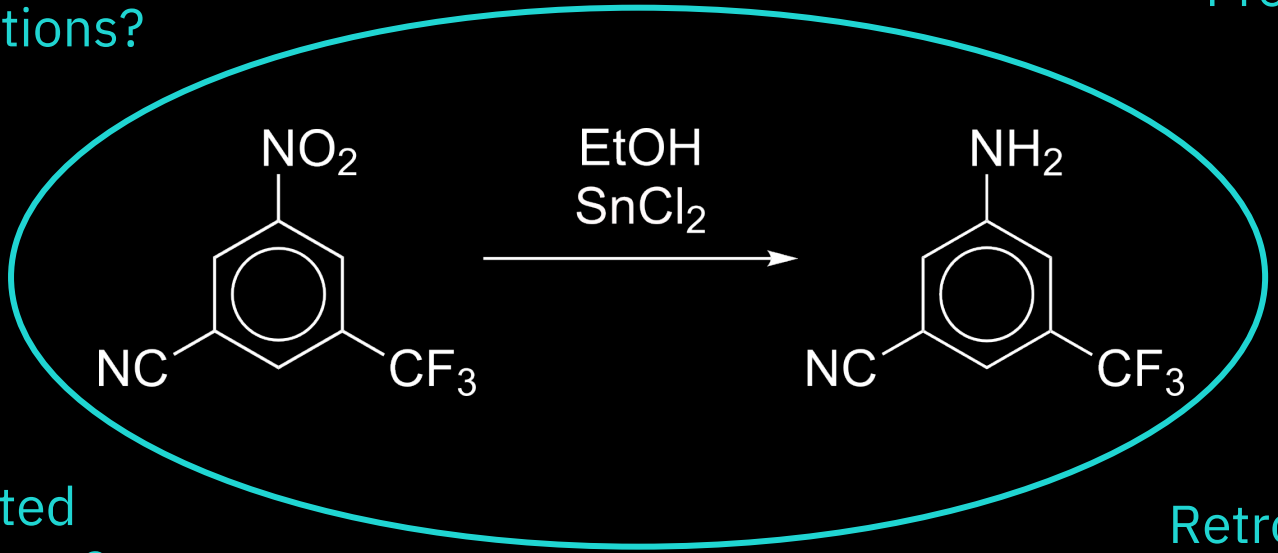
Outline

1. AI models for chemistry
2. Synthesis automation – RoboRXN
3. AI for biocatalysis

AI and chemical reactivity

Experimental conditions?

Product?



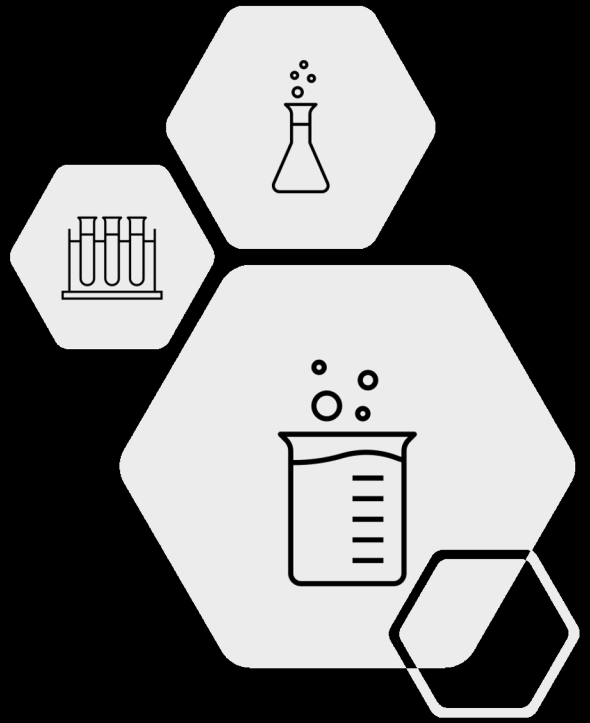
Related reactions?

Retrosynthesis?

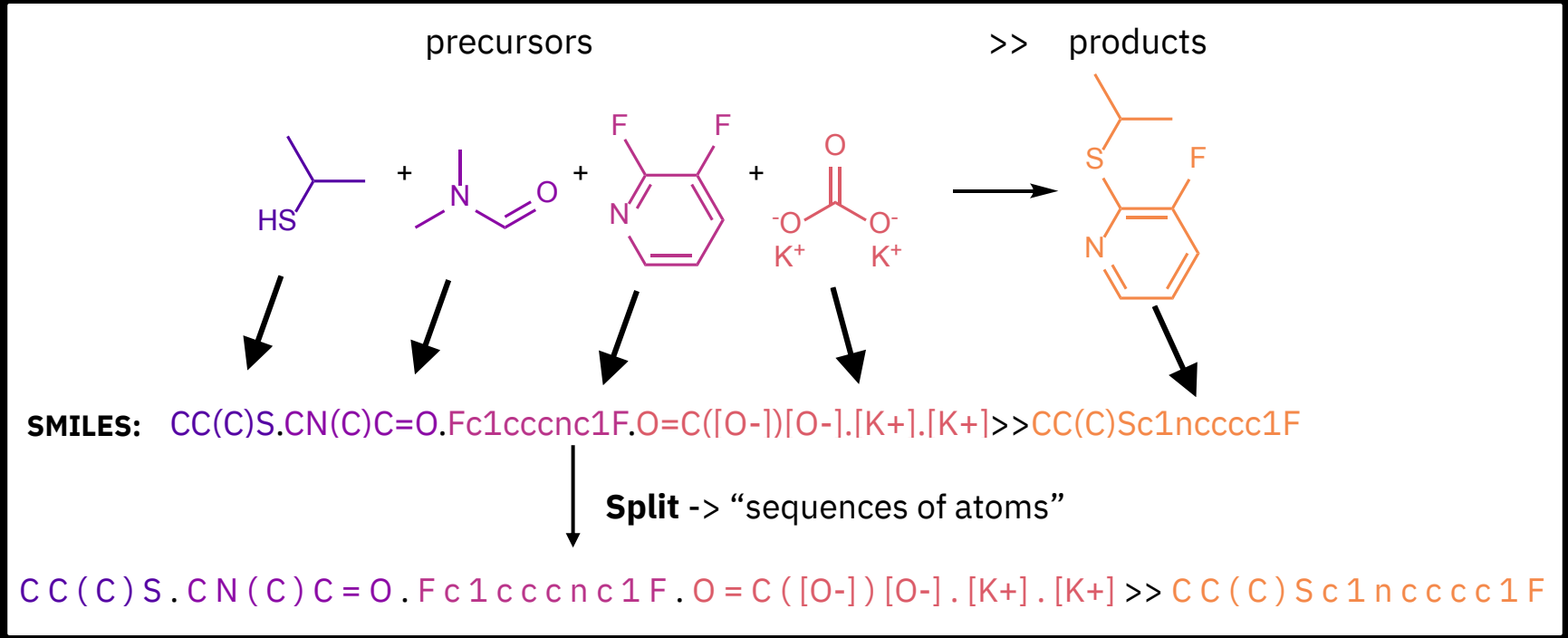
Yield?

Data sources

- Millions of reactions have been reported
- Sources:
  - Publicly available data: patents (USPTO, NextMove's Pistachio, etc.)
  - Scientific publications
  - Proprietary reactions (industry)
  - Publishers
  - Etc.

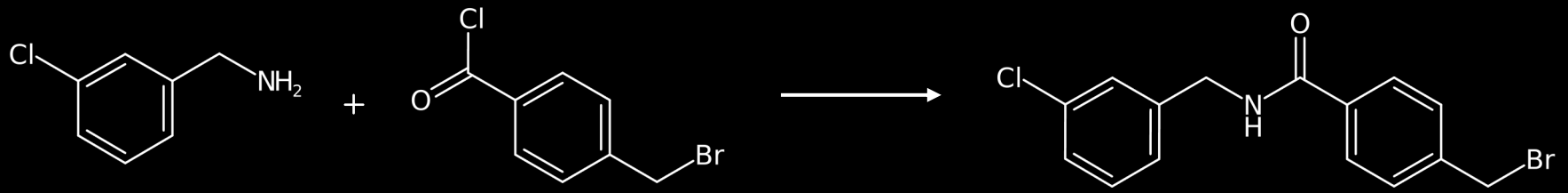


Atoms as letters, molecules as words



→ Borrow methods developed for human languages

Forward reaction prediction



Textual representation (SMILES)

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

“Sentence of atoms”

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

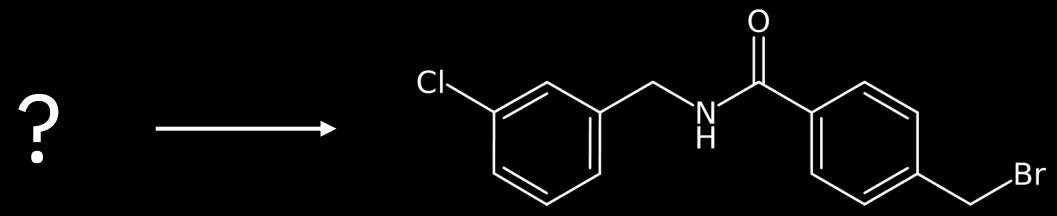
“Translation”

Transformer

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

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

Retrosynthesis

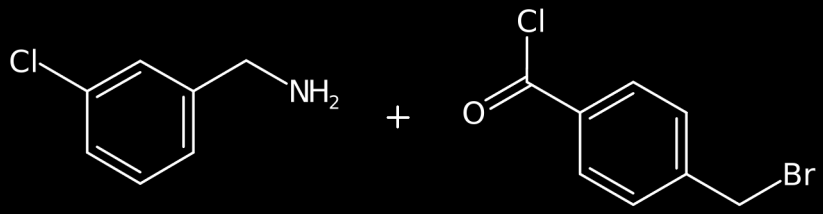


Similar approach, both sides switched

“Translation”

`O=C(NCc1ccccc(Cl)c1)c1ccc(CBr)cc1` ⟹ `NCc1ccccc(Cl)c1.O=C(Cl)c1ccc(CBr)cc1`

Transformer

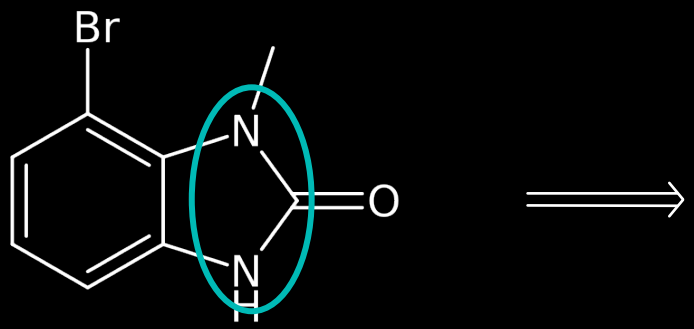


Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A. & Laino, T., *Chem. Sci.*, **2020**, *11*, 3316-3325.

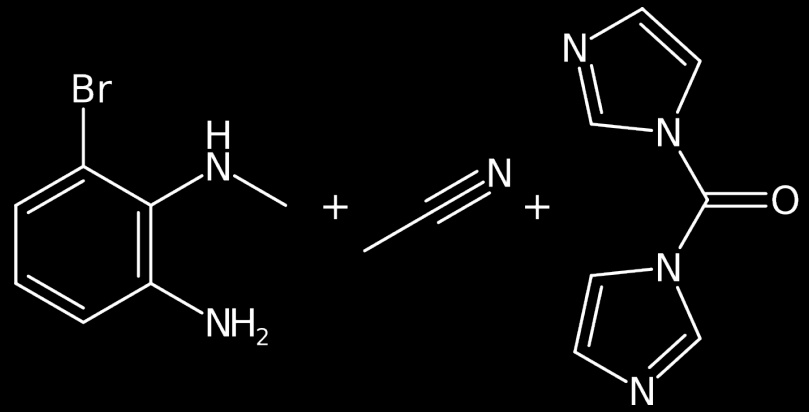
Disconnection-aware retrosynthesis

- Let the chemists decide where to break the compound?

Input (target compound)

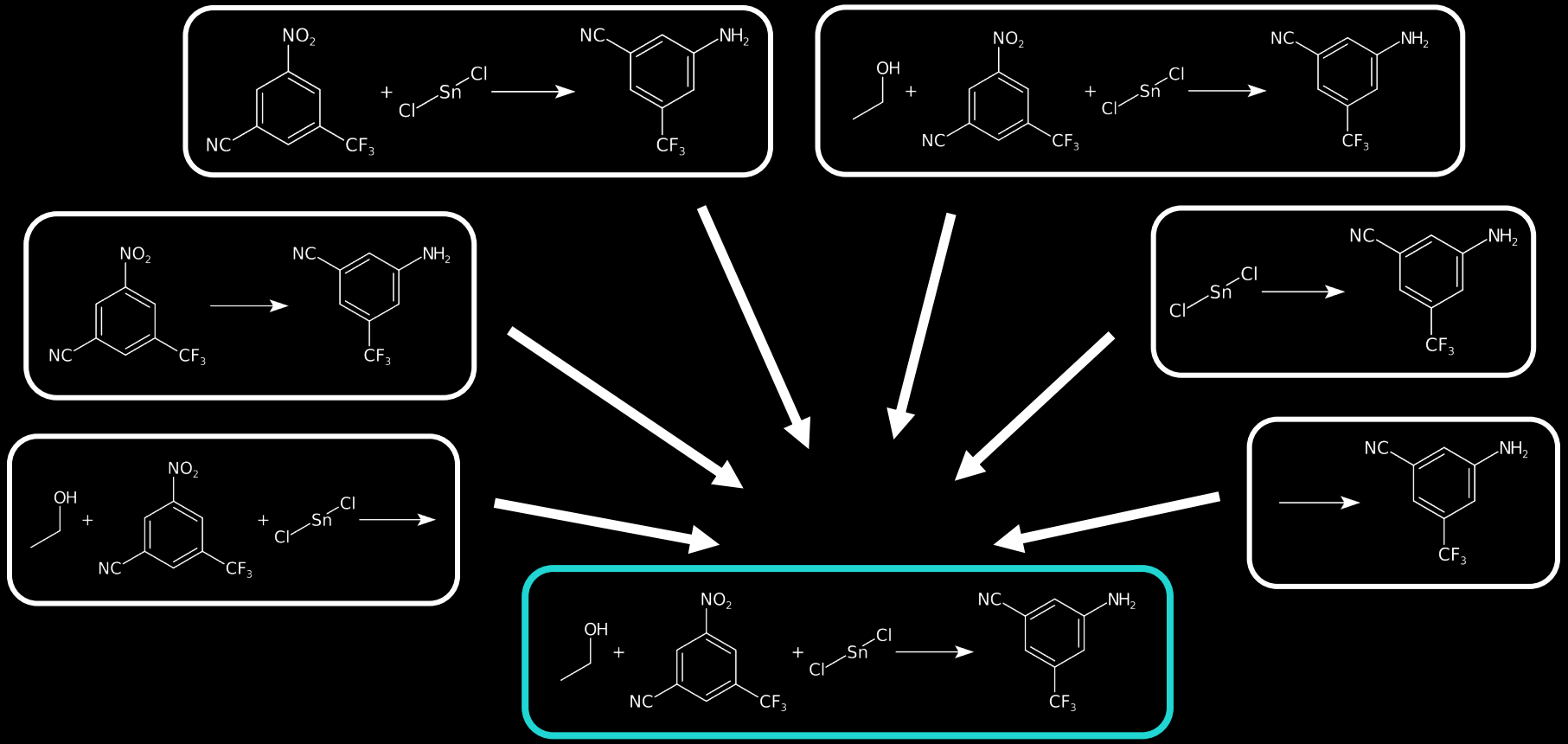


Output (precursors)



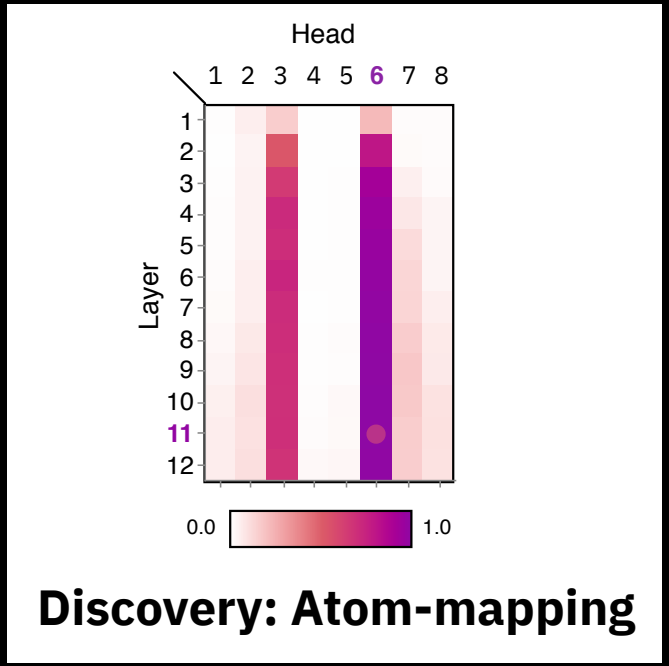
Byekwaso, A.; Schwaller, P.; Vaucher, A. C.; Toniato, A.; Laino, T.; *"AI for Science" workshop @ NeurIPS 2021.*

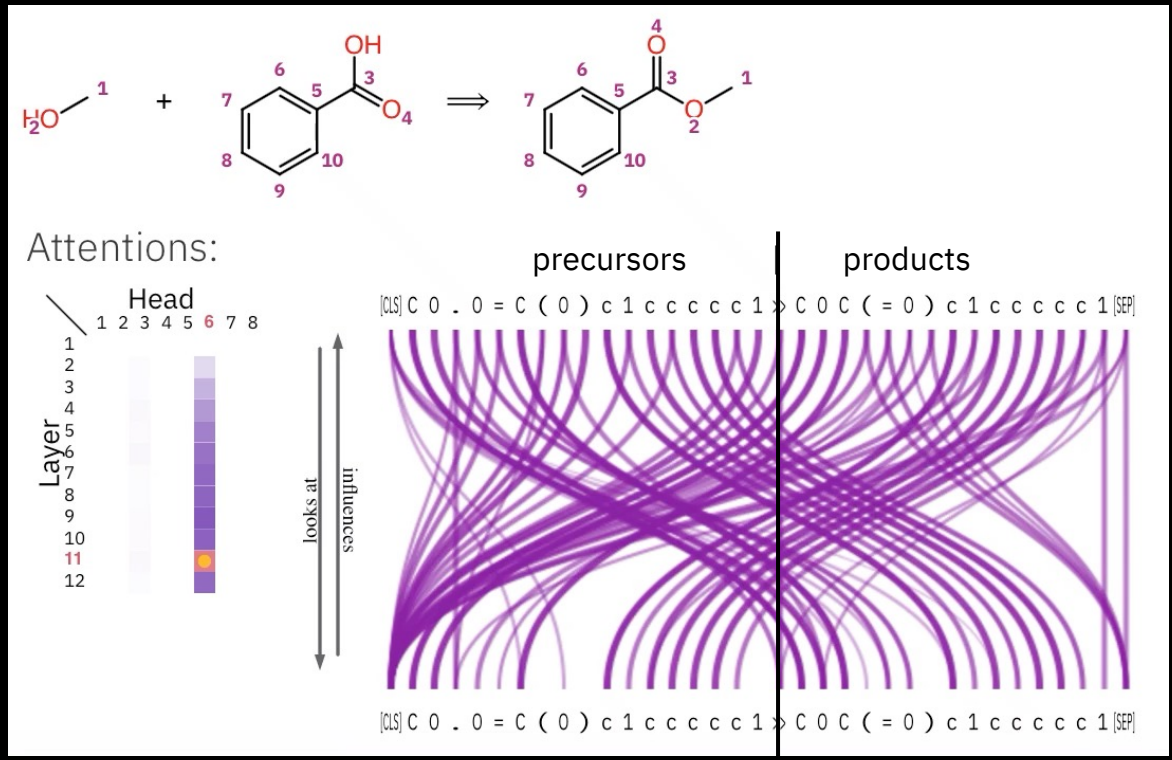
Completing partial chemical equations



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

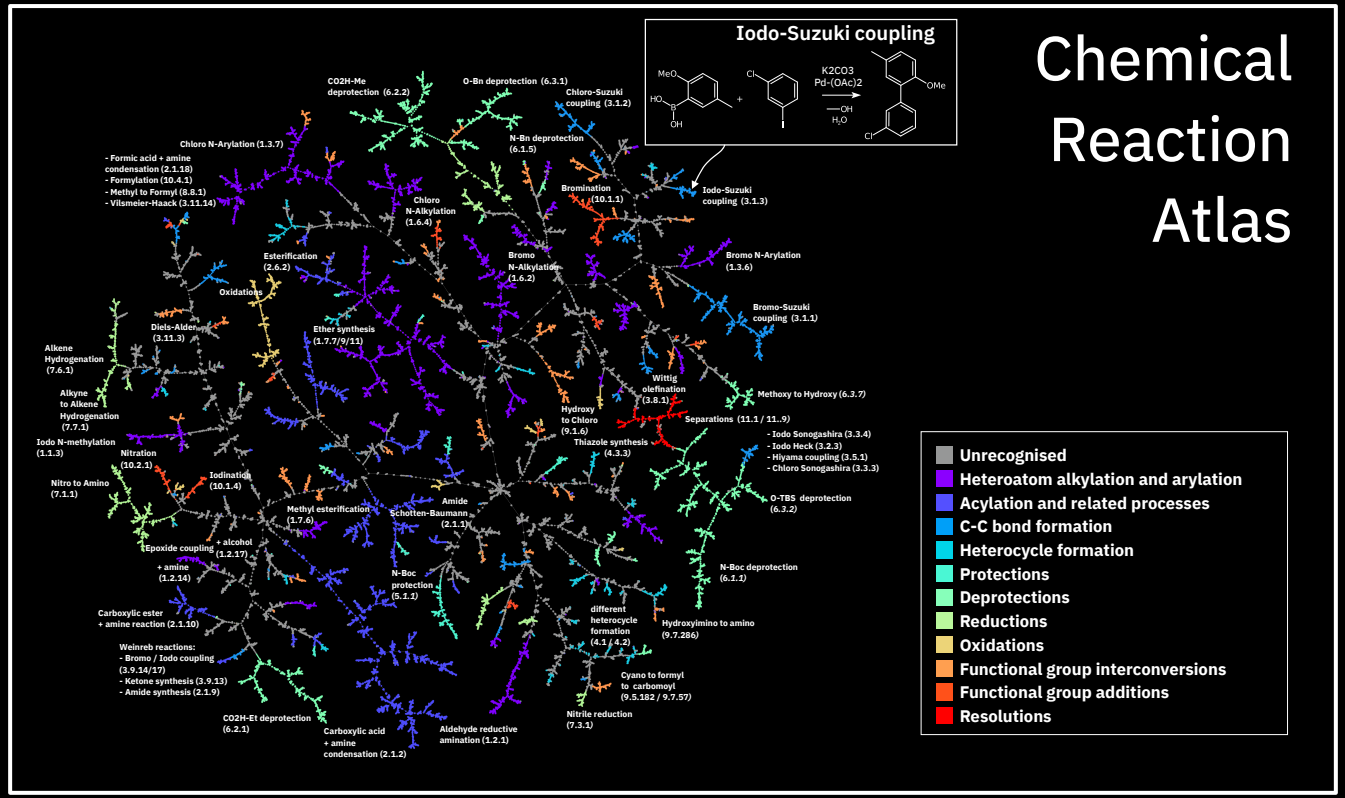
Atom-Mapping and the learning of chemical reaction grammar





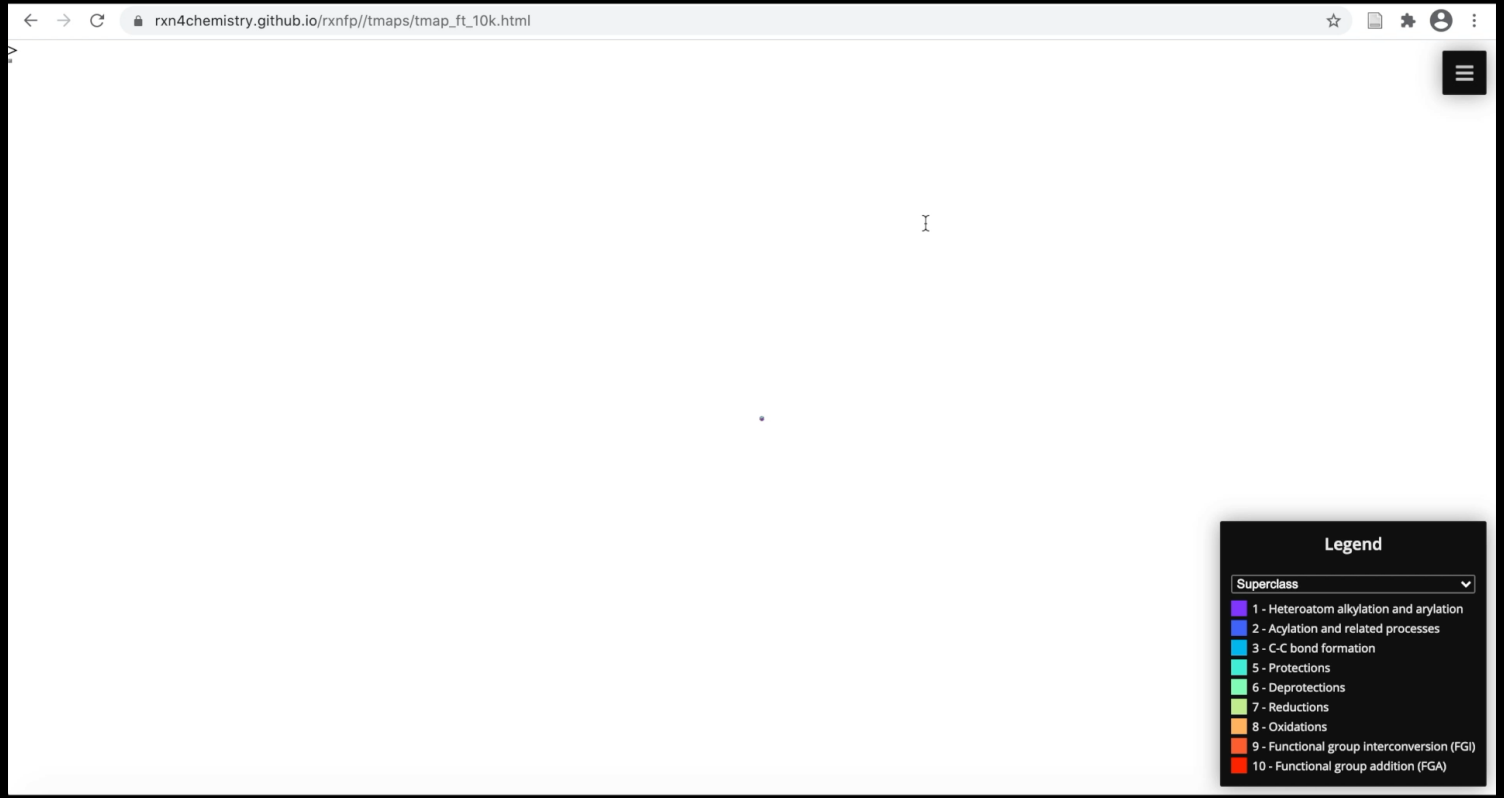
Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A. & Laino, T., *Chem. Sci.*, **2020**, *11*, 3316-3325.

Mapping and classifying chemical reactions



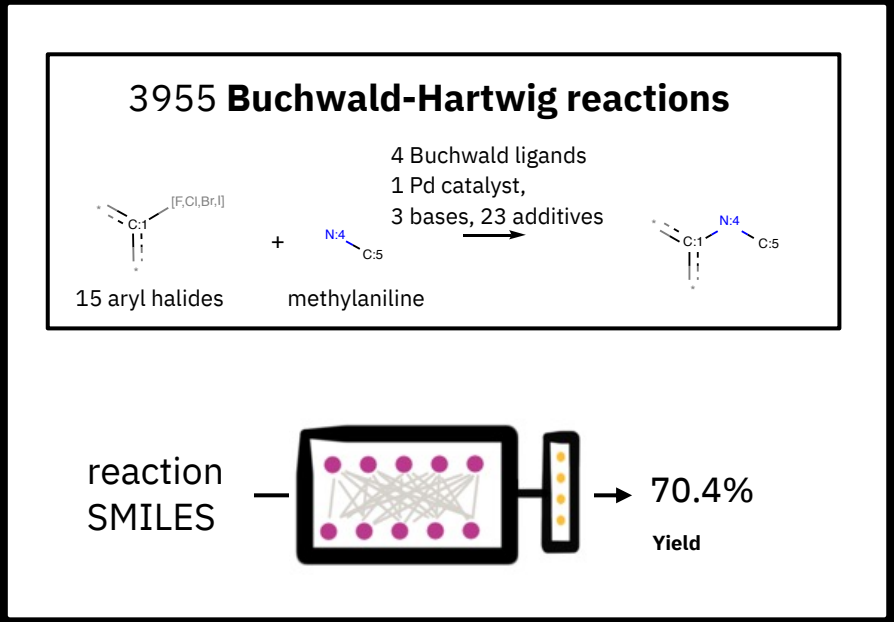
Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T. & Reymond, J.-L., *Nat. Mach. Intell.*, **2021**, *3*, 144-152.

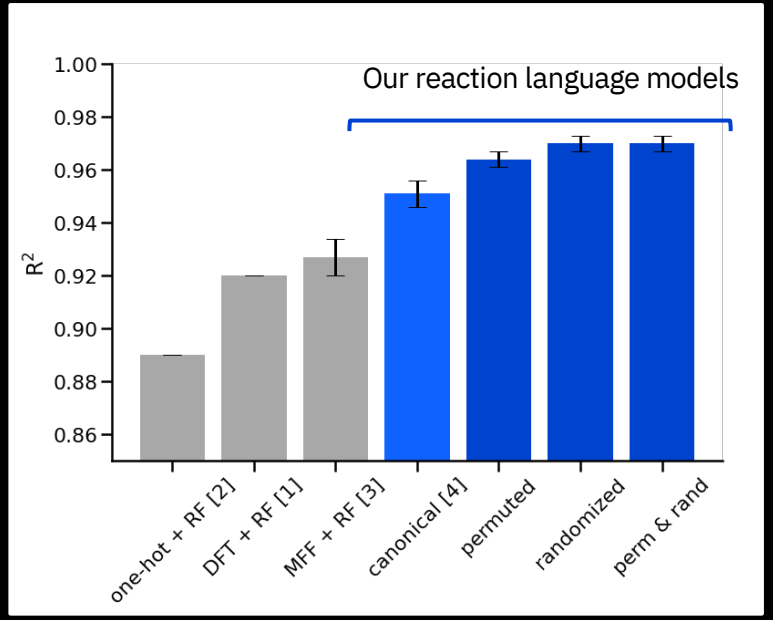
Mapping and classifying chemical reactions



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.
Schwaller, P.; Vaucher, A. C.; Laino, T. & Reymond, J.-L., *Chemrxiv.13286741*, **2020**.

Outline

1. AI models for chemistry
2. Synthesis automation – RoboRXN
3. AI for biocatalysis

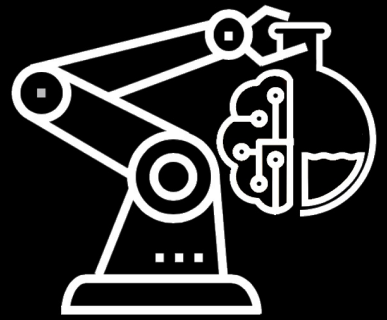
Data and chemical reactions

- Chemists have been doing reactions in roughly the same way for **decades**
- Set of **standard lab operations**
- **Millions** of reactions reported in the literature

How can exploit this **data** to **accelerate discovery**?

- Assist chemists in synthesis planning
- … and run the syntheses for them!

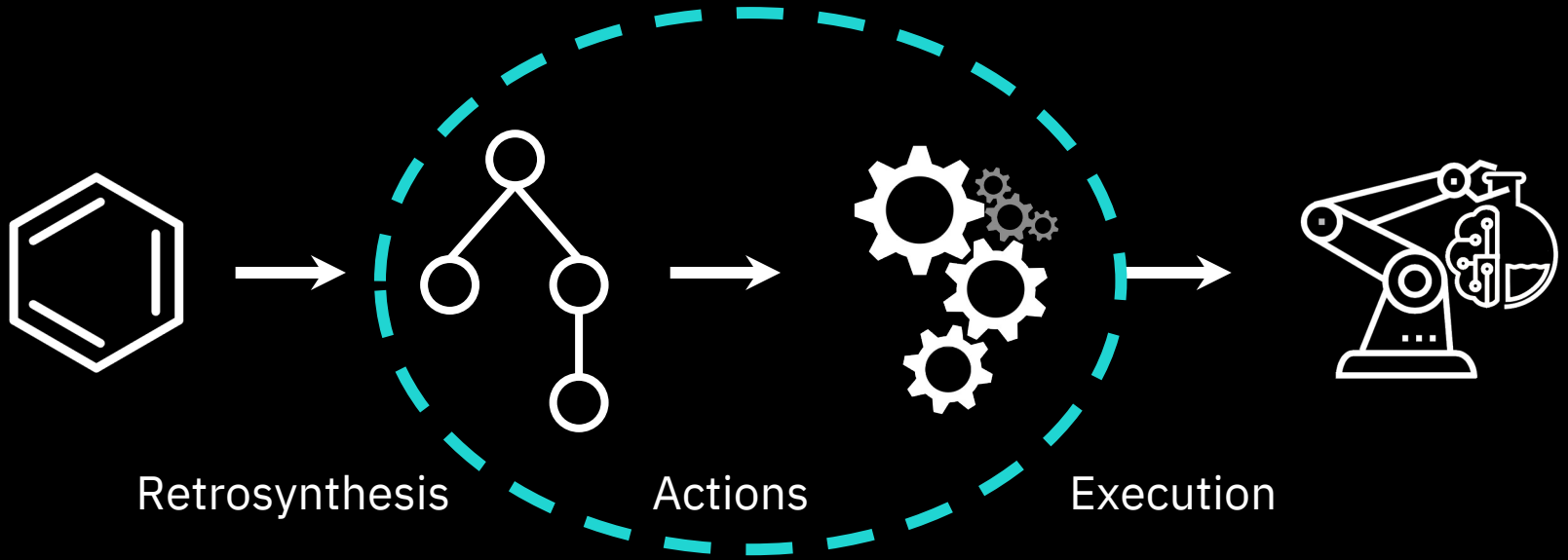
Automated execution on robot



Target molecule

Synthesis execution

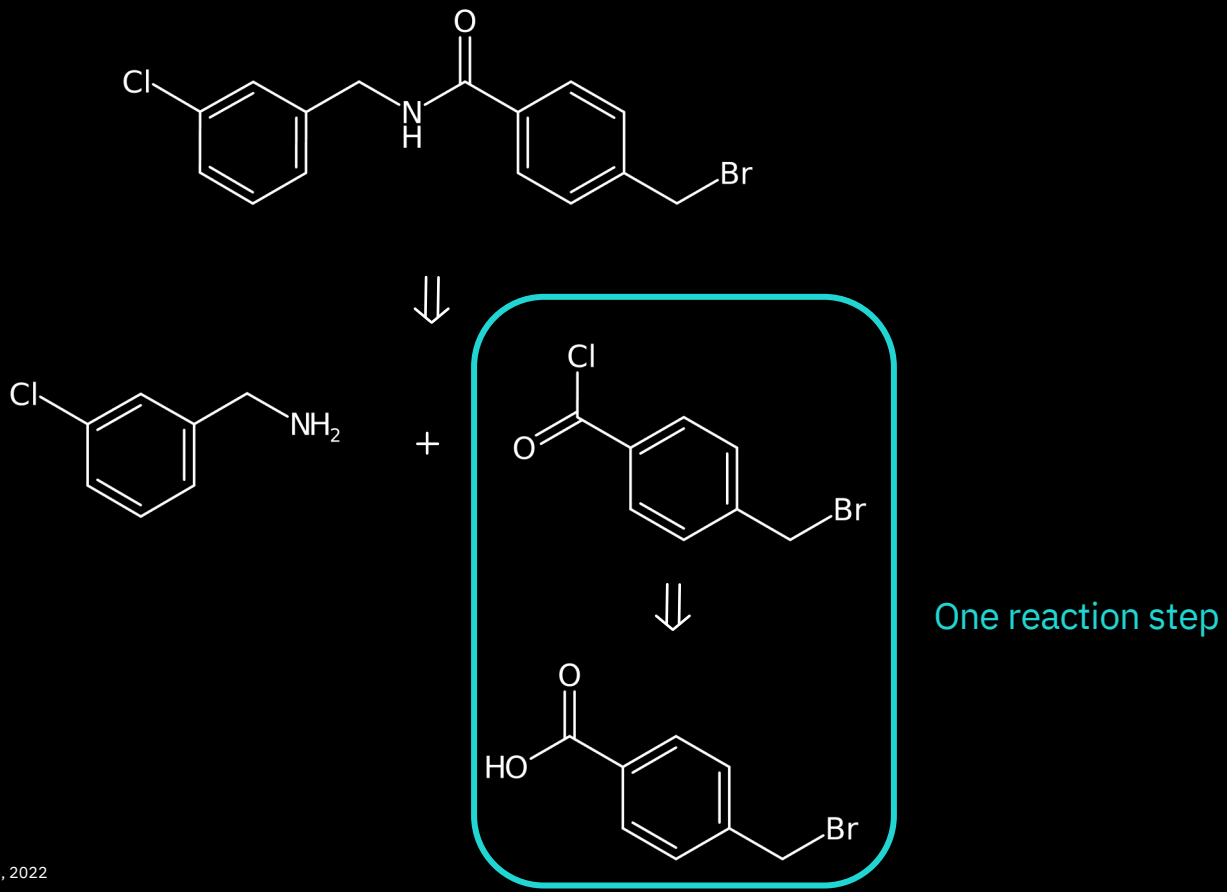
Automated execution on robot



Outline

1. AI models for chemistry
2. Synthesis automation – RoboRXN
    A. Prediction of synthesis actions
    B. Execution on robot
3. AI for biocatalysis

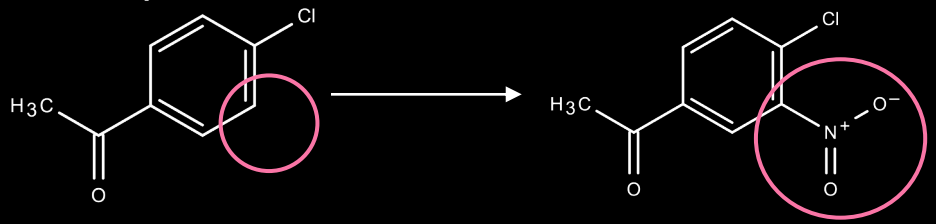
Synthesis actions



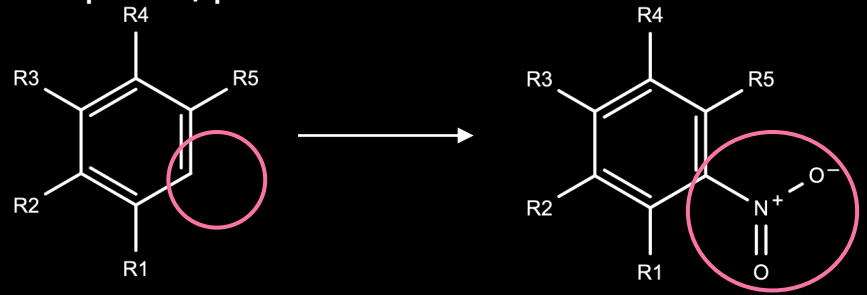
One reaction step

Synthesis actions

Example:

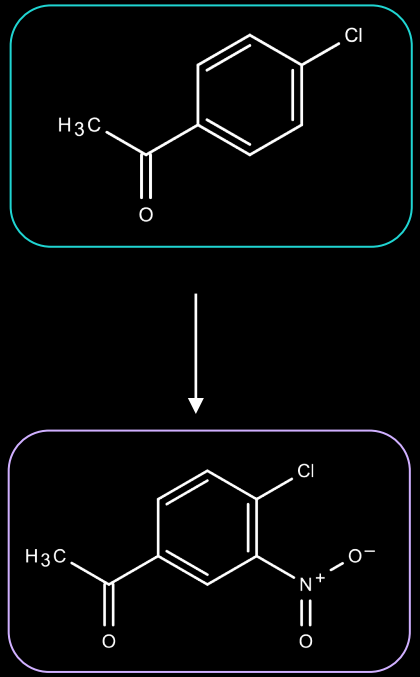


Template/pattern:



- Same template but different synthesis actions!
- Hard to predict
- Ideally: ML model!

Synthesis actions



Operation 1

Operation 2

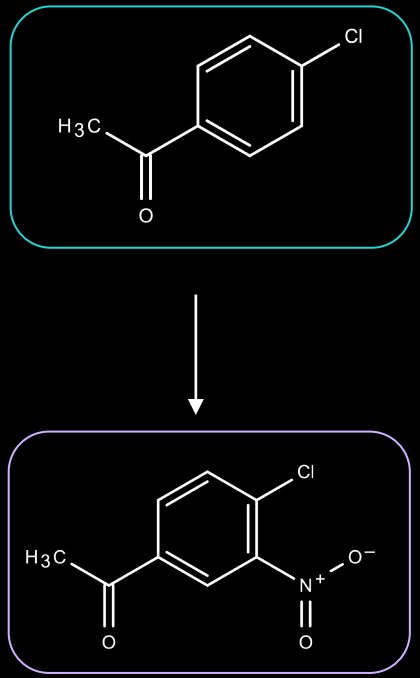
Operation 3

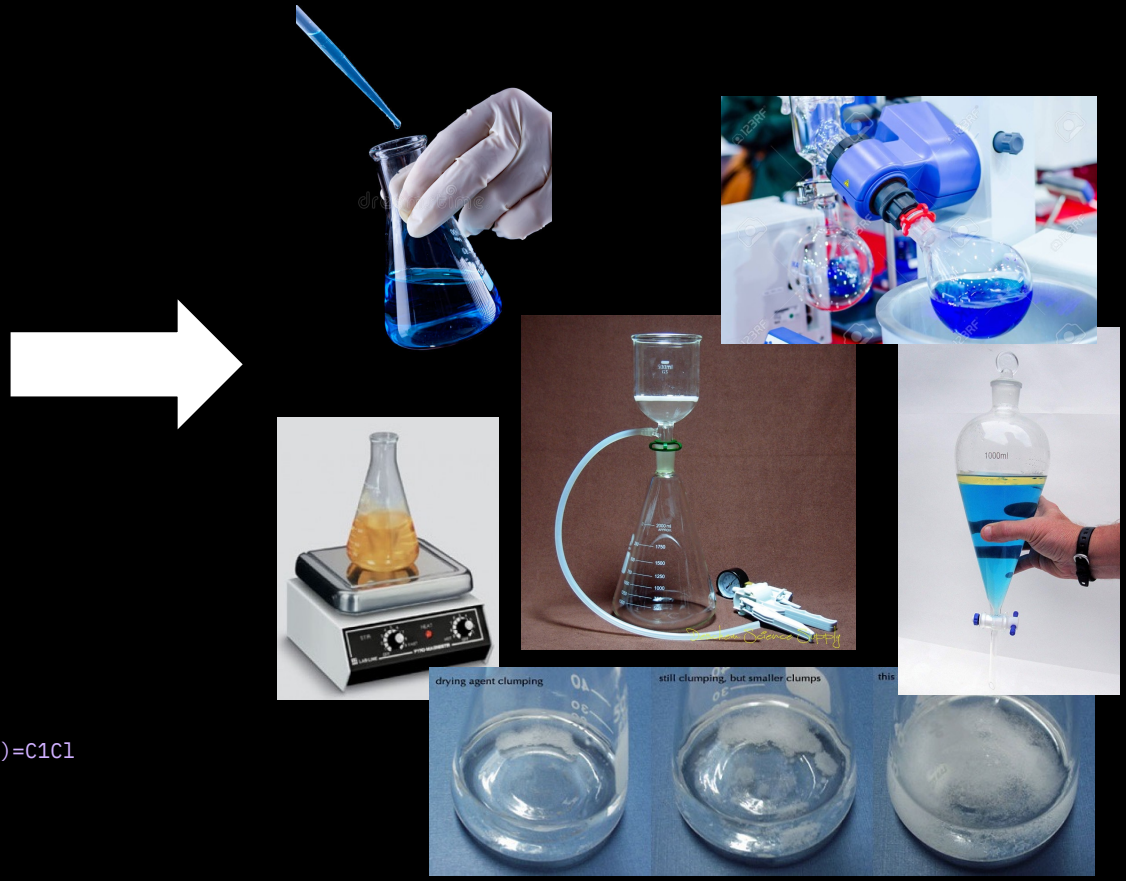
Operation 4

…

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

Synthesis actions





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

Synthesis actions

– No dataset!

– Information is available **indirectly**

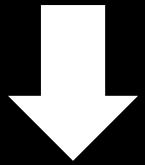
– First: extract actions from text

Example procedure from a patent

A mixture of 1-(4-isopropyl-phenyl)-5-oxo-pyrrolidine-3-carboxylic acid ethyl ester obtained in step 2 (0.7 g, 2.65 mmol) and ethanol were cooled to 10-15° C. Sodium borohydride (0.25 g, 6.6 mmol) was added portion wise over a period of 20 min and the reaction mixture was stirred for 3.5 hrs at 20-25° C. The organic volatiles were evaporated and the residue was taken into brine solution (15 ml). The aqueous layer was extracted with ethyl acetate, dried over Na2SO4 and evaporated to obtain 4-hydroxymethyl-1-(4-isopropyl-phenyl)-pyrrolidin-2-one as an off white solid (0.5 g, 81%).

Models for Paragraph-to-actions

… Sodium borohydride (0.25 g, 6.6 mmol) was added portion wise over a period of 20 min and the reaction mixture was stirred for 3.5 hrs at 20-25° C …



```
Add(name='Sodium borohydride',
    quantity=['0.25 g', '6.6 mmol'],
    duration='20 min')

Stir(temperature='20-25°C',
     duration='3.5 hrs')
```

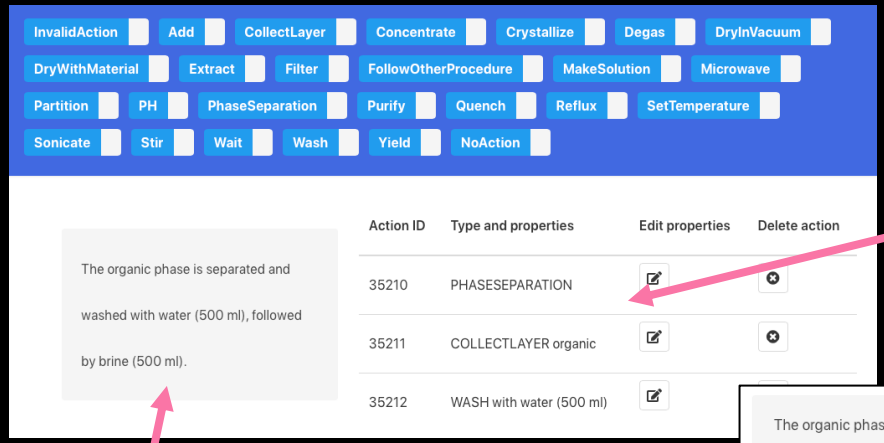
What kind of model?

–Rule-based model?

–Fuly data-driven model?

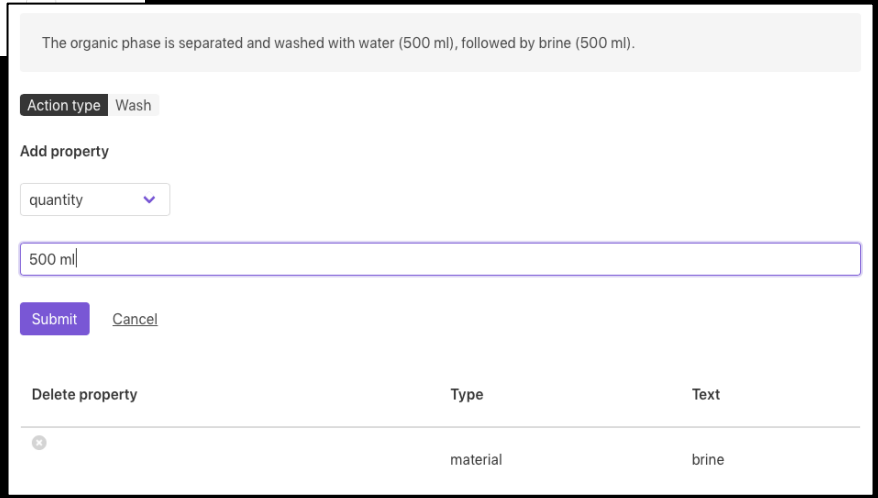
Both!

Hand-annotated data



Initial actions from rule-based model

Sentence to annotate



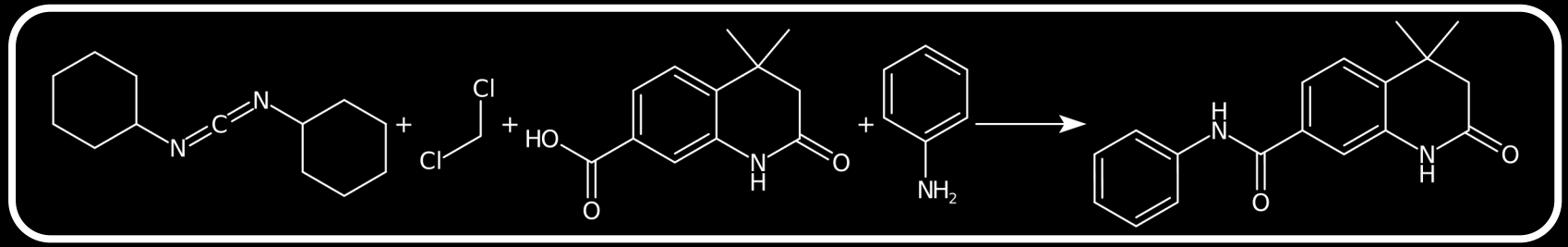
>1700 annotated sentences

Results

Model	100% accuracy
Combined rule-based model	21.9
Pretrained translation model	24.7
Model without pretraining	37.8
Refined translation model	**60.8**

Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T., *Nat. Commun.* **2020**, *11*, 3601.

SMILES-to-actions



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

2.7 g (12.3 mmol) 4,4-Dimethyl-1,2,3,4-tetrahydro-2-oxo-7-quinolinecarboxylic acid were added to a solution of 3.8 g (18.5 mmol) N,N'-dicyclohexylcarbodiimide and 1.1 ml (12.3 mmol) aniline in 80 ml dichloromethane. The reaction mixture was stirred for 4 hours at ambient temperature and the precipitate was filtered off with suction and recrystallised from ethanol. There was obtained 1.2 g of the title compound; m.p. 249-251° C.

ML model

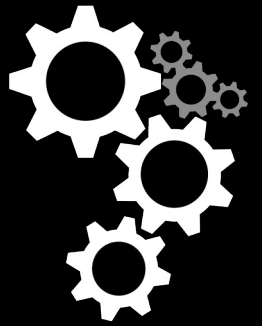
1. MAKESOLUTION with N,N'-dicyclohexylcarbodiimide (3.8 g, 18.5 mmol) and aniline (1.1 ml, 12.3 mmol) and dichloromethane (80 ml)
2. [illegible]
3. ADD 4,4-Dimethyl-1,2,3,4-tetrahydro-2-oxo-7-quinoline[illegible] (2.7 g, 12.3 mmol)
4. STIR for 4 hours at ambient temperature
5. FILTER keep precipitate
6. RECRYSTALLIZE from ethanol
7. YIELD title compound (1.2 g)

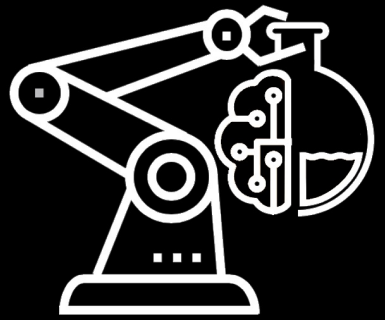
1. ADD 1
2. ADD 4
3. ADD 2
4. ADD 3
5. STIR for @3@ at #4#
6. FILTER keep precipitate
7. RECRYSTALLIZE from ethanol
8. YIELD -1

Outline

1. AI models for chemistry
2. Synthesis automation – RoboRXN
    A. Prediction of synthesis actions
    B. Execution on robot
3. AI for biocatalysis

Execution on chemical robot





Add Concentrate Reflux Filter Extract Stir Wash

Start heating reactor #2

Transfer 3 ml of solvent #5 to reactor #2

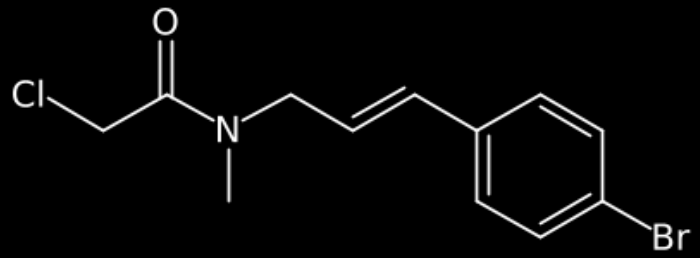
Set pressure

Execution on chemical robot

Cloud-based setup for autonomous synthesis

Hardware @ IBM Research Zurich

DEMO



CN(C/C=C/c1ccc(Br)cc1)C(=O)CCl

Outline

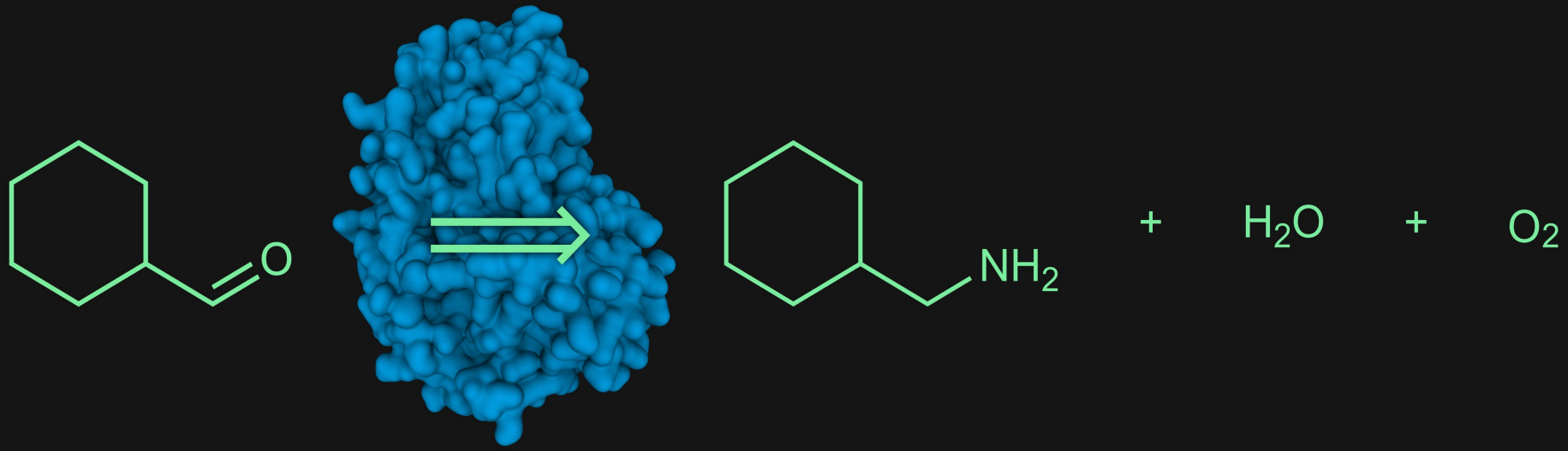
1. AI models for chemistry
2. Synthesis automation – RoboRXN
3. AI for biocatalysis

Outline

1. AI models for chemistry
2. Synthesis automation – RoboRXN
3. AI for biocatalysis
    A. Biocatalyzed synthesis planning
    B. Identification of active sites

Enzymatic catalyst

GreenCatRXN



EC 1.4.3.-

~~TEMPO~~ ~~NaClO~~ ~~DCM~~

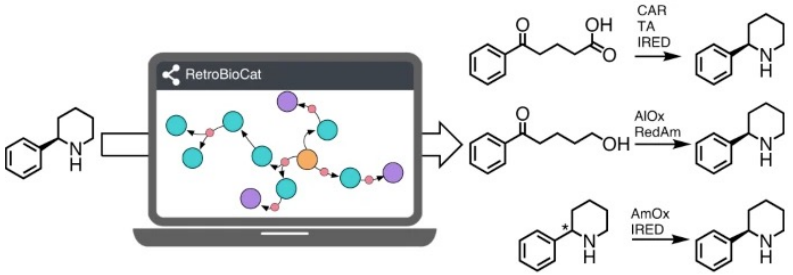
Martinez S. C. et al., Biophys J., 2015

Agarwal P. K., Microb Cell Fact., 2006

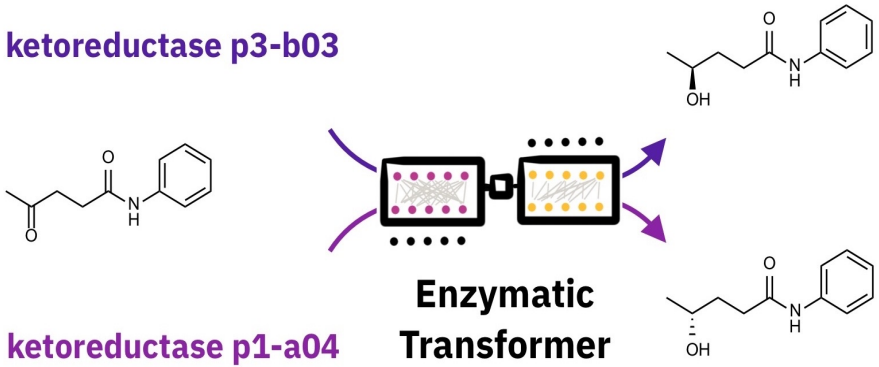
Enzymes in synthesis planning

- CASP (computer-assisted synthetic planning) using enzymes currently relying on rule-based approaches (Finnigan et al., Nat. Catal., 2021)

- Kreutter, Schwaller, and Reymond have shown that including enzyme name information yields good results



Finnigan et al, Nat. Catal., 2021



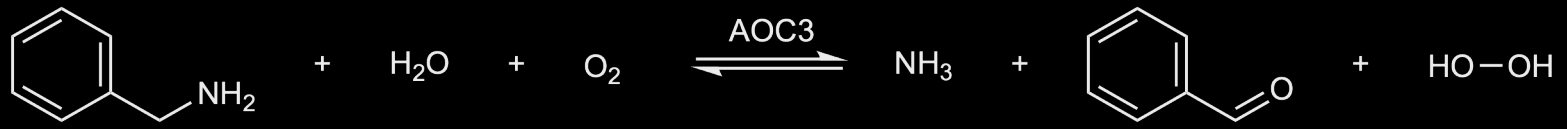
Kreutter et al., Chem. Sci., 2021

Our intuition

cyclohexanemethylamine + water + oxygen = ammonia + cyclohexanecarbaldehyde + hydrogen peroxide

Our intuition

cyclohexanemethylamine + water + oxygen = ammonia + cyclohexanecarbaldehyde + hydrogen peroxide



Class:
1 – Oxidoreductases
2 – Transferases
3 – Hydrolases
4 – Lysases
5 – Isomerases
6 – Ligases
7 – Translocases

Sub-class: Acting on the CH-NH_2 group of donors

Sub-sub-class: with oxygen as acceptor

Primary-amine oxidase

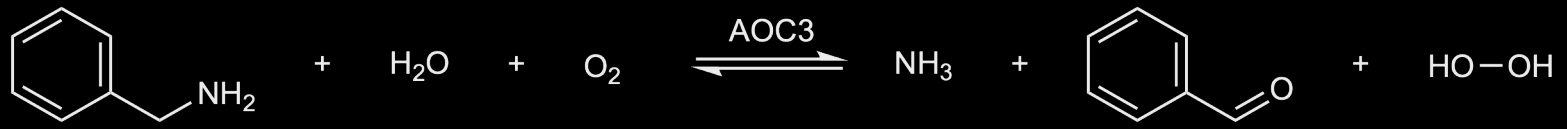
1.4.3.21

Amine Oxidase Copper Containing 3

Bairoch A. et al., Nucleic Acids Res., 2000

Our intuition

cyclohexanemethylamine + water + oxygen = ammonia + cyclohexanecarbaldehyde + hydrogen peroxide

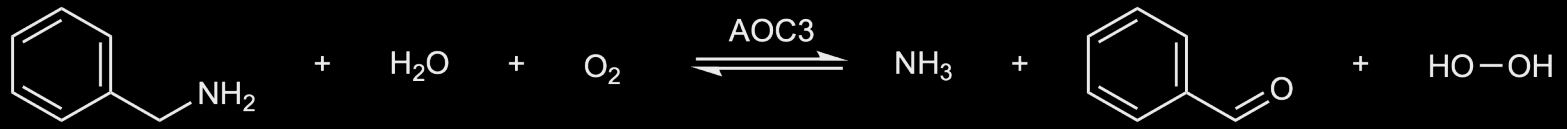


C1CCC(CC1)CN.O. O=O | 1.4.3.21 >>N. C1CCC(CC1)C=O.OO

Enzymatic reaction SMILES

Our intuition

cyclohexanemethylamine + water + oxygen = ammonia + cyclohexanecarbaldehyde + hydrogen peroxide



Enzymatic reaction SMILES:

C1CCC(CC1)CN.O. O=O | 1.4.3.21 >>N. C1CCC(CC1)C=O.OO

Tokenized reaction

C 1 C C C (C C 1) C N . O . O = O | [v1] [u4] [t3] [q21] >> N . C 1 C C C (C C 1) C = O . O O

Schwaller P. et al., ACS Cent. Sci, 2019

Our approach for biocatalyzed synthesis planning

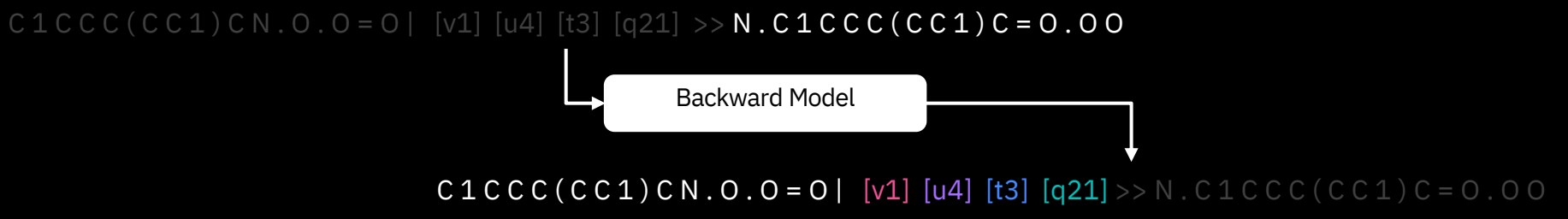
Forward prediction

Schwaller P. et al., ACS Cent. Sci, 2019

Our approach for biocatalyzed synthesis planning

Forward prediction

Backward prediction

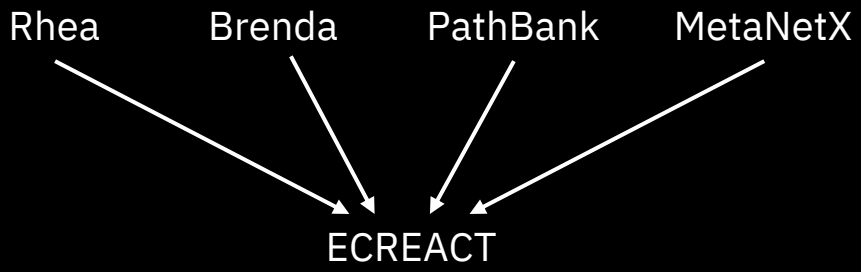


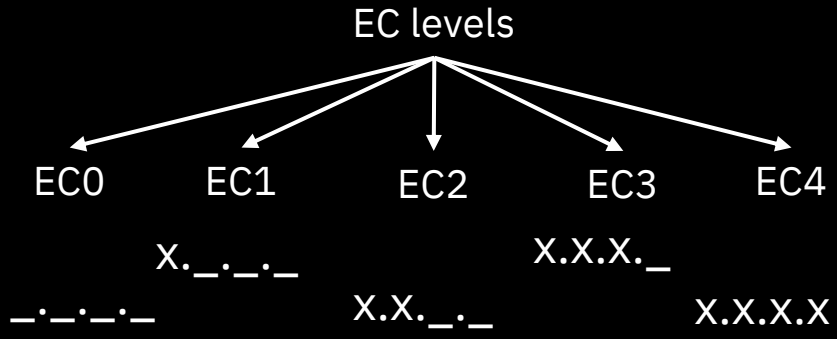
Schwaller et al. Chem. Sci. 2020

Liu et al. ACS Cent. Sci. 2017

Probst, D.; Manica, M.; Nana Teukam, Y. G.; Castrogiovanni, A.; Paratore, F.; Laino, T., *Nat. Commun.* **2022**, *13*, 964.

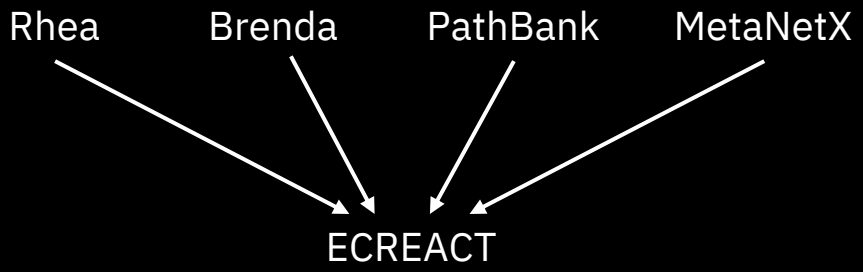
Dataset





Rafael A. et al., Nucleic Acids Res., 2012; Chang A., et al. Nucleic Acids Res., 2021; David S., Nucleic Acids Res., 2020; Thomas N, ACS, 2019

Dataset



Total number of reactions: 63.403

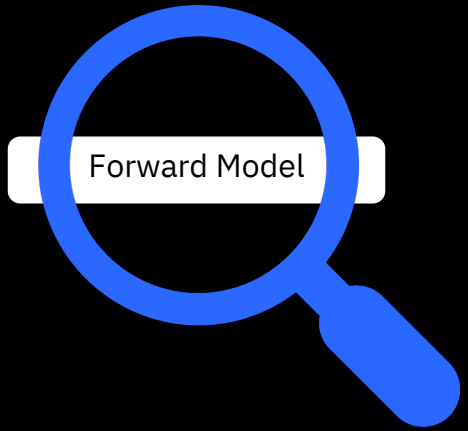
Most represented reaction type: Transferases (EC 2.x.x.x)

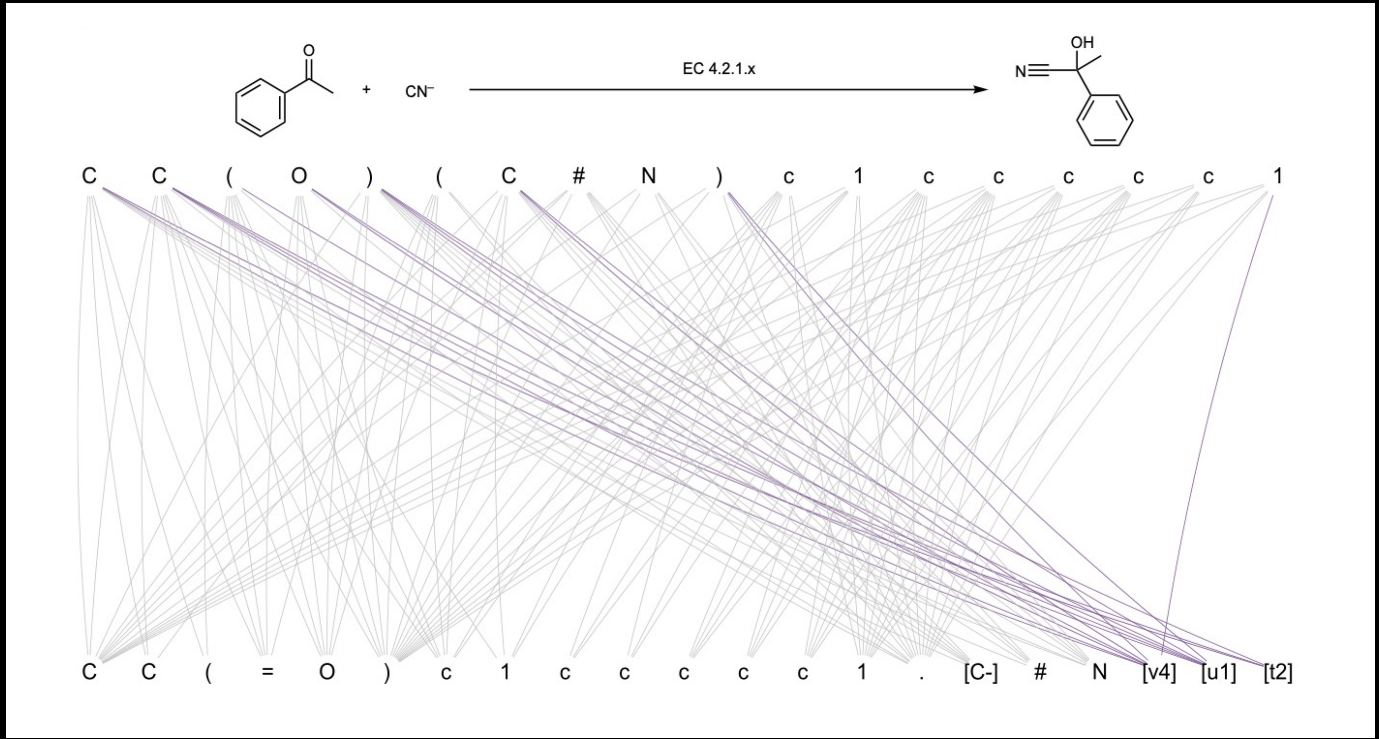
Least represented reaction type: Translocases (EC 7.x.x.x)

Probst D. et al., ChemRxiv, 2021

Rafael A. et al., Nucleic Acids Res., 2012; Chang A., et al. Nucleic Acids Res., 2021; David S., Nucleic Acids Res., 2020; Thomas N, ACS, 2019

Attention Analysis





Schwaller P. et al., Adv. Sci., 2021

Try GreenCatRXN on RXN for Chemistry: https://rxn.res.ibm.com

GitHub: https://github.com/rxn4chemistry/rxn4chemistry

Outline

1. AI models for chemistry
2. Synthesis automation – RoboRXN
3. AI for biocatalysis
    A. Biocatalyzed synthesis planning
    B. Identification of active sites

Active site in proteins

Proteins' activity is directly related to the structure of the active site

Biological function annotation of proteins usually rely on 3D
structural models ([Yousaf et al., 2021, Kozlovskii and Popov, 2020 and 2021])

Some methods identify active sites via sequence similarity (Pfam [Mistry et al., 2020] and PSI-BLAST [Altschul et al., 1997])

PROTEIN STRUCTURE

Scaffold to support and position active site

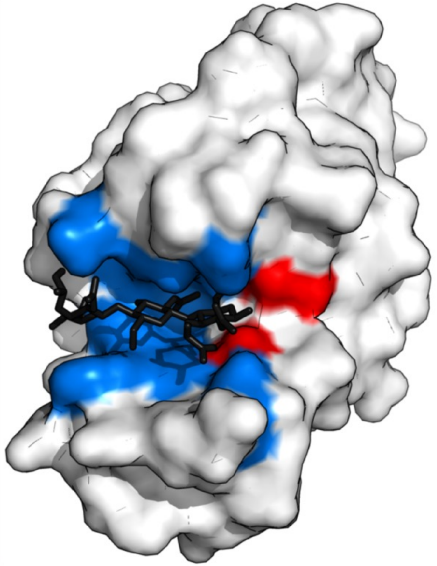
ACTIVE SITE

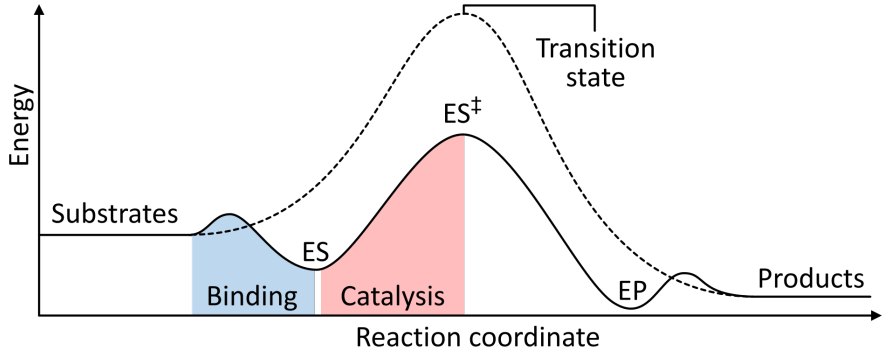
BINDING SITES

Bind and orient substrate(s)

CATALYTIC SITE

Reduce chemical activation energy

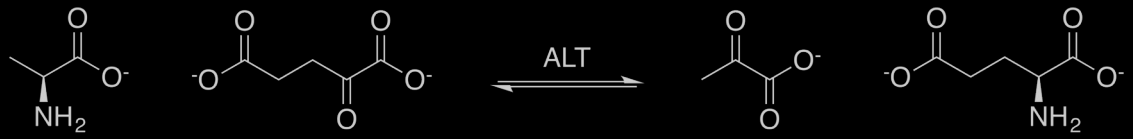




https://en.wikipedia.org/wiki/Enzyme

Biocatalysis as a language

L-alanine + 2-oxoglutarate = pyruvate + L-glutamate



Reaction SMILES

C[C@H](N)C([O-])=O . [O-]C(=O)CCC(=O)C([O-])=O >> CC(=O)C([O-])=O . N[C@@H](CCC([O-])=O)C([O-])=O

Introducing AA sequences

```
10 20 30 40 50
MASSTGDRSQ AVRHGLRAKV LTLDGMNPRV RRVEYAVRGP IVQRALELEQ
60 70 80 90 100
ELRQGVKKPF TEVIRANIGD AQAMGQRPIT FLRQVLALCV NPDLLSSPNF
110 120 130 140 150
PDDAKKRAER ILQACGGHSL GAYSVSSGIQ LIREDVARYI ERRDGGIPAD
160 170 180 190 200
PNNVFLSTGA SDAIVTVLKL LVAGEGHTRT GVLIPIPQYP LYSATLAELG
210 220 230 240 250
AVQVDYYLDE ERAWALDVAE LHRALGQARD HCRPRALCVI NPGNPTGQVQ
260 270 280 290 300
TRECIEAVIR FAFEERLFLL ADEVYQDNVY AAGSQFHSFK KVLMEMGPPY
310 320 330 340 350
AGQQELASFH STSKGYMGEC GFRGGYVEVV NMDAAVQQQM LKLMSVRLCP
360 370 380 390 400
PVPGQALLDL VVSPPAPTDP SFAQFQAEKQ AVLAELAAKA KLTEQVFNEA
410 420 430 440 450
PGISCNPVQG AMYSFPRVQL PPRAVERAQE LGLAPDMFFC LRLLEETGIC
460 470 480 490
VVPGSGFGQR EGTYHFRMTI LPPLEKLRLL LEKLSRFHAK FTLEYS
```

Reaction SMILES

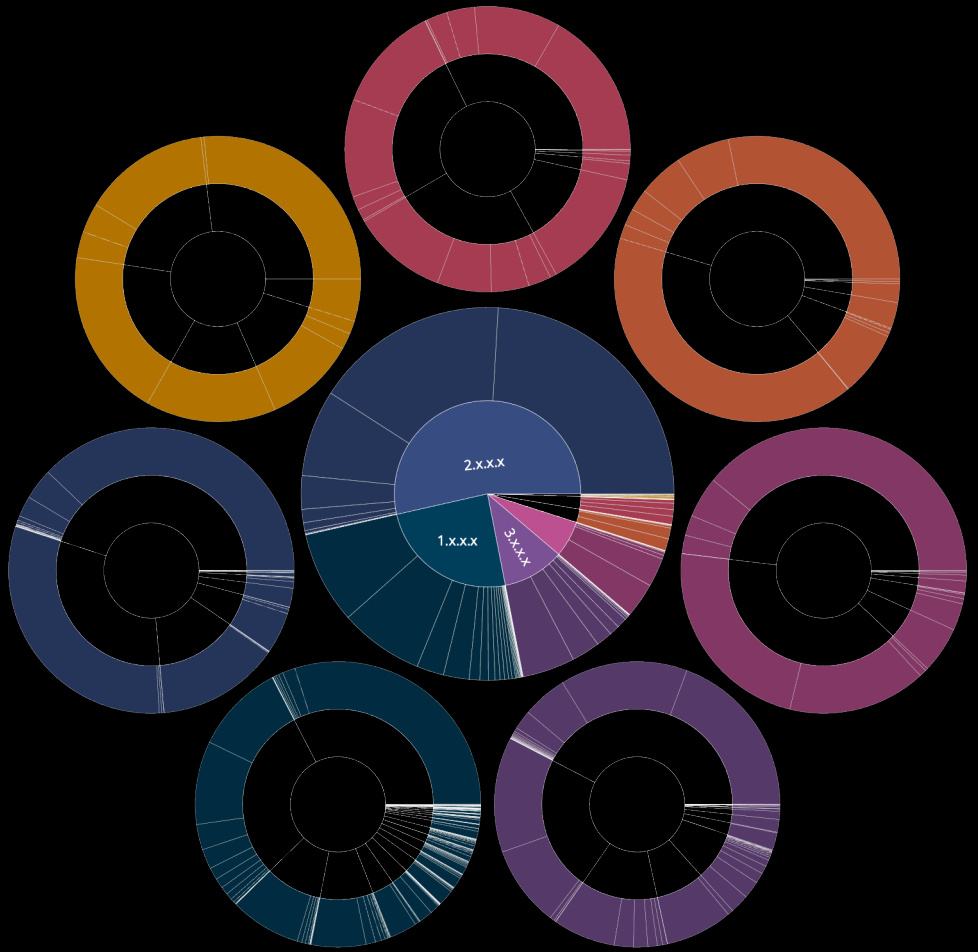
C[C@H](N)C([O-])=O . [O-]C(=O)CCC(=O)C([O-])=O|**MAS...LEYS** >> CC(=O)C([O-])=O . N[C@@H](CCC([O-])=O)C([O-])=O

Dataset preparation

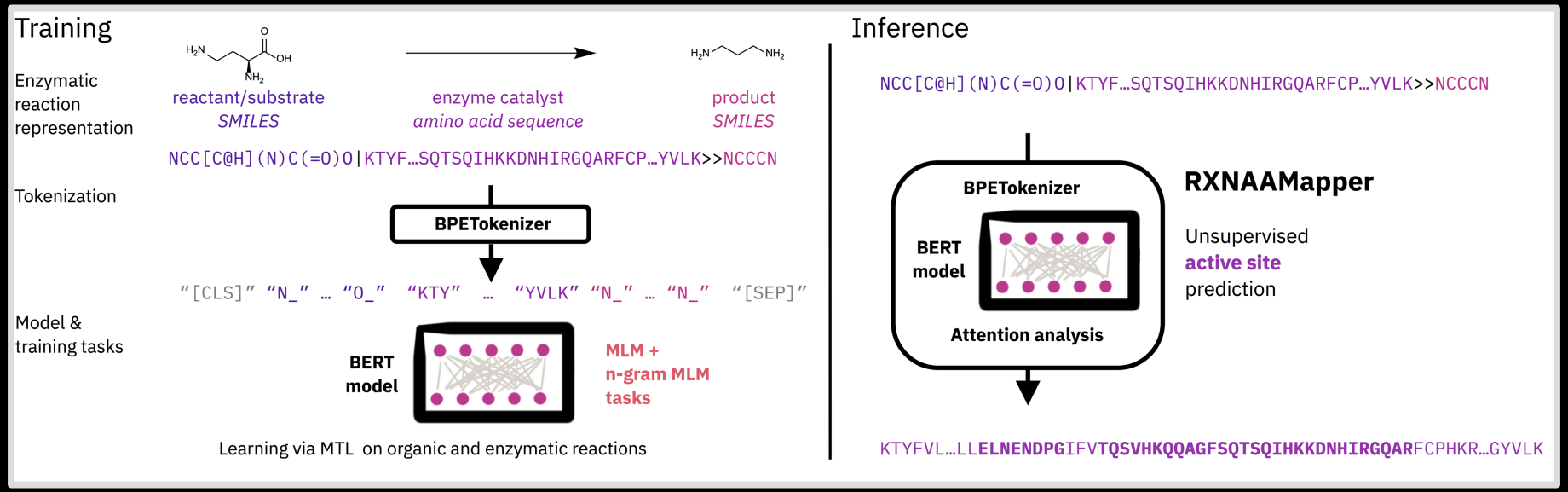
ECREACT: Source from Brenda, MetaNetX, PathBank, and Rhea

Annotated reaction SMILES using AA sequences

Combined with USPTO [Lowe, 2012] (~1M reactions)

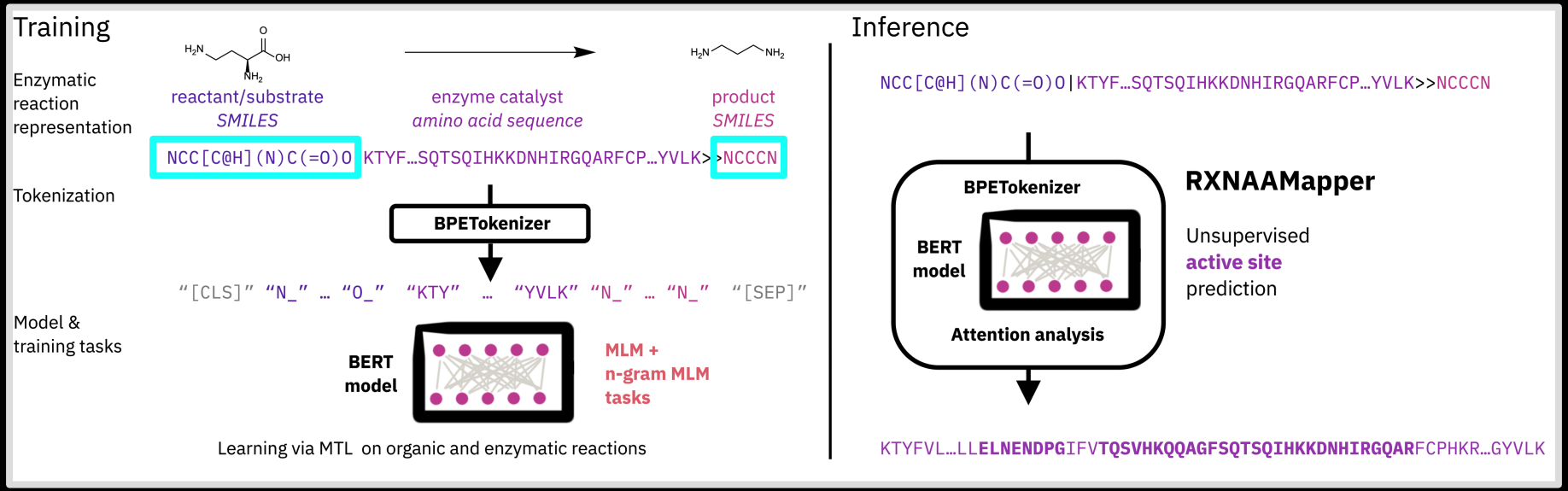


Language modelling for biocatalysed reactions



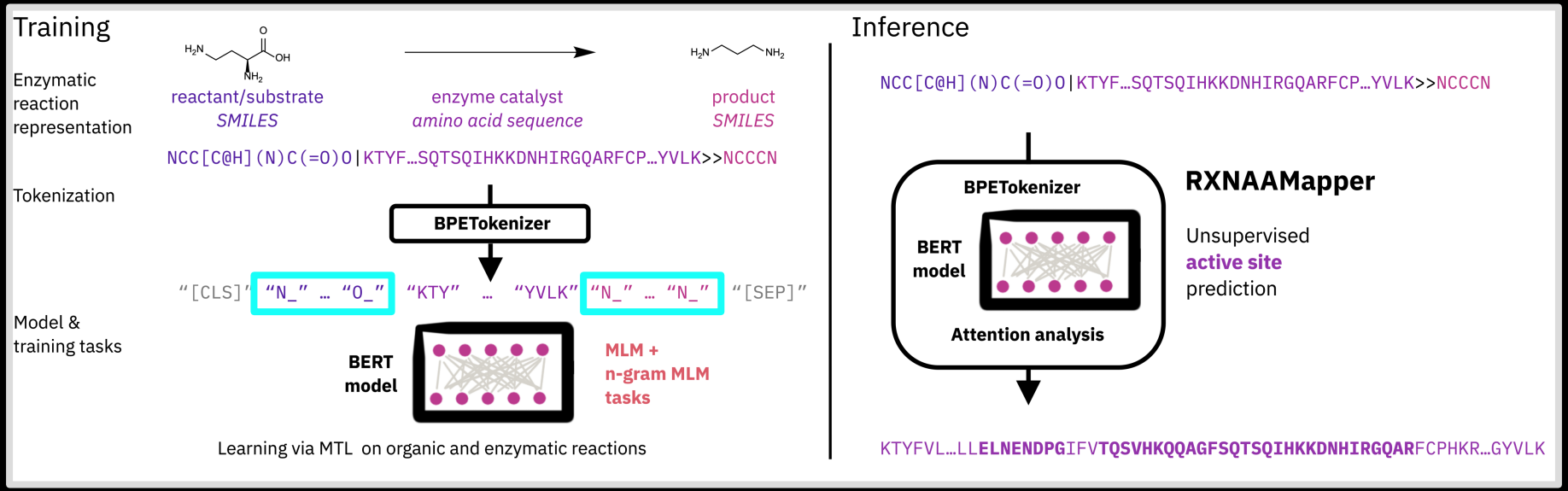
Pesciullesi et al., *Nature Communications*, **2020**
Probst et al., *Nature Communications*, **2022**
Gage, *The C user journal*,

Language modelling for biocatalysed reactions



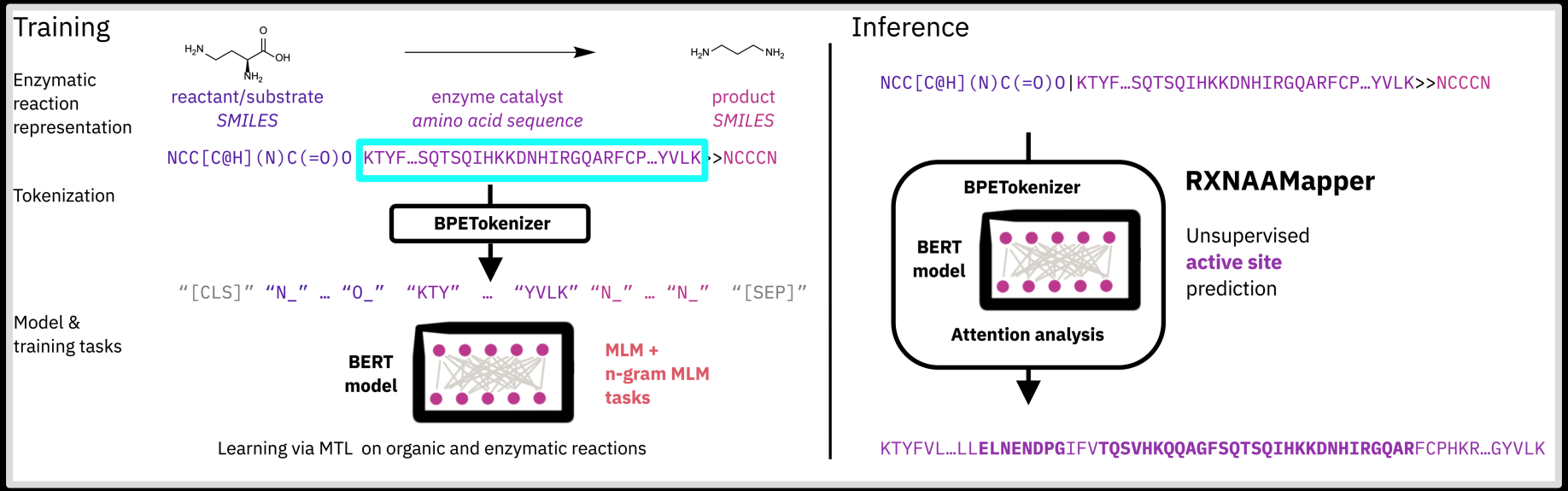
Pesciullesi et al., *Nature Communications*, **2020**
Probst et al., *Nature Communications*, **2022**
Gage, *The C user journal*,

Language modelling for biocatalysed reactions



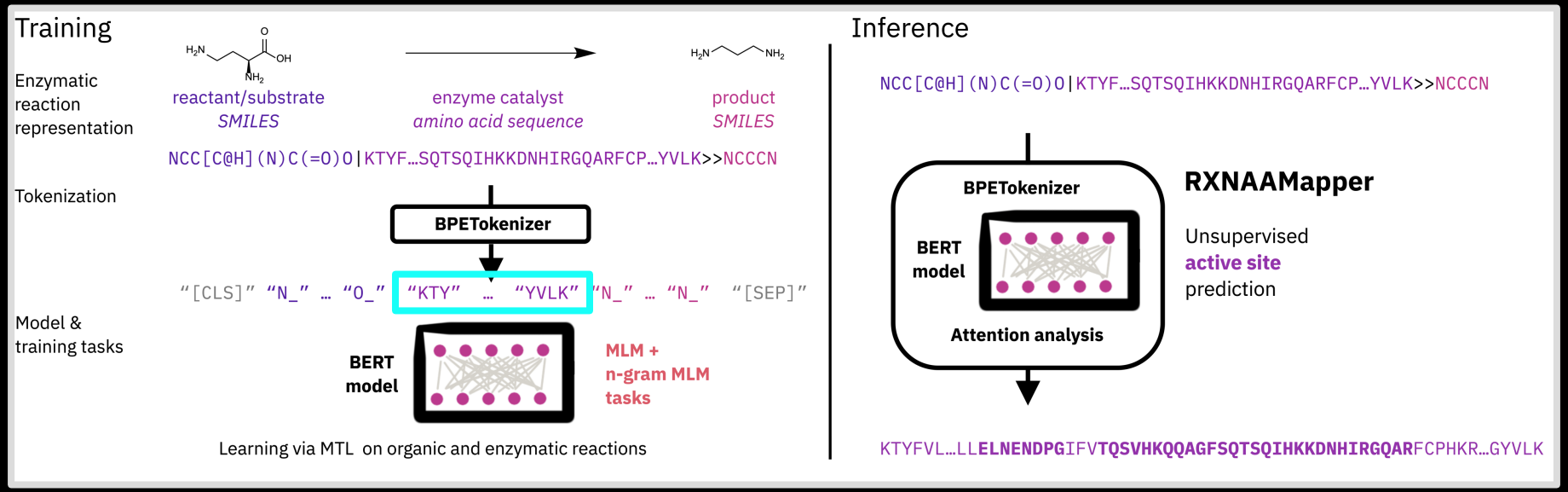
Pesciullesi et al., *Nature Communications*, **2020**
Probst et al., *Nature Communications*, **2022**
Gage, *The C user journal*,

Language modelling for biocatalysed reactions



Pesciullesi et al., *Nature Communications*, **2020**
Probst et al., *Nature Communications*, **2022**
Gage, *The C user journal*,

Language modelling for biocatalysed reactions



Pesciullesi et al., *Nature Communications*, **2020**
Probst et al., *Nature Communications*, **2022**
Gage, *The C user journal*,

Results

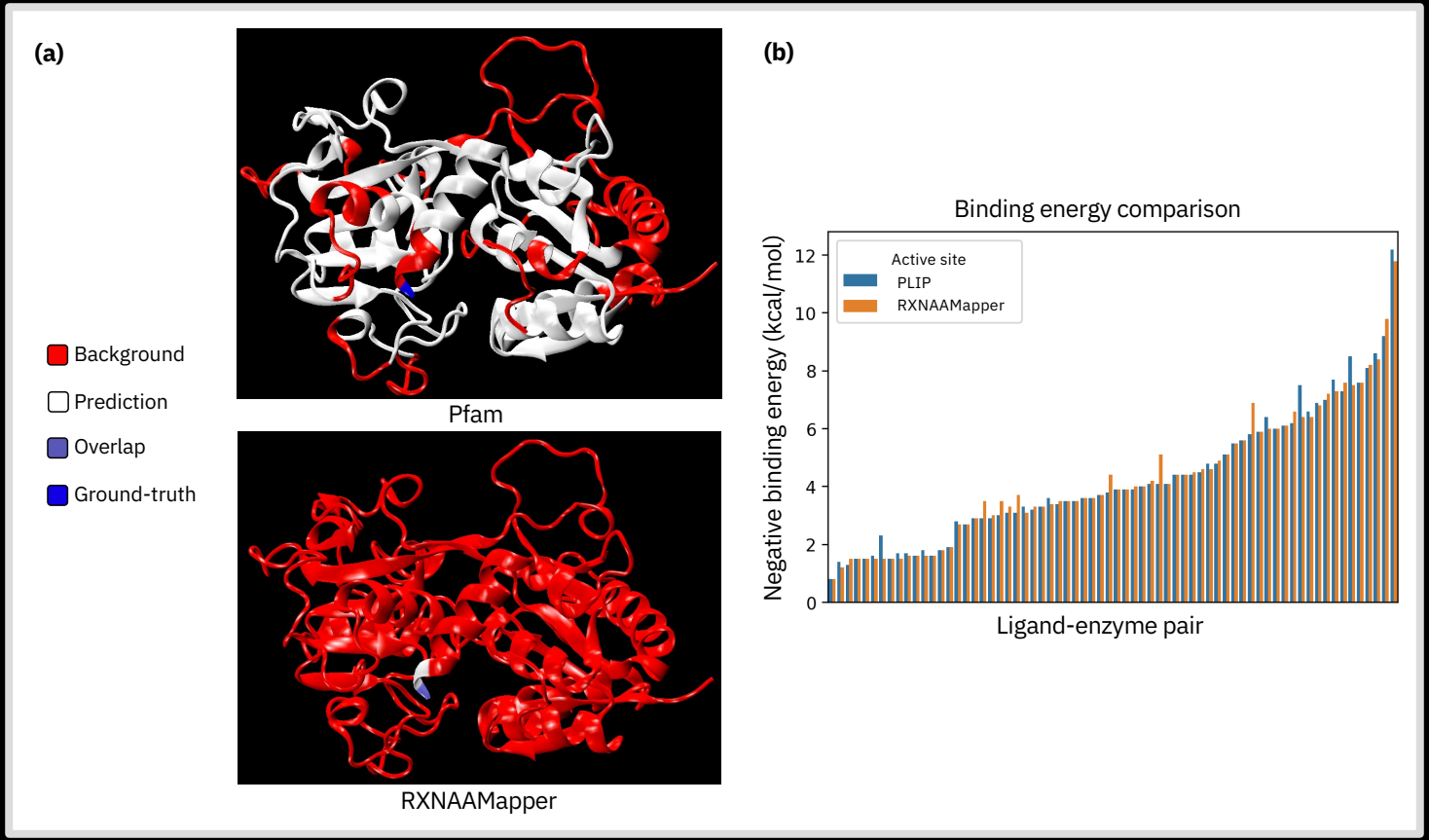
	Overlap Score	False Positive Rate
Random Model	4.98%	84.20%
Pfam	24.01%	78.01%
BERT-base	28.98%	75.56%
RXNAAMapper (ours)	**31.51%**	**66.63%**

Kwate Dassi et al., ChemRxiv , 2021

Mistry et al., *Nucleic Acid Research*, **2020**

Lu et al., *AAAI*, **2022**

Results



Eberhardt et al., *Journal of Chemical Information and Modeling*, **2021**

Trott and Olson, *Journal of Computational Chemistry*, **2010**

Availability

rxn-aa-mapper

Reactions SMILES-AA sequence mapping

setup

```
conda env create -f conda.yml
conda activate rxn_aa_mapper
```

In the following we consider on exar

predict active site

The trained model can used to map reactant atoms to AA sequence locations that potentially represent the active site.

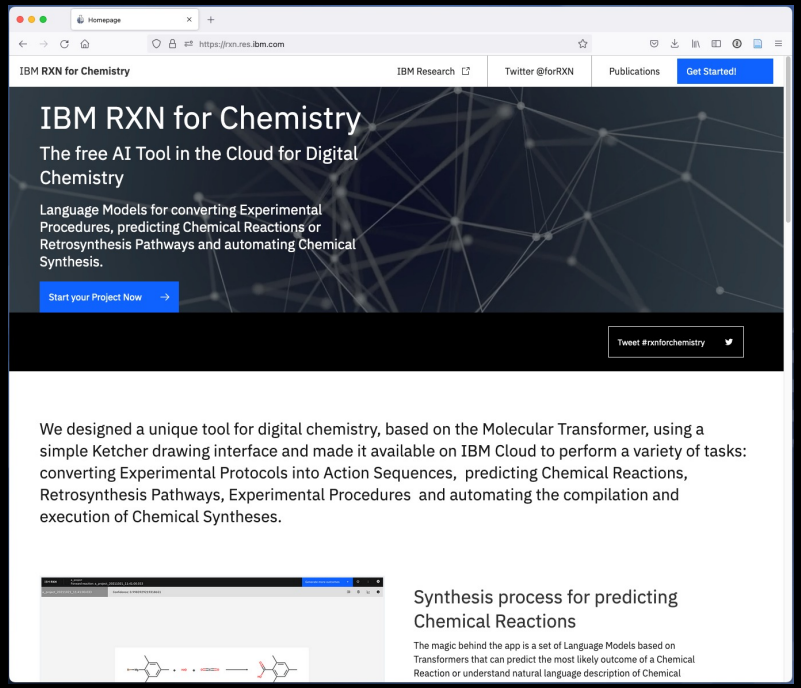
```
from rxn_aa_mapper.aa_mapper import RXNAAMapper

config_mapper = {
    "vocabulary_file": "./examples/vocabulary_token_75K_min_600_max_750_500K.txt",
    "aa_sequence_tokenizer_filepath": "./examples/token_75K_min_600_max_750_500K.json",
    "model_path": "/tmp/rxnaamapper-pretrained-model",
    "head": 3,
    "layers": [11],
    "top_k": 1,
}
mapper = RXNAAMapper(config=config_mapper)
mapper.get_reactant_aa_sequence_attention_guided_maps(["NC(=O)c1ccc[n+]([C@@H]2O[C@H](COP(
```

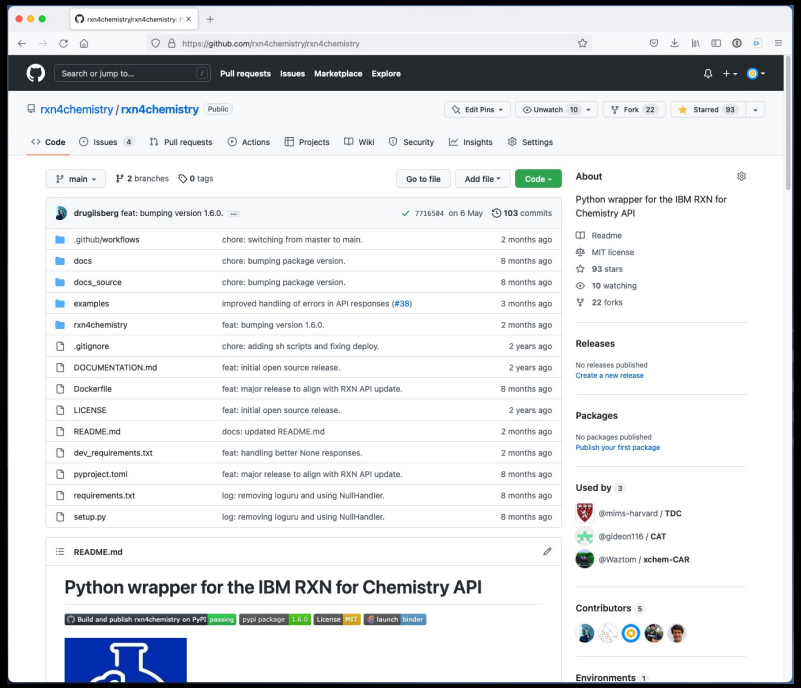
https://github.com/rxn4chemistry/rxnaamapper

Kwate Dassi et al., ChemRxiv , 2021

Outline

1. AI models for chemistry
2. Synthesis automation – RoboRXN
3. AI for biocatalysis





User interface, freely available on:
rxn.res.ibm.com

Access via API / Python wrapper:
github.com/rxn4chemistry/rxn4chemistry

Thank you for your attention!

If you have any questions:

E-mail: ava@zurich.ibm.com

Twitter: @acvaucher

Acknowledgments:

Antonio Cardinale
Alessandro Castrogiovanni
Joppe Geluykens
Teodoro Laino
Matteo Manica
Vishnu H. Nair
Yves Gaetan Nana Teukam
Philippe Schwaller
Aleksandros Sobczyk
Alessandra Toniato
Heiko Wolf
Federico Zipoli
Loïc Kwate Dassi
Daniel Probst