Learning how to do organic chemistry from literature-extracted synthesis actions

1st International Symposium
for Materials R&D Data

8 July 2022

Alain Vaucher
🐦 @acvaucher

IBM **Research**

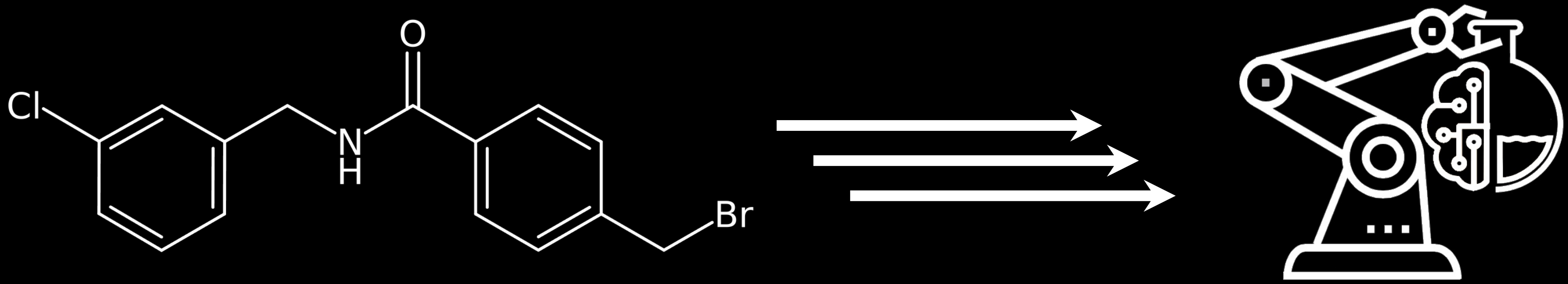
Data and chemical reactions

- 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!

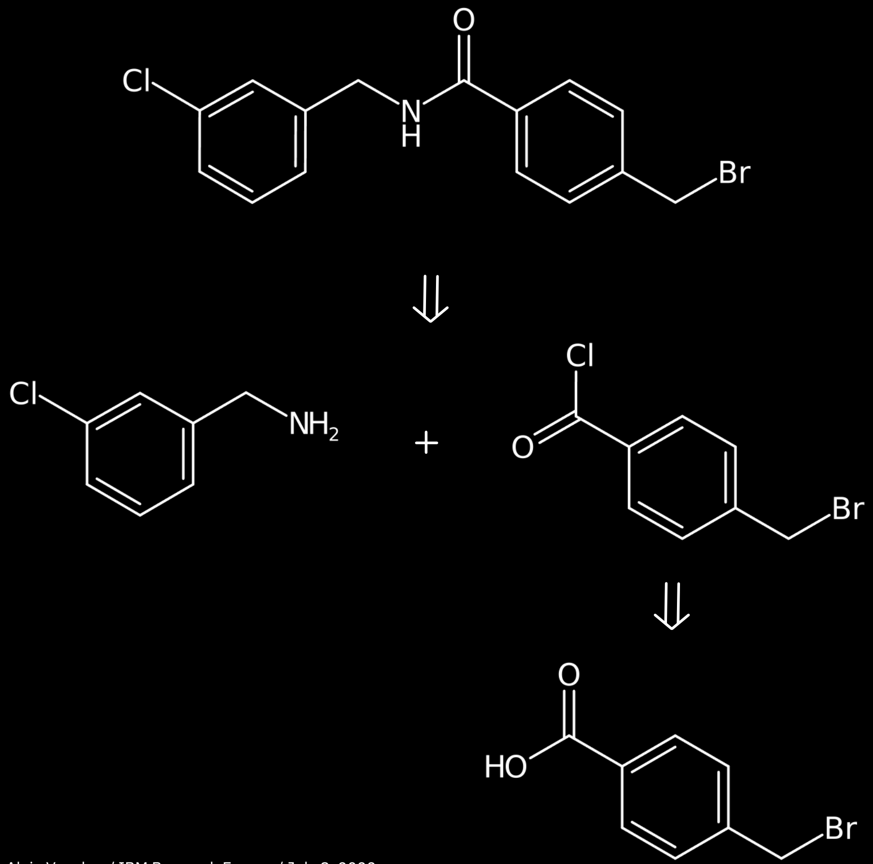
Data and chemical reactions



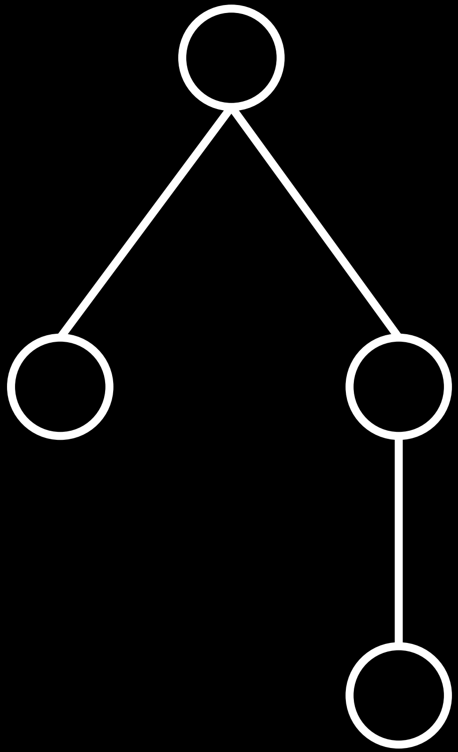
Target molecule

Synthesis execution

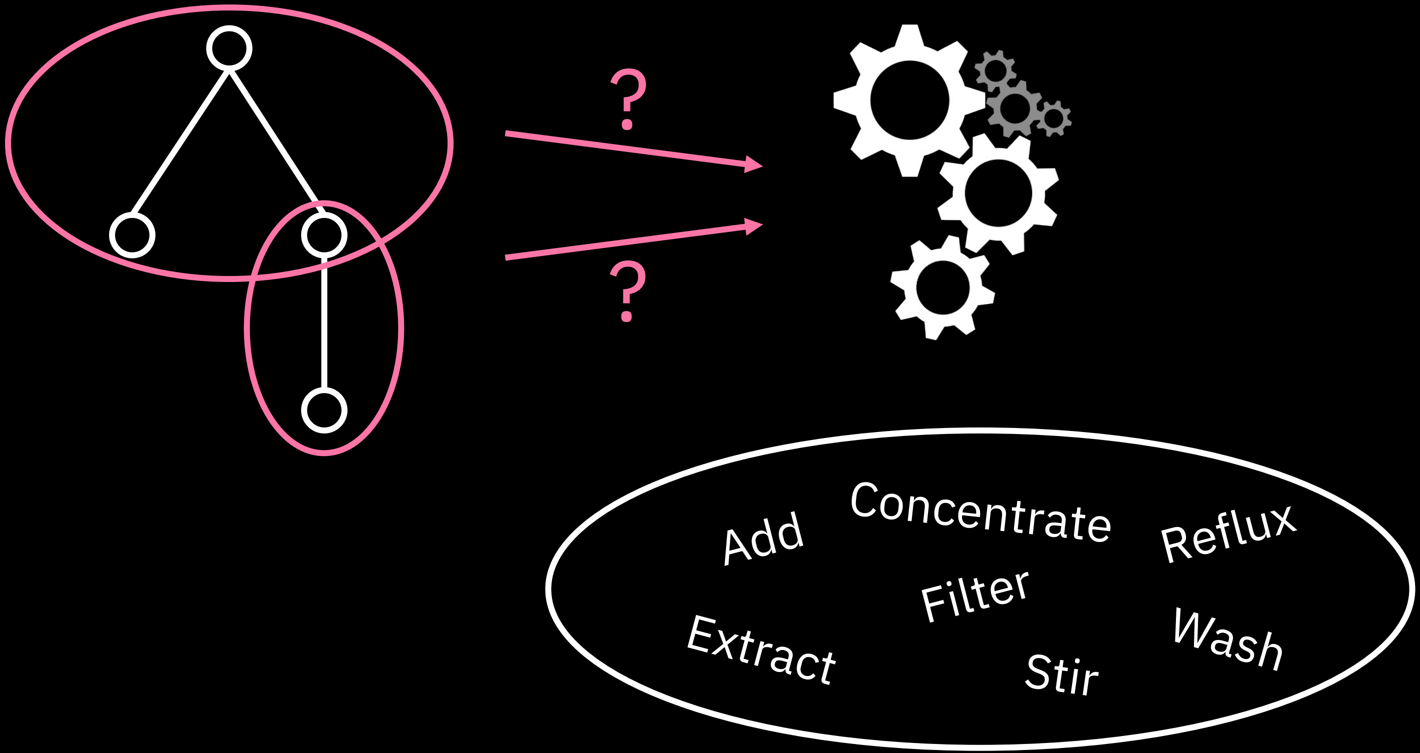
Step 1: retrosynthetic analysis



Retrosynthetic tree



Step 2: experimental steps

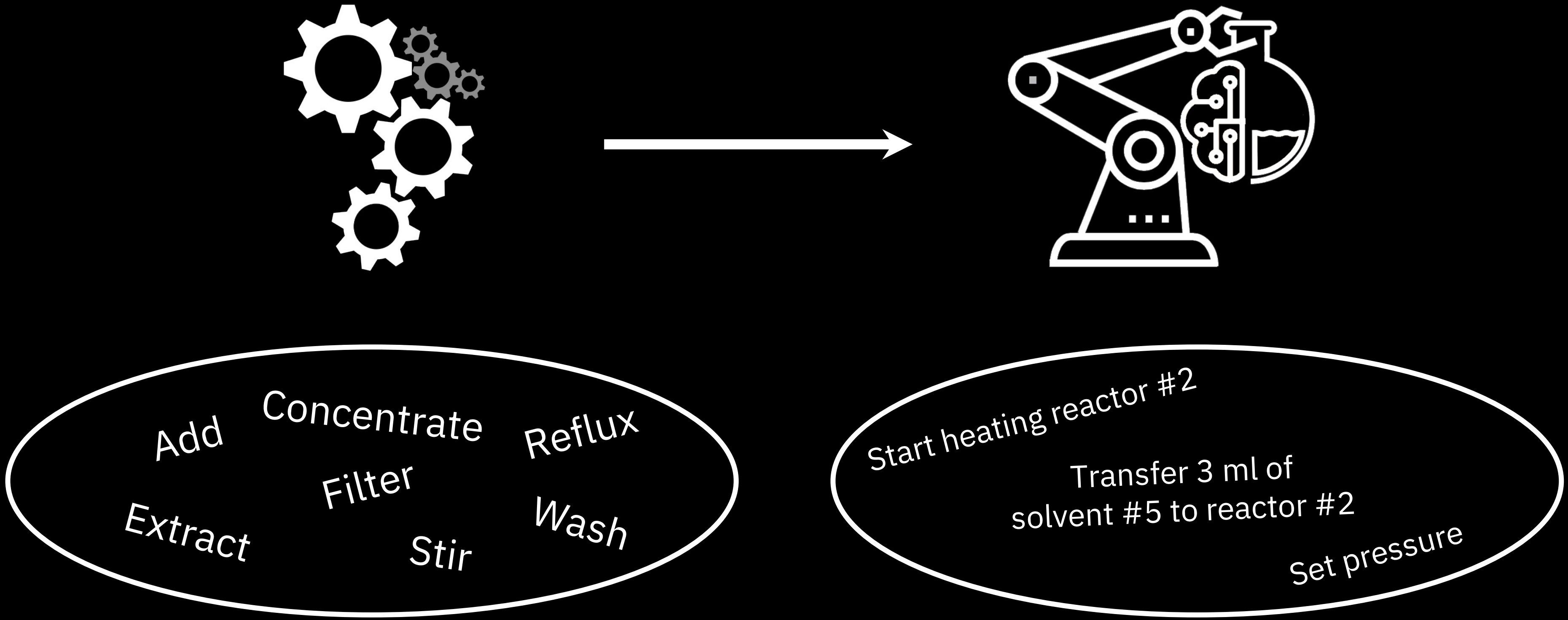


Add Concentrate Reflux

Filter

Extract Stir Wash

Step 3: Execution on robotic system



Add

Concentrate

Reflux

Filter

Extract

Stir

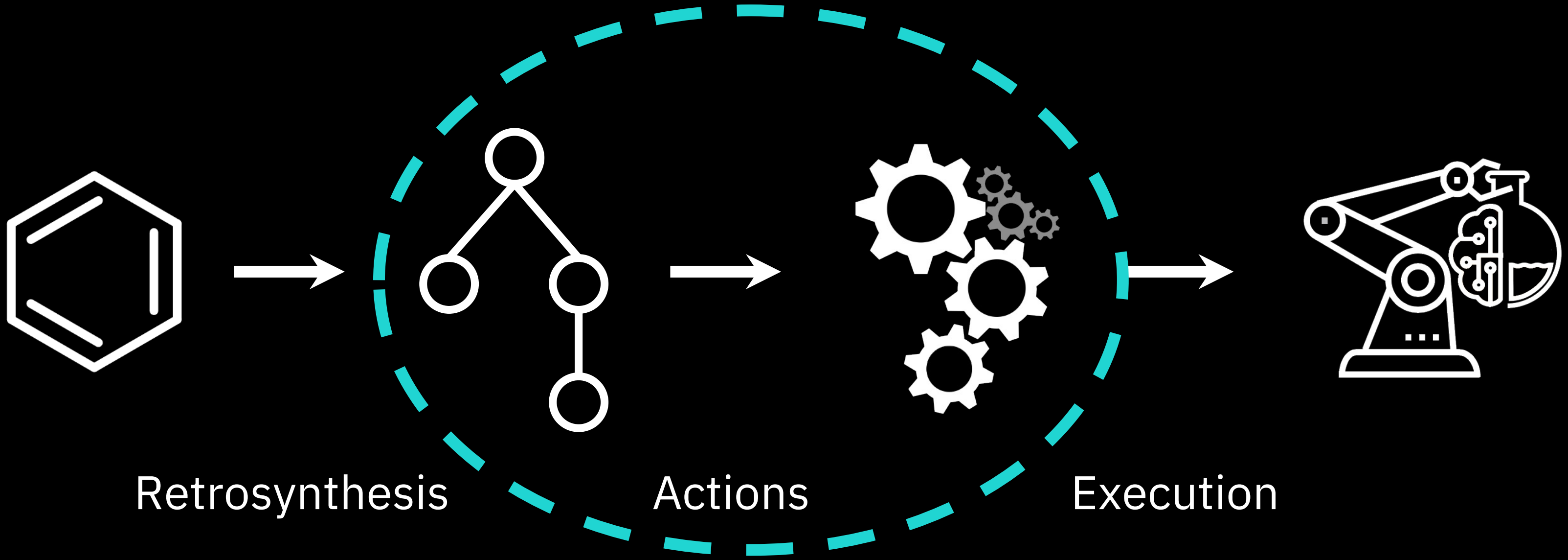
Wash

Start heating reactor #2

Transfer 3 ml of solvent #5 to reactor #2

Set pressure

All together



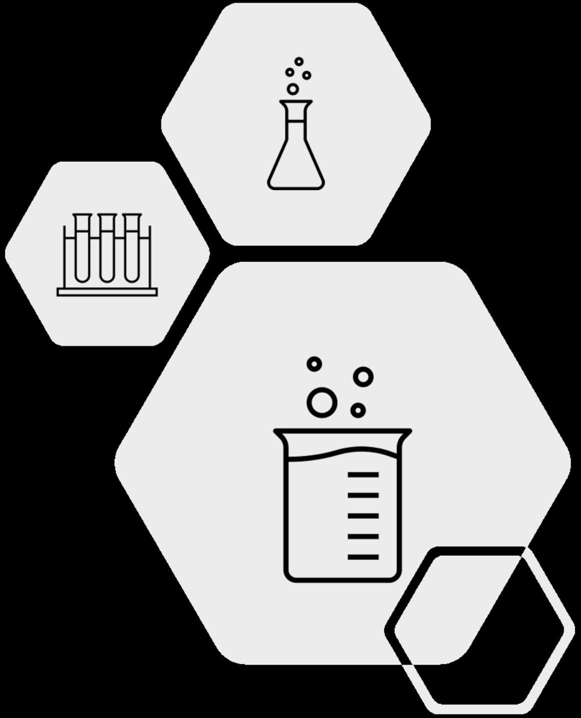
Retrosynthesis

Actions

Execution

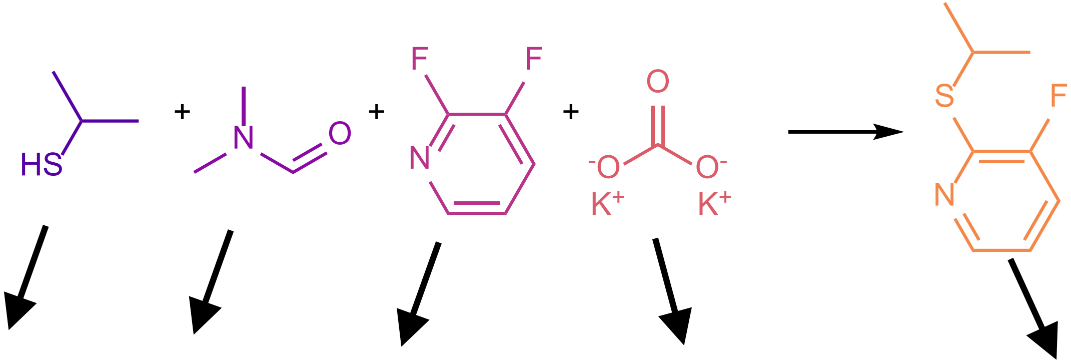
Data sources

- 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

precursors >> products



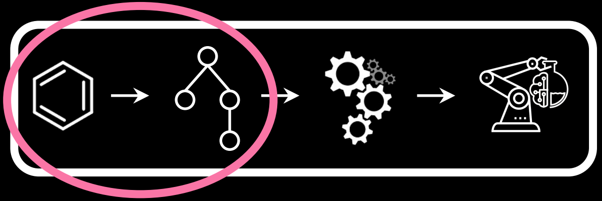
SMILES: CC(C)S.CN(C)C=O.Fc1cccnc1F.O=C([O-])[O-].[K+].[K+]>>CC(C)Sc1ncccc1F

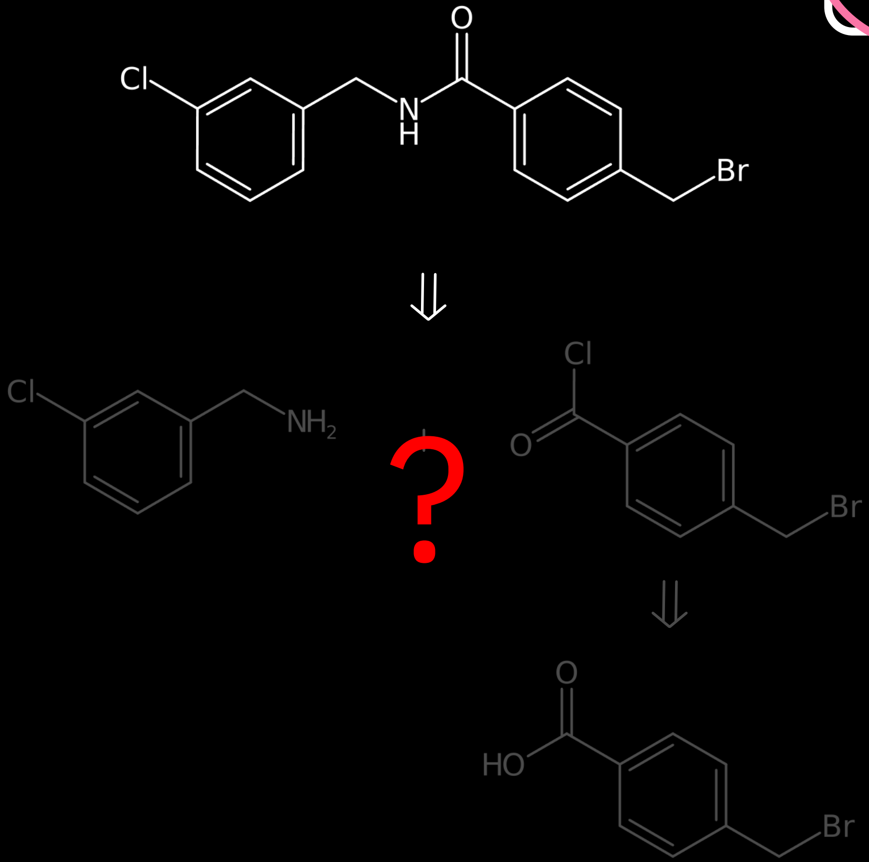
Split -> “sequences of atoms”

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

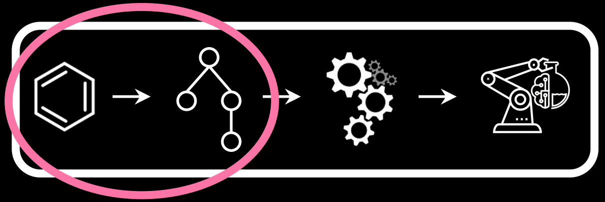
→ Borrow methods developed for human languages

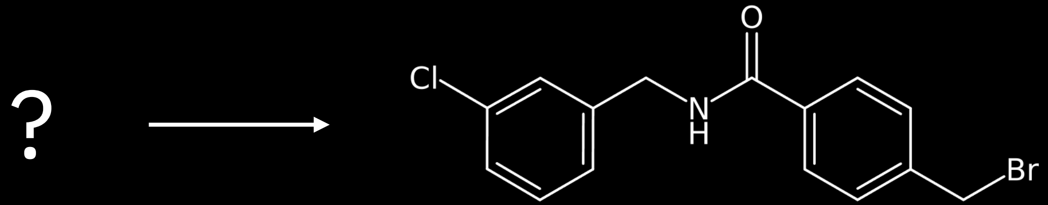
Retrosynthesis





Retrosynthesis





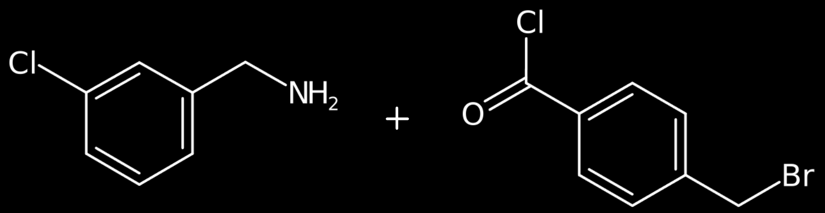
“Translation” from the language of products to the language of precursors

“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

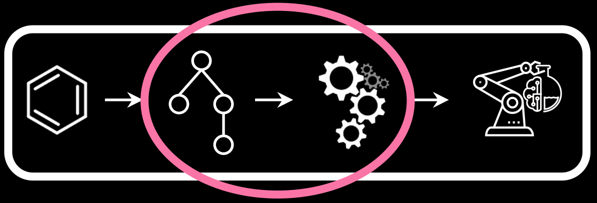
Transformer

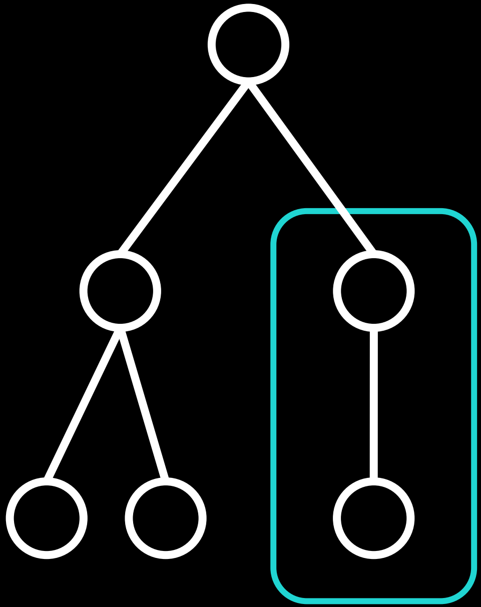
One among many correct sets of precursors →



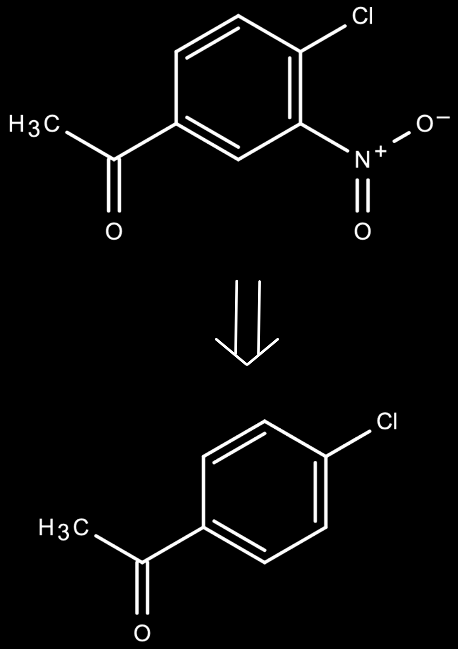
Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A. & Laino, T., *Chem. Sci.*, **2020**, *11*, 3316-3325.

Synthesis actions

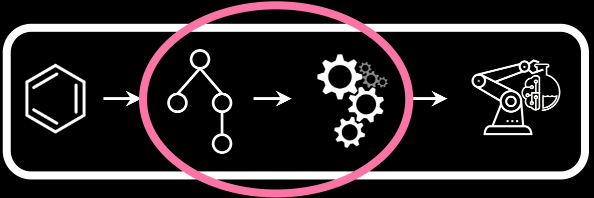




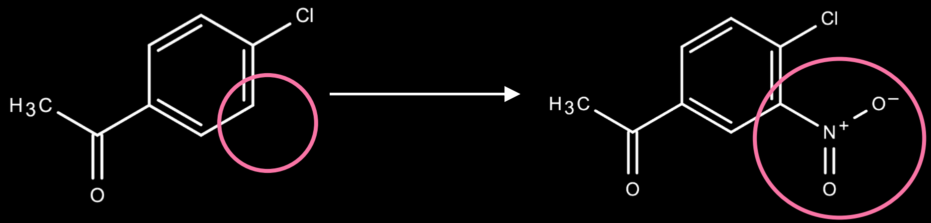
One reaction step



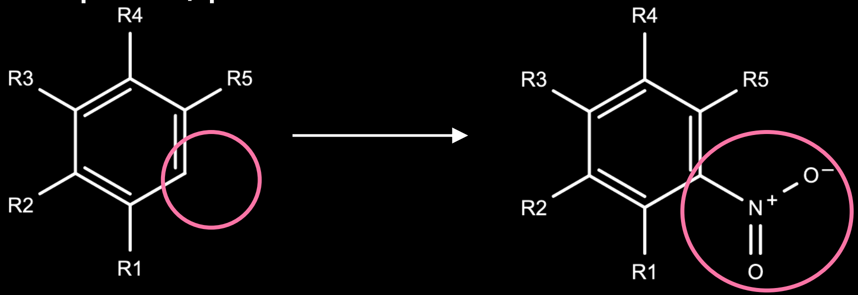
Synthesis actions



Example:

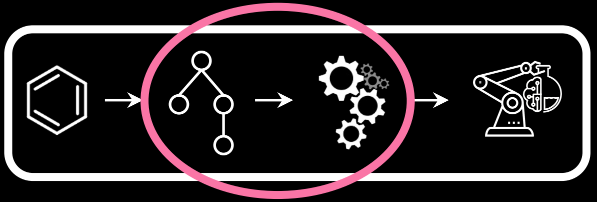


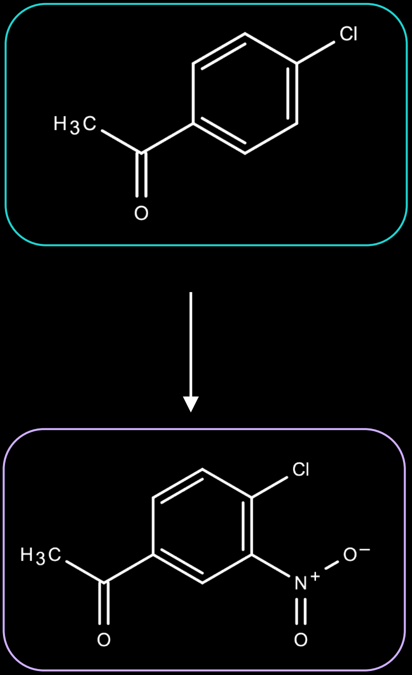
Template/pattern:



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

Synthesis actions

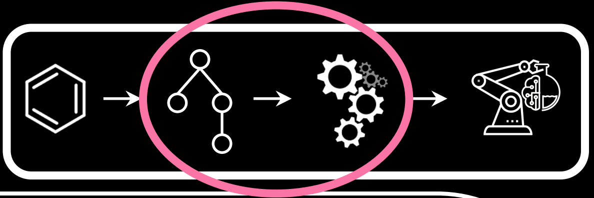




C1=CC(C(=O)C)=CC=C1Cl>>C1=CC(C(=O)C)=CC([N+]([O-])=O)=C1Cl

```
Operation 1

Operation 2

Operation 3

Operation 4

…
```

SMILES-to-actions

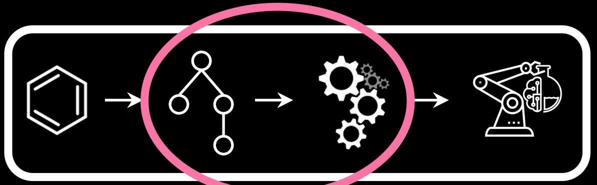


- No dataset!
- Information is available **indirectly**
- First: extract actions from text
- “Paragraph-to-actions” model

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%).

Paragraph-to-actions: Action definition



… 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 …

?

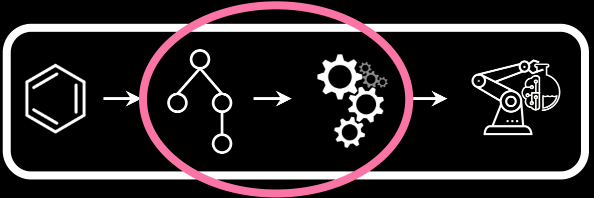
Operation 1

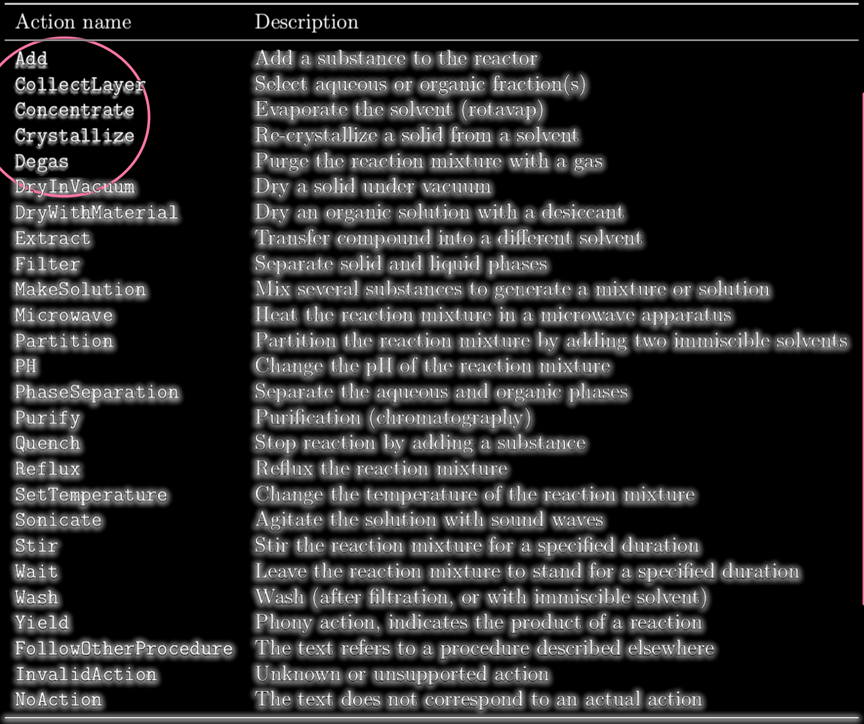
Operation 2

…

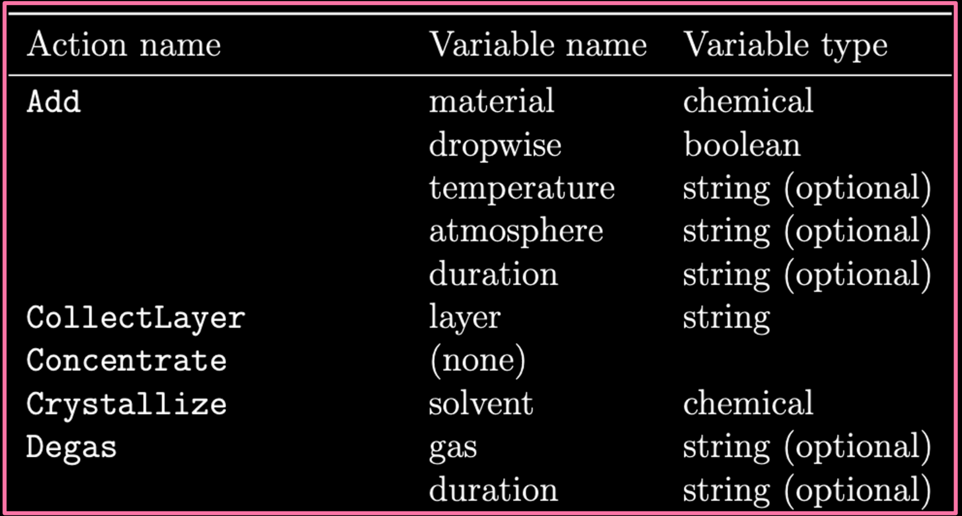
?

Paragraph-to-actions: Action definition



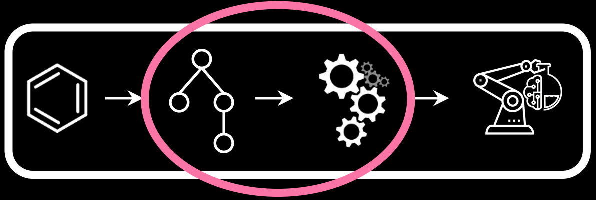


Action name	Description
Add	Add a substance to the reactor
CollectLayer	Select aqueous or organic fraction(s)
Concentrate	Evaporate the solvent (rotavap)
Crystallize	Re-crystallize a solid from a solvent
Degas	Purge the reaction mixture with a gas
DryInVacuum	Dry a solid under vacuum
DryWithMaterial	Dry an organic solution with a desiccant
Extract	Transfer compound into a different solvent
Filter	Separate solid and liquid phases
MakeSolution	Mix several substances to generate a mixture or solution
Microwave	Heat the reaction mixture in a microwave apparatus
Partition	Partition the reaction mixture by adding two immiscible solvents
PH	Change the pH of the reaction mixture
PhaseSeparation	Separate the aqueous and organic phases
Purify	Purification (chromatography)
Quench	Stop reaction by adding a substance
Reflux	Reflux the reaction mixture
SetTemperature	Change the temperature of the reaction mixture
Sonicate	Agitate the solution with sound waves
Stir	Stir the reaction mixture for a specified duration
Wait	Leave the reaction mixture to stand for a specified duration
Wash	Wash (after filtration, or with immiscible solvent)
Yield	Phony action, indicates the product of a reaction
FollowOtherProcedure	The text refers to a procedure described elsewhere
InvalidAction	Unknown or unsupported action
NoAction	The text does not correspond to an actual action



Action name	Variable name	Variable type
Add	material	chemical
	dropwise	boolean
	temperature	string (optional)
	atmosphere	string (optional)
	duration	string (optional)
CollectLayer	layer	string
Concentrate	(none)	
Crystallize	solvent	chemical
Degas	gas	string (optional)
	duration	string (optional)

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 …

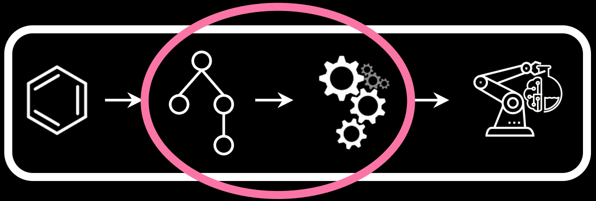
```
Add(name=‘Sodium borohydride’,
    quantity=[‘0.25 g’, ‘6.6 mmol’],
    duration=‘20 min’)

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

What kind of model?

–Rule-based model?

–Fuly data-driven model?

Models for Paragraph-to-actions

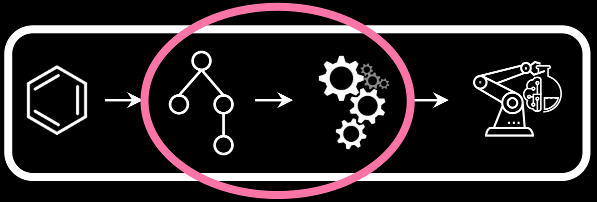


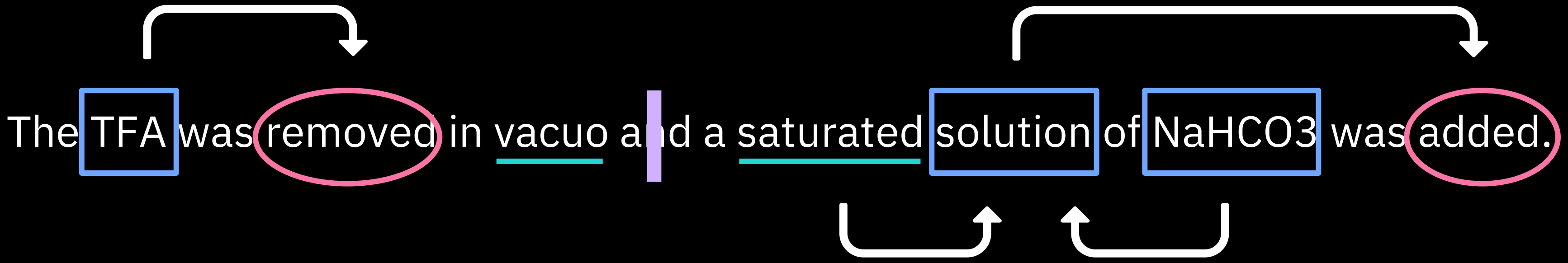
Rule-based model	ML model
Requires no training data	Requires training data
Not very robust, hard to improve	Improve model by improving data

Let’s combine both:

Rule-based training data for ML model

Rule-based model

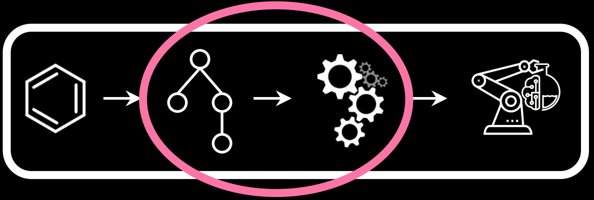


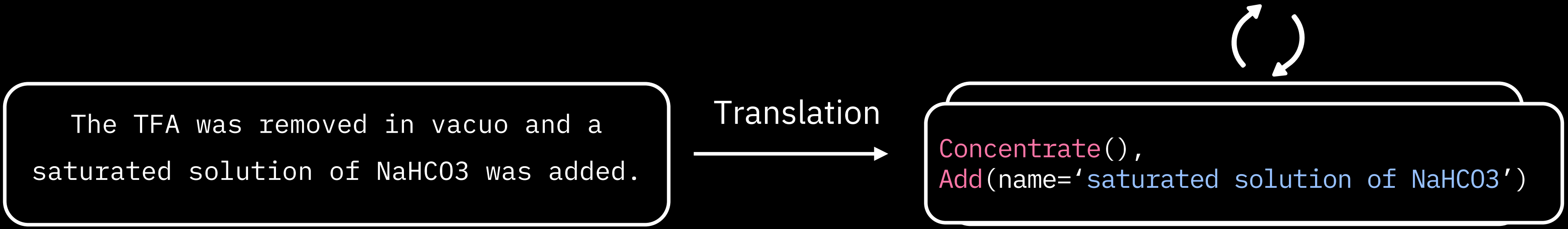


The TFA was removed in vacuo and a saturated solution of $NaHCO_3$ was added.

Generated actions for ~4M sentences

ML model





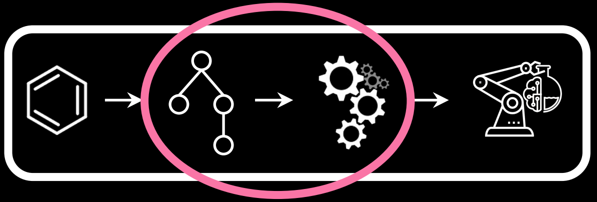
The TFA was removed in vacuo and a saturated solution of NaHCO3 was added.

Translation

```
Concentrate(),
Add(name='saturated solution of NaHCO3')
```

Transformer model, initial training on 4M samples

Rule-based vs ML model



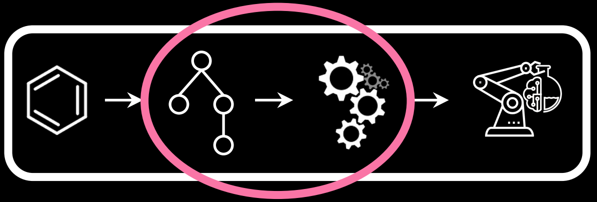
```
Patent sentence:  Diisopropylazodicarboxylate (0.05 ml, 0.302 mmol) was added to the reaction mixture followed by
                  stirring for 3 hours at room temperature.

Rule-based model: ADD Diisopropylazodicarboxylate (0.05 ml, 0.302 mmol); STIR for 3 hours at room temperature.
ML model:         ADD Diisopropylazodicarboxylate (0.05 ml, 0.302 mmol); STIR for 3 hours at room temperature.
```

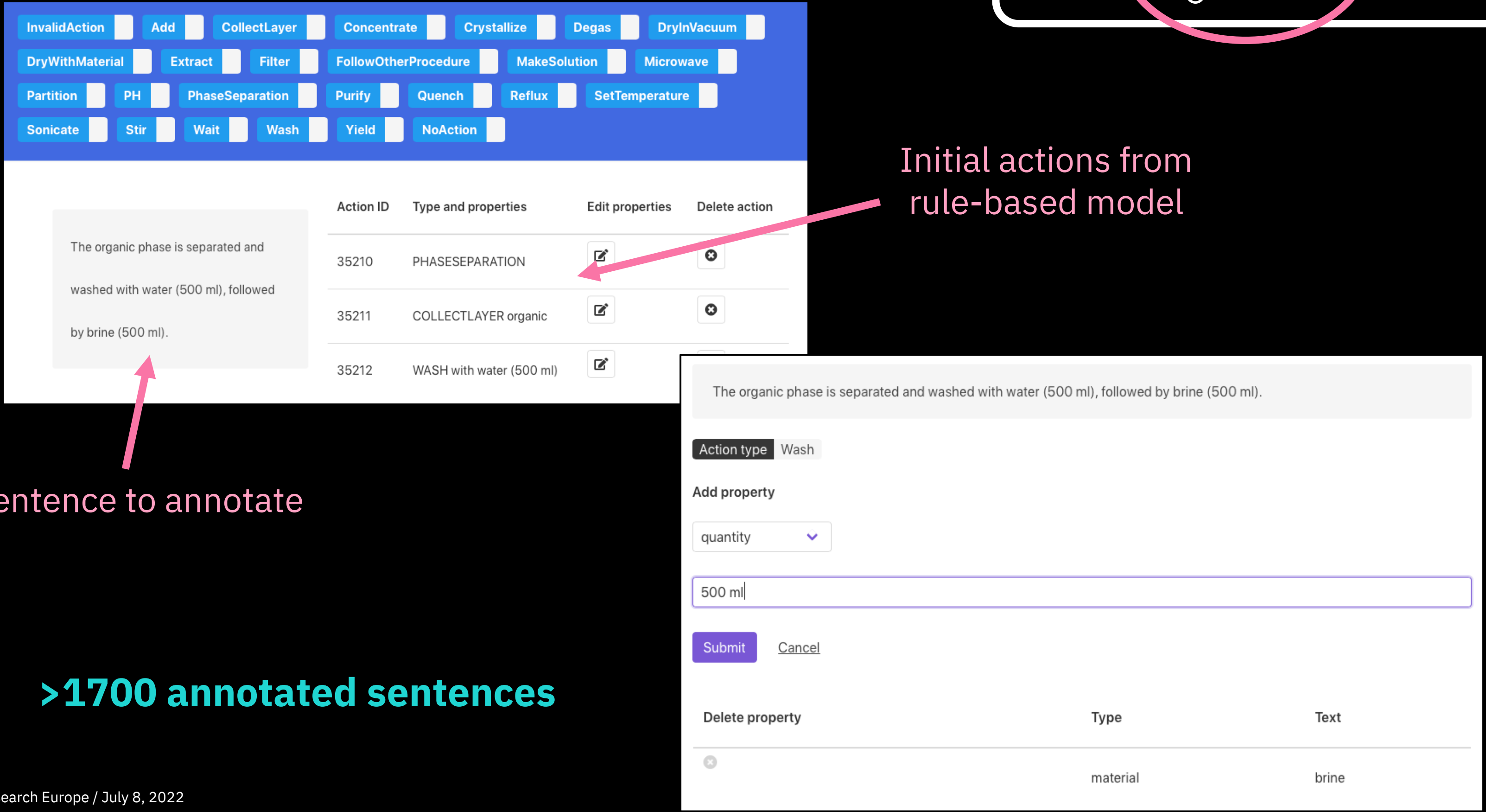
```
Patent sentence:  The reaction mixture was concentrated in vacuo and water was added followed by enough
                  hydrochloric acid (1 M) to acidify the solution.

Rule-based model: CONCENTRATE; ADD water.
ML model:         CONCENTRATE; ADD water.
Ground truth:     CONCENTRATE; ADD water; PH with hydrochloric acid (1 M) to pH acidic.
```

Improving ML model:

- Training data size / quality
- Refine on human annotation

Hand-annotated data





The organic phase is separated and washed with water (500 ml), followed by brine (500 ml).

Action ID	Type and properties	Edit properties	Delete action
35210	PHASESEPARATION		
35211	COLLECTLAYER organic		
35212	WASH with water (500 ml)		

Initial actions from rule-based model

Sentence to annotate

The organic phase is separated and washed with water (500 ml), followed by brine (500 ml).

Action type Wash

Add property

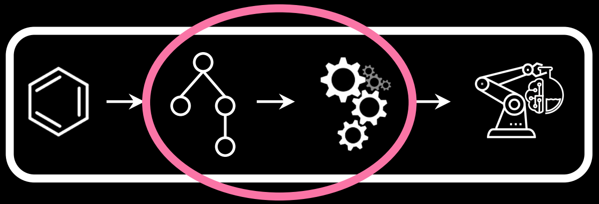
quantity

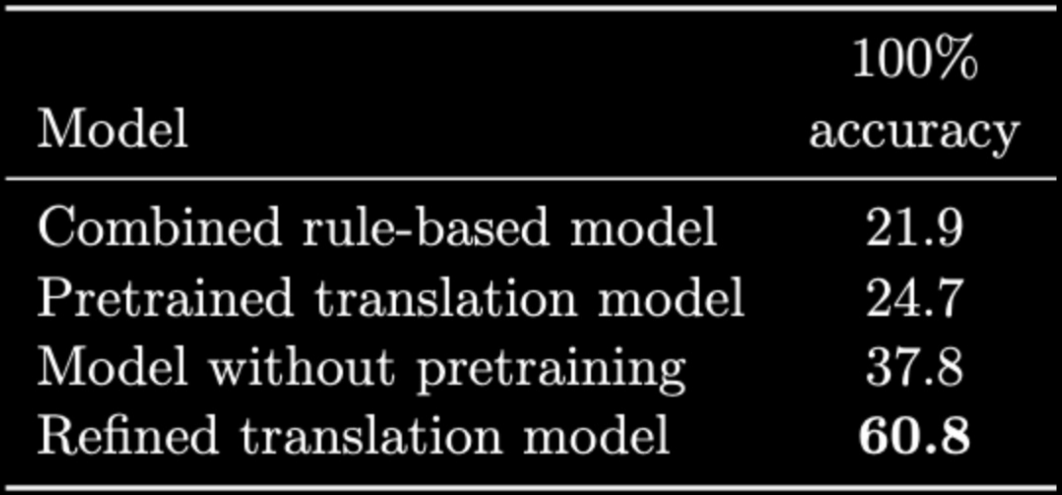
500 ml

Delete property	Type	Text
	material	brine

>1700 annotated sentences

Results

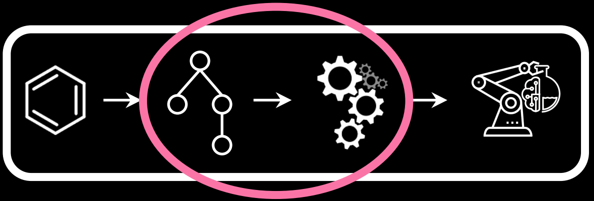


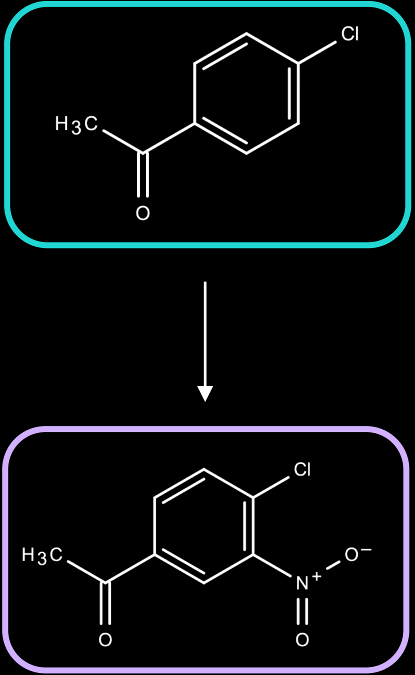


Model	100% accuracy
Combined rule-based model	21.9
Pretrained translation model	24.7
Model without pretraining	37.8
Refined translation model	**60.8**

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

SMILES-to-actions dataset

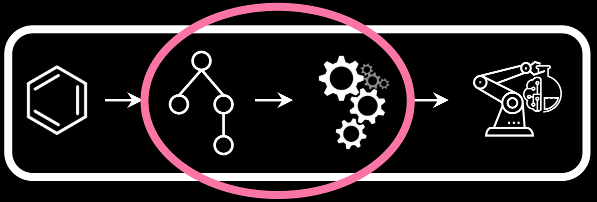


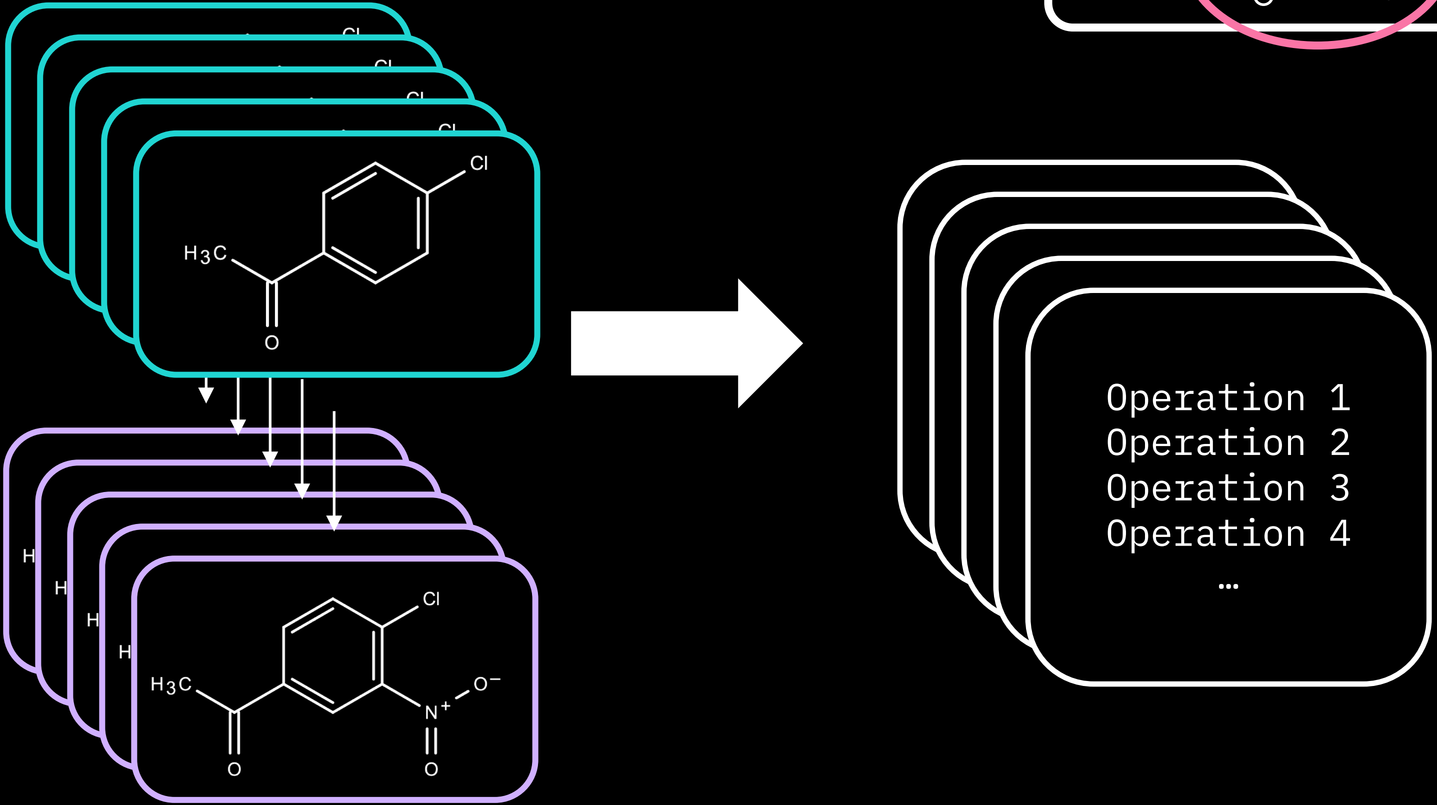


Operation 1
Operation 2
Operation 3
Operation 4
…

Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T., *Nat. Commun.* **2021**, *21*, 2573.

SMILES-to-actions dataset

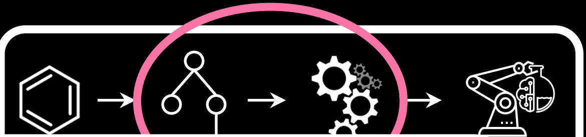


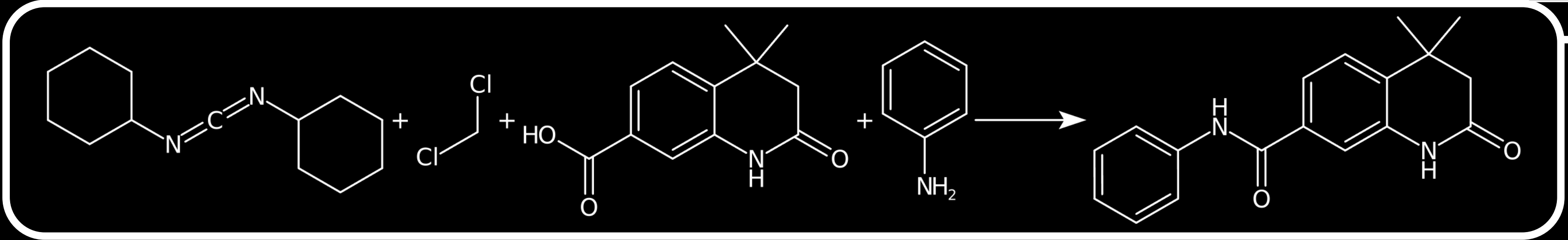


Operation 1
Operation 2
Operation 3
Operation 4
...

Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T., *Nat. Commun.* **2021**, *21*, 2573.

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`

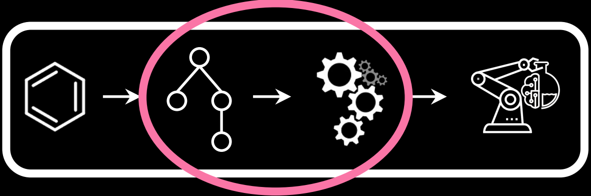
2.7 g (12.3 mmol) 4,4-Dimethyl-1,2,3,4-tetrahydro-2-oxo-7-quinolinecarboxylic acid were added to a solution of 3.8 g (18.5 mmol) N,N'-dicyclohexylcarbodiimide and 1.1 ml (12.3 mmol) aniline in 80 ml dichloromethane. The reaction mixture was stirred for 4 hours at ambient temperature and the precipitate 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

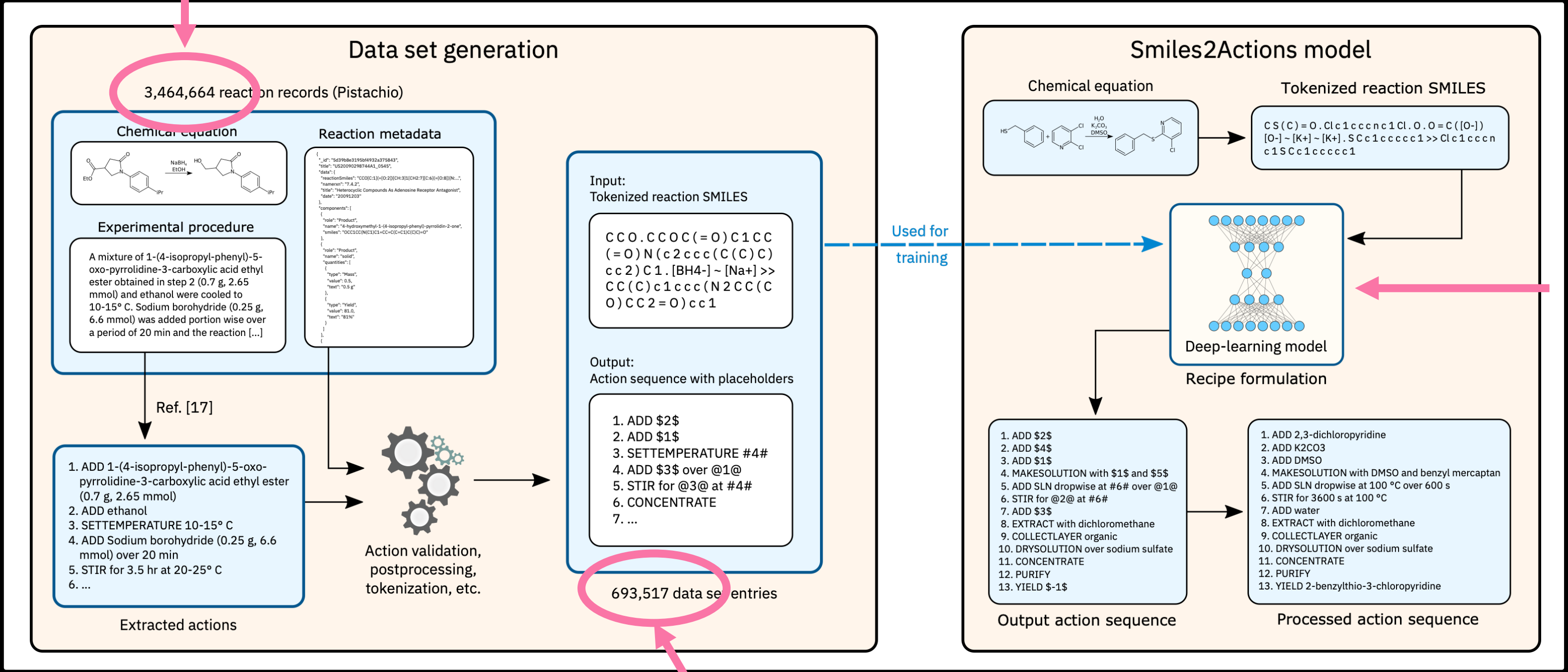
1. MAKESOLUTION ith N,N'-dicyclohexylcar diimide (3.8 g, 18.5 mmol) and anili (1.1 ml, 12.3 mmol) and dichloromethan 80 ml)
2.
3. ADD 4,4-Dimeth 1,2,3,4-tetrahydro-2-oxo-7-quinoline mmol)
4. STIR for 4 hours at ambient temperature
5. FILTER keep precipitate
6. RECRYSTALLIZE from ethanol
7. YIELD title 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\$

SMILES-to-actions



3.5 M



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 [...]

```
{
 "_id": "5d39b8e3195bf4932a375843",
 "title": "US20090298744A1_0545",
 "data": {
  "reactionSmiles": "CCO[C:1](=[O:2])[CH:3]1[CH2:7][C:6](=[O:8])[N:...",
  "namerxn": "7.4.2",
  "title": "Heterocyclic Compounds As Adenosine Receptor Antagonist",
  "date": "20091203"
 },
 "components": [
  {
   "role": "Product",
   "name": "4-hydroxymethyl-1-(4-isopropyl-phenyl)-pyrrolidin-2-one",
   "smiles": "OCC1CC(N(C1)C1=CC=C(C=C1)C(C)C)=O"
  },
  {
   "role": "Product",
   "name": "solid",
   "quantities": [
    {
     "type": "Mass",
     "value": 0.5,
     "text": "0.5 g"
    },
    {
     "type": "Yield",
     "value": 81.0,
     "text": "81%"
    }
   ]
  },
  {
```

1. ADD 1-(4-isopropyl-phenyl)-5-oxo-pyrrolidine-3-carboxylic acid ethyl ester (0.7 g, 2.65 mmol)
2. ADD ethanol
3. SETTEMPERATURE 10-15° C
4. ADD Sodium borohydride (0.25 g, 6.6 mmol) over 20 min
5. STIR for 3.5 hr at 20-25° C
6. ...

```
C C O . C C O C ( = O ) C 1 C C
( = O ) N ( c 2 c c c ( C ( C ) C )
c c 2 ) C 1 . [BH4-] ~ [Na+] >>
C C ( C ) c 1 c c c ( N 2 C C ( C
O ) C C 2 = O ) c c 1
```

1. ADD \$2\$
2. ADD \$1\$
3. SETTEMPERATURE #4#
4. ADD \$3\$ over @1@
5. STIR for @3@ at #4#
6. CONCENTRATE
7. ...

```
C S ( C ) = O . Cl c 1 c c c n c 1 Cl . O . O = C ( [O-] )
[O-] ~ [K+] ~ [K+] . S C c 1 c c c c c 1 >> Cl c 1 c c c n
c 1 S C c 1 c c c c c 1
```

1. ADD \$2\$
2. ADD \$4\$
3. ADD \$1\$
4. MAKESOLUTION with \$1\$ and \$5\$
5. ADD SLN dropwise at #6# over @1@
6. STIR for @2@ at #6#
7. ADD \$3\$
8. EXTRACT with dichloromethane
9. COLLECTLAYER organic
10. DRYSOLUTION over sodium sulfate
11. CONCENTRATE
12. PURIFY
13. YIELD \$-1\$

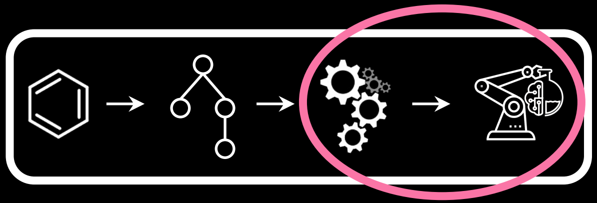
1. ADD 2,3-dichloropyridine
2. ADD K2CO3
3. ADD DMSO
4. MAKESOLUTION with DMSO and benzyl mercaptan
5. ADD SLN dropwise at 100 °C over 600 s
6. STIR for 3600 s at 100 °C
7. ADD water
8. EXTRACT with dichloromethane
9. COLLECTLAYER organic
10. DRYSOLUTION over sodium sulfate
11. CONCENTRATE
12. PURIFY
13. YIELD 2-benzylthio-3-chloropyridine

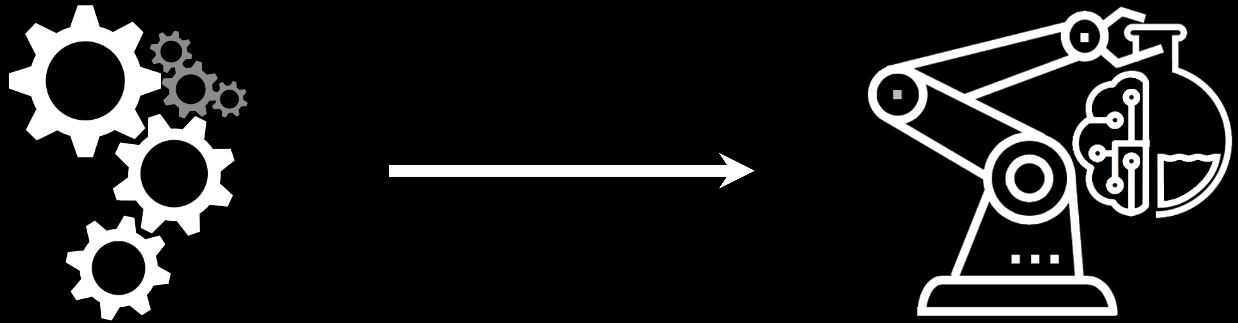
Transformer-based model

0.7 M

Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T., *Nat. Commun.* **2021**, *21*, 2573.

Execution on chemical robot





Add Concentrate Reflux

Filter

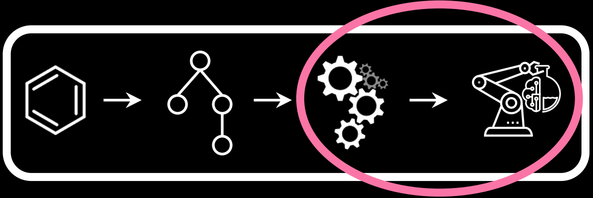
Extract Stir Wash

Start heating reactor #2

Transfer 3 ml of
solvent #5 to reactor #2

Set pressure

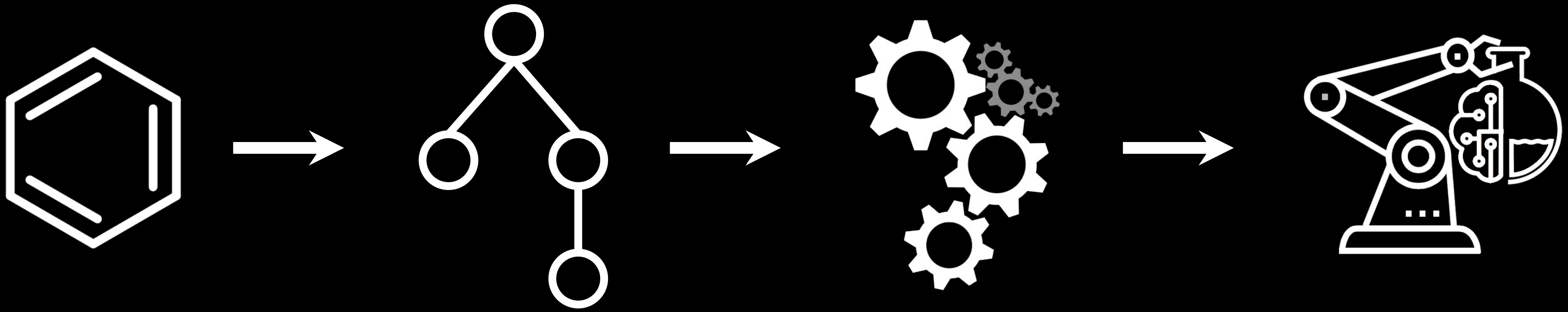
Execution on chemical robot



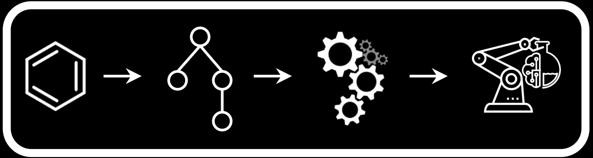
Cloud-based setup for autonomous synthesis

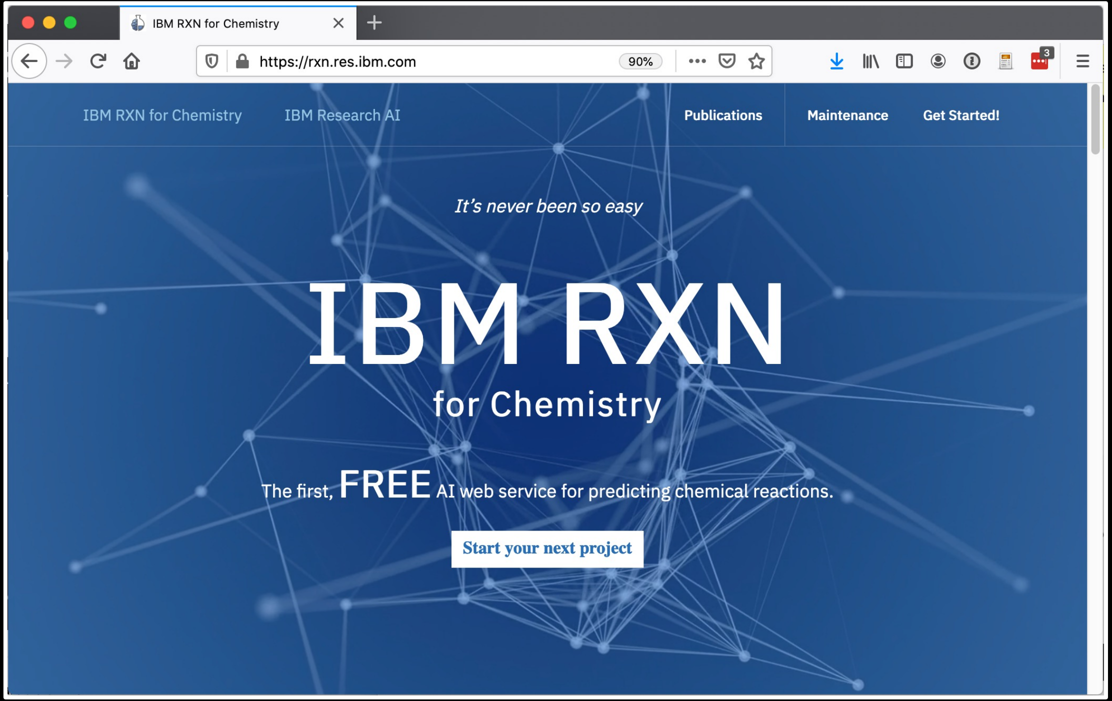
Hardware @ IBM Research Zurich

Summary



IBM RXN





It's never been so easy

IBM RXN

for Chemistry

The first, FREE AI web service for predicting chemical reactions.

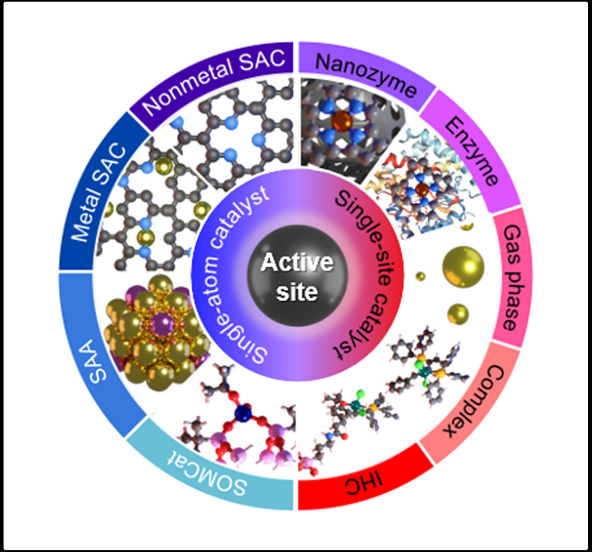
Start your next project

Freely available on:
rxn.res.ibm.com

Demo?

Single Atom Catalysts (SACs)

- Collaboration with aCe group at ETH Zurich (Prof. Pérez-Ramírez)
- Relationship between synthesis and properties?
- Apply action extraction from experimental procedures.
  - New action definitions
  - Fine-tuning of base model
  - Data-driven analysis of relationships



Chem. Rev. **2020**, *120*, 11703–11809 (10.1021/acs.chemrev.0c00576)

Thank you for your attention!

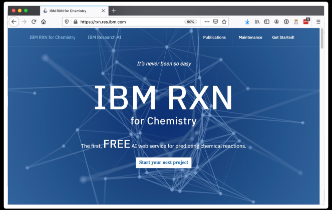
If you have any questions:

E-mail: ava@zurich.ibm.com

Twitter: @acvaucher

Acknowledgments:

Antonio Cardinale
Alessandro Castrogiovanni
Joppe Geluykens
Teodoro Laino
Matteo Manica
Vishnu H. Nair
Philippe Schwaller
Aleksandros Sobczyk
Alessandra Toniato
Heiko Wolf
Federico Zipoli



It's never been so easy

IBM RXN

for Chemistry

The first, FREE AI web service for predicting chemical reactions.

Freely available on:
rxn.res.ibm.com