Fueling the Digital Chemistry Revolution with Language Models

Teodoro Laino
IBM Research Europe – Zurich
NCCR Catalysis

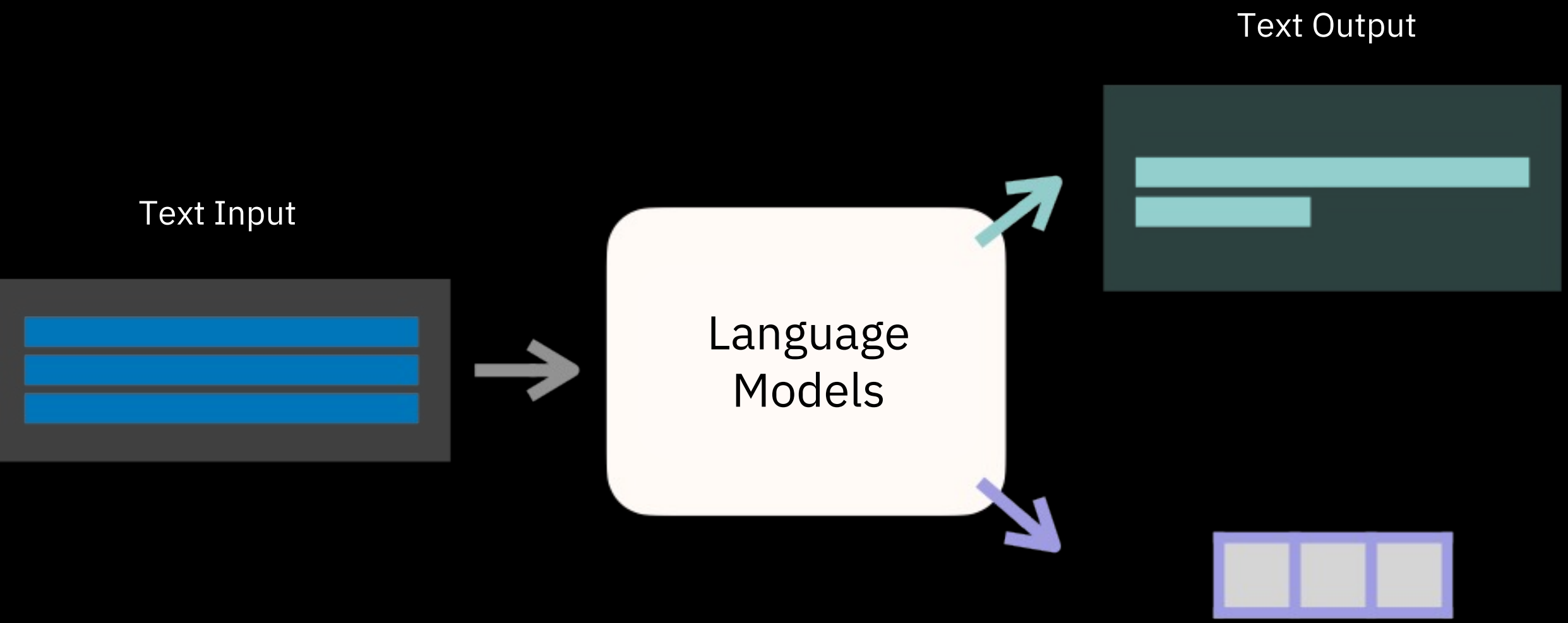
@teodorolaino

Credits

Current Members

Previous Members

What are language models?


Text Output
Text Input
Language Models
Numeric representation of text

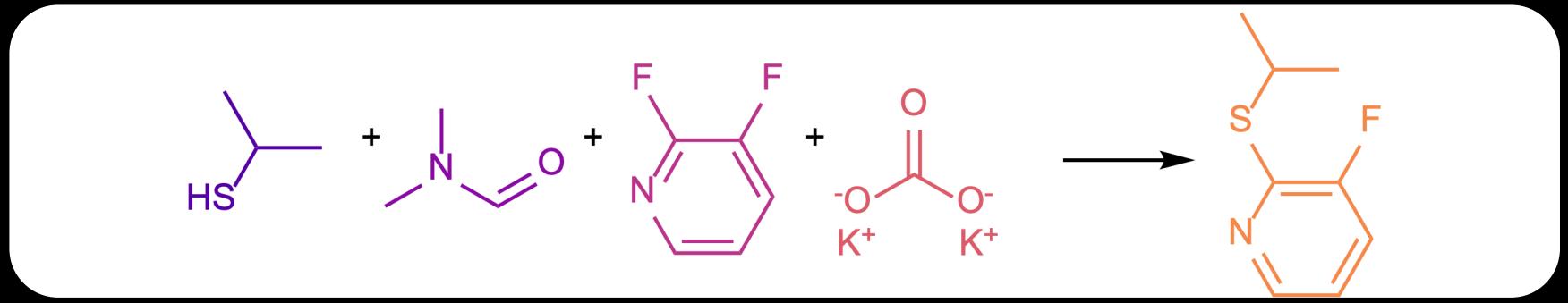
What is a language?

What is a language?

“A language is a purely human and non-intrinsic method of communicating ideas, emotions, and desires by means of voluntarily produced symbols” (Sapir, 1921)

Language and Chemical Tasks

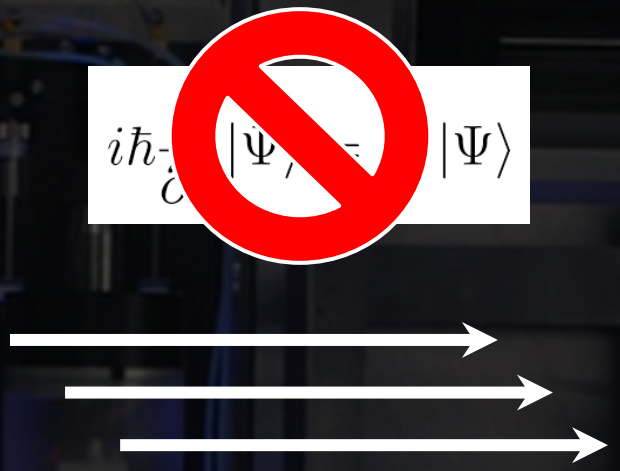
2.7 g (12.3 mmol) 4,4-Dimethyl-1,2,3,4-tetrahydro-2-oxo-7-quinolinecarboxylic acid were added to a solution of 3.8 g (18.5 mmol) N,N'-dicyclohexylcarbodiimide and 1.1 ml (12.3 mmol) aniline in 80 ml dichloromethane. The reaction mixture was stirred for 4 hours at ambient temperature and the precipitate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



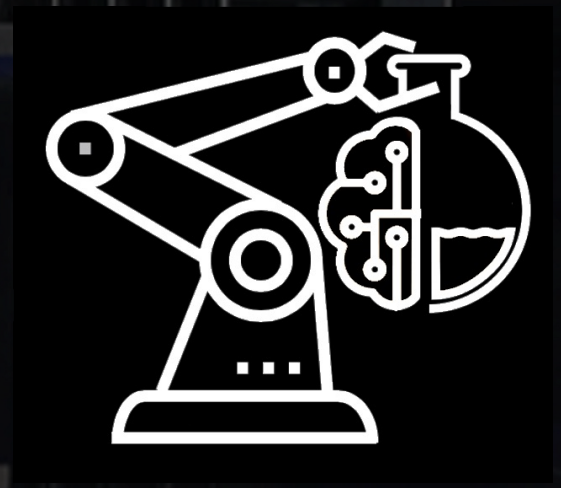
Chemical Synthesis

Data and chemical reactions

Target molecule



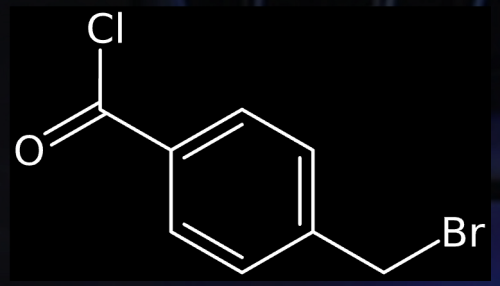
2.7 g (12.3 mmol) 4,4-Dimethyl-1,2,3,4-tetrahydro-2-oxo-7-quinolinecarboxylic acid were added to a solution of 3.8 g (18.5 mmol) N,N'-dicyclohexylcarbodiimide and 1.1 ml (12.3 mmol) aniline in 80 ml dichloromethane. The reaction mixture was stirred for 4 hours at ambient temperature and the precipitate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

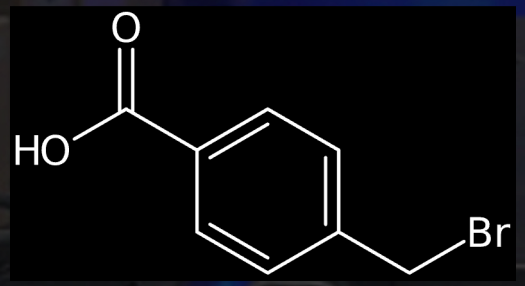


Synthesis execution

Synthesis Design

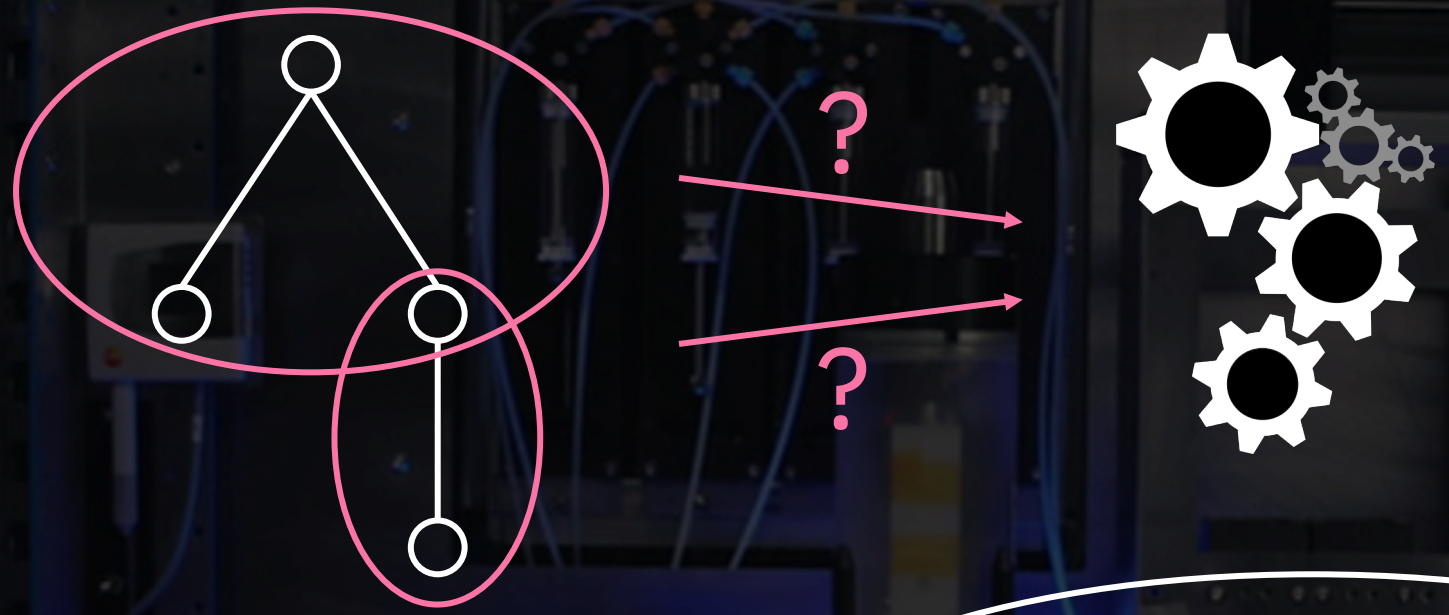
+





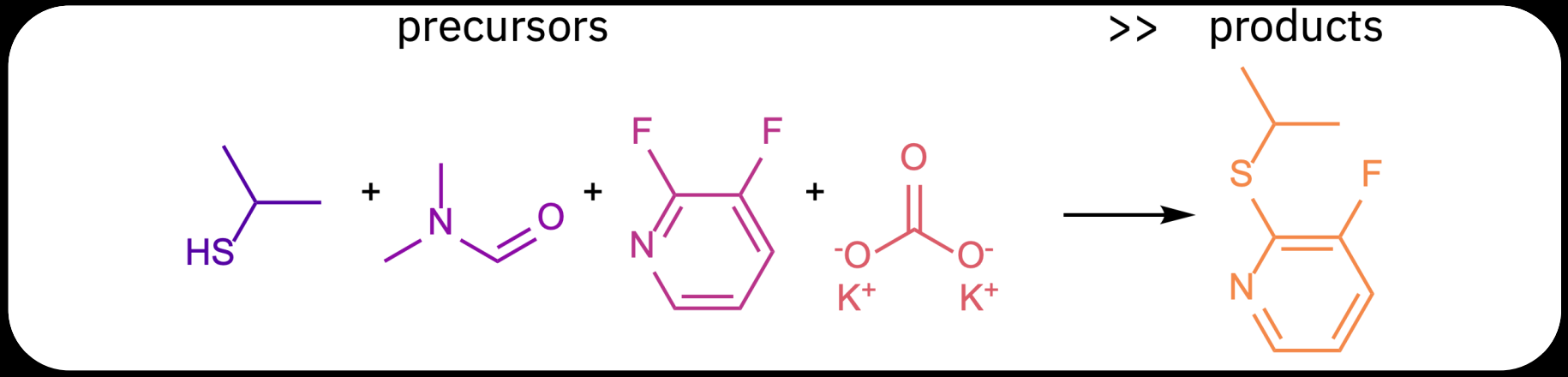
Retrosynthetic tree

Synthesis Execution



Add
Concentrate
Reflux
Filter
Extract
Stir
Wash

Atoms as *letters*, molecules as *words*



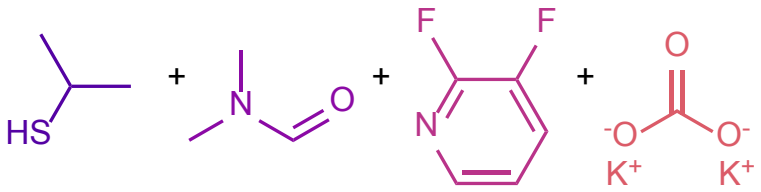
CC(C)S.CN(C)C=O.Fc1cccnc1F.O=C([O-])[O-].[K+].[K+]>>CC(C)Sc1ncccc1F

Cast reaction prediction as translation task

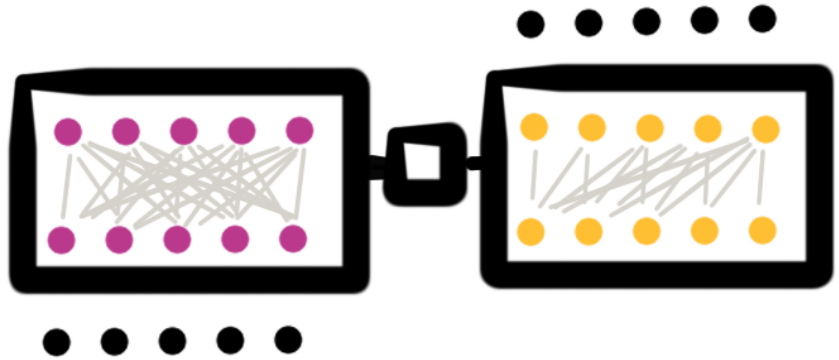
Chem. Sci., 2018, 9, 6091-6098 ; ACS Cent. Sci. 2019, 5, 9, 1572-1583

Molecular Transformer

precursors

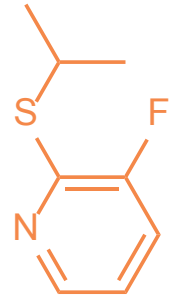


CC(C)S.CN(C)C=O.Fc1ccccnc1F.
O=C([O-])[O-].[K+].[K+]



encoder **decoder**

products

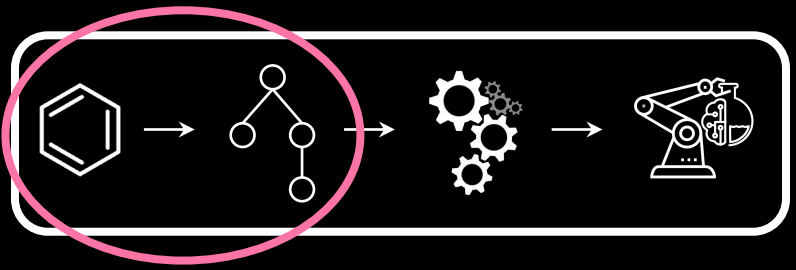


CC(C)Sc1ncccc1F

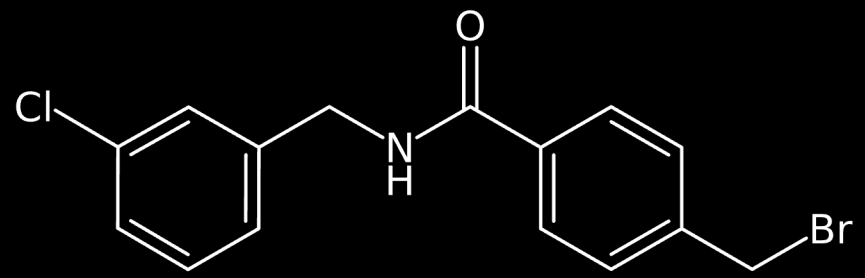
- **No rules** integrated / no chemical knowledge
- **Accurate predictions** on unseen reactions (**>90% accuracy** on benchmark)
- Better than rule and graph-based approaches

Chem. Sci., 2018, 9, 6091-6098 ; ACS Cent. Sci. 2019, 5, 9, 1572-1583

Synthesis Design



?

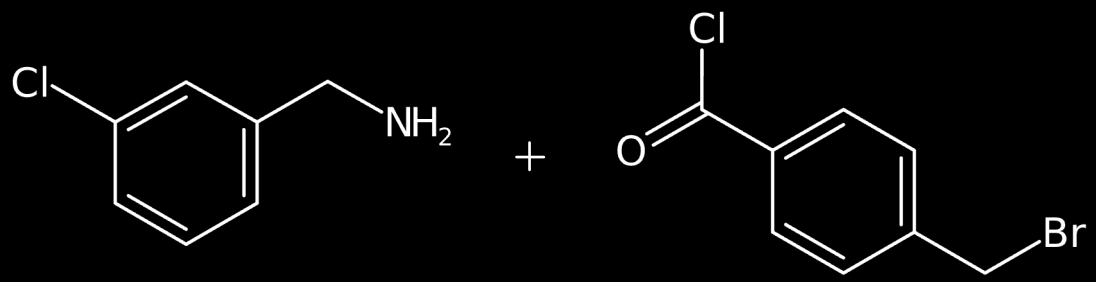


Similar approach, both sides switched

“Translation”

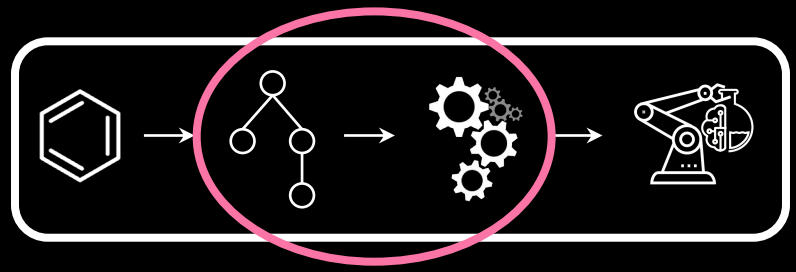
Transformer

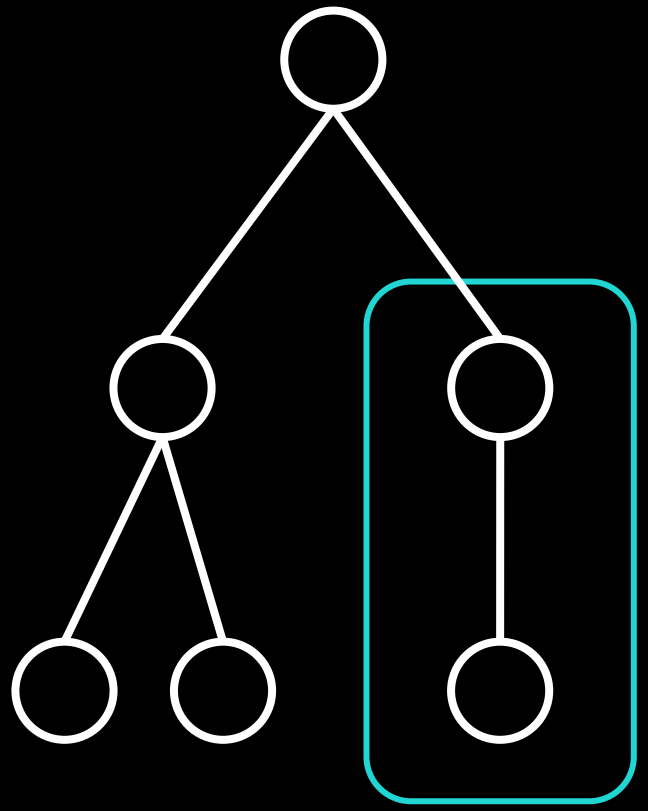
O = C (N C c 1 c c c c (Cl) c 1) c 1 c c c (C Br) c c 1 ⟹ N C c 1 c c c c (Cl) c 1 . O = C (Cl) c 1 c c c (C Br) c c 1



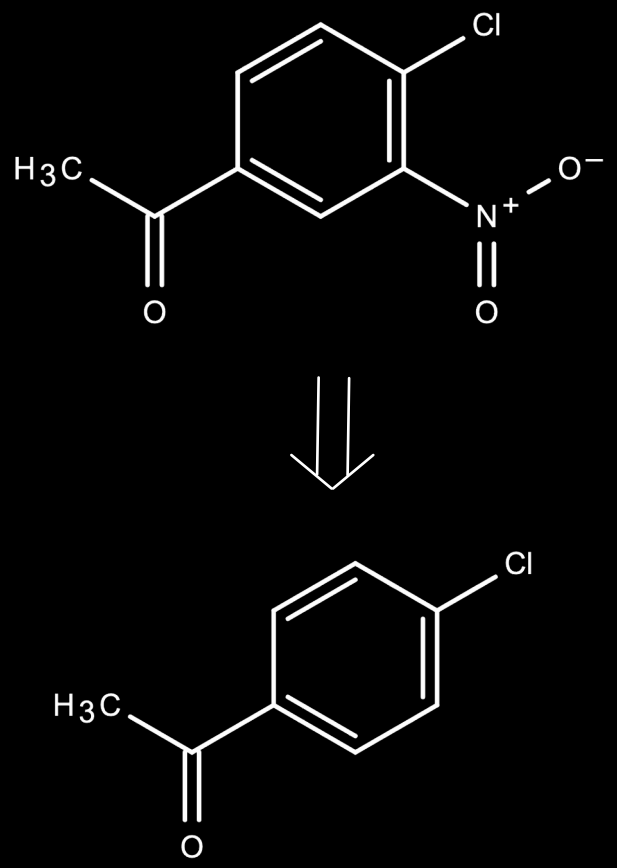
Chem. Sci., 2020, 11, 3316-3325

Synthesis actions

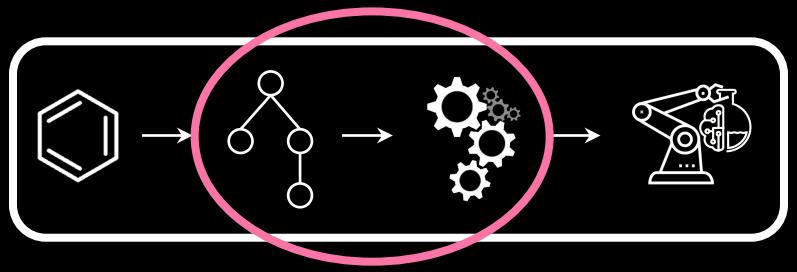




One reaction step



Building a dataset for ML model



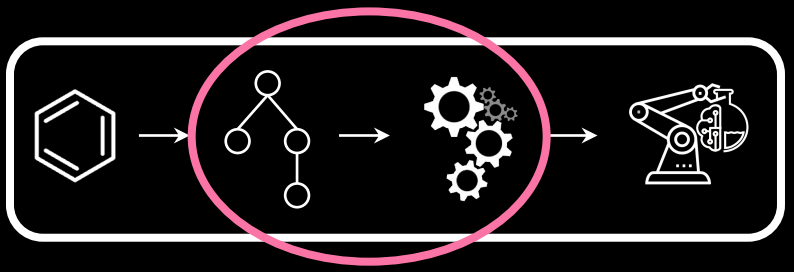
The TFA was removed in vacuo and a saturated solution of NaHCO3 was added.

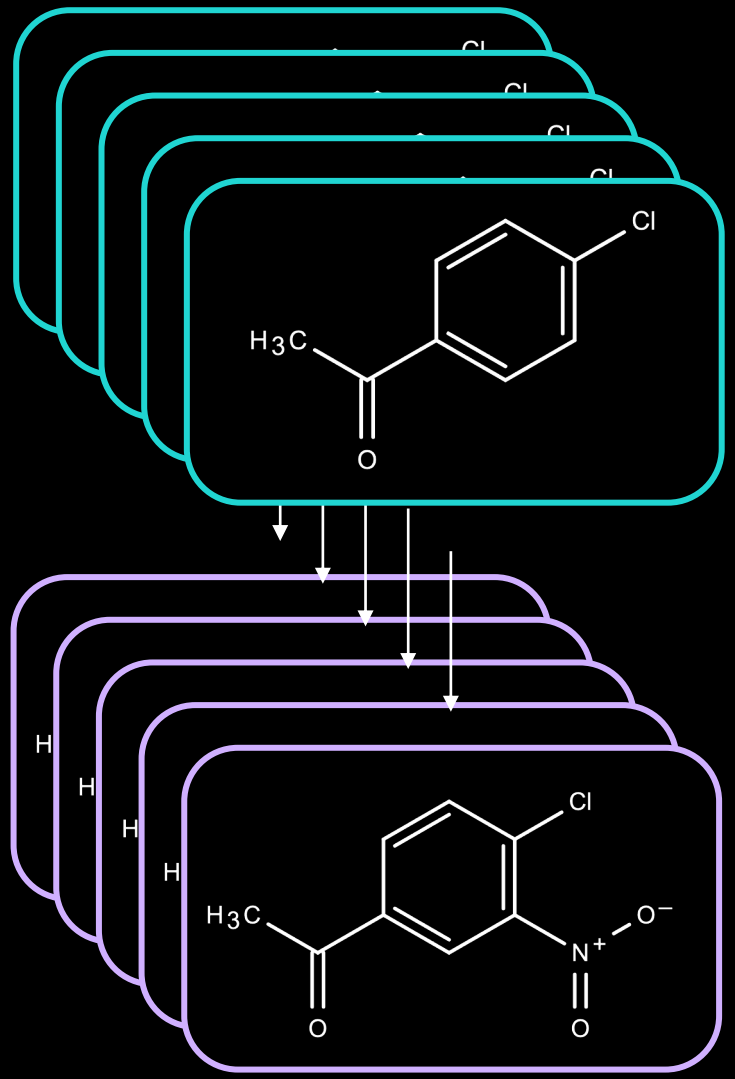
Translation

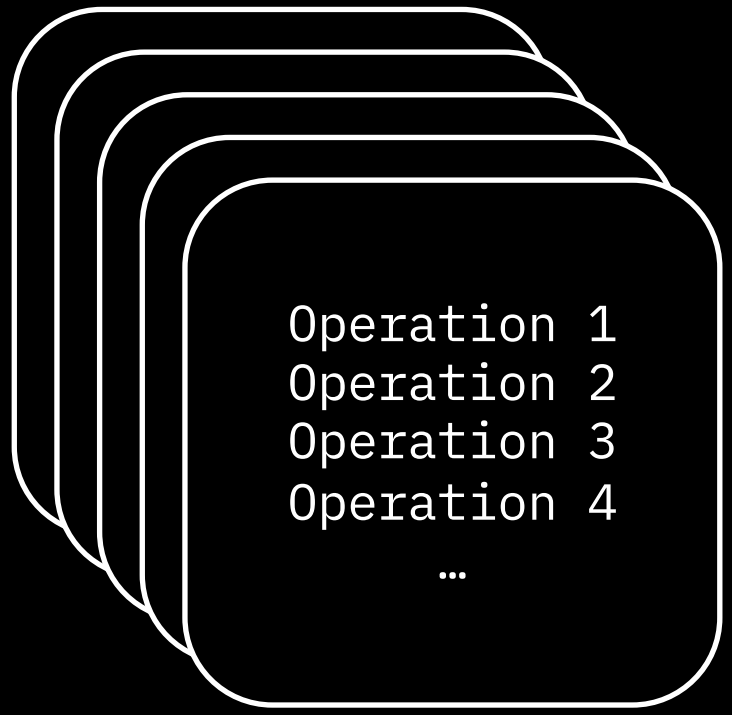
```
Concentrate(),
Add(name='saturated solution of NaHCO3')
```

Nat. Comm., 2020, 11, 3601

SMILES-to-actions dataset

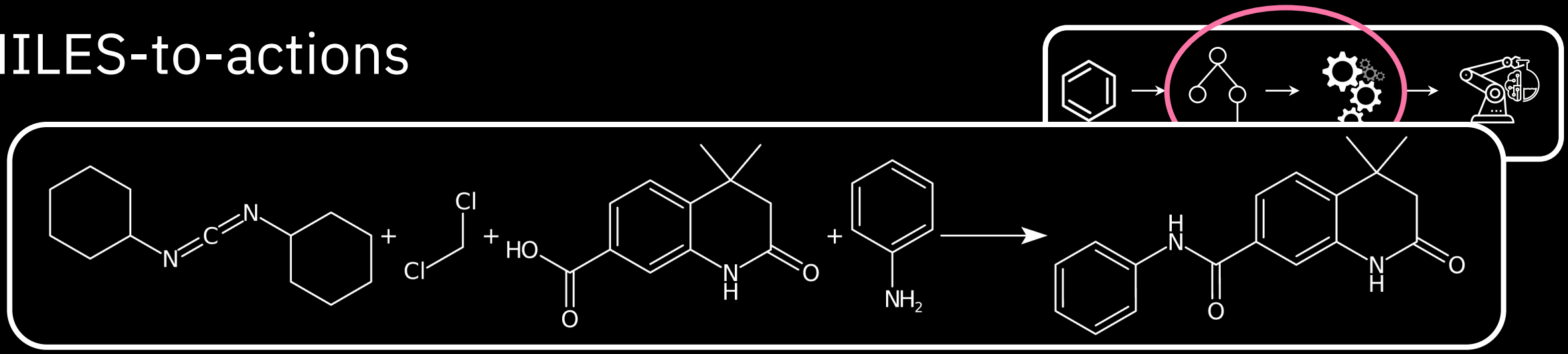






Nat. Comm., 2021, 12, 2573

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.

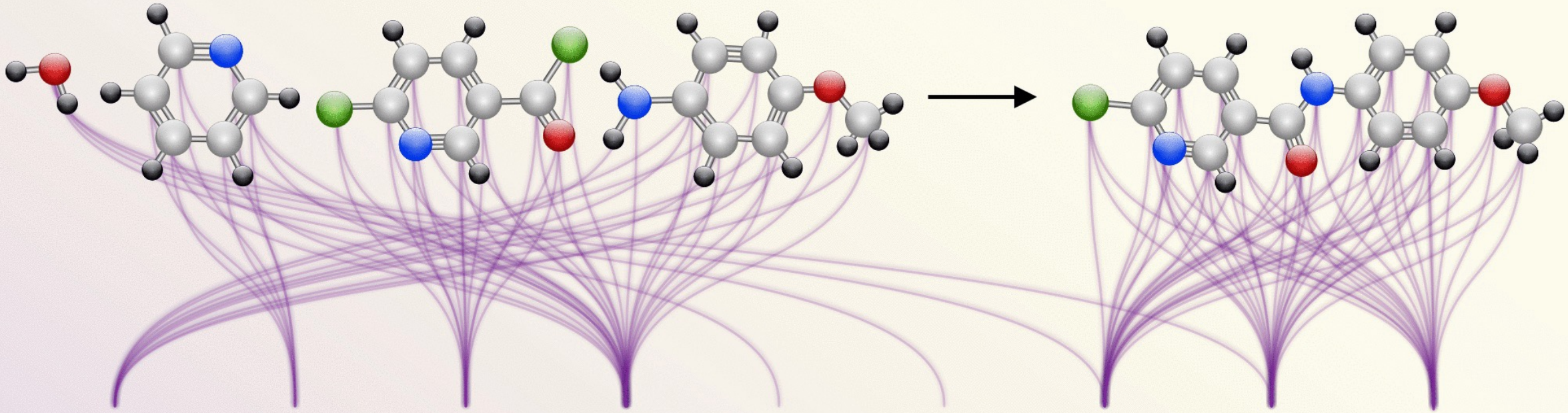
ML model

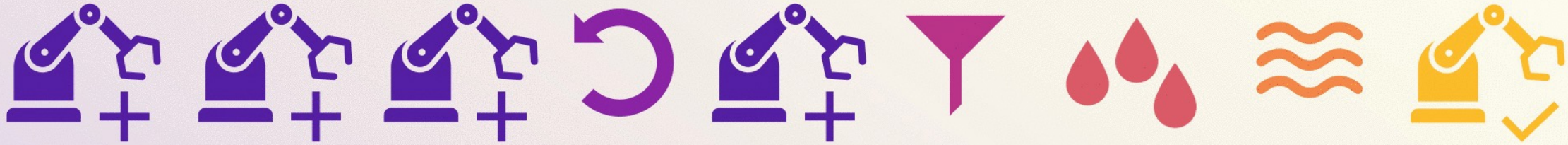
1. MAKESOLUTION ith N,N'-
   dicyclohexylcarb iimide (3.8 g, 18.5 mmol)
   and aniline (1.1 12.3 mmol) and
   dichloromethan 0 ml)
2. [illegible]
3. ADD 4,4-Dimeth 1,2,3,4-tetrahydro-2-
   oxo-7-quinoline
   mmol)
4. STIR for 4 hours at ambient temperature
5. FILTER keep precipitate
6. RECRYSTALLIZE 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

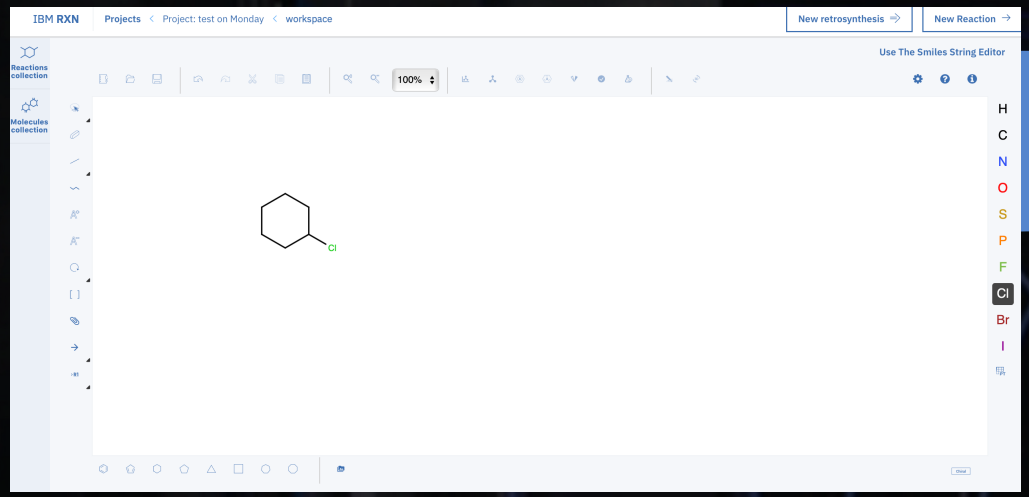
Nat. Comm., 2021, 12, 2573

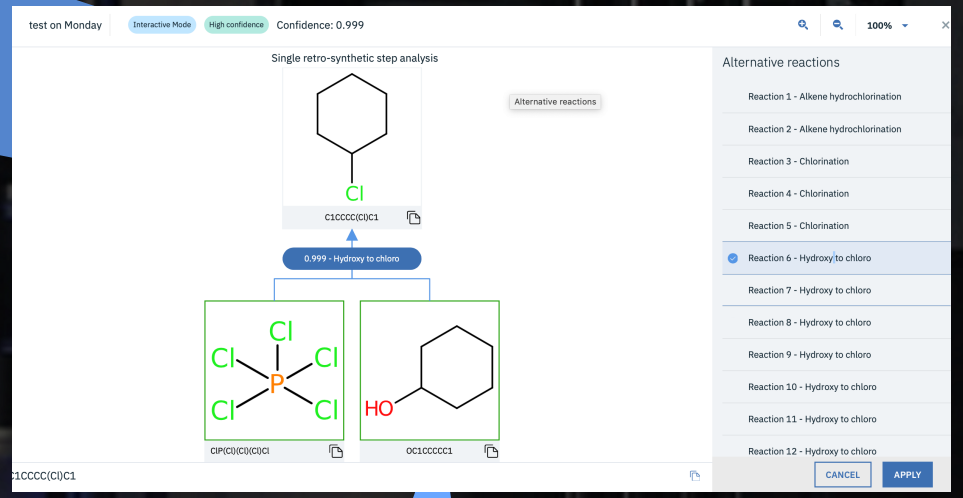
O . c1ccncc1 . O=C(Cl)c1ccc(Cl)nc1 . COc1ccc(N)cc1 >> COc1ccc(NC(=O)c2ccc(Cl)nc2)cc1

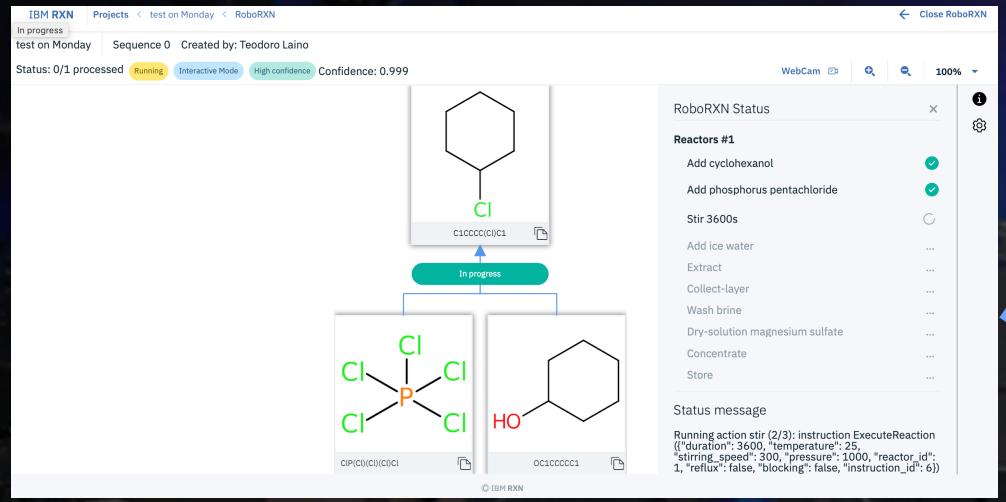


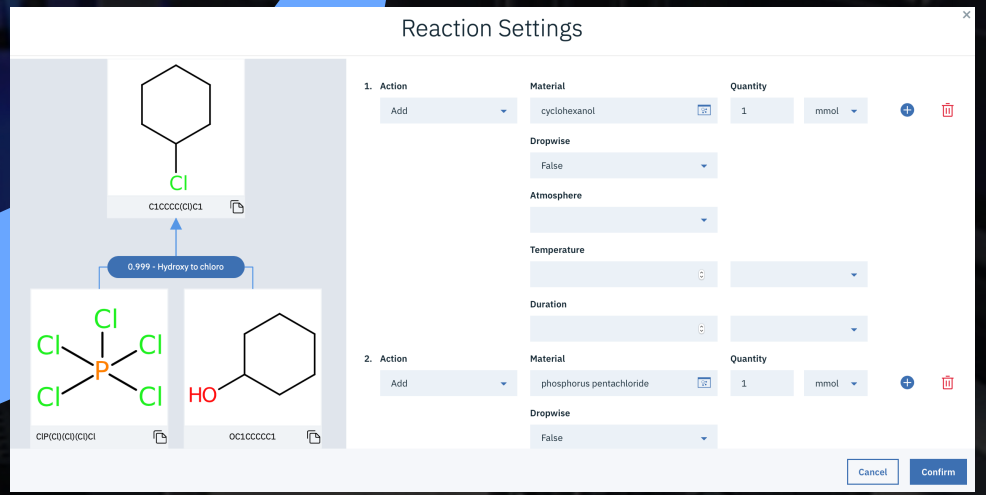


SMILES-2-ACTIONS

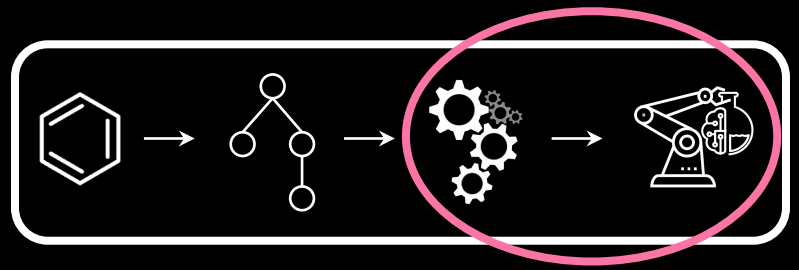

IBM RXN
Projects
Project: test on Monday
workspace
New retrosynthesis
New Reaction
Use The Smiles String Editor
Reactions collection
Molecules collection
100%
H
C
N
O
S
P
F
Cl
Br
I
Cl

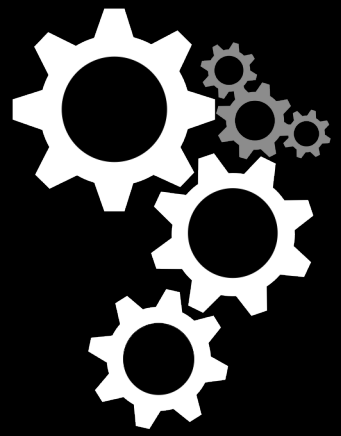

test on Monday
Interactive Mode
High confidence
Confidence: 0.999
100%
Single retro-synthetic step analysis
Alternative reactions
Alternative reactions
Reaction 1 - Alkene hydrochlorination
Reaction 2 - Alkene hydrochlorination
Reaction 3 - Chlorination
Reaction 4 - Chlorination
Reaction 5 - Chlorination
Reaction 6 - Hydroxy to chloro
Reaction 7 - Hydroxy to chloro
Reaction 8 - Hydroxy to chloro
Reaction 9 - Hydroxy to chloro
Reaction 10 - Hydroxy to chloro
Reaction 11 - Hydroxy to chloro
Reaction 12 - Hydroxy to chloro
C1CCCC(Cl)C1
0.999 - Hydroxy to chloro
ClP(Cl)(Cl)(Cl)Cl
OC1CCCCC1
CANCEL
APPLY

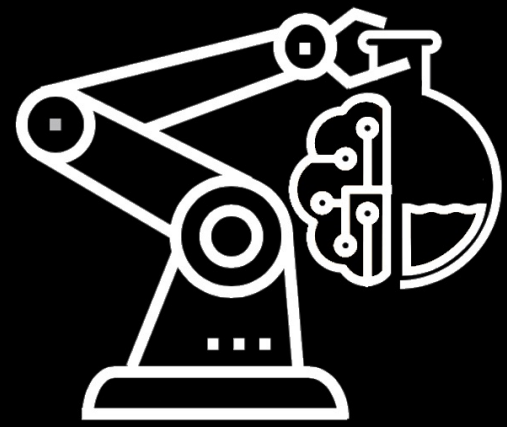

IBM RXN
Projects
test on Monday
RoboRXN
Close RoboRXN
In progress
test on Monday
Sequence 0
Created by: Teodoro Laino
Status: 0/1 processed
Running
Interactive Mode
High confidence
Confidence: 0.999
WebCam
100%
C1CCCC(Cl)C1
In progress
ClP(Cl)(Cl)(Cl)Cl
OC1CCCCC1
RoboRXN Status
Reactors #1
Add cyclohexanol
Add phosphorus pentachloride
Stir 3600s
Add ice water
Extract
Collect-layer
Wash brine
Dry-solution magnesium sulfate
Concentrate
Store
Status message
Running action stir (2/3): instruction ExecuteReaction ({"duration": 3600, "temperature": 25, "stirring_speed": 300, "pressure": 1000, "reactor_id": 1, "reflux": false, "blocking": false, "instruction_id": 6})
IBM RXN


Reaction Settings
C1CCCC(Cl)C1
0.999 - Hydroxy to chloro
ClP(Cl)(Cl)(Cl)Cl
OC1CCCCC1
1. Action
Add
Material
cyclohexanol
Quantity
1
mmol
Dropwise
False
Atmosphere
Temperature
Duration
2. Action
Add
Material
phosphorus pentachloride
Quantity
1
mmol
Dropwise
False
Cancel
Confirm

Execution on chemical robot







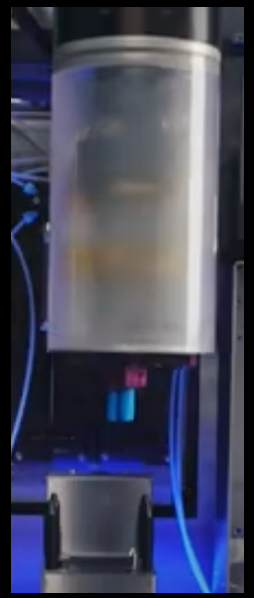
Add
Concentrate
Reflux
Filter
Extract
Stir
Wash

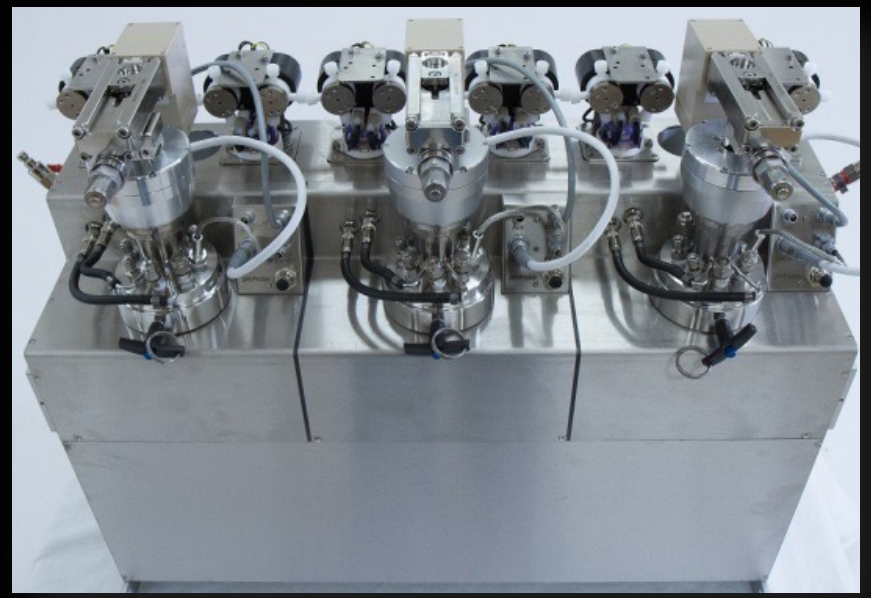
Start heating reactor #2

Transfer 3 ml of
solvent #5 to reactor #2

Set pressure

Flex Autoplant





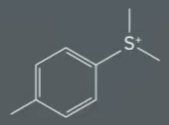
FLEX AUTOPLANT robotic platform

Analytics


Agilent
1D DAD
192.168.254.13
Deck Switch
LC–MS DAD
192.168.254.11
1D Preparative
2D Pump
1D Pump
LC–MS Pump


Dauerversuch!
Bei
Gefahr abschalten.
192.168.254.12

Synthesizing new molecule



Started: Nov 30 2020, 6:49am PT

Live from IBM RoboRXN

Action 2

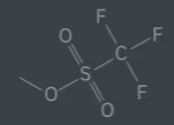
Overview

Adding $C_2H_3F_3O_3S$

In this action, the molecule methyl trifluoromethane sulfonate is added to Reactor 2.

Methyl trifluoromethane
$C_2H_3F_3O_3S$

2D 3D



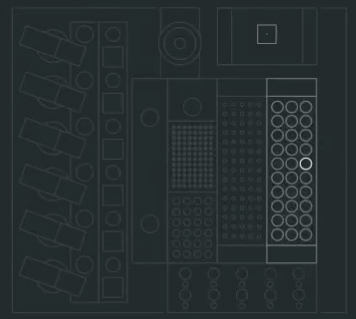
Methyl trifluoromethane sulfonate is a brown liquid. Insoluble in water. This material is a very reactive methylating agent, also known as methyl triflate.

NOW

10 ml of reagent containing methyl trifluoromethane sulfonate is being moved from Vial 61 and added to Reactor 2.

Position of the robot arm

Moving to Vial 61



Live view module

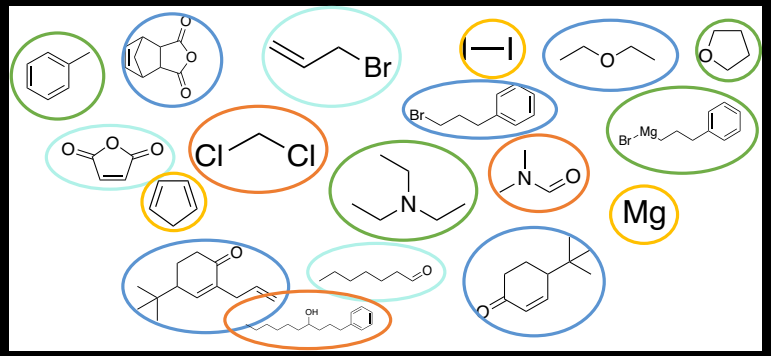
2 Adding $C_2H_3F_3O_3S$

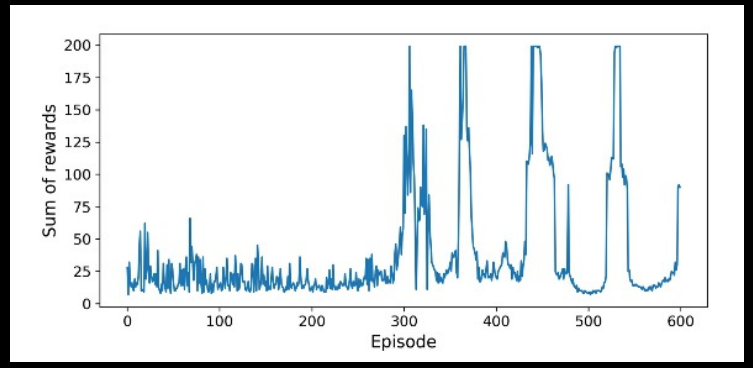
00:06:00

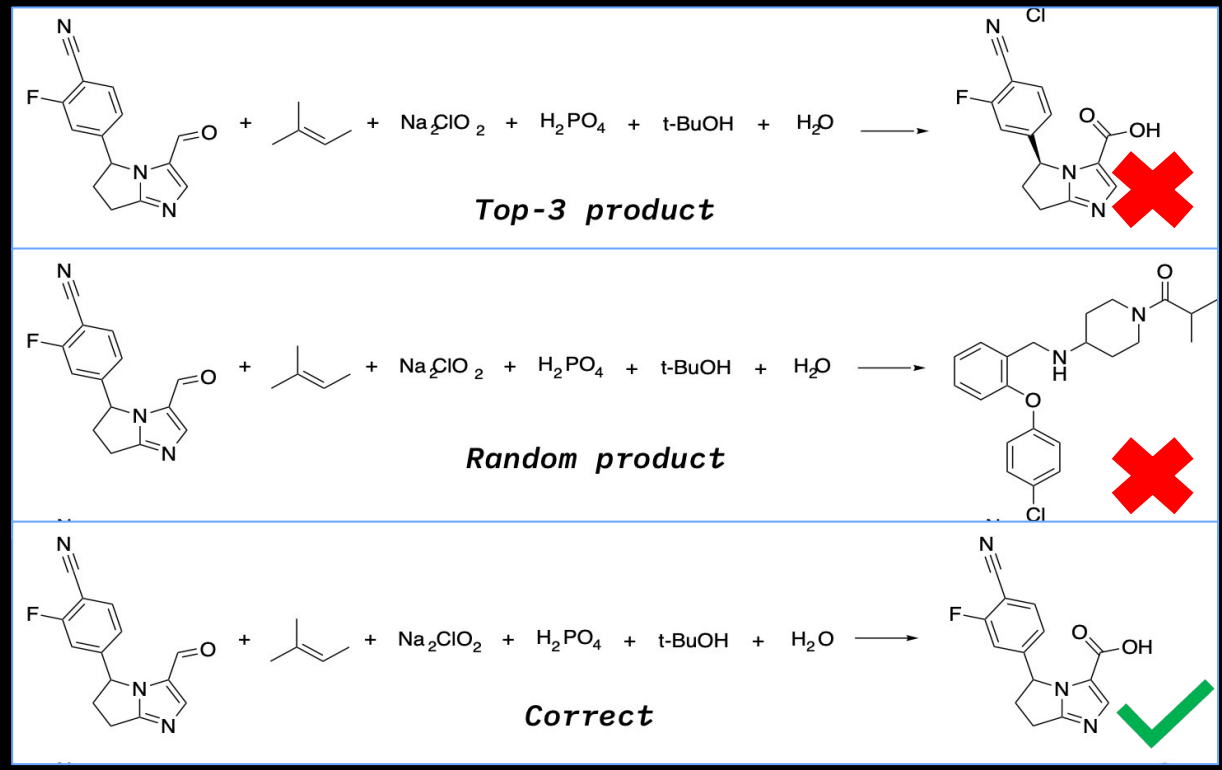
LIVE

Unassisted *data sets curation*

The most **difficult examples to learn** during training of reactions prediction models = likely **examples of wrong chemistry**.

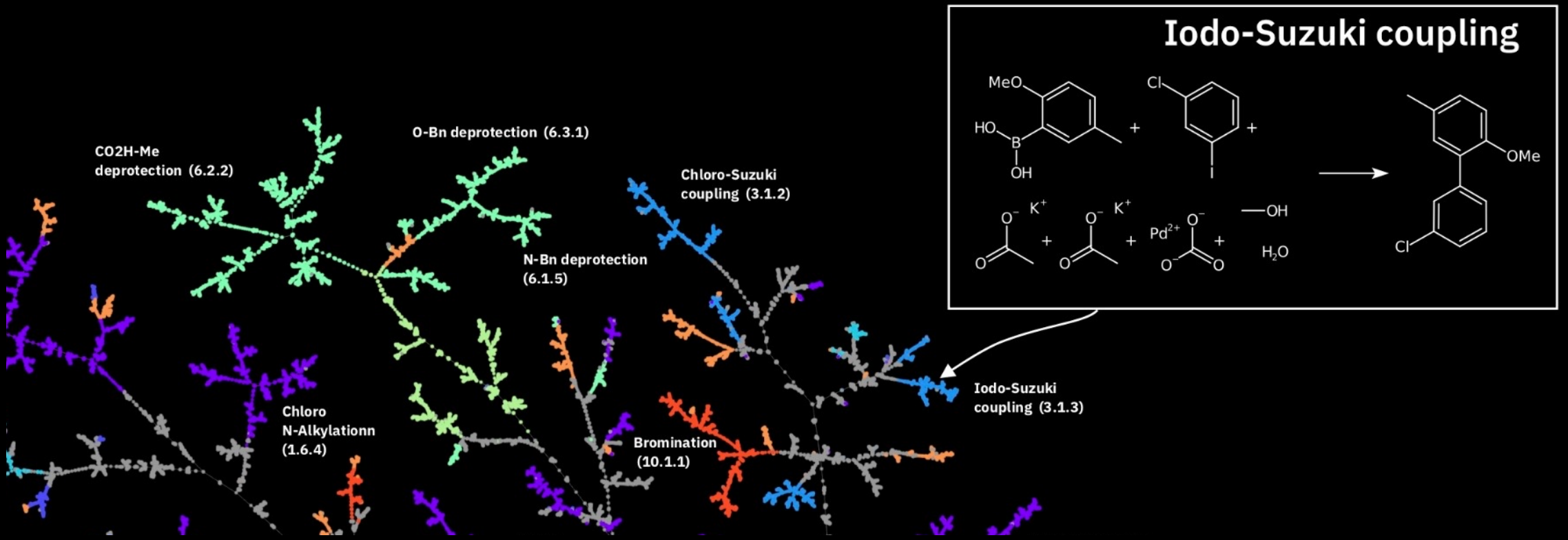






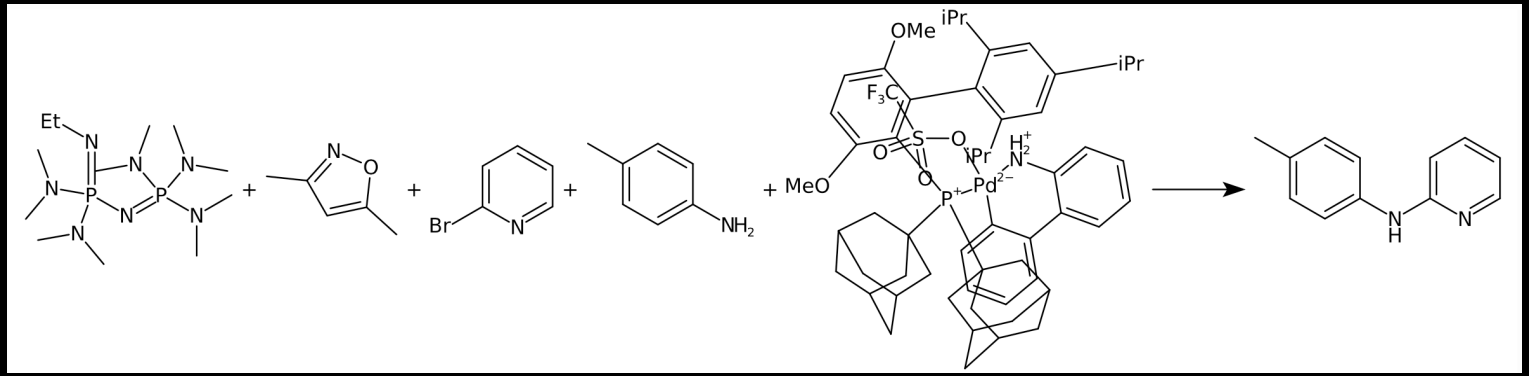
Nat. Mach. Intell. 2021, 3, 485–494

Reaction Classification, atlases and fingerprints



Nat. Mach. Intell., 2021, 3, 144–152

Prediction of reaction yields



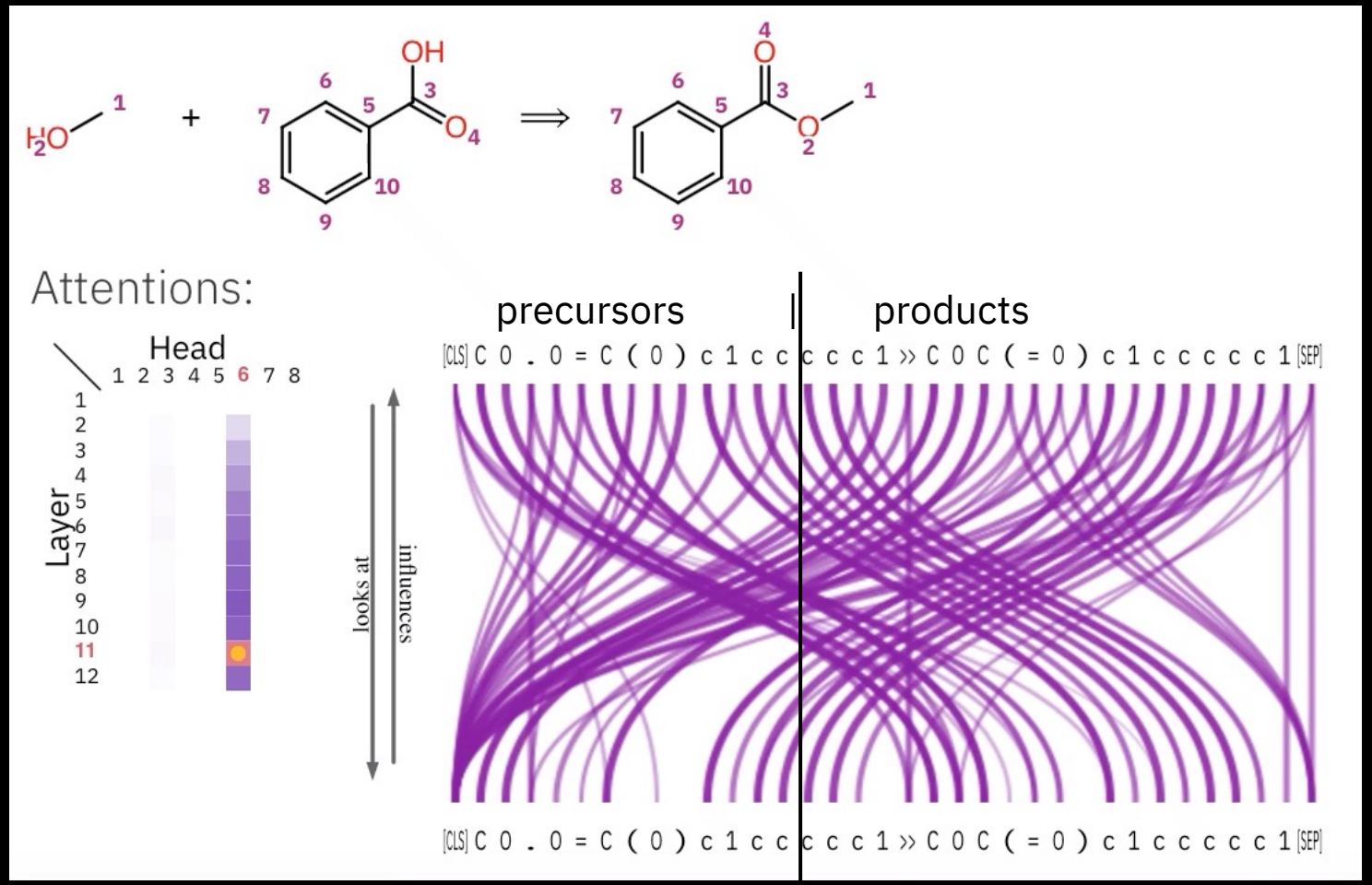
Brc1ccccn1.CCN=P(N=P(N(C)C)(N(C)C)N(C)C)(N(C)C)N(C)C.COc1ccc(OC)c([P+](C23CC4CC(CC(C4)C2)C3)(C23CC4CC(CC(C4)C2)C3)[Pd-2]2(OS(=O)(=O)C(F)(F)F)[NH2+]c3ccccc3-c3ccccc32)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C.Cc1cc(C)on1.Cc1ccc(N)cc1>>Cc1ccc(Nc2ccccn2)cc1

77.6 %

Mach. Learn.: Sci. Technol., 2021, 2, 015016

Atom-Mapping and the learning of chemical reaction grammar (RXNMapper)

CO.O=C(O)c1ccccc1>>COC(=O)c1ccccc1

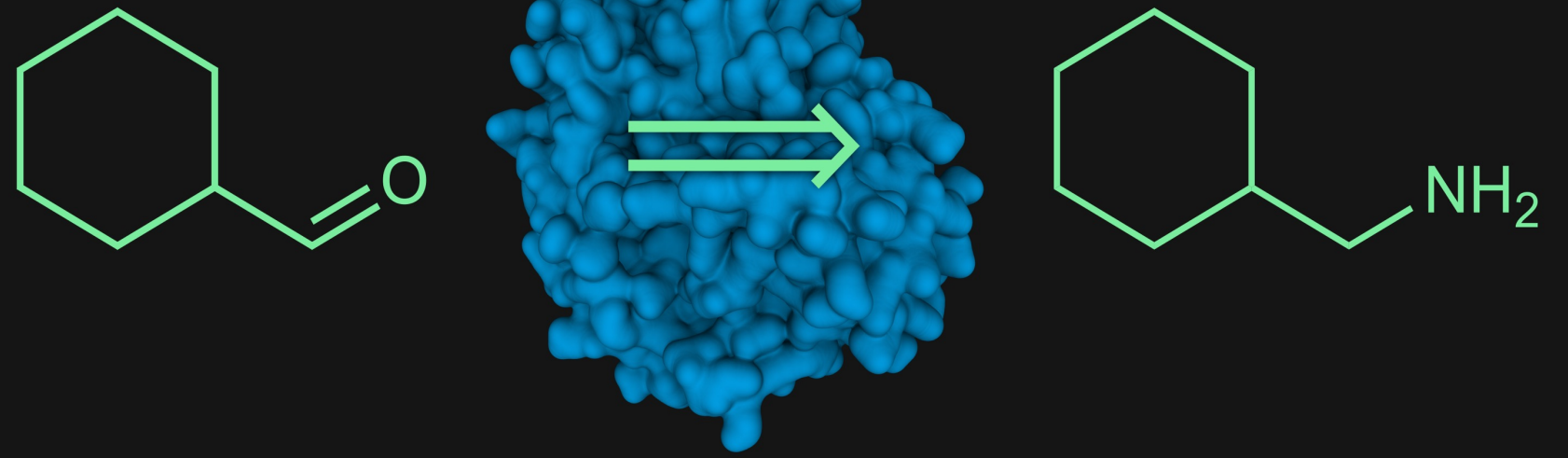


Adv. Science, 2021, 7, 15, eabe4166

Enzymatic catalysis

NCCR Catalysis

GreenCatRXN

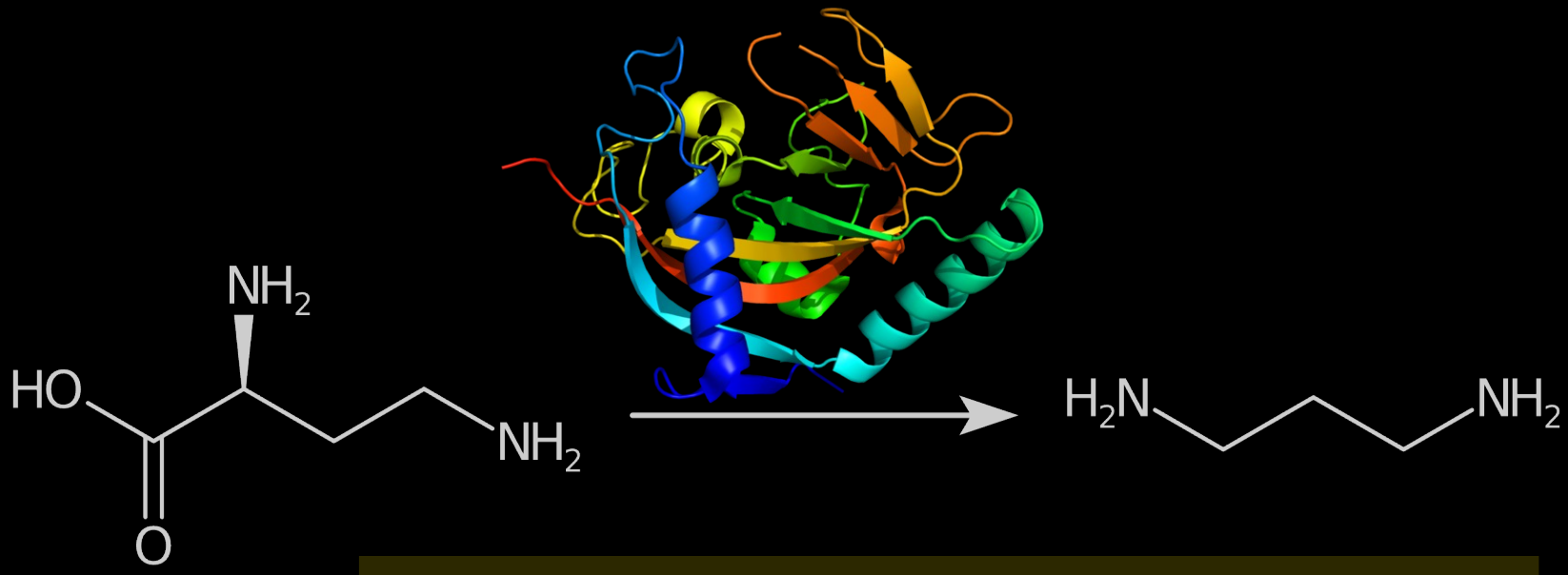


EC 1.4.3.-

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

Nat. Commun. 2022, 13, 964

Recovering active sites (3D) from sequential data (1D)



NCC[C@H](N)C(=O)O | KTYF…SQTSQIHKKDNHIRGQARFCP…YVLK | NCCCN

KTYFVL…LL**ELNENDPG**IFV**TQSVHKQQAGFSQTSQIHKKDNHIRGQAR**FCPHKR…GYVLK

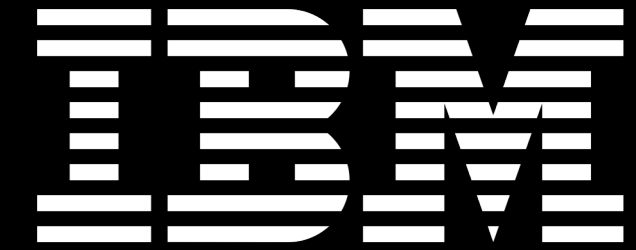
References

Chem. Sci., 2018, 9, 6091-6098
ACS Cent. Sci. 2019, 5, 9, 1572-1583
Chem. Sci., 2020, 11, 3316-3325
Nat. Commun., 2020, 11, 3601
Nat. Commun., 2020, 11, 4874
Nat. Mach. Intell., 2021, 3, 144–152
Nat. Mach. Intell. 2021, 3, 485–494
Adv. Science, 2021, 7, 15, eabe4166
Mach. Learn.: Sci. Technol., 2021, 2, 015016
Nat. Commun., 2021, 12, 2573
Nat. Commun. 2022, 13, 964

Collaborators:

Watch the story of RoboRXN (short): https://youtu.be/ewE1wh7sTUE
Watch the story of RoboRXN (long): https://youtu.be/i2-LgHjgDTs

More information and access/test: https://rxn.res.ibm.com


IBM