Fueling the Digital Chemistry Revolution with Language and Multimodal Foundation Models

Teodoro Laino
IBM Research Europe – Zurich

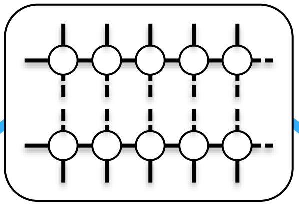
@teodorolaino

IBM

AI breakthroughs for language are changing scientific discovery

Generative modeling and transformers are achieving new breakthroughs in chemistry

Natural Language Processing (NLP)



Transformers

AI

Language

↑

Sentences

↑

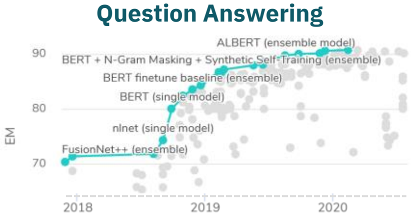
Words

↑

Letters

Data

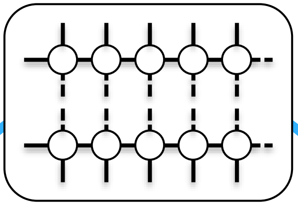
AI Models



Improved Language Tasks

→

Accelerated Chemistry



Transformers

AI

Chemistry

↑

Properties/Reactions

↑

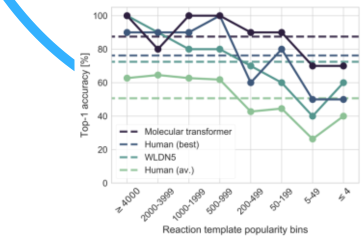
Molecules

↑

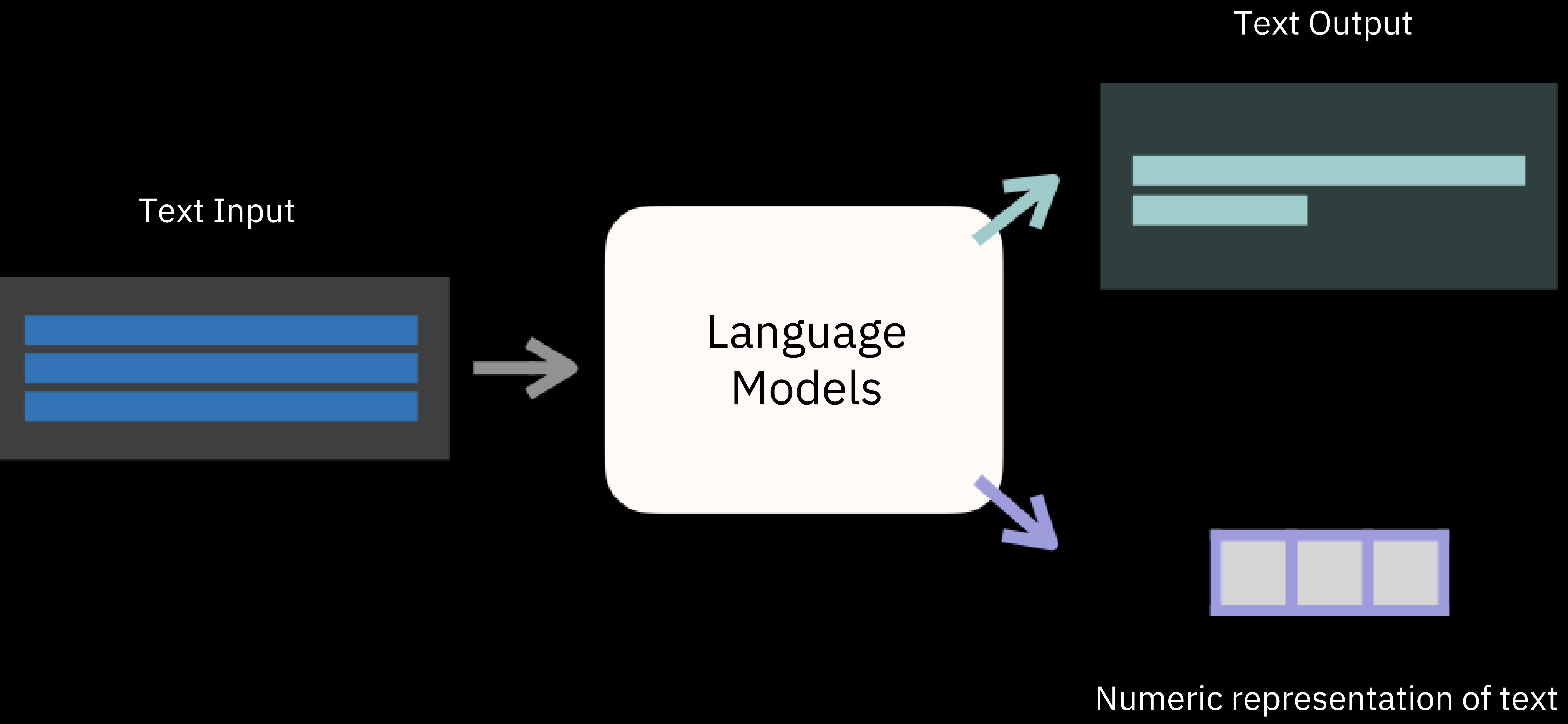
Atoms

Data

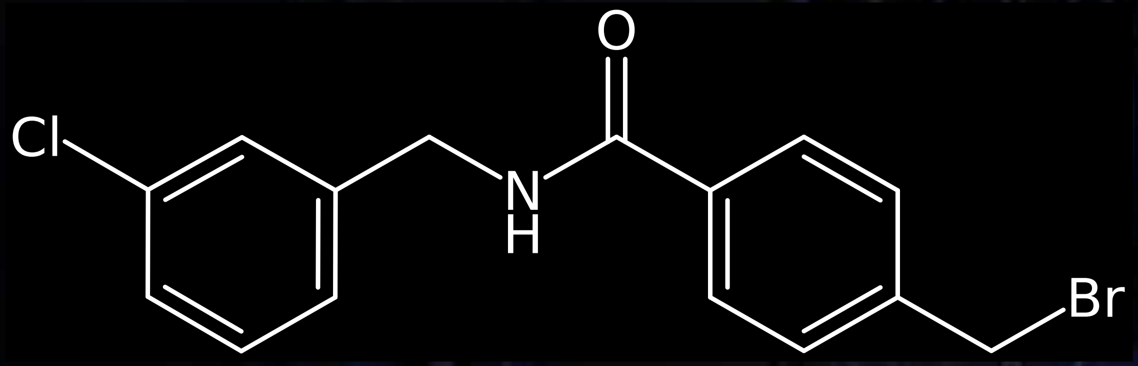
AI Models



Improved Chemistry Tasks


Text Output
Text Input
Language Models
Numeric representation of text

Data and chemical reactions

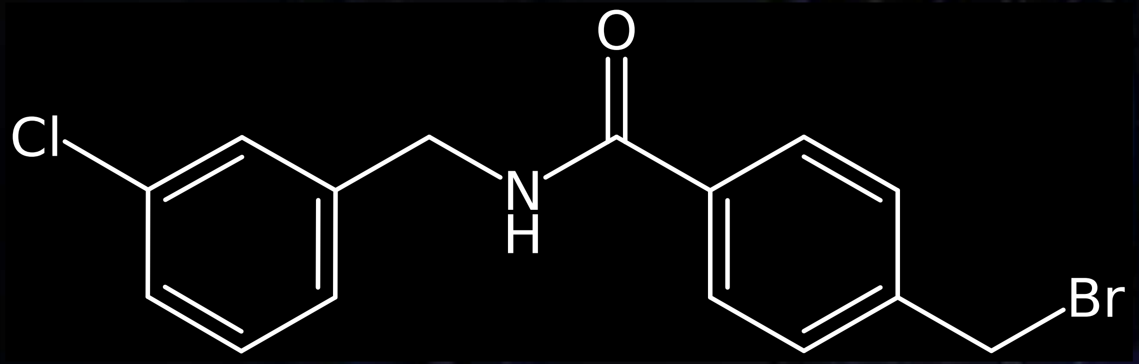


Target molecule

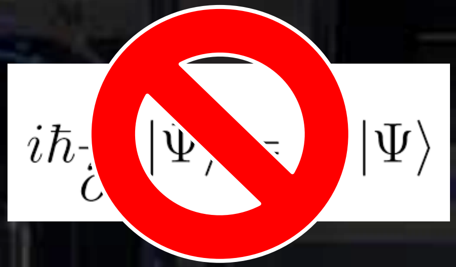
2.7 g (12.3 mmol) 4,4-Dimethyl-1,2,3,4-tetrahydro-2-oxo-7-quinolinecarboxylic acid were added to a solution of 3.8 g (18.5 mmol) N,N'-dicyclohexylcarbodiimide and 1.1 ml (12.3 mmol) aniline in 80 ml dichloromethane. The reaction mixture was stirred for 4 hours at ambient temperature and the precipitate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

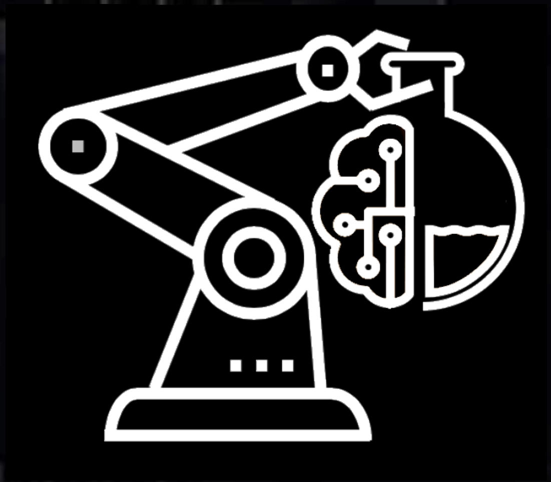
Synthesis execution

Data and chemical reactions



Target molecule

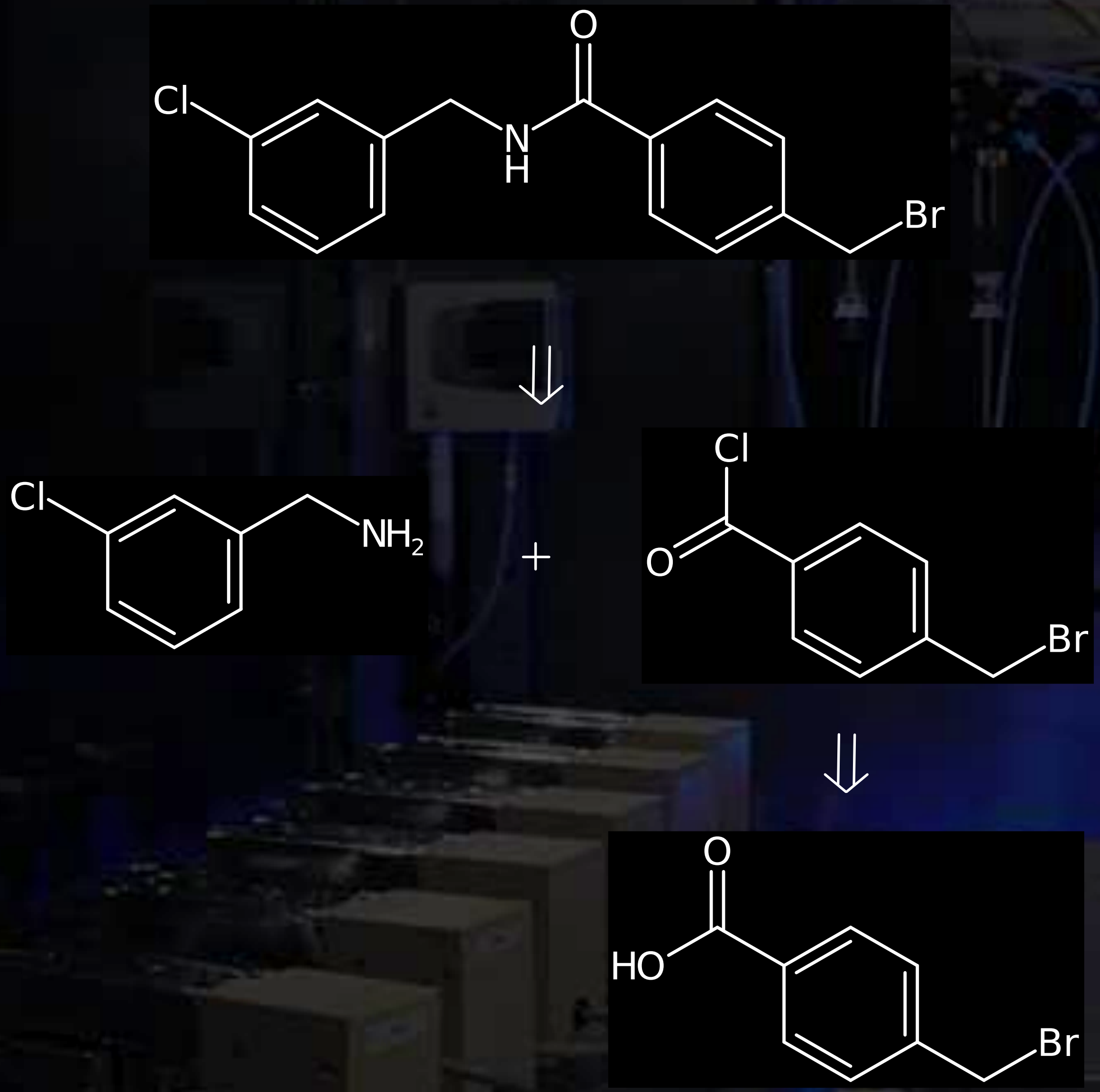




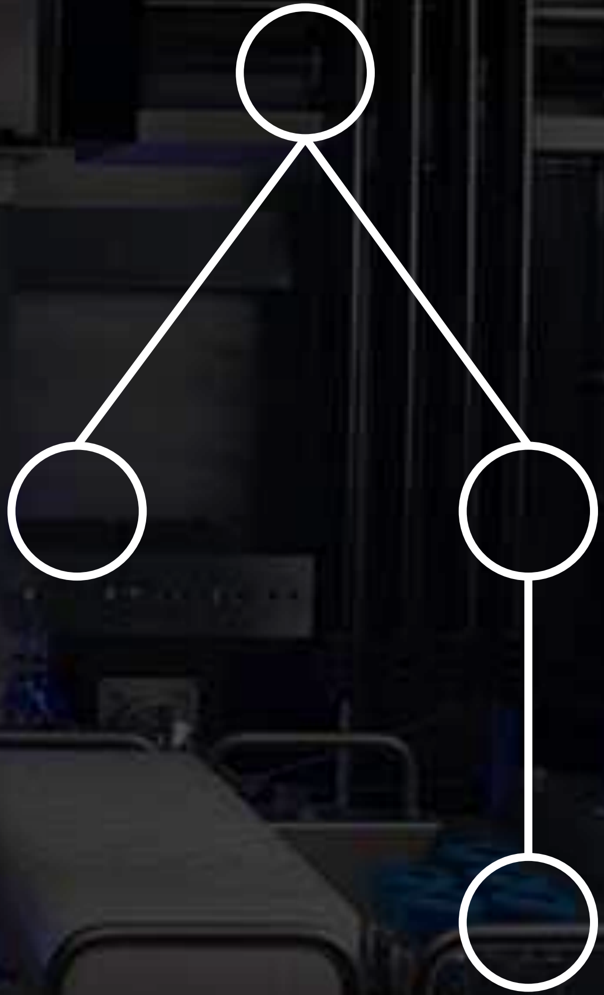
Synthesis execution

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.

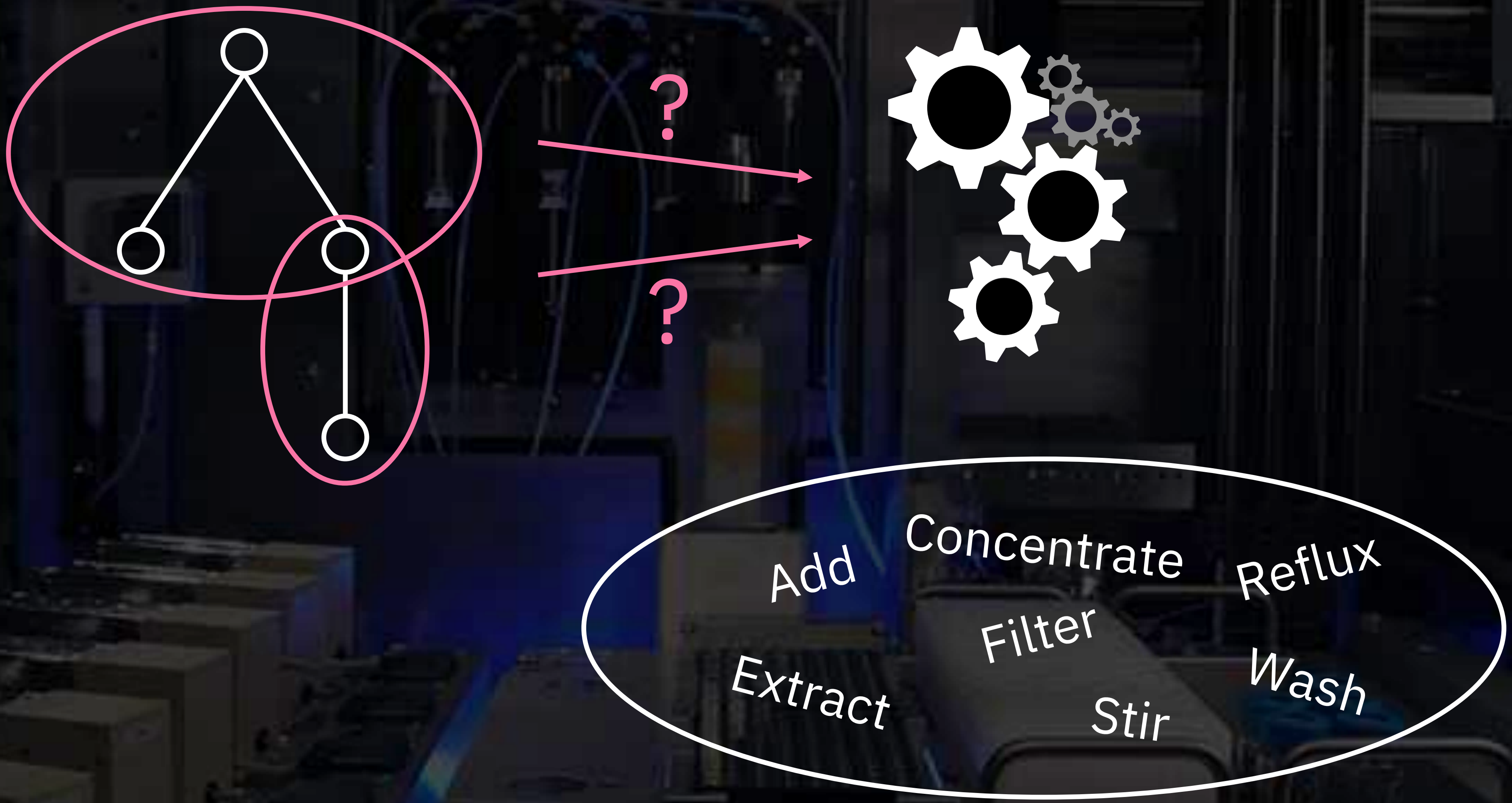
Synthesis Design



Retrosynthetic tree



Synthesis Execution



Add

Concentrate

Reflux

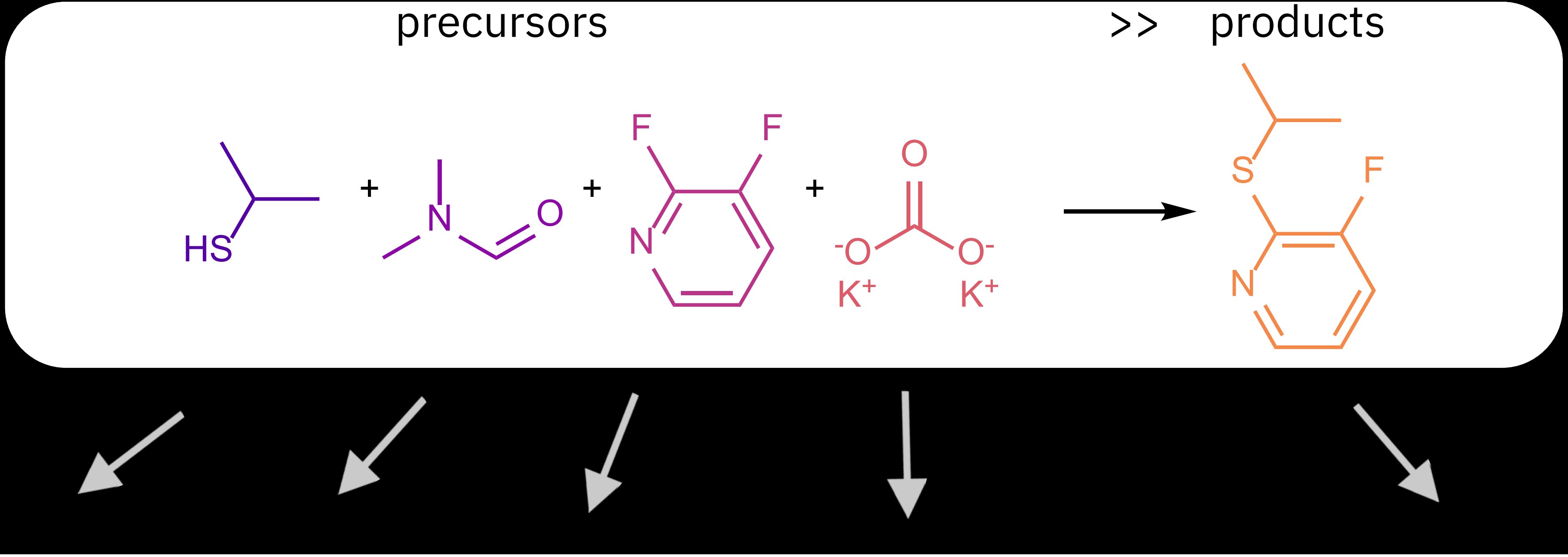
Filter

Extract

Wash

Stir

Atoms as *letters*, molecules as *words*

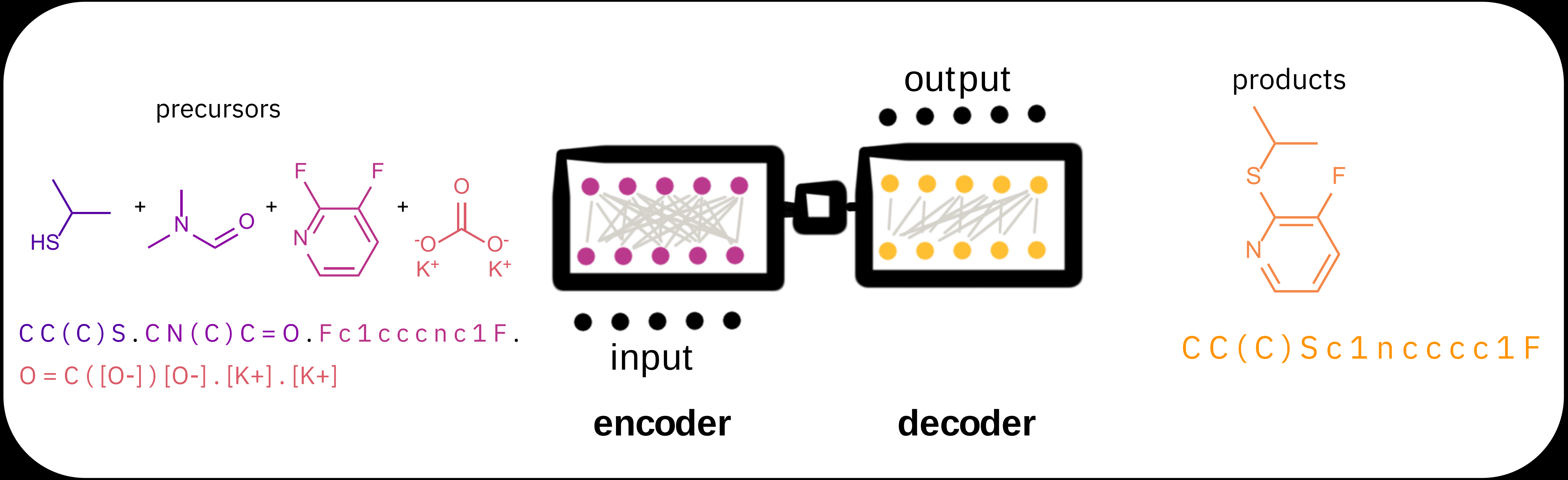


CC(C)S.CN(C)C=O.Fc1cccnc1F.O=C([O-])[O-].[K+].[K+]>>CC(C)Sc1ncccc1F

Cast reaction prediction as translation task

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

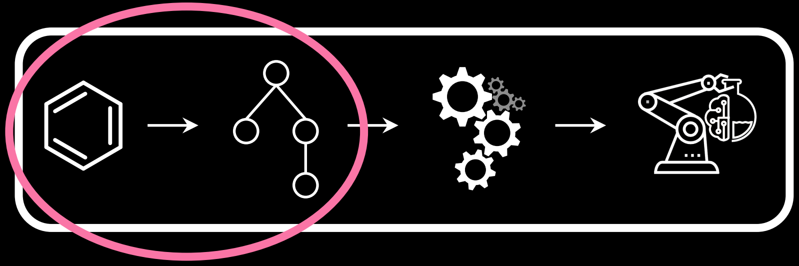
Molecular Transformer

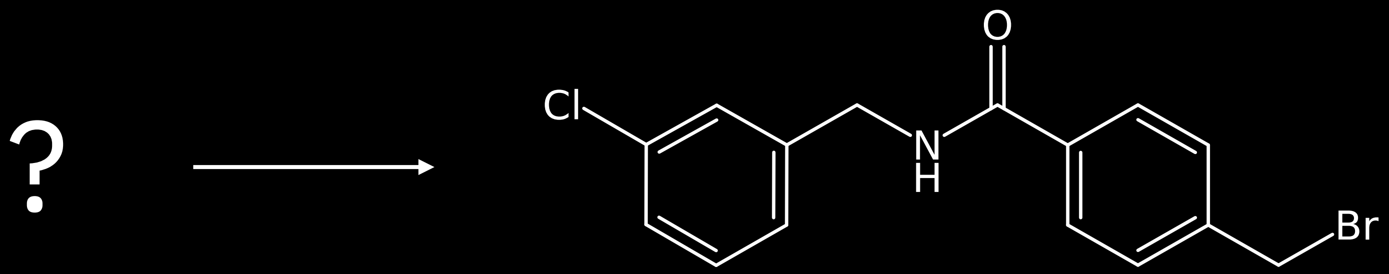


- **No rules** integrated / no chemical knowledge
- **Accurate predictions** on unseen reactions (**>90% accuracy** on benchmark)
- Better than rule and graph-based approaches

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

Synthesis Design



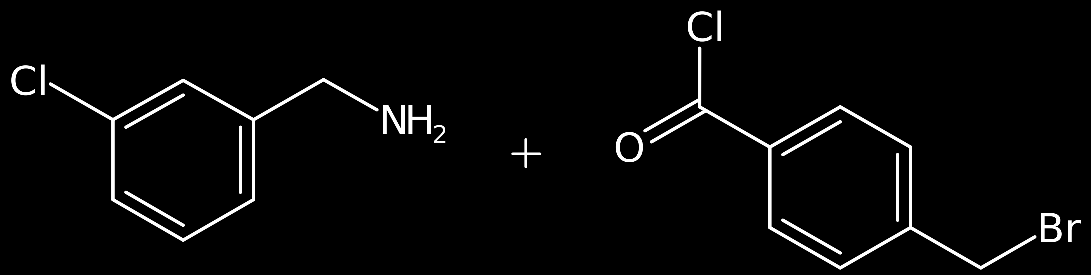


Similar approach, both sides switched

“Translation”

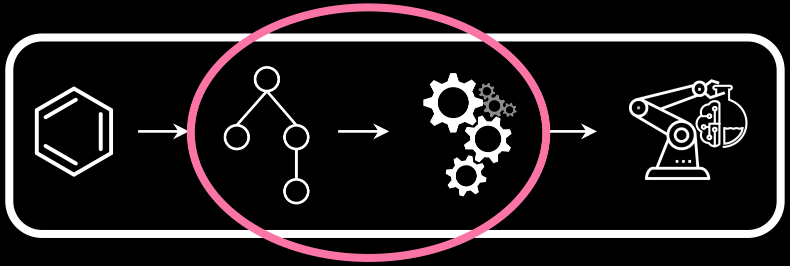
O=C(NCc1cccc(Cl)c1)c1ccc(CBr)cc1 ⟹ NCc1cccc(Cl)c1.O=C(Cl)c1ccc(CBr)cc1

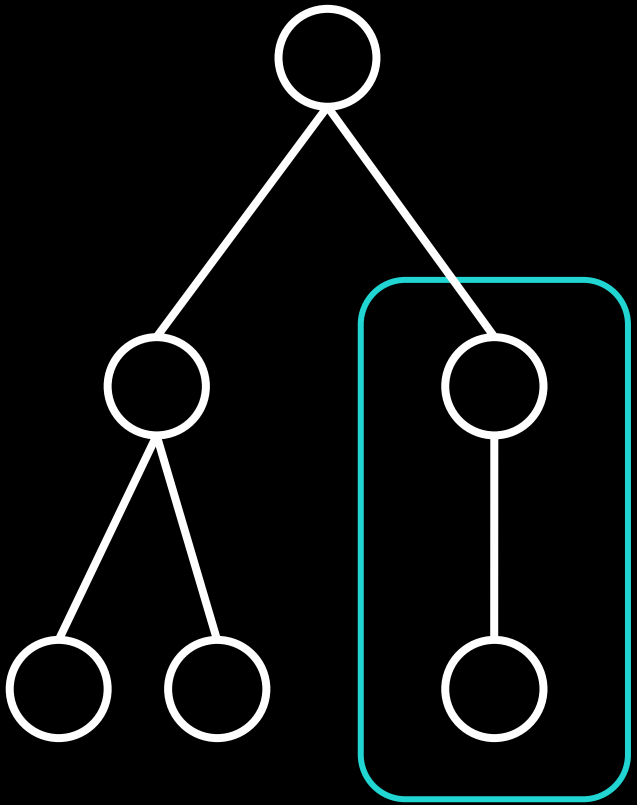
Transformer



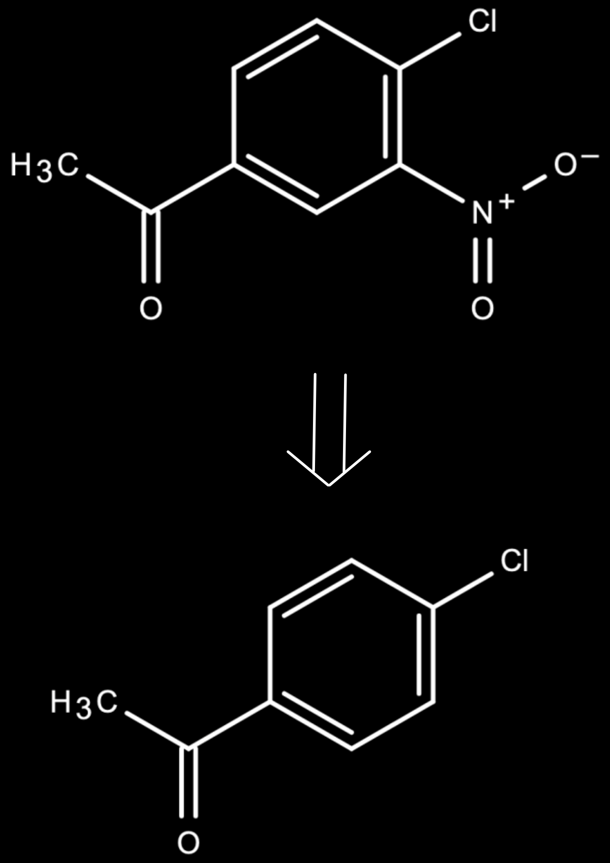
Chem. Sci., 2020, 11, 3316-3325

Synthesis actions

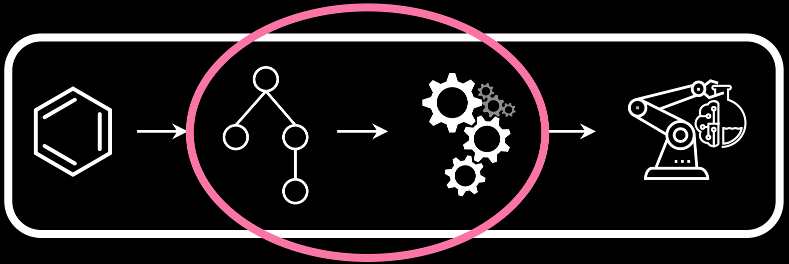


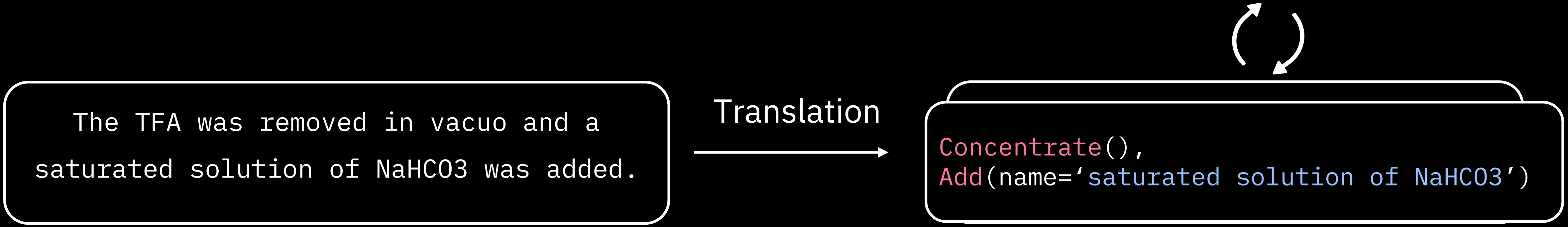


One reaction step



Building a dataset for ML model





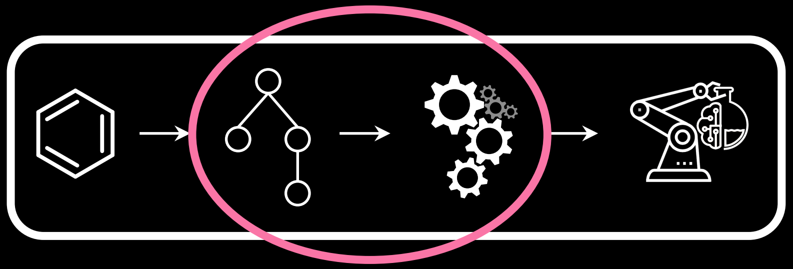
The TFA was removed in vacuo and a saturated solution of NaHCO3 was added.

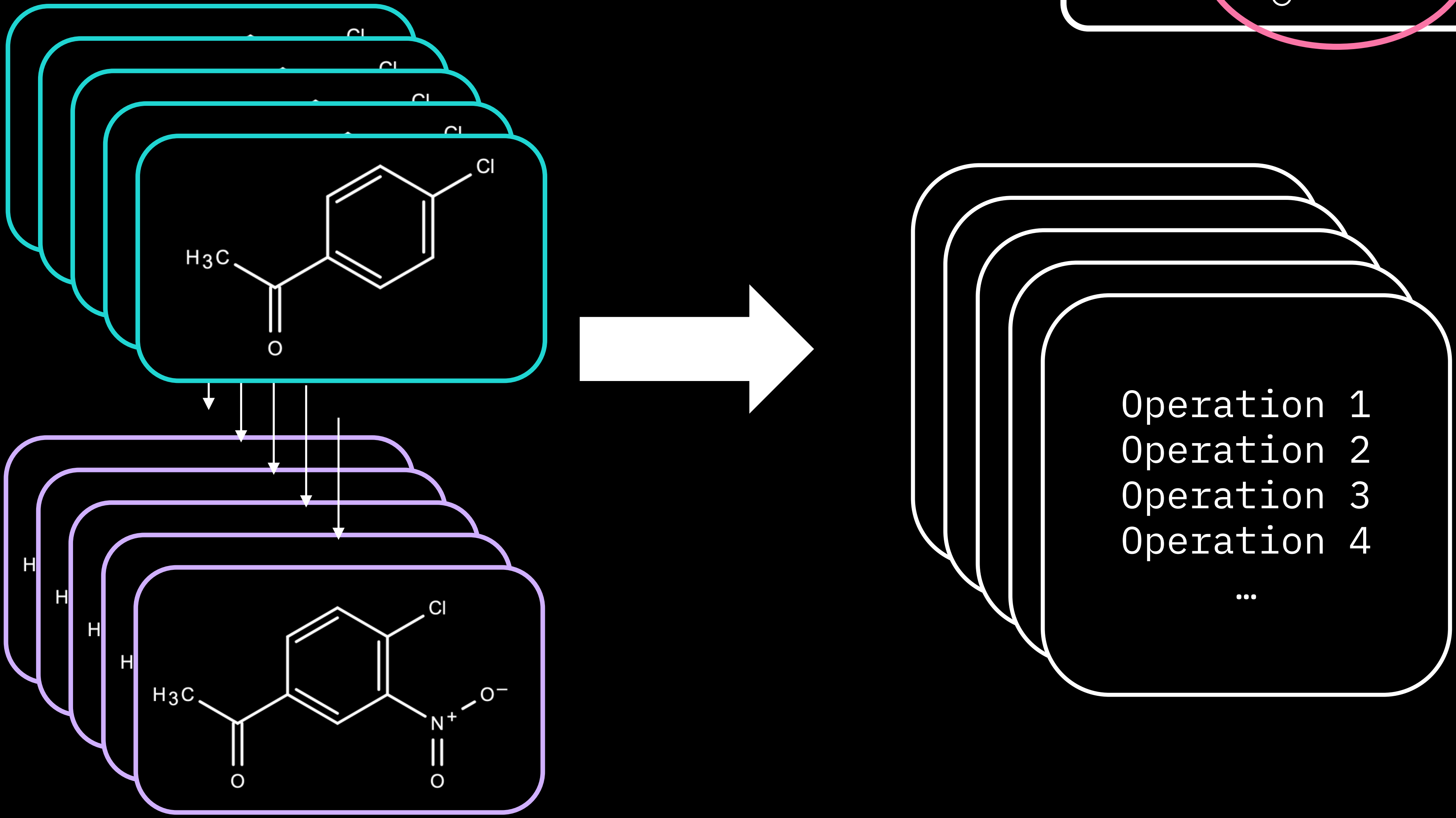
Translation

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

Nat. Comm., 2020, 11, 3601

SMILES-to-actions dataset

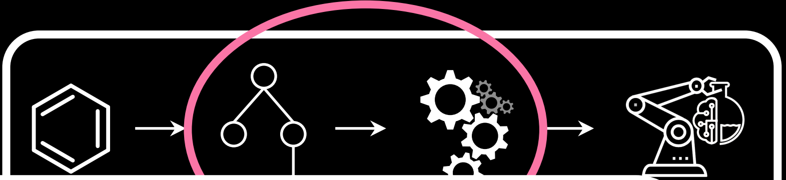


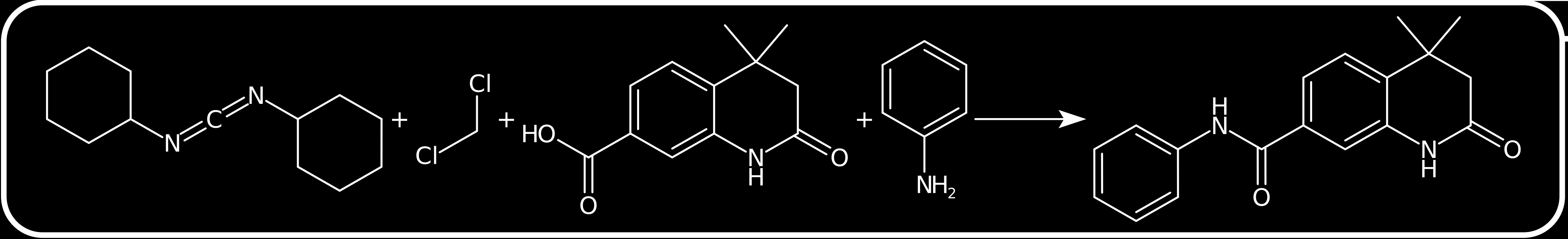


```
Operation 1
Operation 2
Operation 3
Operation 4
…
```

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.

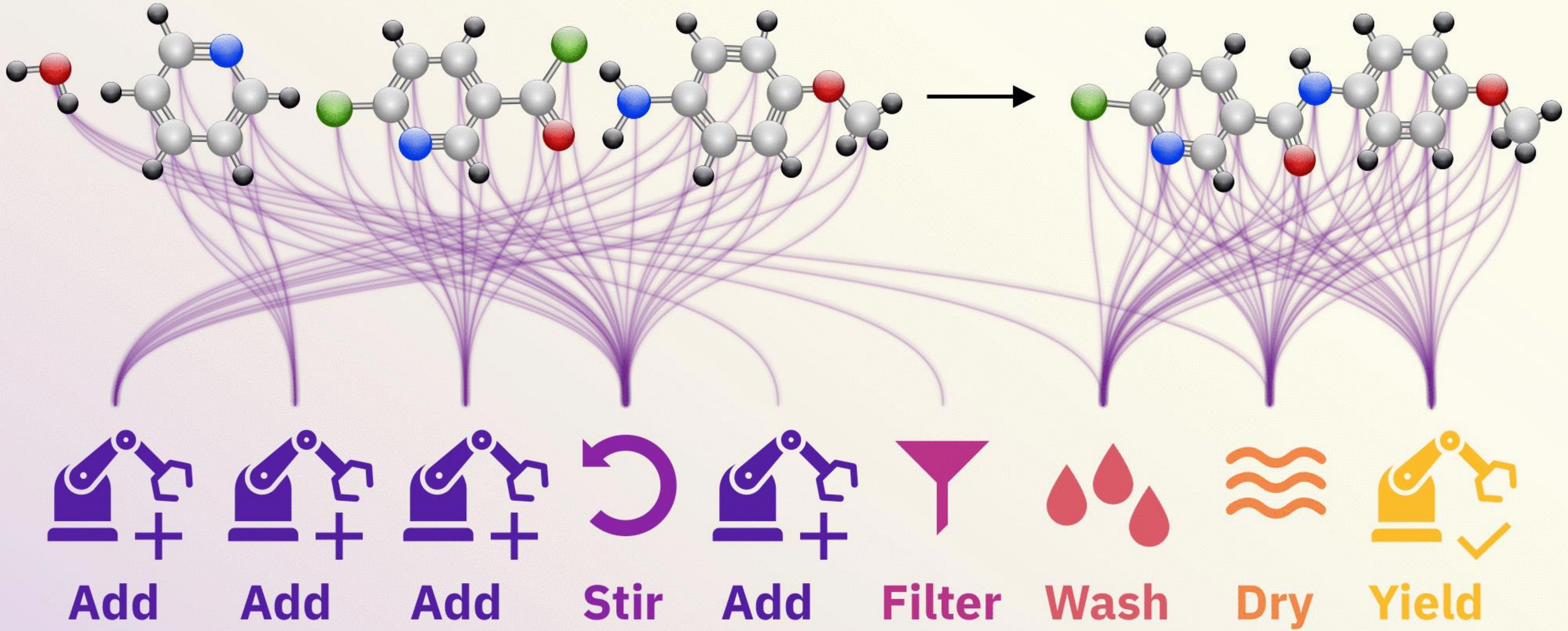
ML model

1. MAKESOLUTION ith N,N'-dicyclohexylcarb iimide (3.8 g, 18.5 mmol) and aniline (1.1 12.3 mmol) and dichloromethane 80 ml)
2. ADD SLN
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)

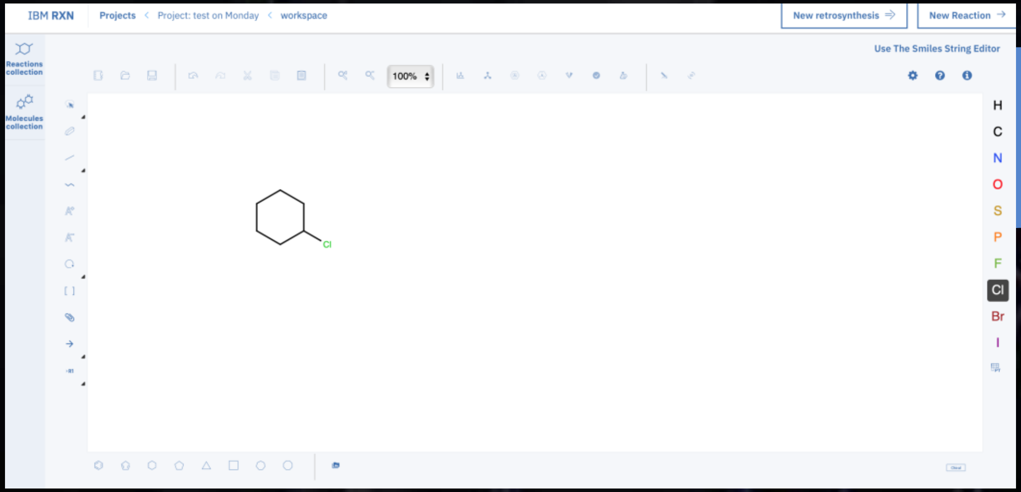
1. ADD \$1\$
2. ADD \$4\$
3. ADD \$2\$
4. ADD \$3\$
5. STIR for @3@ at #4#
6. FILTER keep precipitate
7. RECRYSTALLIZE from ethanol
8. YIELD \$-1\$

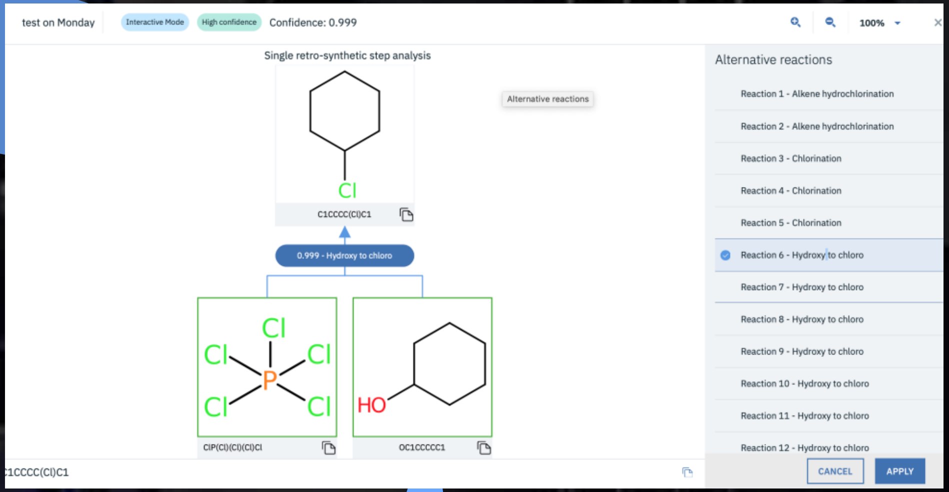
Nat. Comm., 2021, 12, 2573

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



SMILES-2-ACTIONS




test on Monday

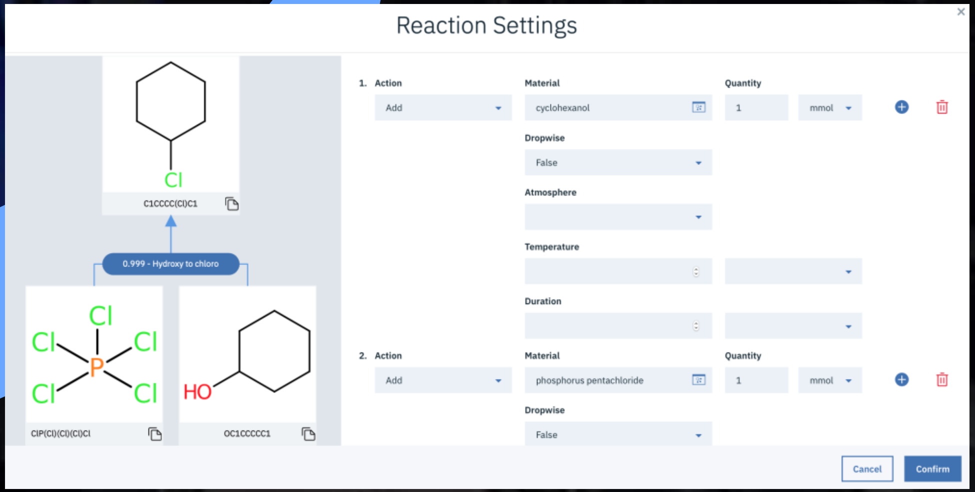
Confidence: 0.999

Single retro-synthetic step analysis

C1CCCC(Cl)C1

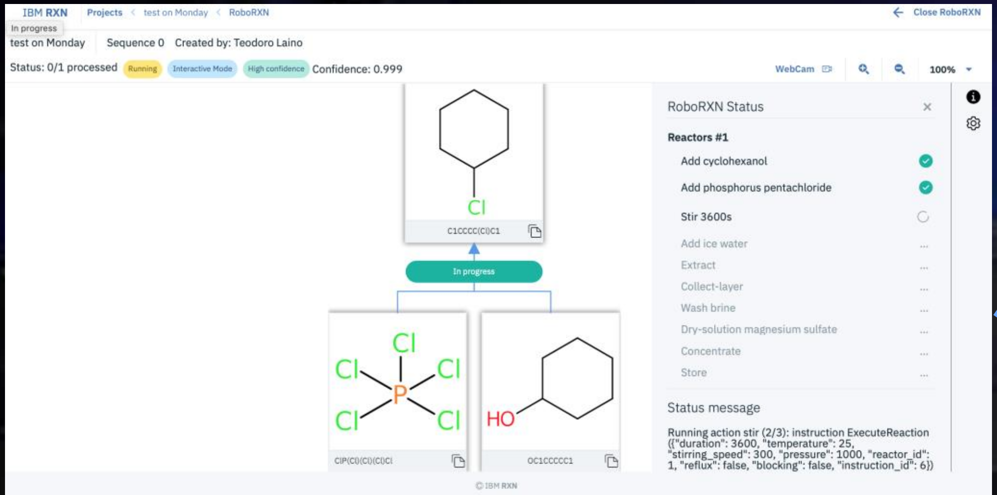
Alternative reactions

- 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


Reaction Settings

1. Action: Add
Material: cyclohexanol
Quantity: 1 mmol
Dropwise: False
Atmosphere:
Temperature:
Duration:

2. Action: Add
Material: phosphorus pentachloride
Quantity: 1 mmol
Dropwise: False



test on Monday Sequence 0 Created by: Teodoro Laino

Status: 0/1 processed

Confidence: 0.999

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

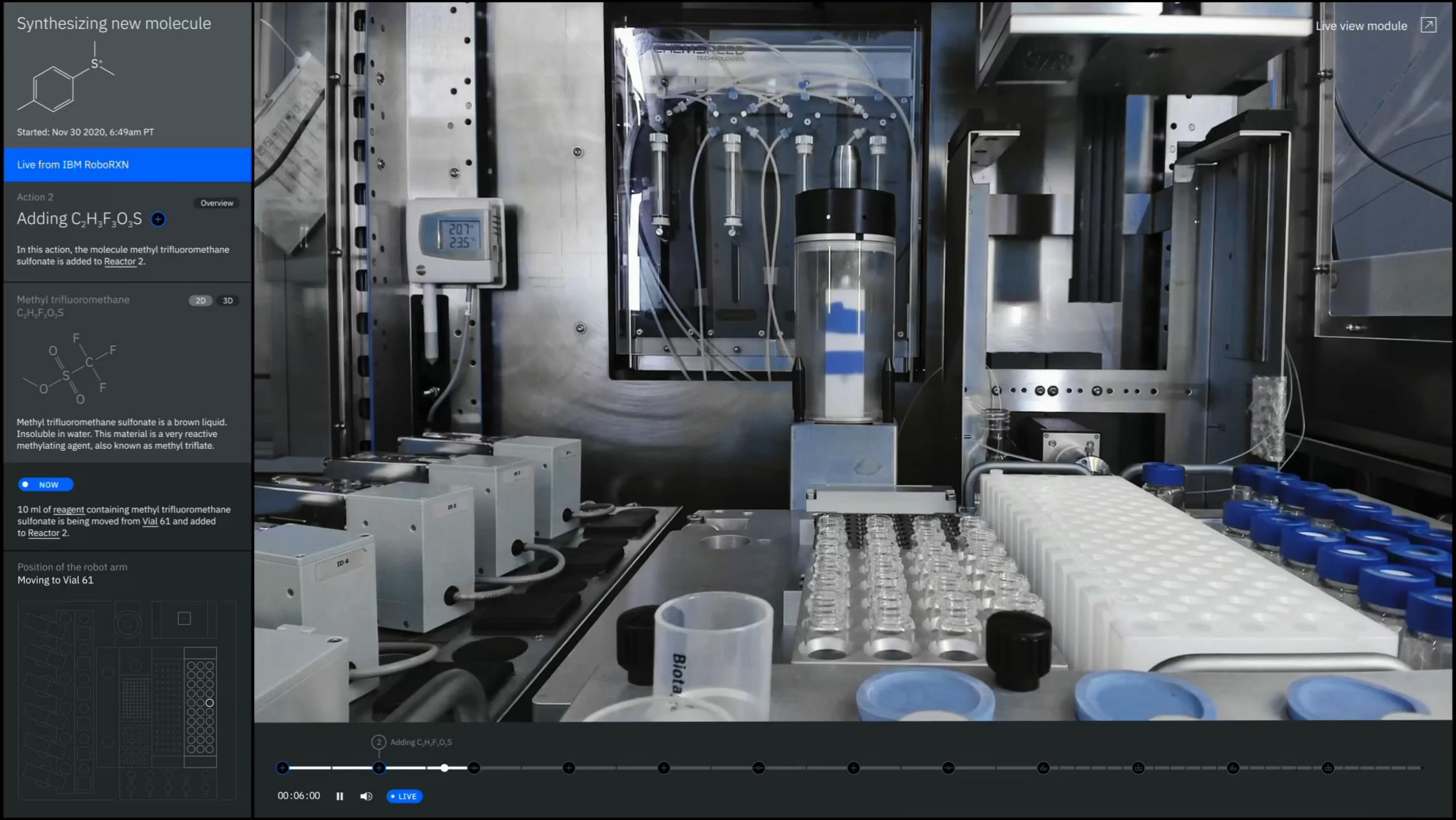
Flex Autoplant

FLEX AUTOPLANT robotic platform

Analytics

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

Dauerversuch!
Bei
Gefahr abschalten.
Agilent
6230 LC/TOF
192.168.254.12
Spinsolve 80 ULTRA
Carbon



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 $C_2H_3F_3O_3S$

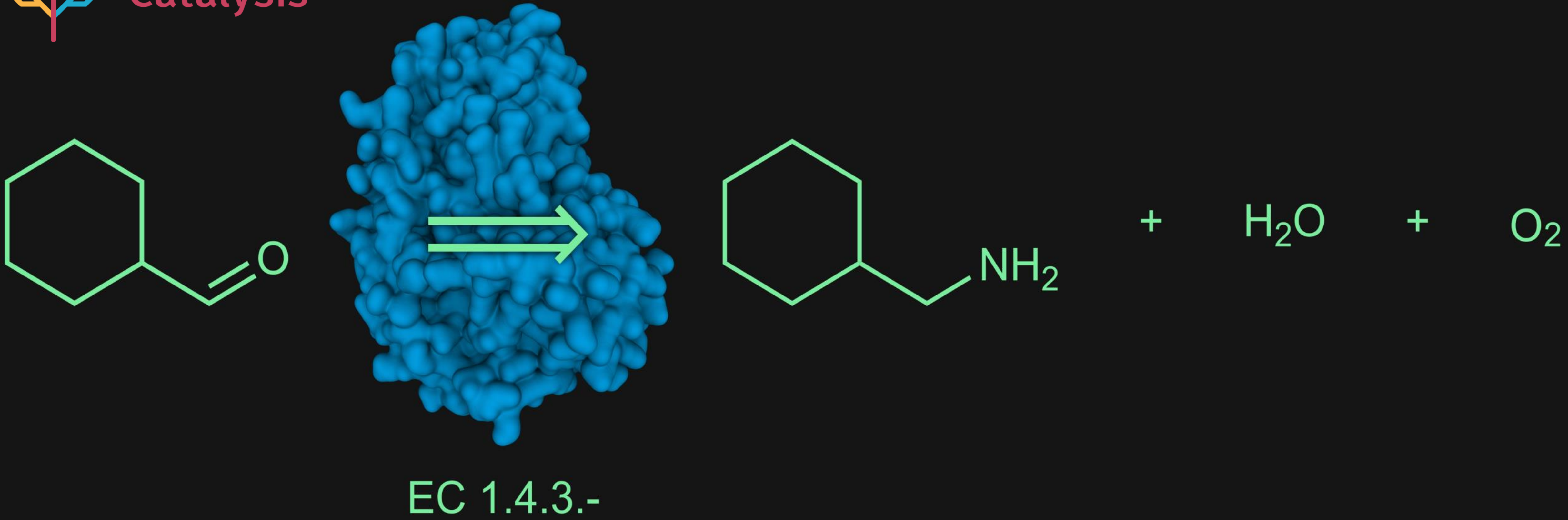
00:06:00

LIVE

Enzymatic catalysis

GreenCatRXN

NCCR Catalysis



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

Nat. Commun. 2022, 13, 964

The research lab is the central element in scientific discovery

Up to 70% of experimentation is not reproducible because of flawed experimental data or metadata[1]

Only one-third to one-half of original findings are confirmed in replication studies[2]

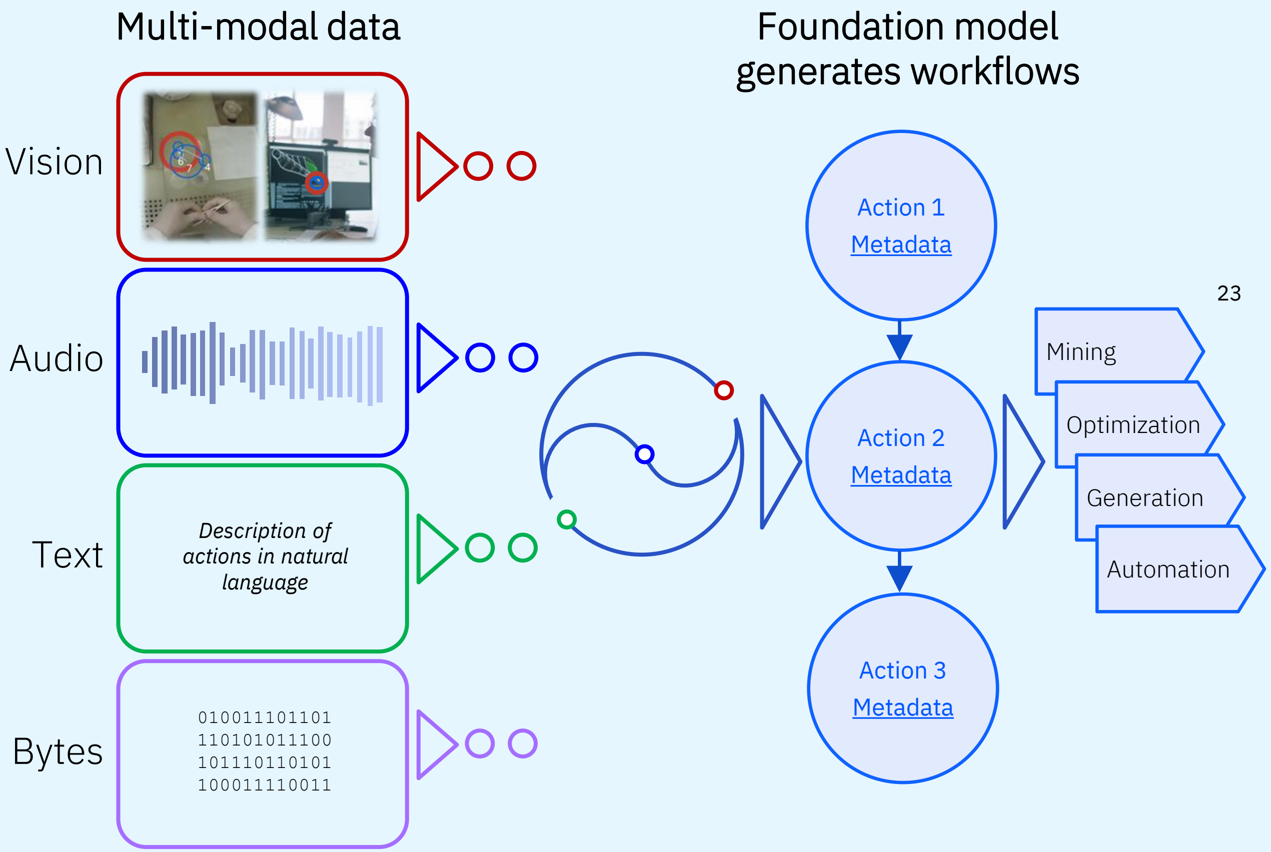
Studies can last years instead of months due to small differences in protocols[3]

More than half of researchers report being unable to reproduce their own experiments[4]

[1] Goncalves, R.S., Musen M.A., *Scientific Data* **6**, 190021 (2019)
[2] Aarts, A. A. *et al.*, *Science* **349**, 943 (2015)
[3] Hines, W. C. *et al.*, *Cell Rep.* **6**, 779–781 (2014)
[4] Baker, M., *Nature* **533**, 452-454 (2016)

Multi-modal foundation models

Capturing end-to-end business workflows for mining, optimization, generation, and automation



Key innovations

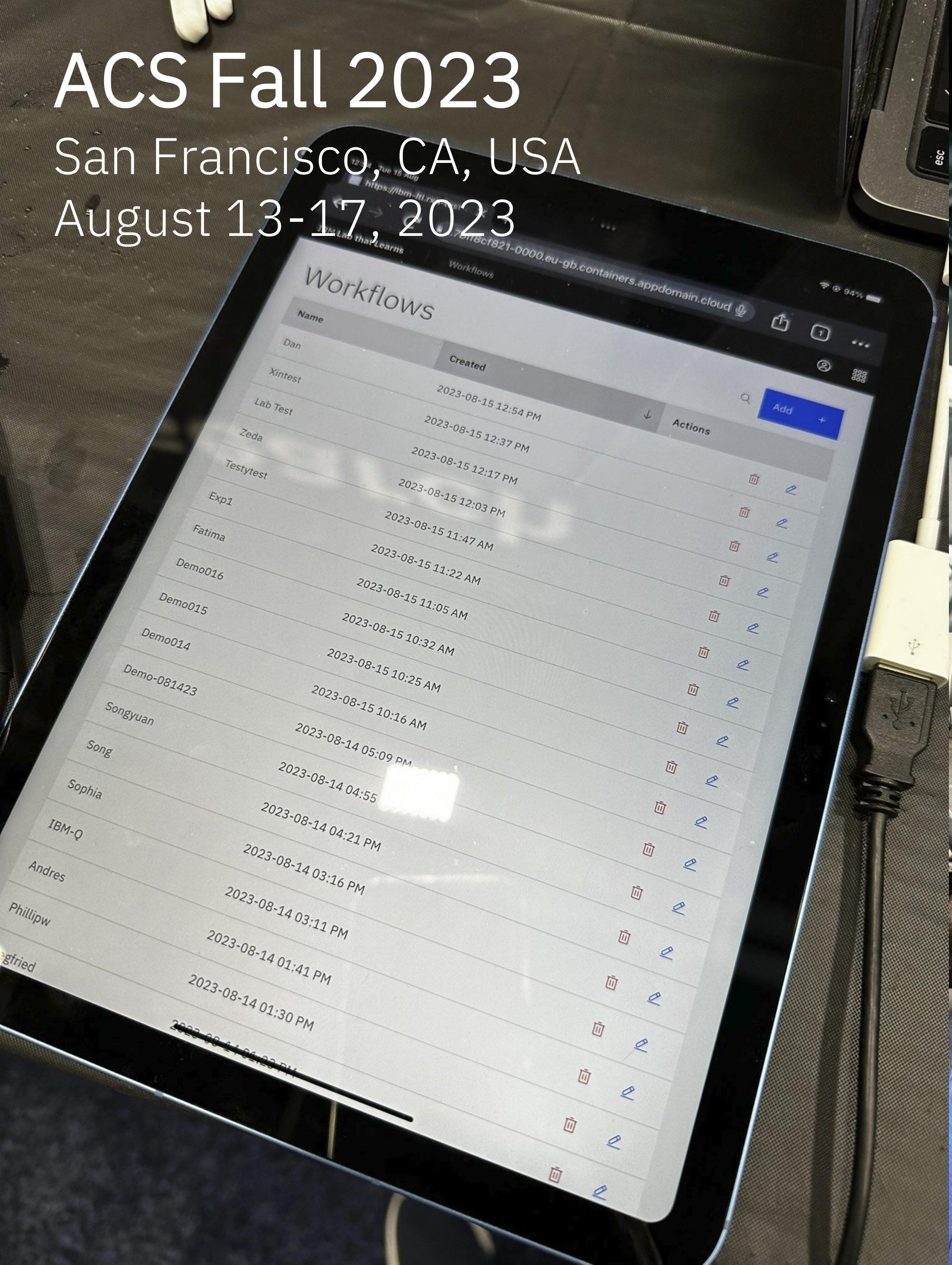
- AI foundation models for automatic documentation of manual procedures and validation of outcomes
- Hybrid and multi-cloud computing to automatically integrate all data and metadata of any experiment

Benefits for the lab

- Capture all details needed to fully capture and describe an experiment
- Minimize the time wasted using different tools to organize data
- Reproduce any version of an experiment at any point in time
- Discover patterns by continuously learning over all experimental data

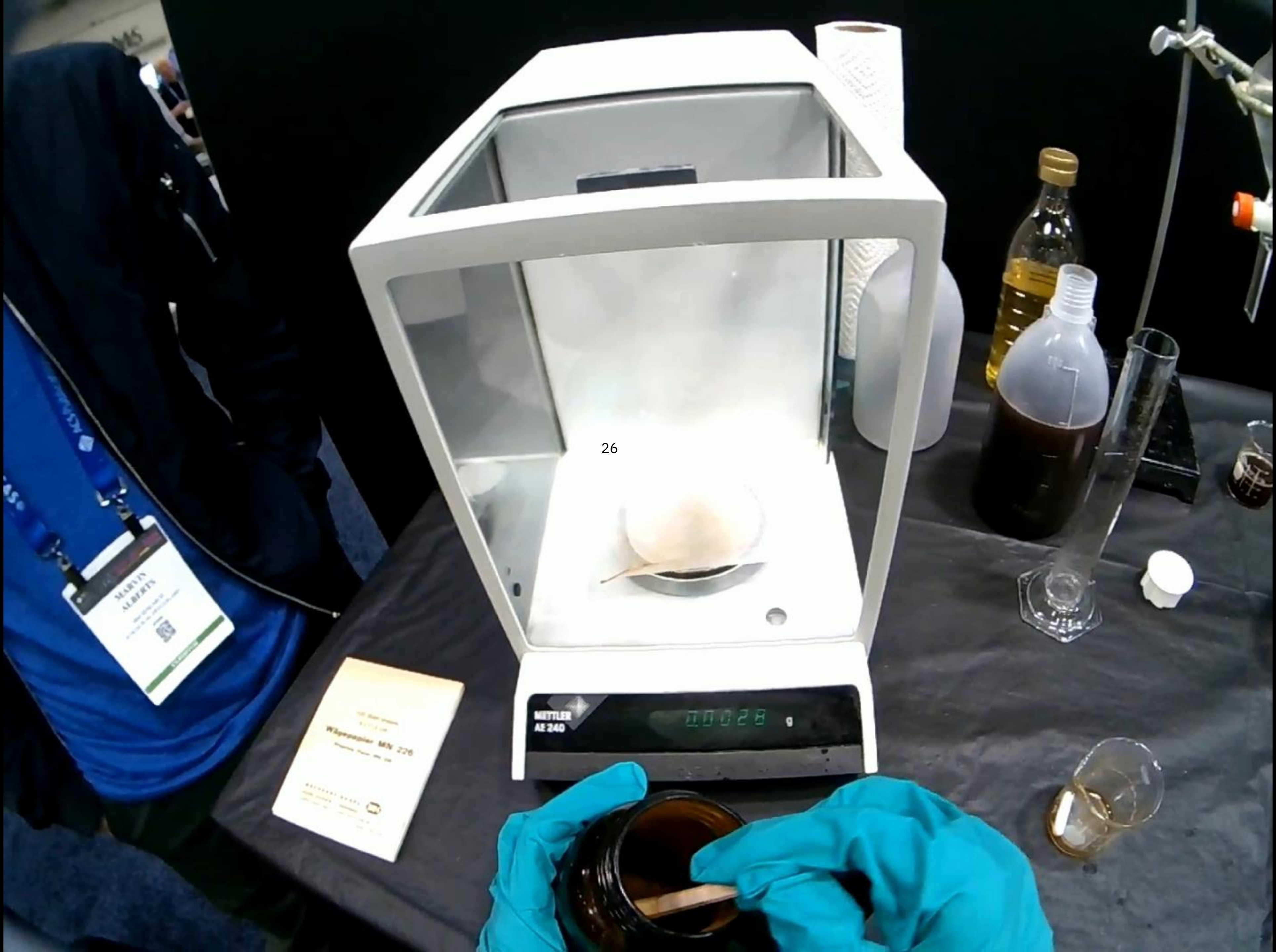
ACS Fall 2023

San Francisco, CA, USA
August 13-17, 2023

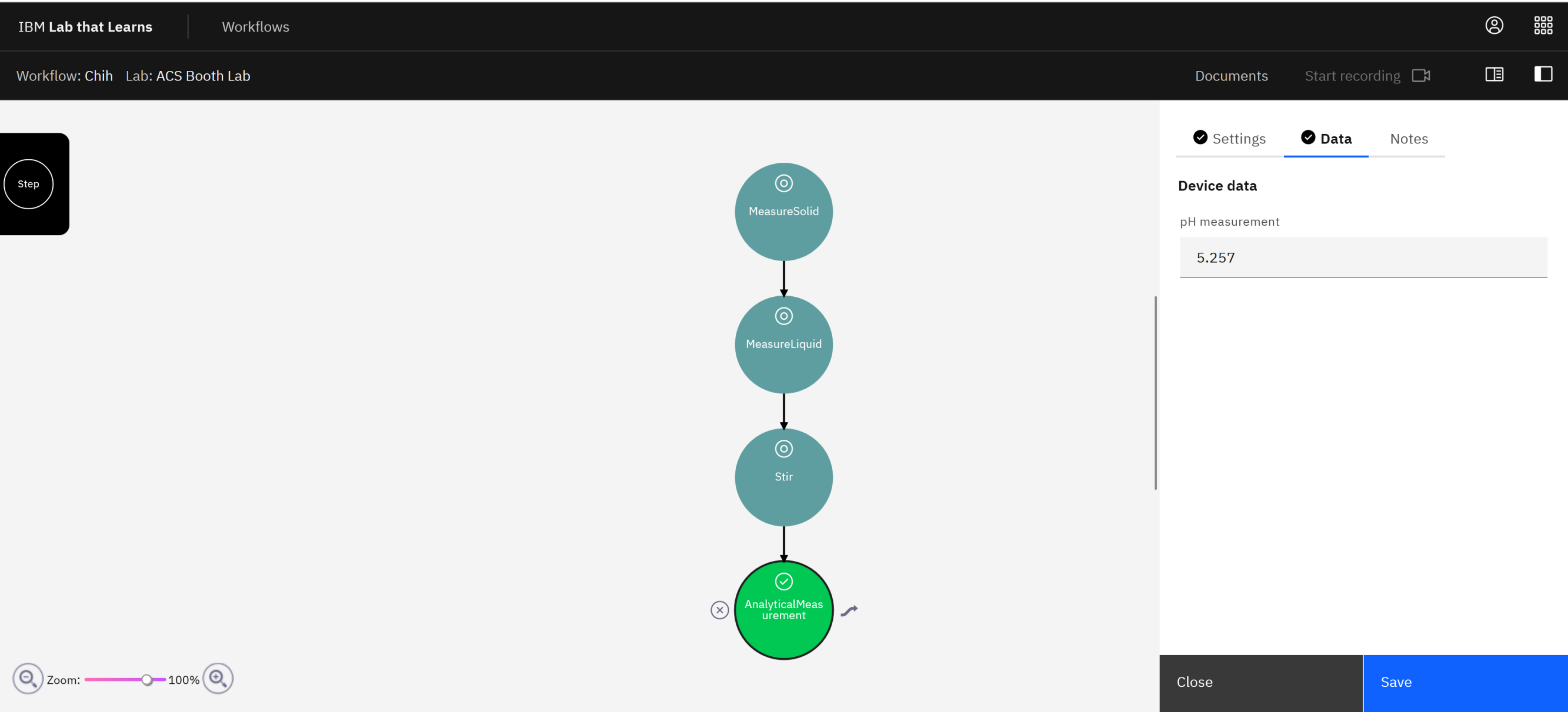


Workflows

Name	Created	Actions
Dan	2023-08-15 12:54 PM	
Xintest	2023-08-15 12:37 PM	
Lab Test	2023-08-15 12:17 PM	
Zeda	2023-08-15 12:03 PM	
Testytest	2023-08-15 11:47 AM	
Exp1	2023-08-15 11:22 AM	
Fatima	2023-08-15 11:05 AM	
Demo016	2023-08-15 10:32 AM	
Demo015	2023-08-15 10:25 AM	
Demo014	2023-08-15 10:16 AM	
Demo-081423	2023-08-14 05:09 PM	
Songyuan	2023-08-14 04:55	
Song	2023-08-14 04:21 PM	
Sophia	2023-08-14 03:16 PM	
IBM-Q	2023-08-14 03:11 PM	
Andres	2023-08-14 01:41 PM	
Phillipw	2023-08-14 01:30 PM	
gfried	[illegible]	


MARVIN ALBERTS
METTLER
AE 240
0.0028 g
Wägepapier MN 226

Workflow captured in the Lab that Learns



Device data

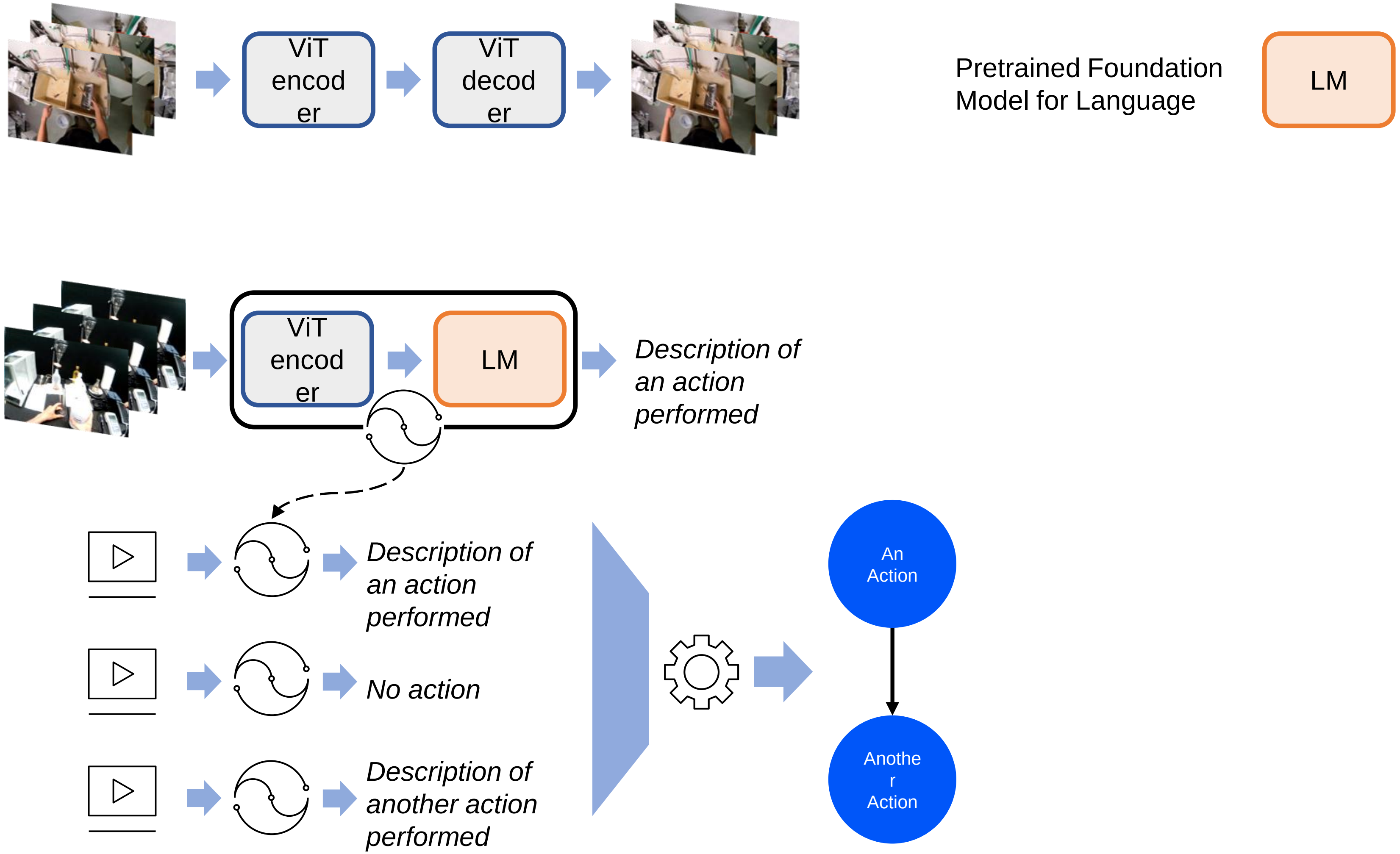
pH measurement: 5.257

Multimodal foundation models for the Lab that Learns

Pretraining Foundation Model for Egocentric Videos

Fine-tuning Vision-Language Foundation Model for Laboratory Procedures

Workflow Consolidation


ViT encod er
ViT decod er
Pretrained Foundation Model for Language
LM
ViT encod er
LM
Description of an action performed
Description of an action performed
No action
Description of another action performed
An Action
Anothe r Action

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

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

Collaborators:

UNIVERSITÄT BERN

IN SUPREMÆ DIGNITATIS 1343

NCCR Catalysis

IBM