# IBM RXN for Chemistry: A Chemical Reaction Prediction Platform

**Philippe Schwaller** (@pschwllr)

http://dx.doi.org/10.26434/chemrxiv.7297379

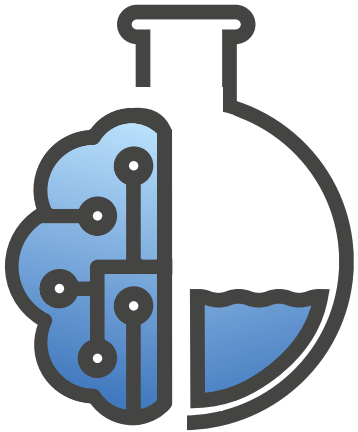
IBM **Research** AI

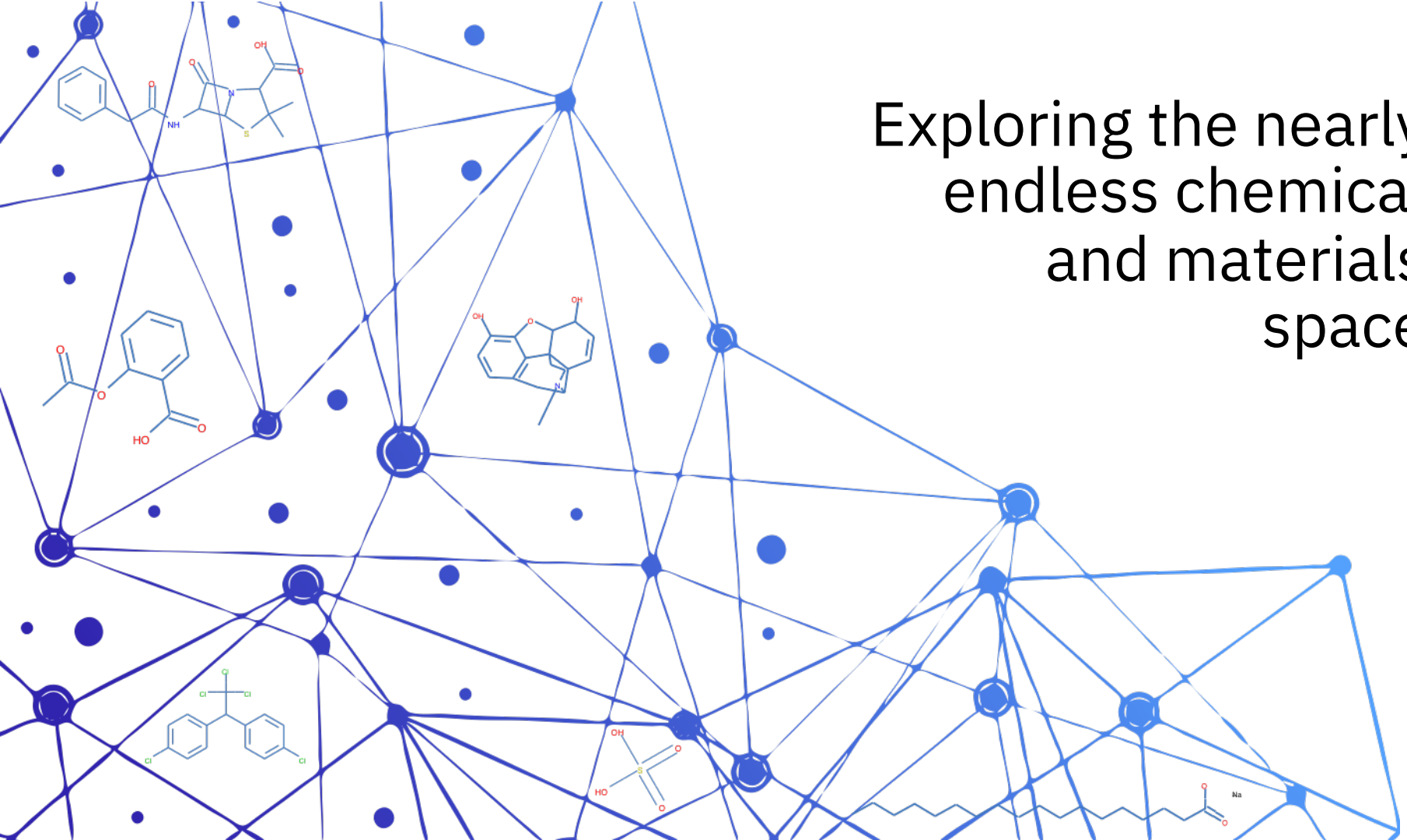
Theophile Gaudin, Teodoro Laino & Costas Bekas
IBM Research Zurich, Switzerland

Peter Bolgar & Alpha Lee
University of Cambridge

UNIVERSITY OF CAMBRIDGE

THE WINTON PROGRAMME FOR THE Physics of Sustainability





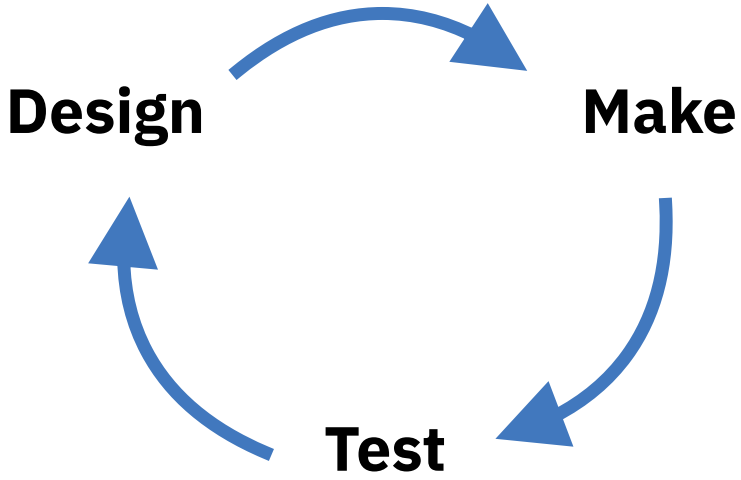
# Exploring the nearly endless chemical and materials space

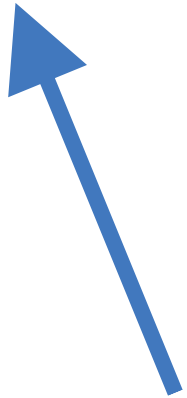
In silico molecules / materials
VAE / GAN / RL

What to make?

Synthetic route planning
Reaction outcome prediction

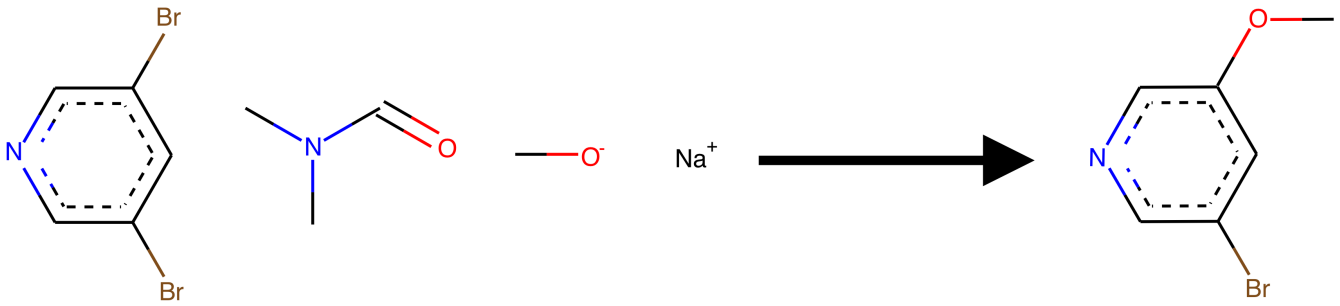
How to make them?



Experimental verification

# Data-Driven Chemical Reaction Prediction

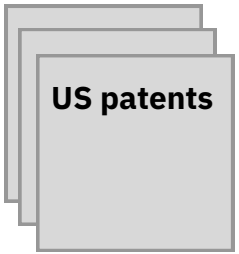



reactants & reagents

products

Data > Representation > Model > Results

# Data




**Lowe (2012, 2017)**

**text-mining**

**Reaction SMILES**
- **reactants >> products**

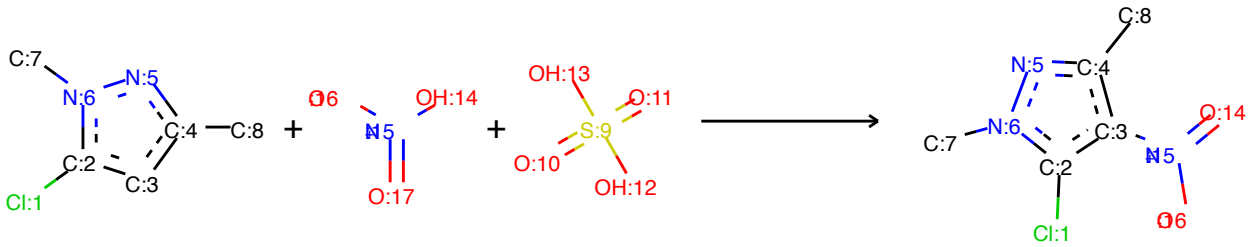
**CML files**

filtered

**Benchmark sets**

| | |
|---|---|
| Jin et al. NeurIPS 2017 | **USPTO_MIT** |
| Schwaller et al. NeurIPS 2017 | **USPTO_STEREO** |
| Bradshaw et al. ArXiv 2018 | **USPTO_LEF** |

[Cl:1][c:2]1[cH:3][c:4]([CH3:8])[n:5][n:6]1[CH3:7].[OH:14][N+:15]([O-:16])=[O:17].[S:9](=[O:10])(=[O:11])([OH:12])[OH:13]>>[Cl:1][c:2]1[c:3]([N+:15](=[O:14])[O-:16])[c:4]([CH3:8])[n:5][n:6]1[CH3:7]

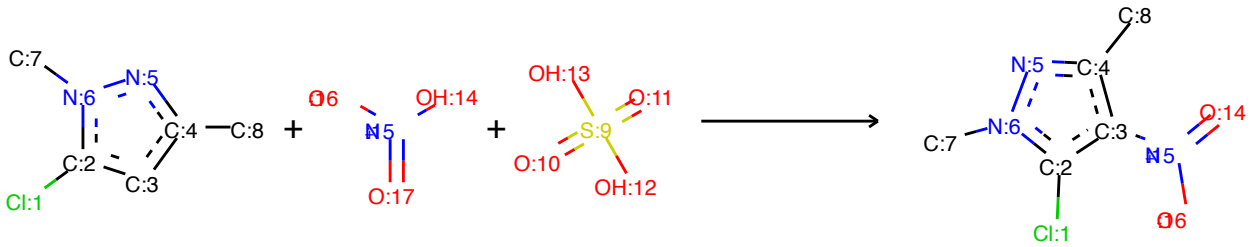


# Atom-mapping from Indigo TK

- USPTO dataset description:

  While typically correct, the **atom-maps are wrong in many cases** and hence **should not be entirely relied on**.

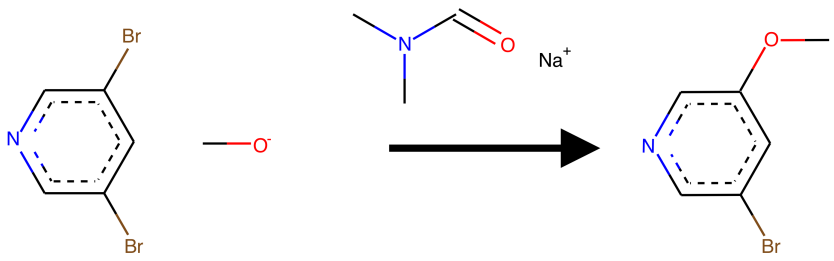
[Cl:1][c:2]1[cH:3][c:4]([CH3:8])[n:5][n:6]1[CH3:7].[OH:14][N+:15]([O-:16])=[O:17].[S:9](=[O:10])(=[O:11])([OH:12])[OH:13]>>[Cl:1][c:2]1[c:3]([N+:15](=[O:14])[O-:16])[c:4]([CH3:8])[n:5][n:6]1[CH3:7]



Dataset Bias in the Natural Sciences: A Case Study in Chemical Reaction Prediction and Synthesis Design
2018-11-21T14:10:55Z (GMT) by Ryan-Rhys Griffiths , Philippe Schwaller , Alpha Lee

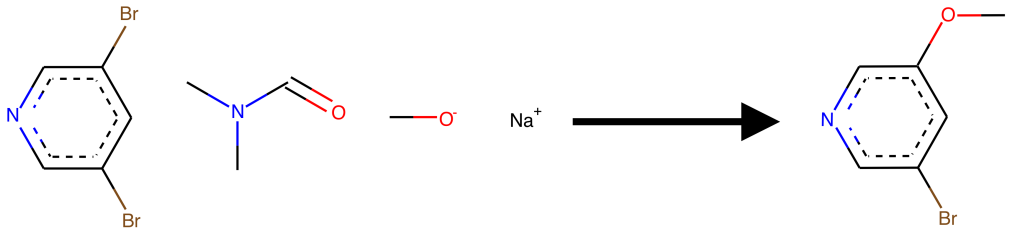
# Separated vs mixed reagents

**Separated**




reactants reagents products

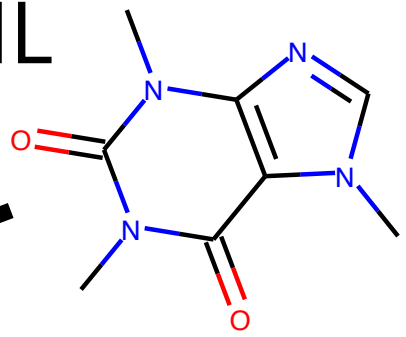
**Mixed**




reactants & reagents products

No distinction between reactants and reagents

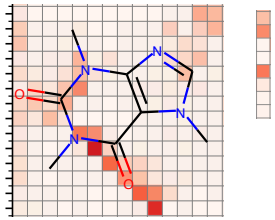
# Representing molecules for ML



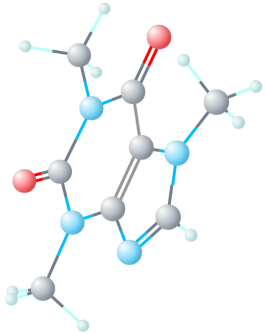

**Molecular fingerprints**
000010000....0100

**Text-based representations**
- SMILES / INCHI
- CN1C=NC2=C1C(=O)N(C(=O)N2C)C

**Graph-based**

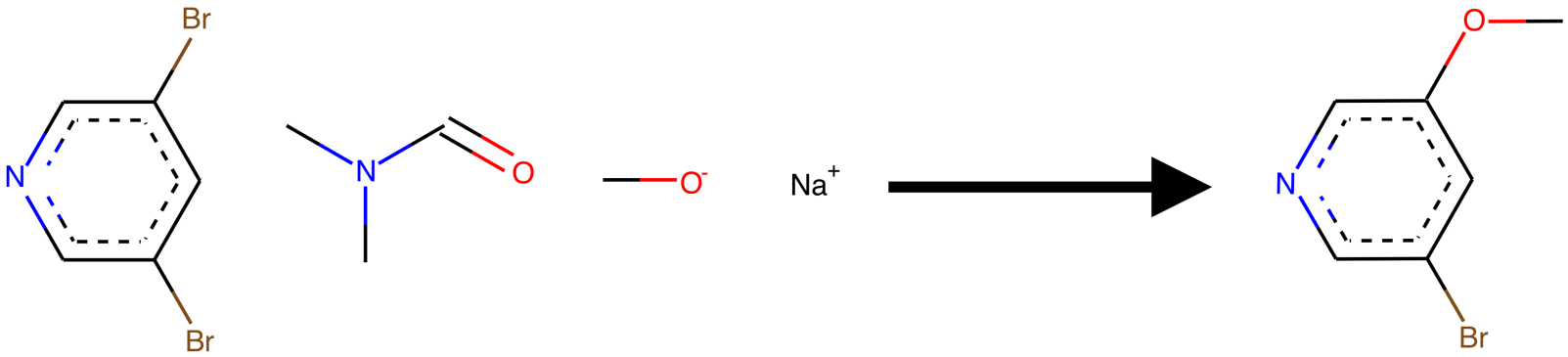


**3D structure & surface**



**1D** **2D** **3D**

# Atoms as letters, molecules as words



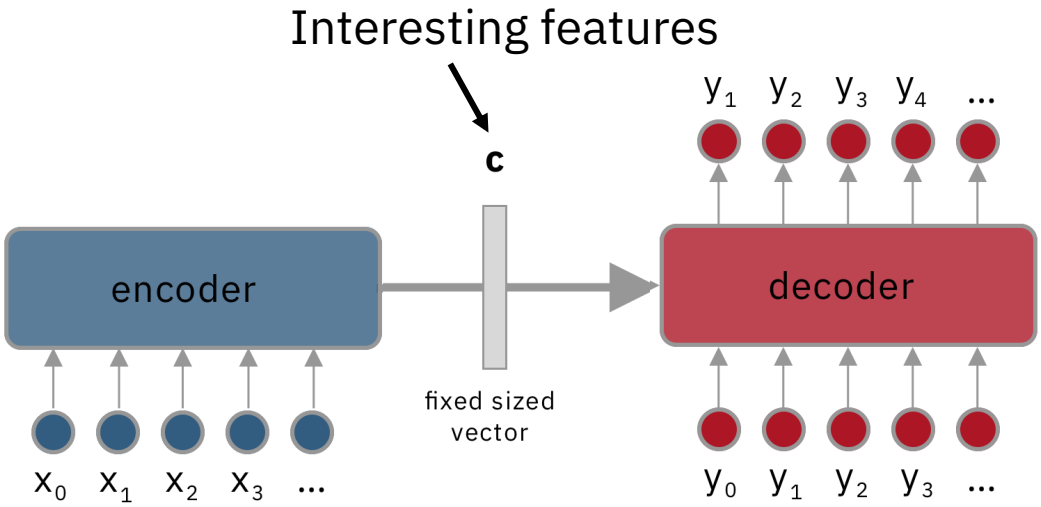

Br c 1 c n c c ( Br ) c 1 . C N ( C ) C = O . C [O-] . [Na+]

C O c 1 c n c c ( Br ) c 1

**SMILES to SMILES** prediction with sequence-2-sequence models

Schwaller et. al. Chem. Sci., 2018, 9, 6091-6098 / Nam & Kim **arXiv:1612.09529**

# Model: Standard Seq-2-Seq




**Problem: fixed size**

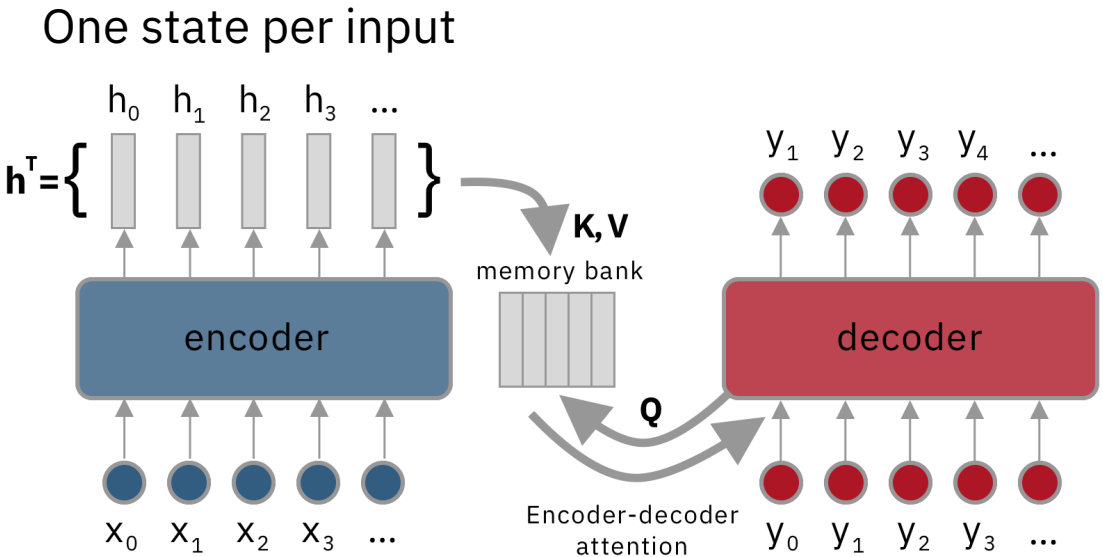
INPUTS = reactants + reagents

Br c 1 c n c c ( Br ) c 1 . C N ( C ) C = O . C [O-] . [Na+]

OUTPUTS = products

C O c 1 c n c c ( Br ) c 1

# Seq-2-Seq with Attention




Attention = ability to selectively concentrate on one aspect of context

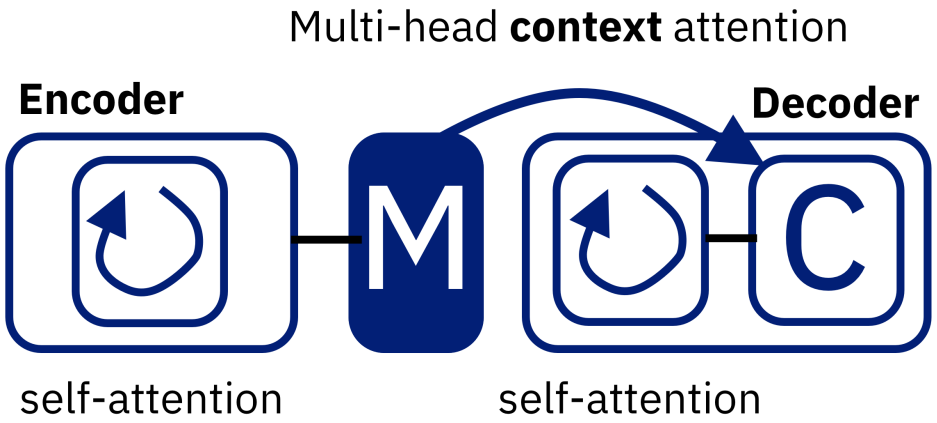
# Transformer Architecture

INPUTS = reactants + reagents

OUTPUTS = products

Br c 1 c n c c ( Br ) c 1 . C N ( C ) C = O . C [O-] . [Na+]

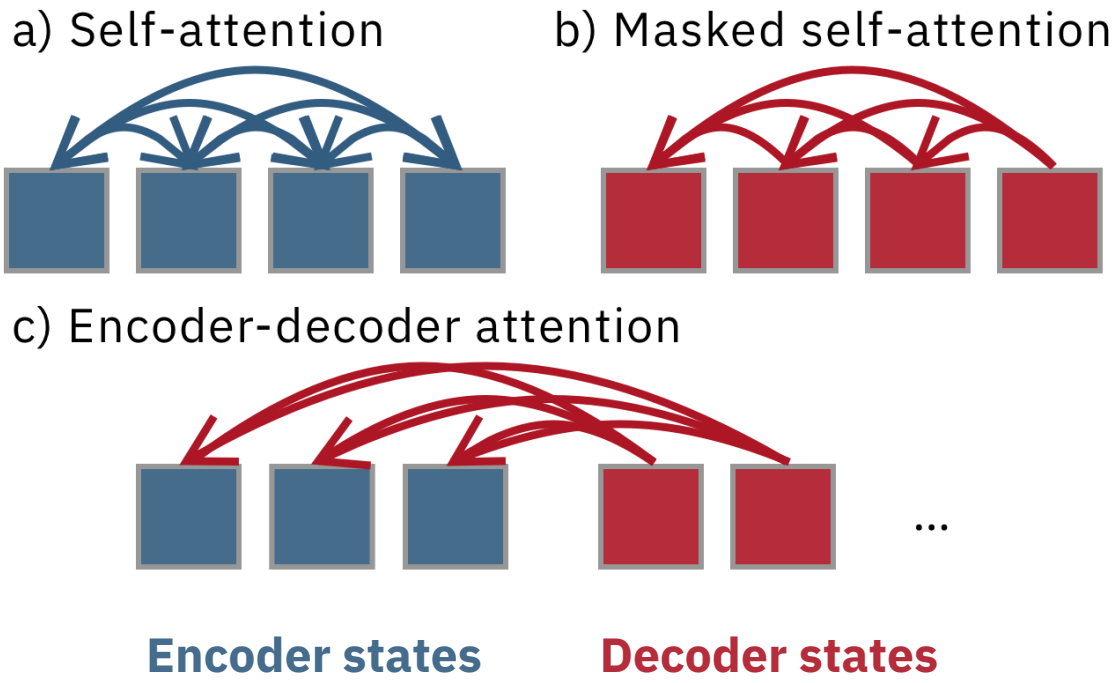
C O c 1 c n c c ( Br ) c 1



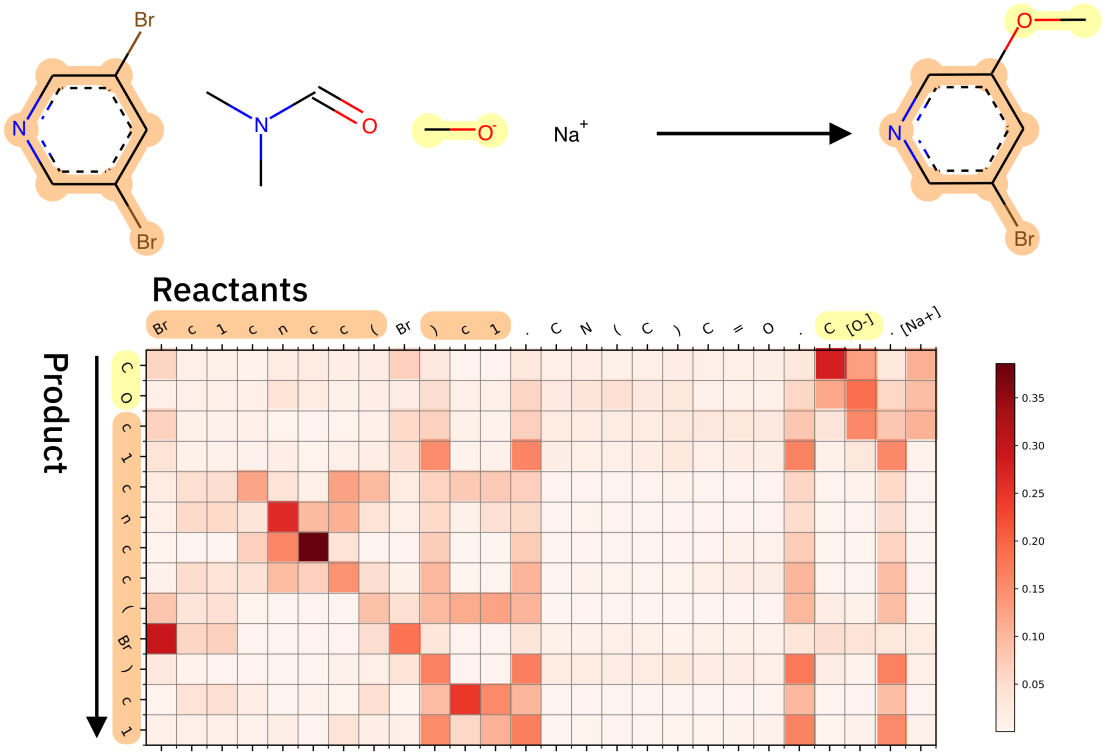

- NO recurrent neural networks
- Stacks of **attention layers**
- **Multi-head** attention

**Attention Is All You Need – Vaswani et al. NeurIPS 2017**

# Attention types in Transformer

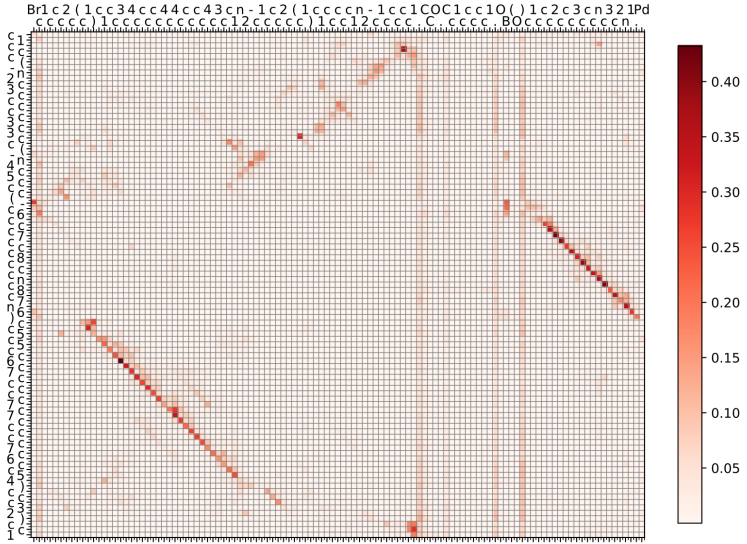


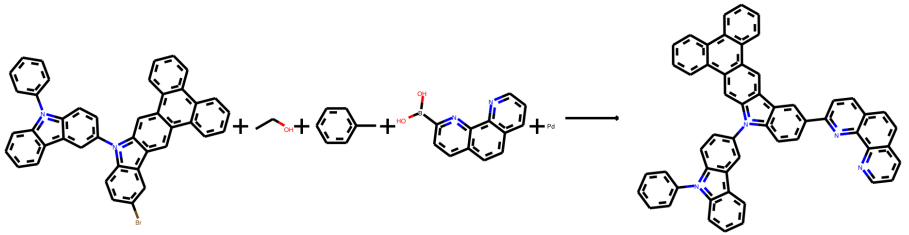
# Visualizing encoder-decoder attention




Plotting the attention weights at every decoder time step.

# Bromo-Suzuki coupling reaction



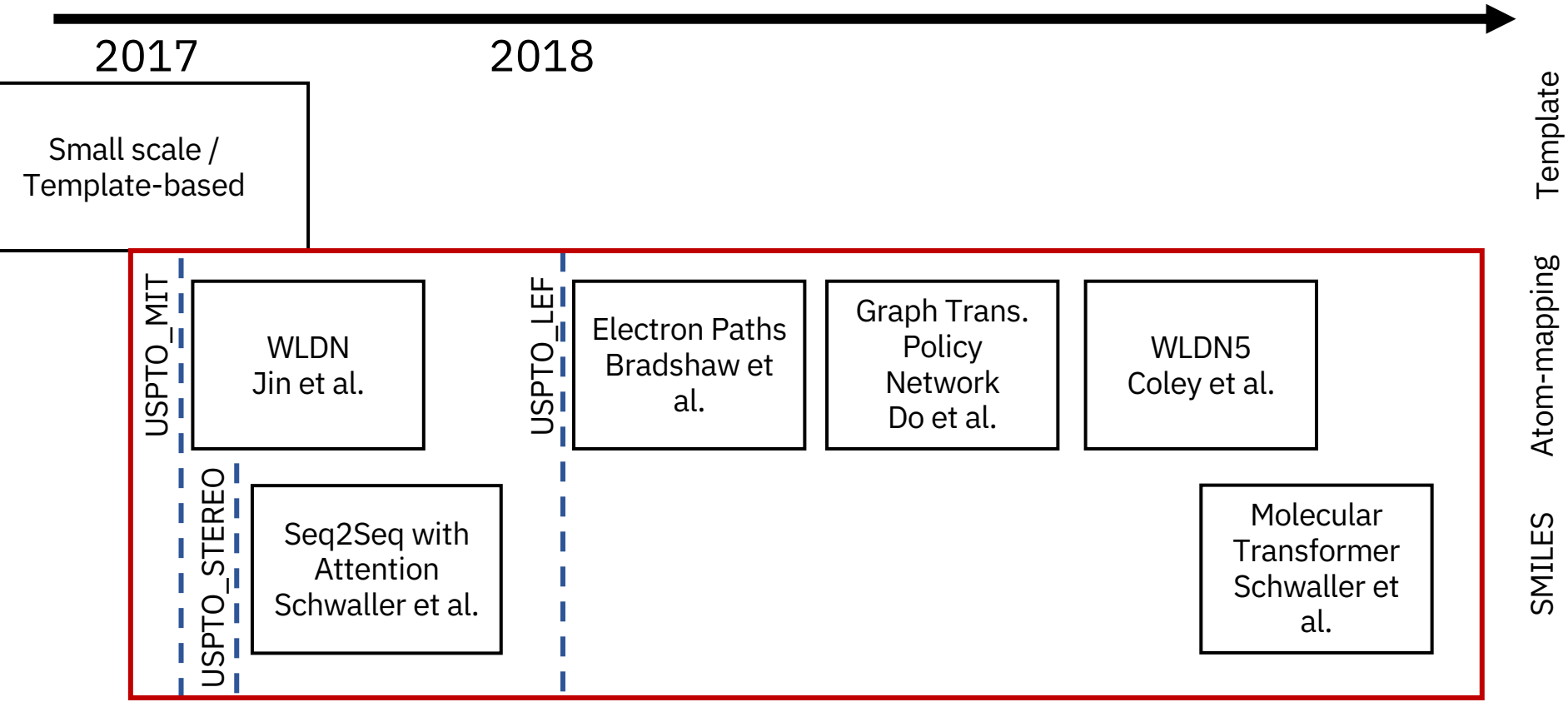
INPUT:
Brc1ccc2c(c1)c1cc3c**4**ccccc**4**c**4**ccccc**4**c3cc1n2-c1ccc2c(c1)c1ccccc1n2-c1ccccc1.CCO.Cc1ccccc1.OB(O)c1ccc2ccc3cccnc3c2n1.[Pd]

OUTPUT:
c1ccc(-n2c3ccccc3c3cc(-n4c5ccc(c6ccc**7**ccc**8**cccnc**8**c**7**n6)cc5c5cc6c**7**ccccc**7**c**7**ccccc7c6cc54)ccc32)cc1

# Features of Smiles-2-Smiles Approach

- Trained end-2-end
- Fully data-driven
- Rule / template-free
- Physics agnostic model
- Attention weights & confidence score

# Recent Data-Driven Reaction Prediction on open-source USPTO dataset



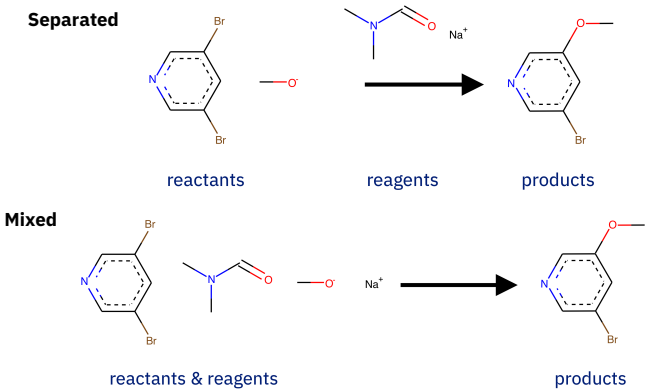
# Top-1 accuracy [%] comparison

**require atom-mapping in training set**

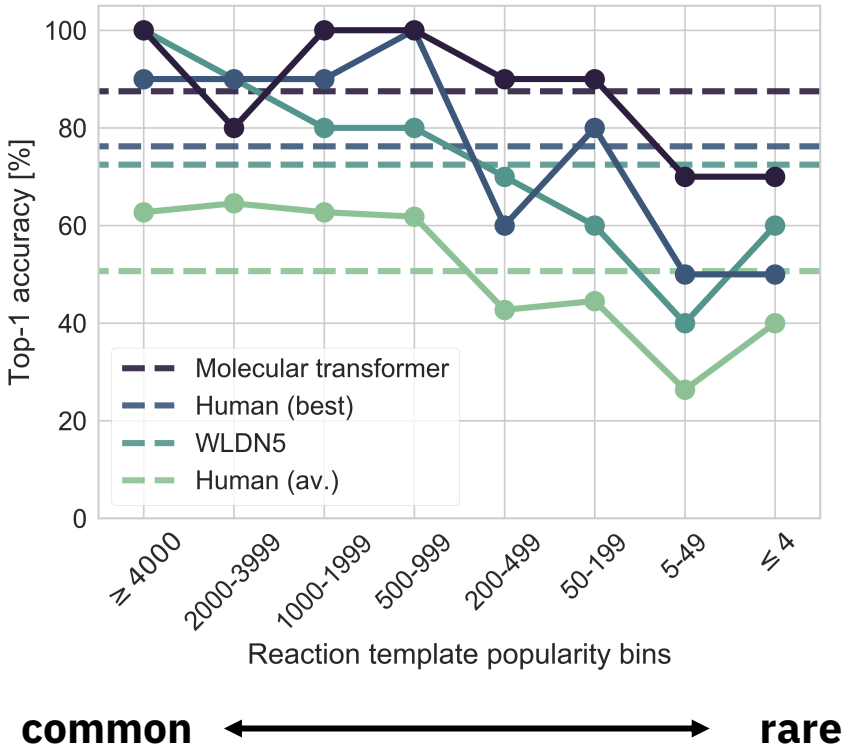
| USPTO* | | S2S [1] | WLDN[2] | ELEC [3] | GTPN [4] | WLDN5 [5] | **MT** |
|---|---|---|---|---|---|---|---|
| _MIT_sep | 500k | 80.3 | 79.6 | | 82.4 | 85.6 | **90.4** |
| _MIT_mixed | | | 74.0 | | | | **88.6** |
| _LEF_sep | 350k | | 84.0 | 87.0 | 87.4 | 88.3 | **92.0** |
| _LEF_mixed | | | | | | | **90.3** |


[1] Schwaller et al.: Chem. Sci., 2018, 9, 6091-6098
[2] Jin et al.: NIPS, 2017, 30, 2607-2616
[3] Bradshaw et al.: arXiv:1805.10970
[4] Do et al.: arXiv:1812.09441
[5] Coley et al. : Chem. Sci., 2019, 10, 370-377


**MT** **Molecular Transformer for Chemical Reaction Prediction and Uncertainty Estimation**



# Human prediction benchmark




**87.5 % Molecular Transformer**

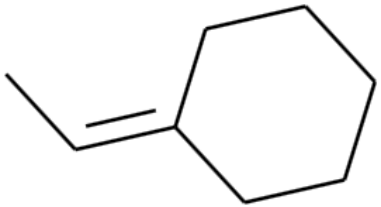
**76.5 % best human**
**72.5 % Coley et al. model**

**50.6 % average human**

- **80 reactions** (10 reactions per bin)
- Given to **11 chemists**
- **Mixed** input setting for all

Original study performed by Coley et al. 2018 – WLDN5

# What else can we do? Reaction scoring



+ HCl

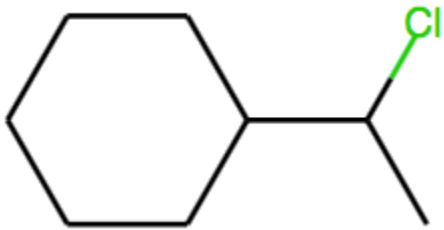
Markovnikov

CC=C1CCCCC1.Cl>>CCC1(Cl)CCCCC1


Score: 0.99

Anti-Markovnikov

CC=C1CCCCC1.Cl>>CC(Cl)C1CCCCC1

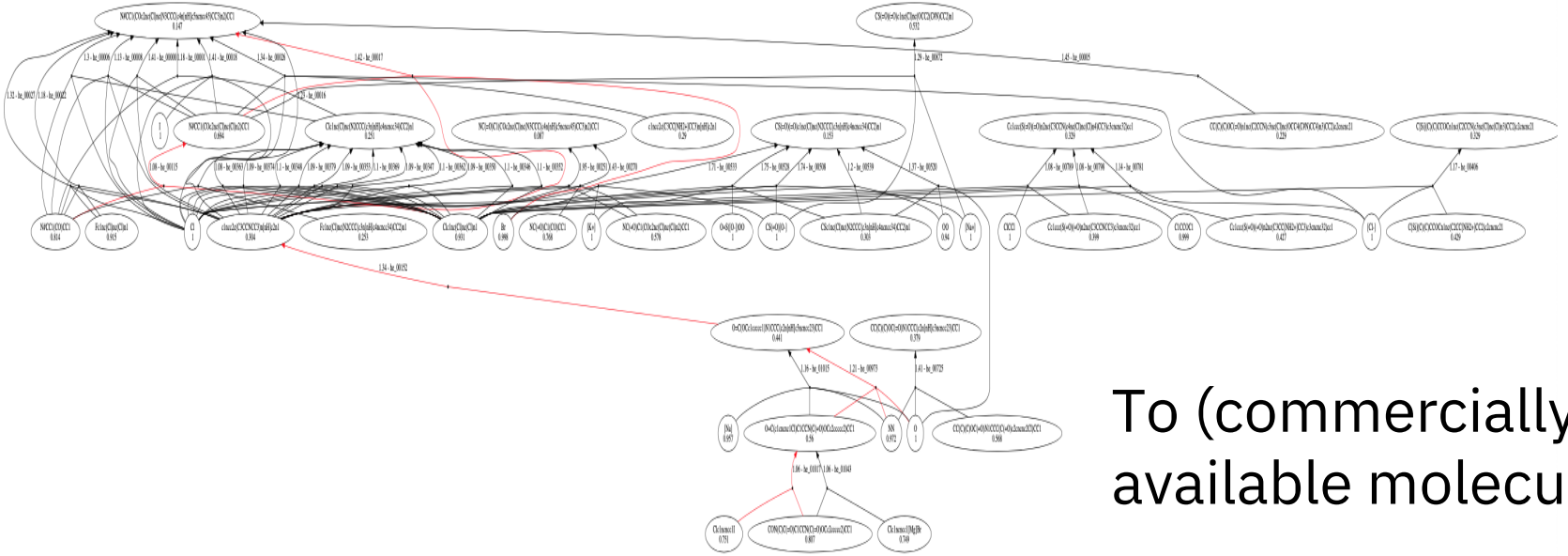



Score: 0.001

When we provide the model with **reactants>>products**

# Single step validation in retrosynthesis

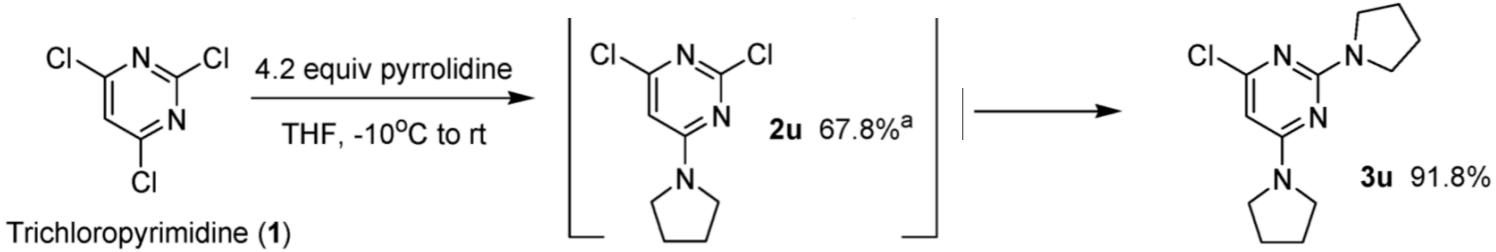
Target



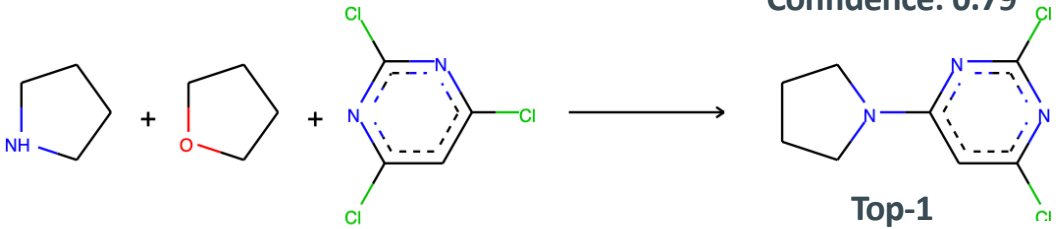
To (commercially) available molecules

# Regioselectivity – Halogenated Pyrimidines

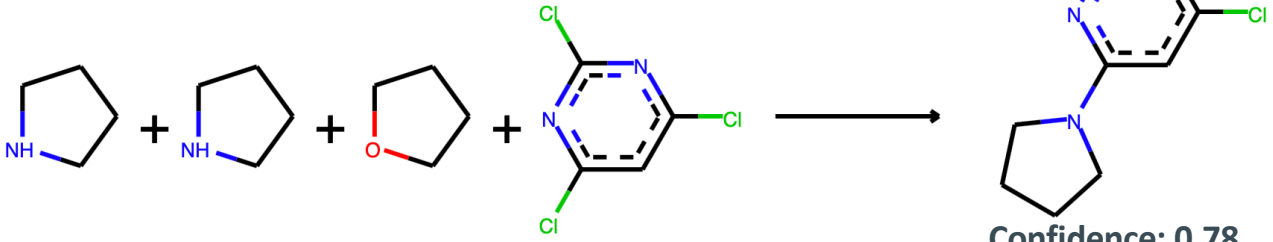
*Org. Process Res. Dev.*, **2006**, *10* (5), pp 921–926 **DOI:** 10.1021/op060093q




Confidence: 0.79




Top-1

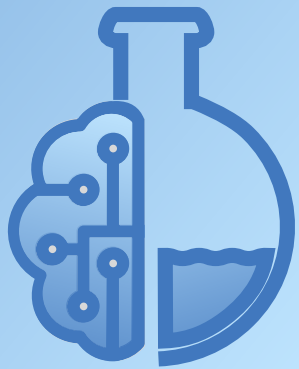


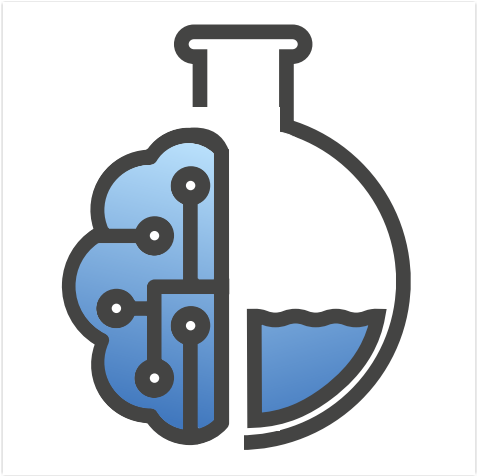

Confidence: 0.78

# IBM **RXN** for Chemistry

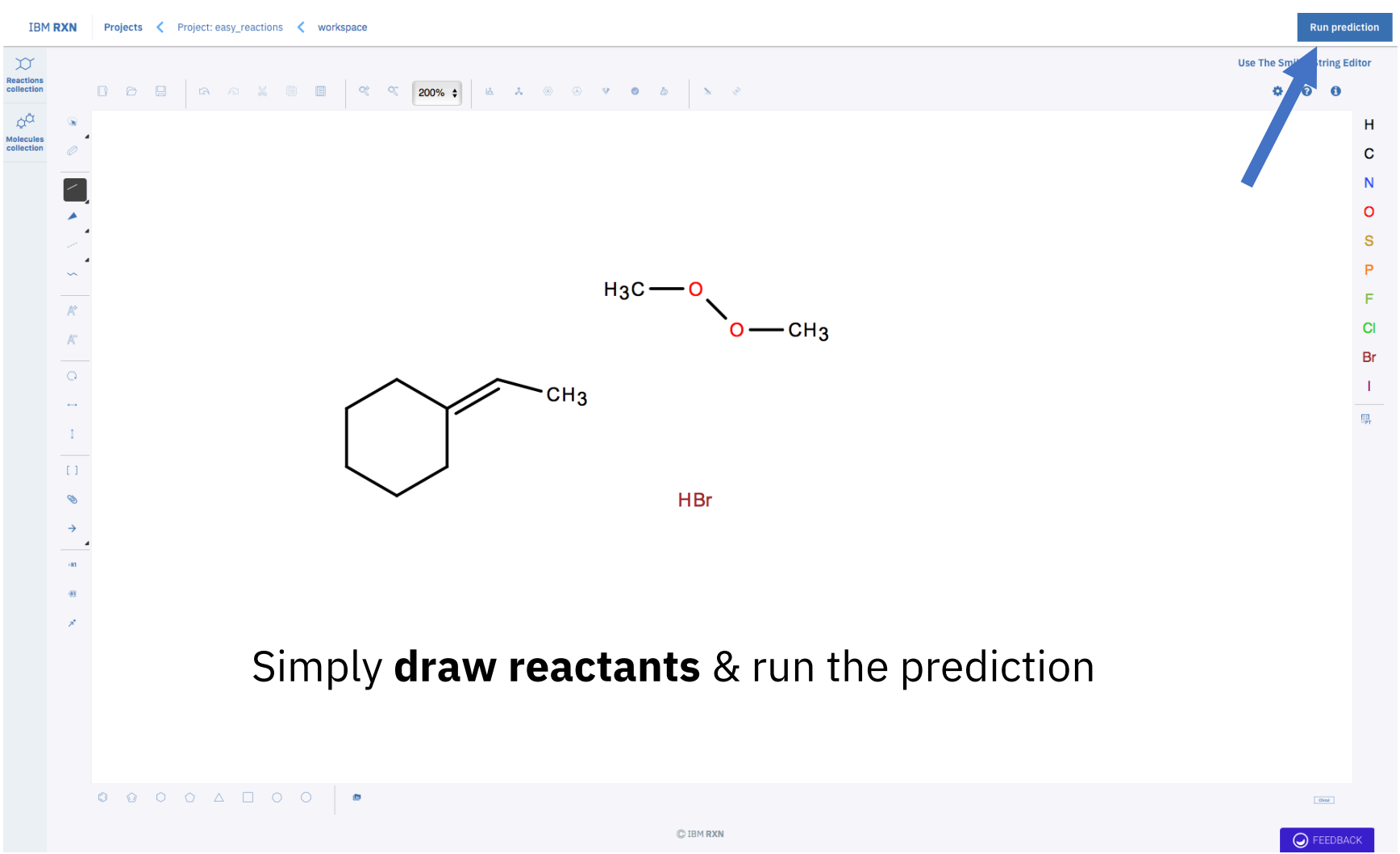
#RXNFORCHEMISTRY

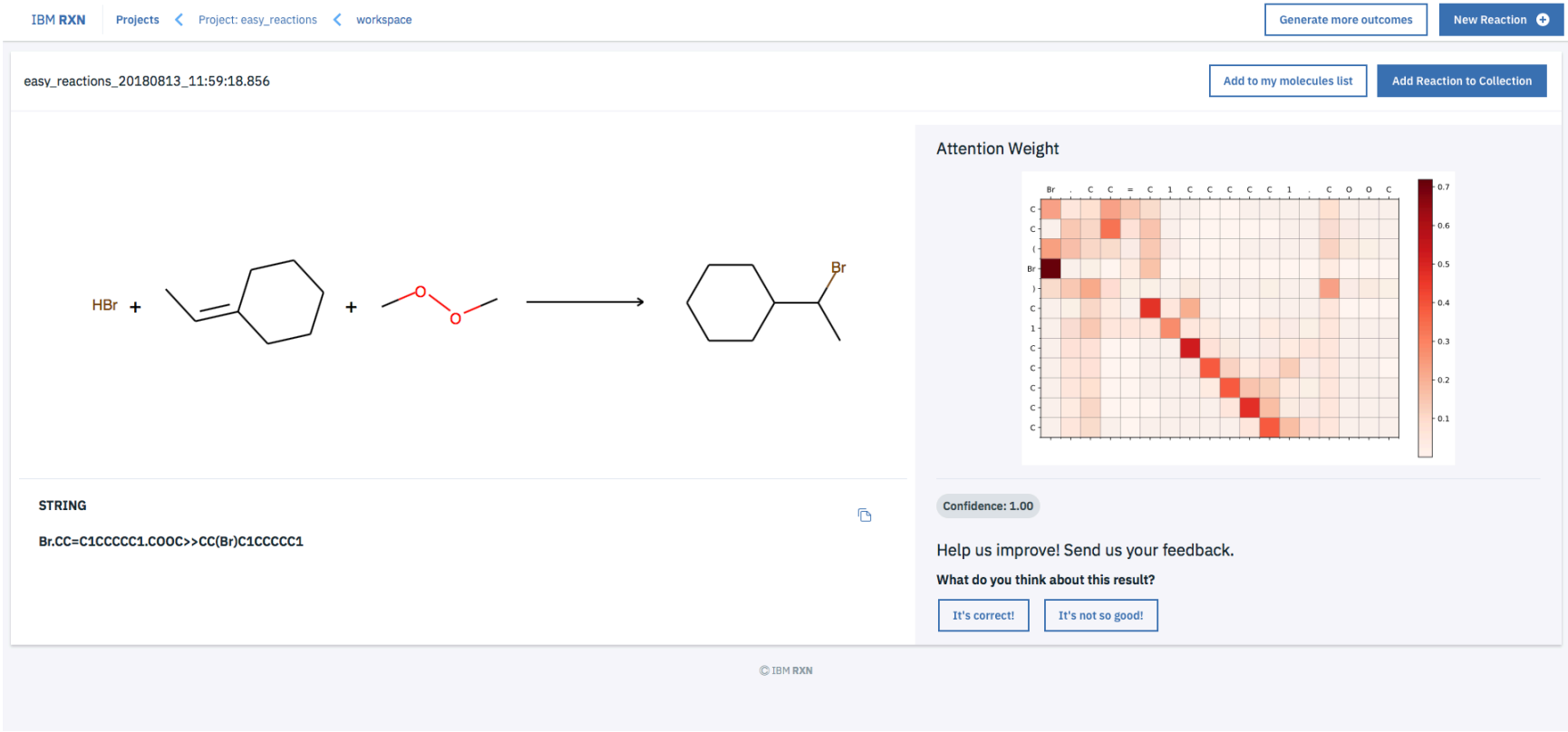
Freely available on:
**rxn.res.ibm.com**



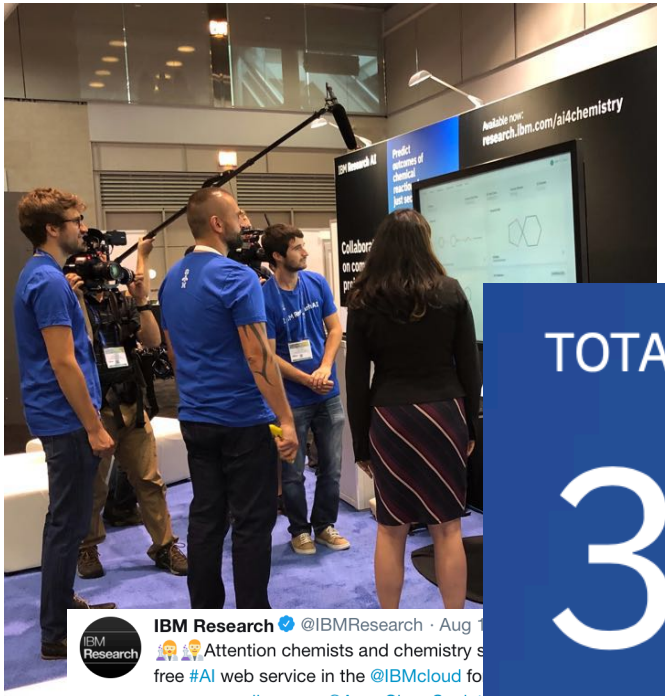


# Live DEMO


IBM RXN
Projects
Project: easy_reactions
workspace
Run prediction
Reactions collection
Molecules collection
200%
H
C
N
O
S
P
F
Cl
Br
I
H3C
O
O
CH3
CH3
HBr
Simply draw reactants & run the prediction
© IBM RXN
FEEDBACK




Get back the **product**, the **attention** weights and the **confidence**




452

Posted by u/ibmzrl 2 days ago

# IBM launches the first, FREE AI web service for predicting chemical reactions

rxn.res.ibm.com/

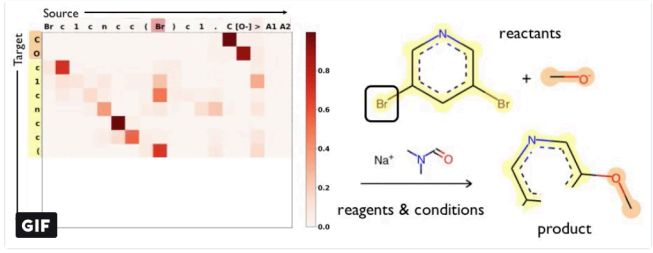
79 Comments Share Save •••

## TOTAL PREDICTED REACTIONS

# 37442

**IBM Research** @IBMResearch · Aug

Attention chemists and chemistry s
free #AI web service in the @IBMcloud fo
now rxn.res.ibm.com @AmerChemSociety




21 656 942


BOSTON
ACS
Chemistry for Life®
true
science
AMERICAN CHEMICAL SOCIETY
HOUSING HELP DESK

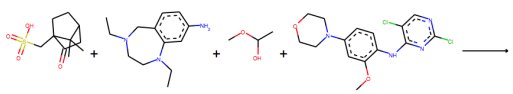
# Human vs AI challenge

| Human Total Time | AI Total Time | Human Correct Answers | AI Correct Answers | Total Challenges |
|---|---|---|---|---|
| 02:40:58.3 | 00:02:36.0 | 35 | 64 | 86 |

Your challenges results

| Challenge | Human Total Time | AI Total Time | Human Answer | AI Answer |
|---|---|---|---|---|
| Chloro N-arylation | 00:02:30.2 | 00:00:04.3 | Correct | Correct |

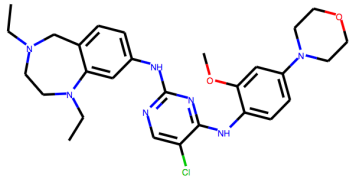
**Reactants**



**STRING**

CC1(C)C2CCC1(CS(=O)(=O)O)C(=O)C2.CCN1CCN(CC)c2cc(N)ccc2C1.COC(C)O.COc1cc(N2CCOCC2)ccc1Nc1nc(Cl)ncc1Cl

**Ground Truth**



**STRING**

CCN1CCN(CC)c2cc(Nc3ncc(Cl)c(Nc4ccc(N5CCOCC5)cc4OC)n3)ccc2C1

# API access & open-source code

IBM RXN API - https://rxn.res.ibm.com

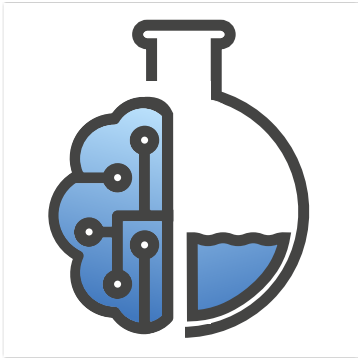
```
curl --data '{ "reactants" :
"C(O)1=C(O)C=C(C2=C(O)C=C([H])C([H])=C2C2=C([H])C=C([H])C=C2)C=C1O.[H].[H].
[H]",  "mol":"" }'  --header "Content-Type: application/json" --header "
Authorization:apk-have-here-your-own-key"  -X POST
https://rxn.res.ibm.com/rxn/api/api/v1/predictions/pr?projectId=5c532f56d6cb7
600019ea342
```

Code, trained models, data:
https://github.com/pschwllr/MolecularTransformer

# Key Messages

- **SMILES** (and other line notations) might be an alternative representation → **N**atural **L**anguage **P**rocessing methods
- The **Molecular Transformer** is **universally applicable** across existing reaction datasets
  - No distinction between reactants-reagents required
  - Predictions in few hundreds of ms
  - Current state of the art in data-driven reaction prediction
- Have a look at **IBM RXN for Chemistry (it's free!)**, give us feedback and help us improve

# Thank you for your attention!



IBM **Research** AI

UNIVERSITY OF CAMBRIDGE

THE WINTON PROGRAMME FOR THE Physics of Sustainability

Philippe Schwaller

phs@zurich.ibm.com / Twitter: @pschwllr

**- "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models (https://doi.org/10.1039/C8SC02339E)**

**- Molecular Transformer for Chemical Reaction Prediction and Uncertainty Estimation (https://dx.doi.org/10.26434/chemrxiv.7297379)**

NextMove Software

docker

RDKit
Open-Source Cheminformatics and Machine Learning

NMT
An open source neural machine translation system.

PyTorch