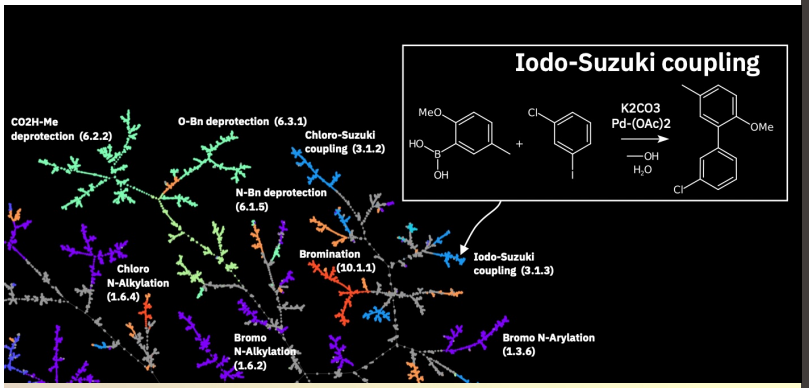
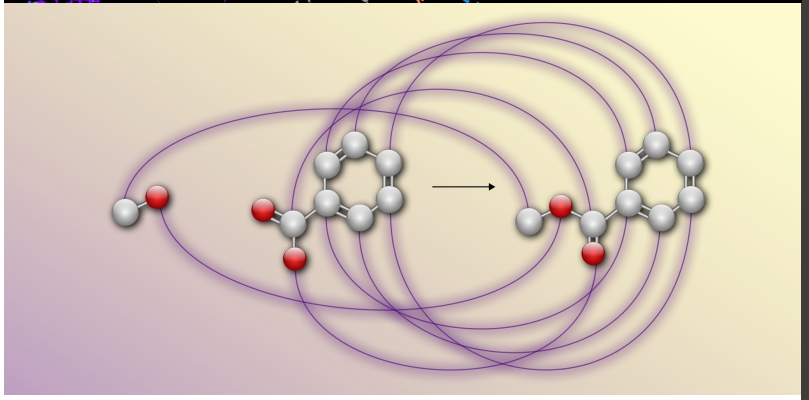
09.05.2022





Machine Learning for Synthesis Planning

Philippe Schwaller (he/him/his)

Laboratory of Artificial Chemical Intelligence (LIAC)

@SchwallerGroup

philippe.schwaller@epfl.ch

IBM **Research**

UNIVERSITÄT BERN

NCCR Catalysis

Grew up in Fribourg, Switzerland
- French
- Swiss German / German

NCCR Catalysis

TT Assistant Prof
in Digital Chemistry
since Feb 2022

EPFL LIAC

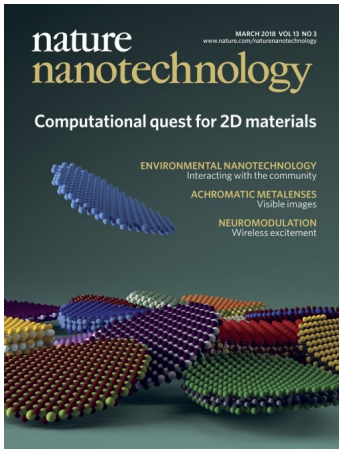
Collaboration with synthetic chemists

b UNIVERSITÄT BERN

PhD in Chemistry
and Molecular Sciences (‘21)
Prof Jean-Louis Reymond

EPFL

MANCHESTER 1824
The University of Manchester

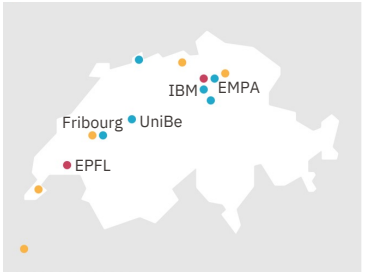


Materials Science &
Engineering

Virtual screening &
simulation workflows
Prof Nicola Marzari

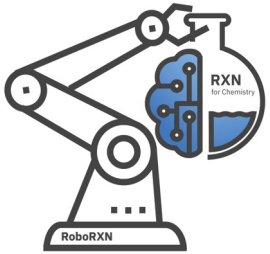
BSc (‘14)/MSc (‘16)





MPhil in Physics (’19)
Dr Alpha Lee

UNIVERSITY OF CAMBRIDGE





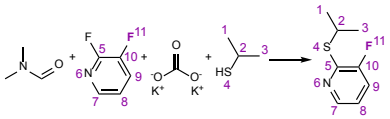
Empa
Materials Science and Technology

Lab work on ternary polymer
blends for organic solar cells
Prof Frank Nüesch

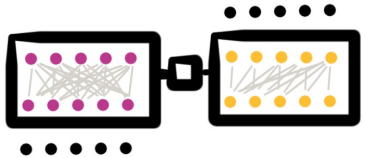
Machine learning for chemical synthesis
Intern/PhD/Postdoc
Dr Teodoro Laino

IBM **Research** Europe

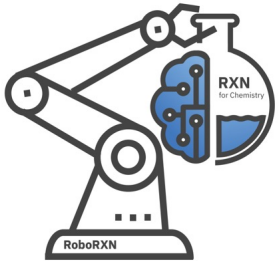
Outline



CC(C)S.CN(C)C=O.Fc1cccnc1
F.O=C([O-])[O-].[K+].[K+]
>>CC(C)Sc1ncccc1F

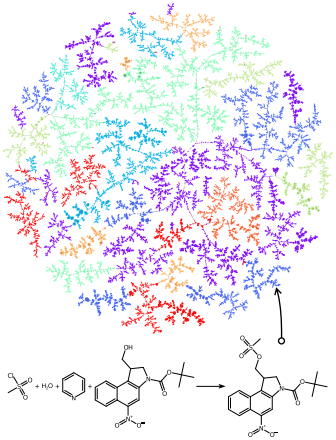


Data Representation Models

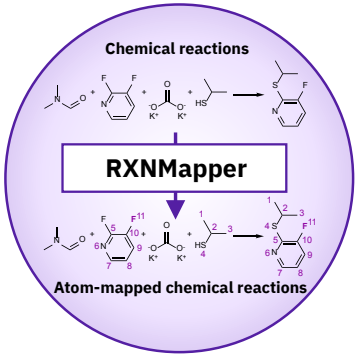




Real-world applications



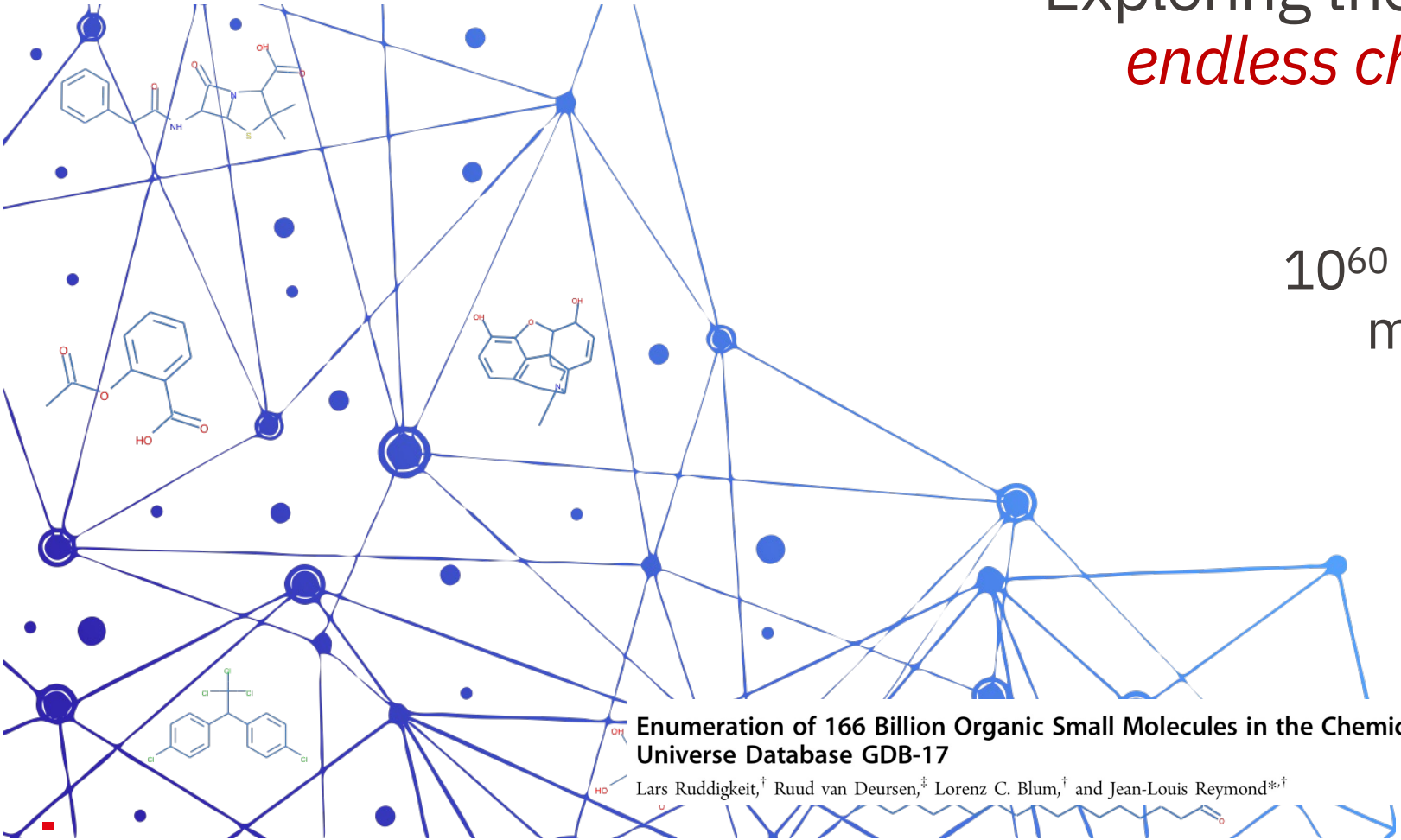
Explore Chemical Reaction Space



What chemistry do the models learn?

Exploring the nearly *endless chemical space*

10^{60} drug-like molecules



Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17

Lars Ruddigkeit,[†] Ruud van Deursen,[‡] Lorenz C. Blum,[†] and Jean-Louis Reymond*[,†]

What molecule to make?

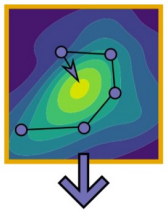
Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli[†#], Jennifer N. Wei[‡#], David Duvenaud[¶#], José Miguel Hernández-Lobato[§#], Benjamín Sánchez-Lengeling[‡], Dennis Sheberla[‡], Jorge Aguilera-Iparraguirre[†], Timothy D. Hirzel[†], Ryan P. Adams[¶⊥], and Alán Aspuru-Guzik[*‡⊥]

[Submitted on 5 Jan 2017]

Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks

Marwin H.S. Segler, Thierry Kogej, Christian Tyrchan, Mark P. Waller

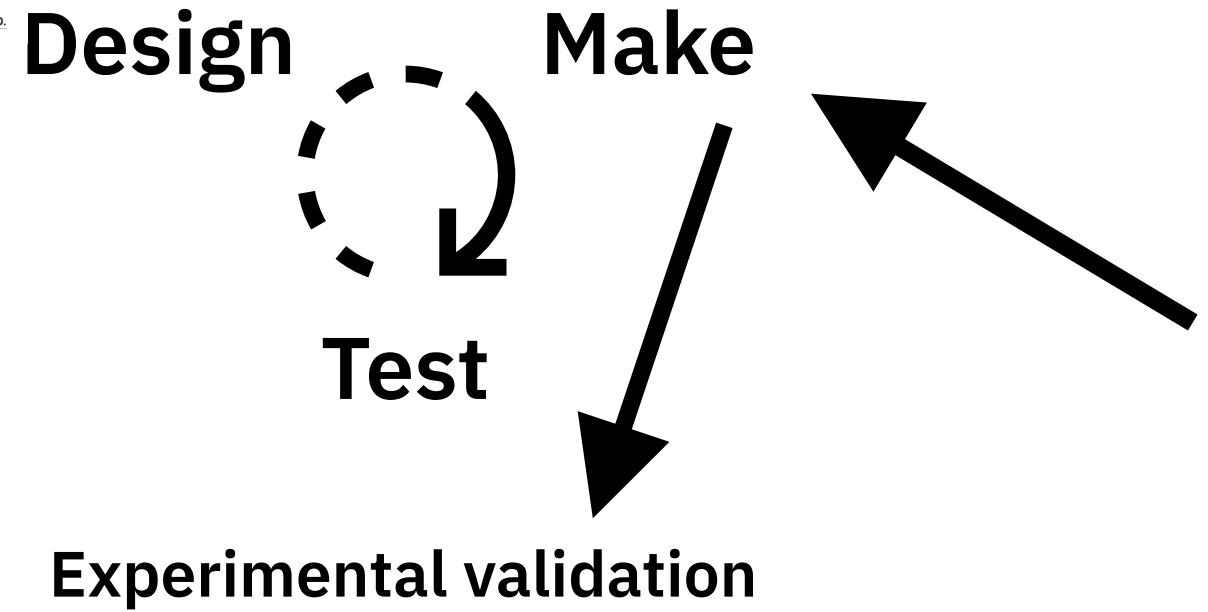


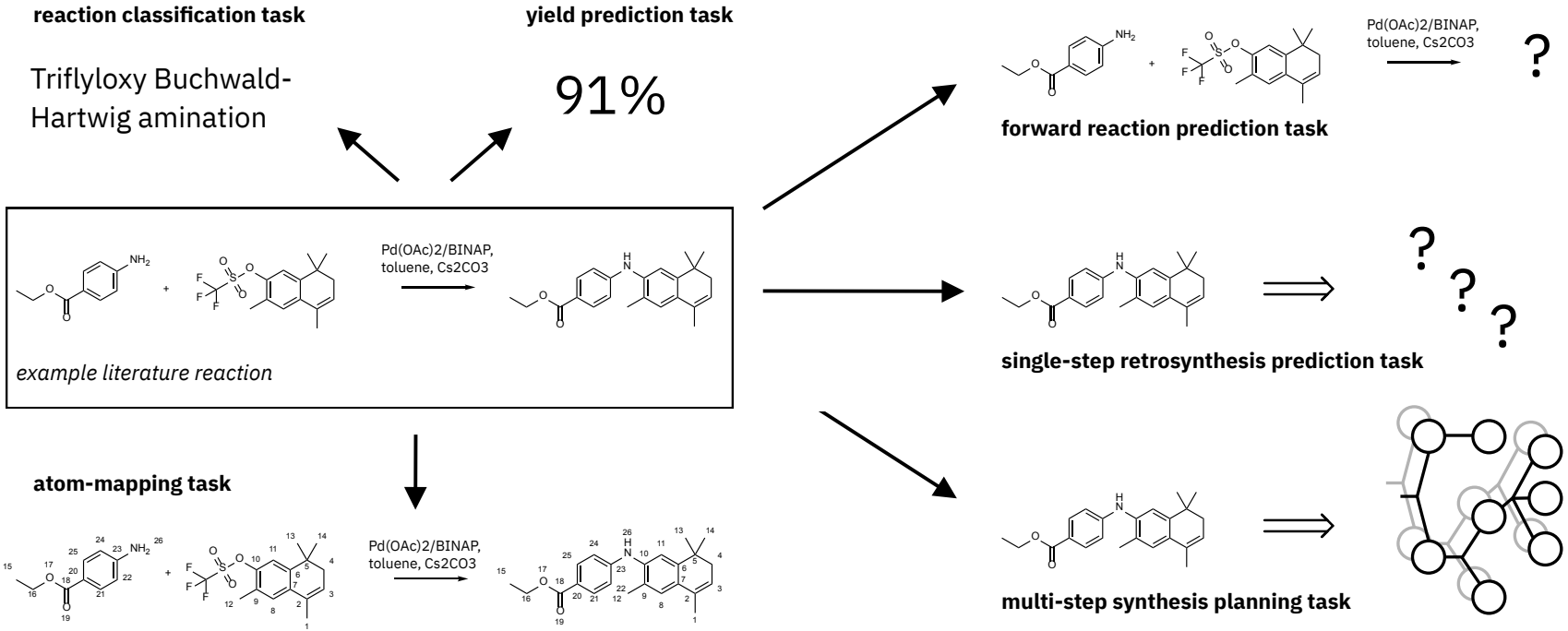
Inverse:
Optimization, evolutionary strategies, generative models

How to make it?

Reaction prediction

Synthesis planning

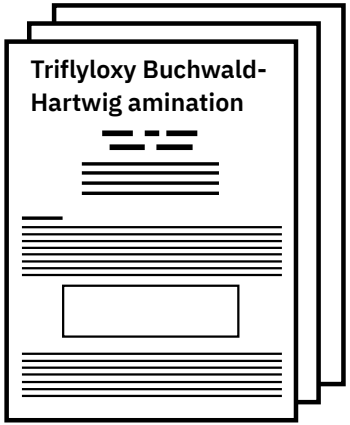




Machine Intelligence for Chemical Reaction Space. Schwaller et al. 2022
https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wcms.1604

Data sources for ML?

Literature (broad, e.g, Thieme)

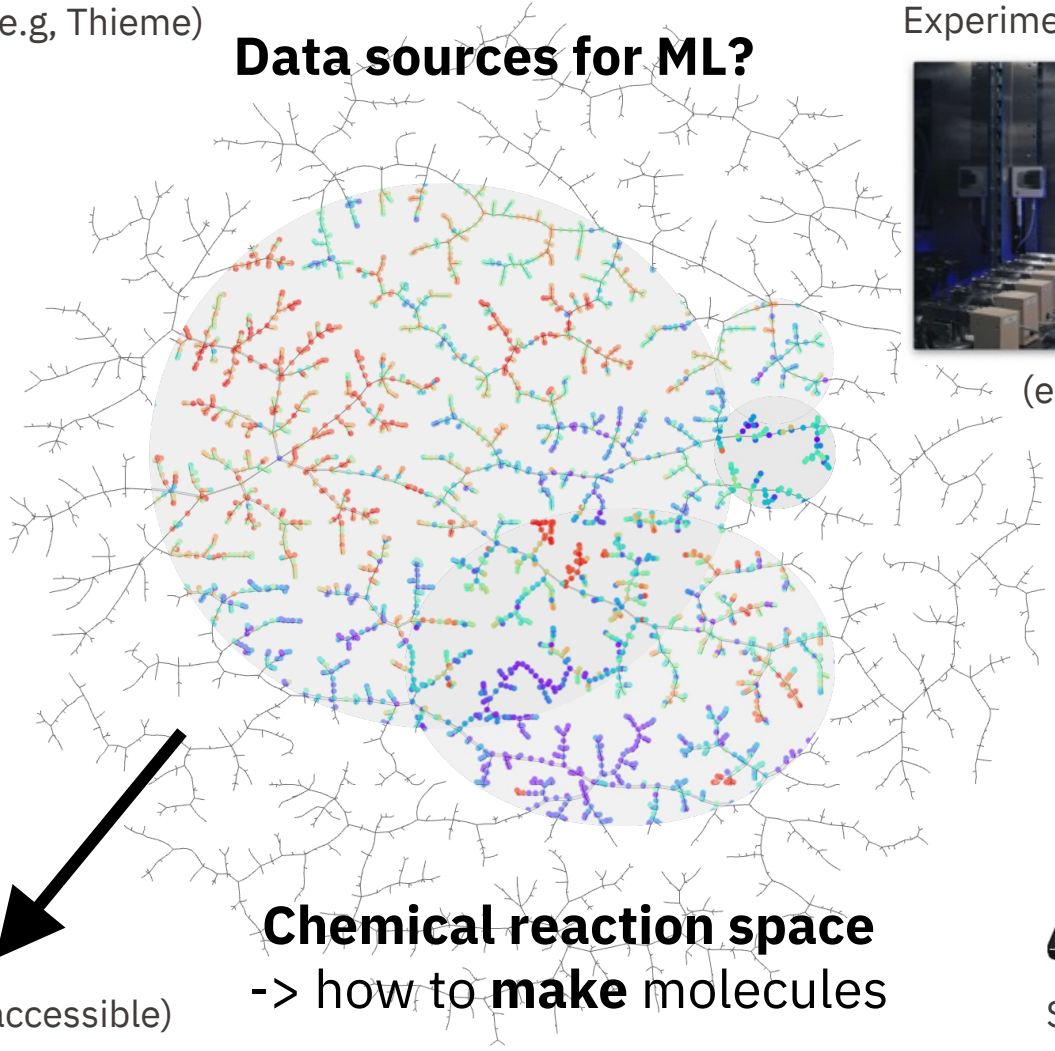


Experiments ELN/HTE (narrow)



(e.g. ORD, Kearnes et al.)

US20030166932A1: General Procedure H A solution of trifluoromethanesulfonic acid 3,5,8,8-tetramethyl-7,8-dihydronaphthalen-2-yl ester (Compound 35, 0.41 g, 1.2 mmol), Pd(OAc)2 (0.027 g, 0.12 mmol), BINAP (0.11 g, 0.18 mmol), Cs2CO3 (0.56 g, 1.72 mmol), ethyl 4-aminobenzoate (0.25 g, 1.5 mmol) and 5 mL of toluene was flushed with argon for 10 min, then stirred at 100° C. in a sealed tube for 48 h. After the reaction mixture had been cooled to room temperature, the solvent was removed, and the residue was purified by flash column (hexane:ethyl acetate=4:1) to give 0.34 g (80%) of the title compound as a yellowish solid.



Chemical reaction space
-> how to **make** molecules

$$\hat{H}\Psi = E\Psi$$

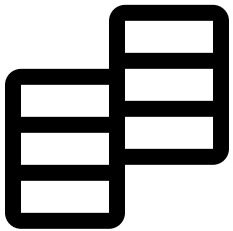


Simulations (narrow)

- Patents (broad, accessible)

Chemical reaction data

US Patents



Text-mining (Lowe 2012/17)

Millions of reactions

BrC(Br)(Br)Br.CC...>>...
CO.Nc1cccc([N+]...>>...
CC(=O)O[BH-]...>>...
(OC(C)=O)OC(C)=O..>>...
...
precursors>>products

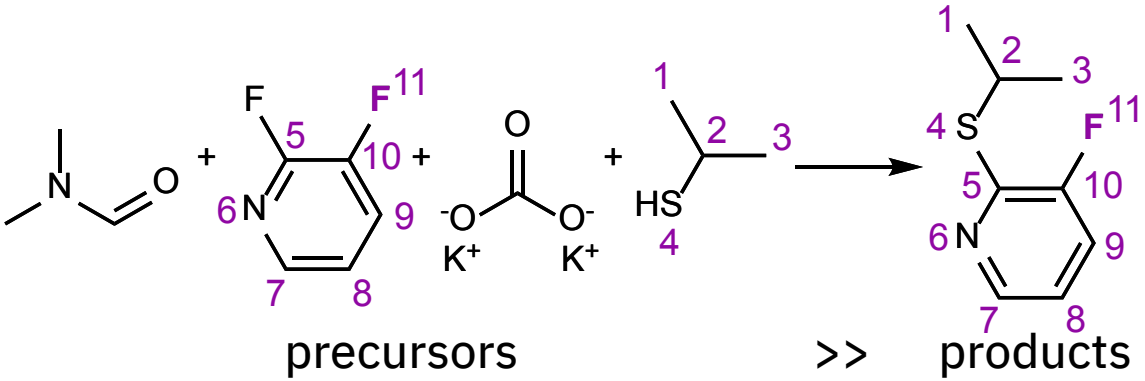
Benchmark sets

USPTO_MIT

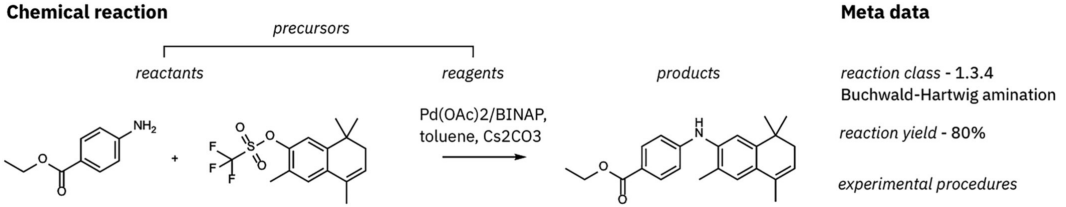
USPTO_STEREO

Reaction SMILES

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



Reaction *representations*

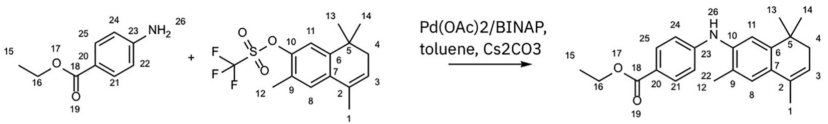


Reaction SMILES (text-based reaction representation, precursors>>products)

CC(=O)[O-].CC(=O)[O-].CC1=CCC(C)(C)c2cc(OS(=O)(=O)C(F)(F)F)c(C)cc21.CCOC(=O)c1ccc(N)cc1.Cc1ccccc1.O=C([O-])[O-].[Cs+].[Cs+].
[Pd+2].c1ccc(P(c2ccccc2)c2ccc3ccccc3c2-c2c(P(c3ccccc3)c3ccccc3)ccc3ccccc23)cc1>>CCOC(=O)c1ccc(Nc2cc3c(cc2C)C(C)=CCC3(C)C)cc1

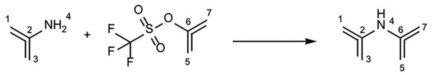
Atom-mapping (e.g. RXNMapper)

Atom-mapped reaction (required for reaction template, centre and bond change extraction)

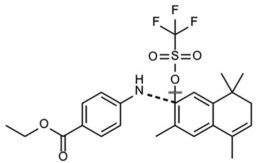


CC(=O)[O-].CC(=O)[O-].Cc1ccccc1.O=C([O-])[O-].O=S(=O)(O[c:11]1[cH:12][c:13]2[c:14]([cH:15][c:16]1[CH3:17])[C:18]([CH3:19])=[CH:20]
[CH2:21][C:22]2([CH3:23])[CH3:24])C(F)(F)F.[CH3:1][CH2:2][O:3][C:4](=[O:5])[c:6]1[cH:7][cH:8][c:9]([NH2:10])[cH:25][cH:26]1.[Cs+].[Cs+].
[Pd+2].c1ccc(P(c2ccccc2)c2ccc3ccccc3c2-c2c(P(c3ccccc3)c3ccccc3)ccc3ccccc23)cc1>>[CH3:1][CH2:2][O:3][C:4](=[O:5])[c:6]1[cH:7][cH:8]
[c:9]([NH:10][c:11]2[cH:12][c:13]3[c:14]([cH:15][c:16]2[CH3:17])[C:18]([CH3:19])=[CH:20][CH2:21][C:22]3([CH3:23])[CH3:24])[cH:25][cH:26]1

Reaction template



Condensed Graph of Reaction



Chemical reactions from US patents (1976-Sep2016)

Dataset posted on 13.06.2017, 18:49 by Daniel Lowe

12748 views | 8727 downloads

Reactions extracted by text-mining from United States patents published between 1976 and September 2016. The reactions are available as CML or reaction SMILES. Note that the reactions SMILES are derived from the CML. The files can be unzipped using a program like 7-Zip.

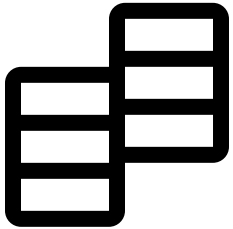


"While **typically correct**, the **atom-maps** are **wrong in many cases** and hence should not be entirely relied on."

https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873

Chemical reaction data

US Patents



Text-mining (Lowe 2012/17)

Millions of reactions

BrC(Br)(Br)Br.CC...>>...
CO.Nc1cccc([N+]...>>...
CC(=O)O[BH-]...>>...
(OC(C)=O)OC(C)=O..>>...
...
precursors>>products

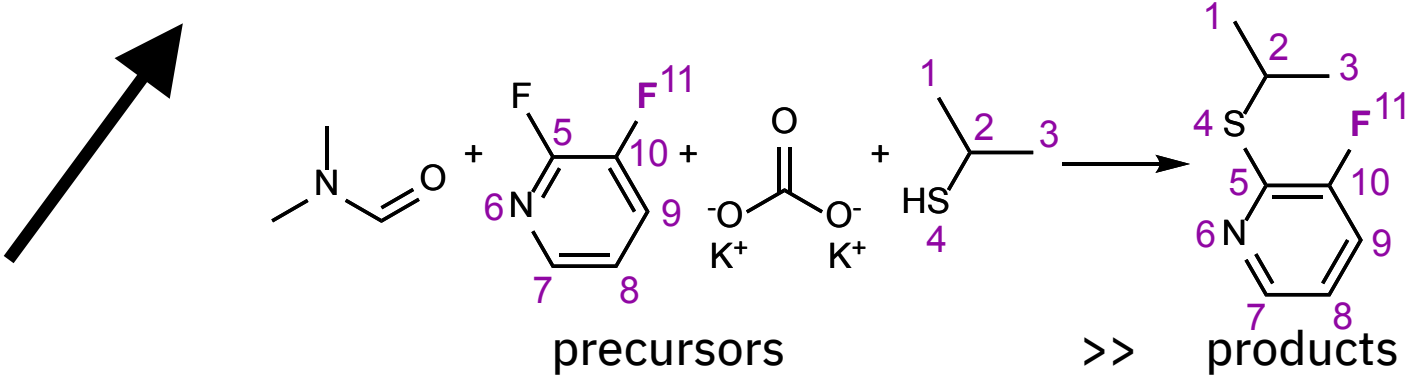
Benchmark sets

USPTO_MIT

USPTO_STEREO

Reaction SMILES

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

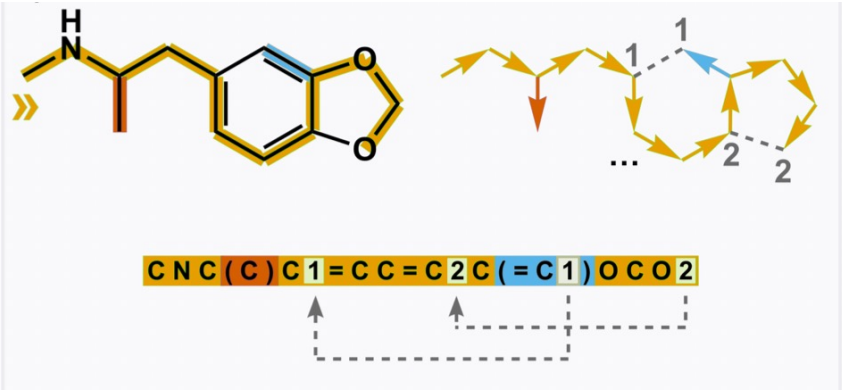


SMILES

SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules

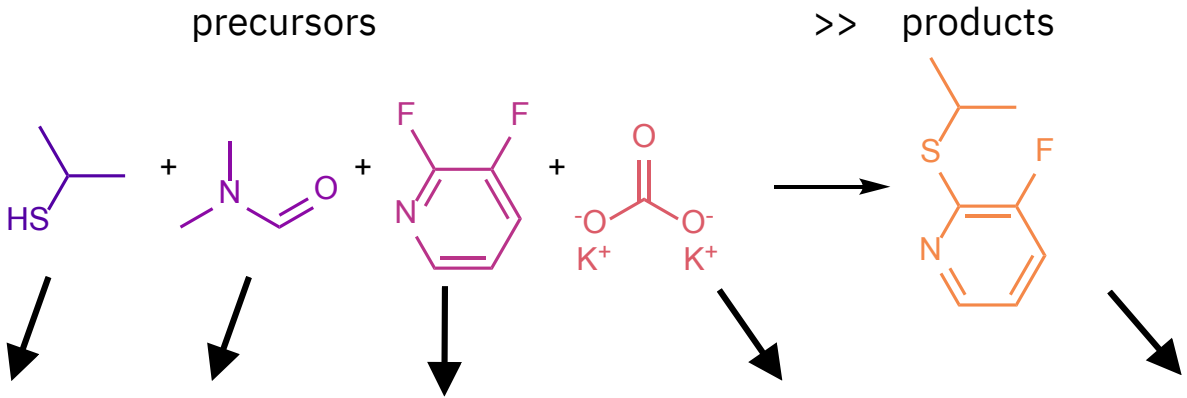
DAVID WEININGER
Medicinal Chemistry Project, Pomona College, Claremont, California 91711

Received June 17, 1987



Krenn, Mario, et al. "SELFIES and the future of molecular string representations." *arXiv preprint arXiv:2204.00056* (2022).

Atoms as *letters*, molecules as *words*



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

Split -> “sequences of atoms” called tokens

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

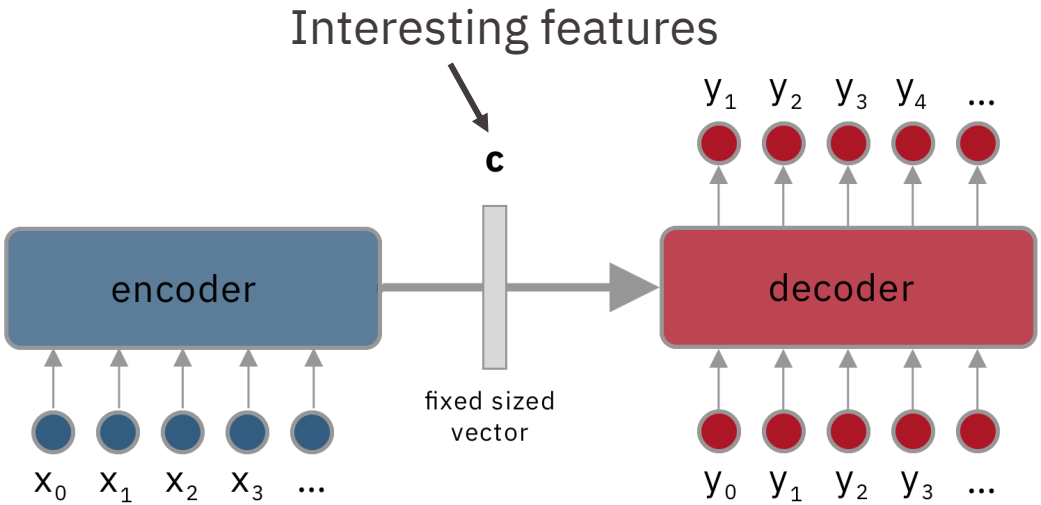
→ Borrow methods developed for human languages

Nam & Kim, arXiv:1612.09529; Liu et al., ACS Centr. Sci. 2017; Schwaller et al., Chem. Sci, 2018

Sequence-2-sequence models

French: Le chat est noir.

German: Die Katze ist schwarz.



Problem: fixed size

INPUTS = reactants + reagents

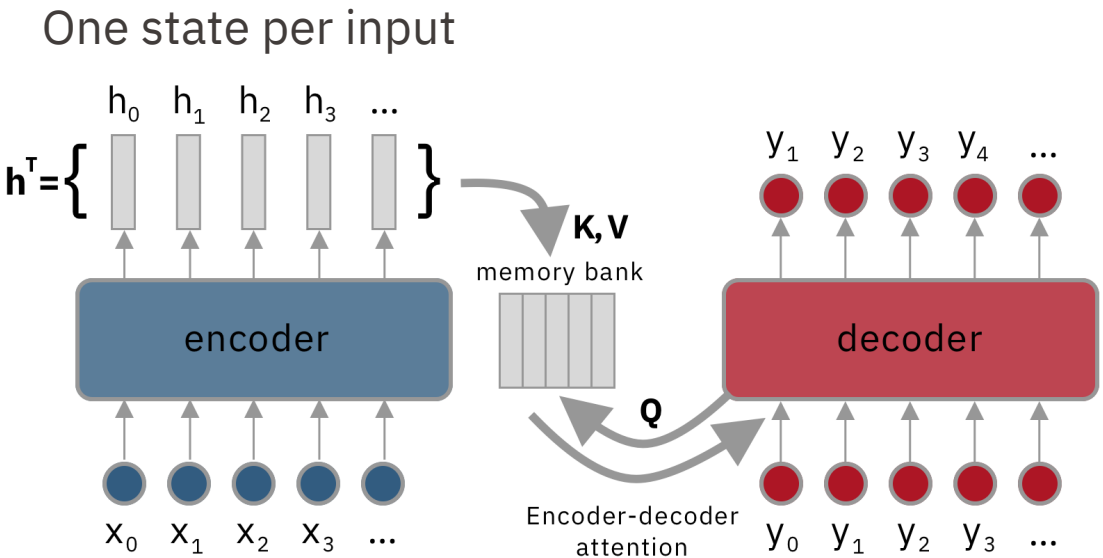
Br c 1 c n c c (Br) c 1 . C N (C) C = O . C [O-] . [Na+]

OUTPUTS = products

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

Sutskever et al., Sequence to Sequence Learning with Neural Networks. NeurIPS, 2014.
Cho et al. , Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP, 2014.

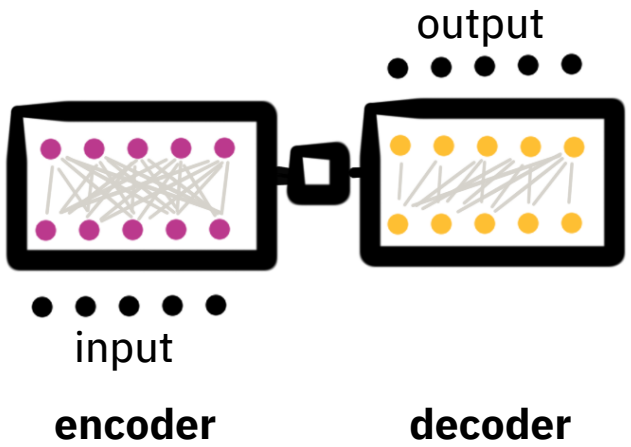
Sequence-2-sequence models *with attention*



Attention = ability to focus on most important features

Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015
Luong et al., Effective approaches to attention-based neural machine translation, EMNLP, 2015

Transformer architecture

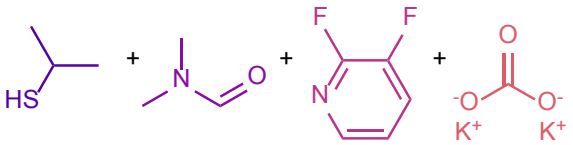


- Stacks of attention layers
- Multi-head attention

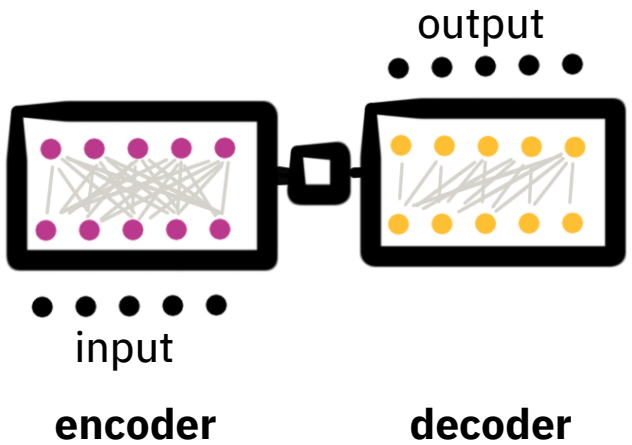
Vaswani et al., Neural Machine Translation by Jointly Learning to Align and Translate. NeurIPS, 2017

Molecular Transformer

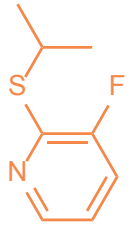
precursors



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



products



CC(C)Sc1ncccc1F

- **No rules** integrated / no chemical knowledge
- **Accurate predictions** on unseen reactions
- Better than rule and graph-based approaches

Schwaller et al., Molecular Transformer – A Model for Uncertainty-Calibrated Chemical Reaction Prediction. ACS Central Science, 2019

UNIVERSITY OF CAMBRIDGE

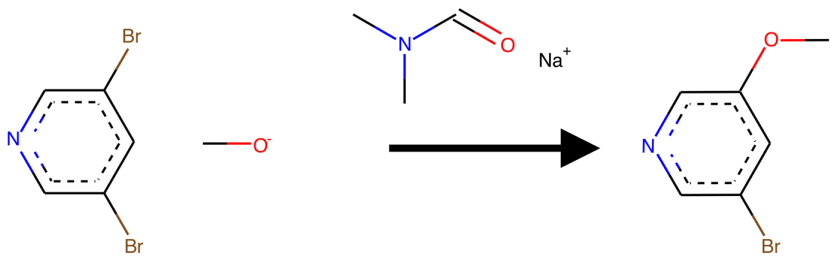
IBM **Research**

USPTO-MIT benchmark (no stereochemistry)

	2018		2020	
Top-1 Acc. [%]	**WLDN5 Coley et al.**	**Molecular Transformer**	**Graph-NN Qian et a.**	**Augmented MT. Tetko et al.**
separated	85.6	90.4	90	91.9
mixed	74 (earlier version)	88.6	Not possible	90.4

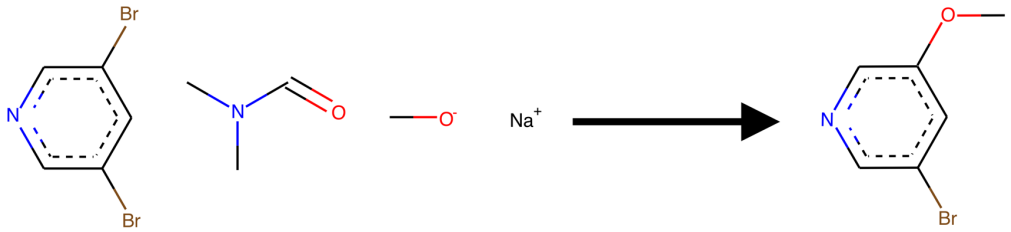
Separated vs mixed setting

Separated



reactants reagents products

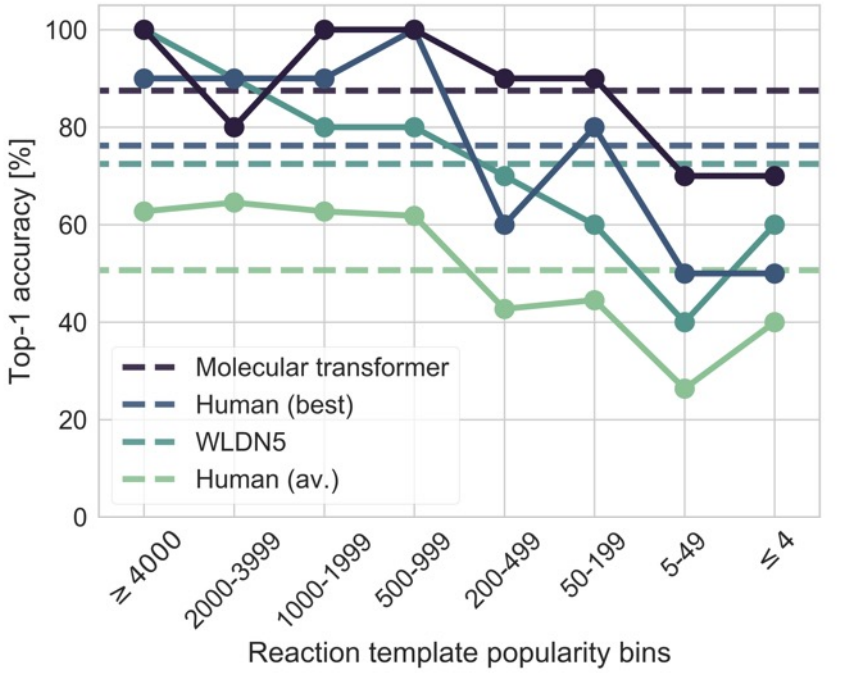
Mixed



reactants & reagents products

No distinction between reactants and reagents

Human prediction benchmark



common ⟷ **rare**

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**

[HTML] A graph-convolutional neural network model for the prediction of chemical reactivity
CW **Coley**, W Jin, L Rogers, TF Jamison… - Chemical …, 2019 - pubs.rsc.org
We present a supervised learning approach to predict the products of organic reactions given their reactants, reagents, and solvent (s). The prediction task is factored into two stages comparable to manual expert approaches: considering possible sites of reactivity …
☆ 99 Cited by 132 Related articles All 8 versions

Graph-edit-based, atom-mapping dependent

Methods	Top-n accuracy (%)			
	1	3	5	10
USPTO_480k_mixed				
MEGAN (Sacha et al., 2021)	86.3	92.4	94.0	95.4
Molecular Transformer (Schwaller et al., 2019)	88.6	93.5	94.2	94.9
Graph2SMILES (D-GCN) (*ours*)	90.3	94.0	94.6	95.2
Graph2SMILES (D-GAT) (*ours*)	90.3	94.0	94.8	95.3
Augmented Transformer (Tetko et al., 2020)	90.6	-	**96.1**	-
Chemformer (Irwin et al., 2021)	**91.3**	-	93.7	94.0

Chemformer: a pre-trained transformer for computational chemistry

Ross Irwin[1], Spyridon Dimitriadis[1,2], Jiazhen He[1] and Esben Jannik Bjerrum[3,1]

State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis

Igor V. Tetko, Pavel Karpov, Ruud Van Deursen & Guillaume Godin

Augmented Transformer

PERMUTATION INVARIANT GRAPH-TO-SEQUENCE MODEL FOR TEMPLATE-FREE RETROSYNTHESIS AND REACTION PREDICTION

Graph2SMILES

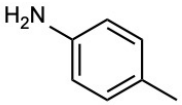
Zhengkai Tu[1,2] and Connor W. Coley[1,3]

Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits

Mikołaj Sacha, Mikołaj Błaż, Piotr Byrski, Paweł Dąbrowski-Tumański, Mikołaj Chromiński, Rafał Loska, Paweł Włodarczyk-Pruszyński, and Stanisław Jastrzębski*

MEGAN

Extensive data augmentations



=

c1c(N)ccc(C)c1	Cc1ccc(N)cc1
c1cc(C)ccc1N	c1(N)ccc(C)cc1
c1c(C)ccc(N)c1	Nc1ccc(C)cc1
c1(C)ccc(N)cc1	c1cc(N)ccc1C

Molecule SMILES randomizations

{aryl_halide}.{methylaniline}.{pd_catalyst}.{ligand}.{base}.{additive}>>{product}
{ligand}.{base}.{methylaniline}.{additive}.{pd_catalyst}.{aryl_halide}>>{product}
{base}.{methylaniline}.{pd_catalyst}.{aryl_halide}.{additive}.{ligand}>>{product}
{additive}.{base}.{aryl_halide}.{ligand}.{methylaniline}.{pd_catalyst}>>{product}
{aryl_halide}.{pd_catalyst}.{base}.{ligand}.{methylaniline}.{additive}>>{product}

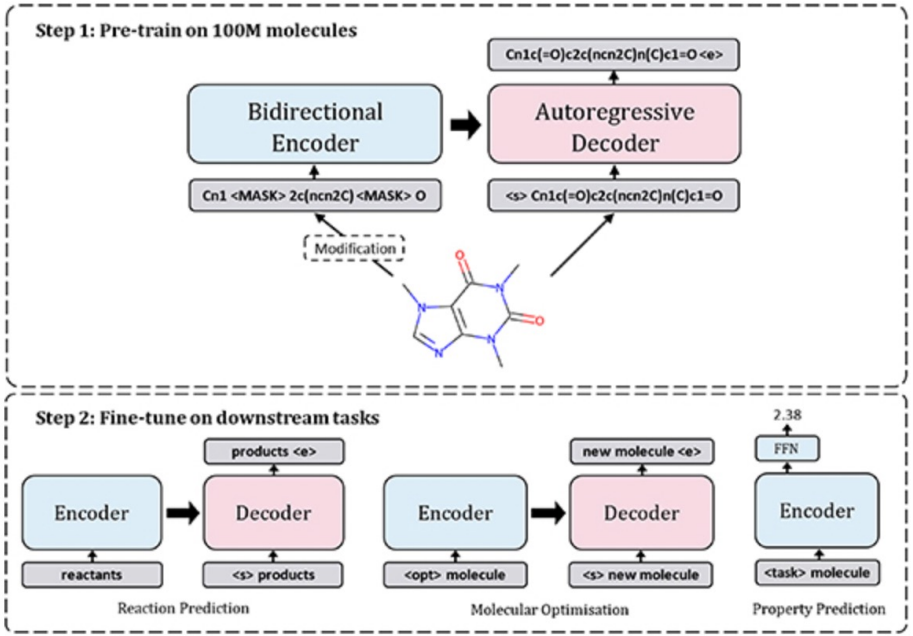
Molecule permutations

Bigger model (6 layers of size 512) compared to Molecular Transformer (4 layers of size 256) and 100x test-time augmentation -> much slower at inference time

State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis

Igor V. Tetko, Pavel Karpov, Ruud Van Deursen & Guillaume Godin

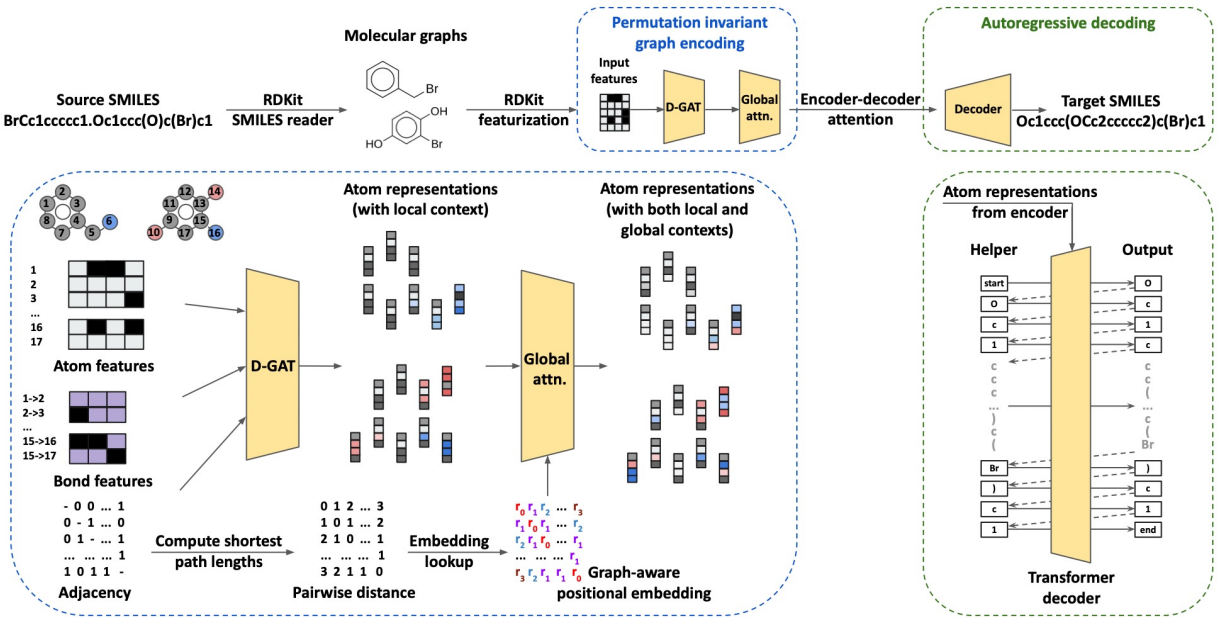
Large-scale pretraining



Chemformer: a pre-trained transformer for computational chemistry

Ross Irwin[1], Spyridon Dimitriadis[1,2], Jiazhen He[1] and Esben Jannik Bjerrum[3,1]

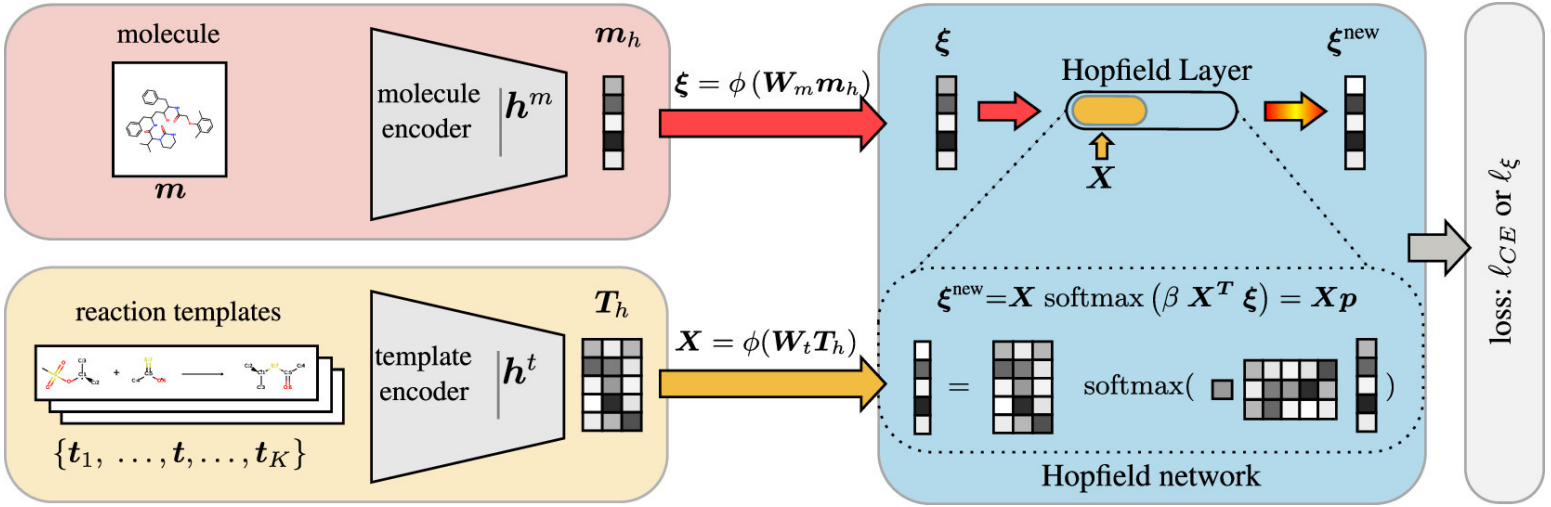
Graph2SMILES -> Graph encoder with a SMILES decoder



PERMUTATION INVARIANT GRAPH-TO-SEQUENCE MODEL FOR TEMPLATE-FREE RETROSYNTHESIS AND REACTION PREDICTION

Zhengkai Tu[1,2] and Connor W. Coley[1,3]

Improving Few- and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks

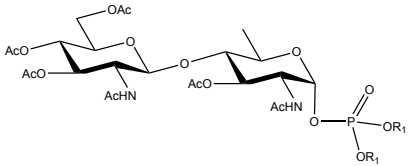


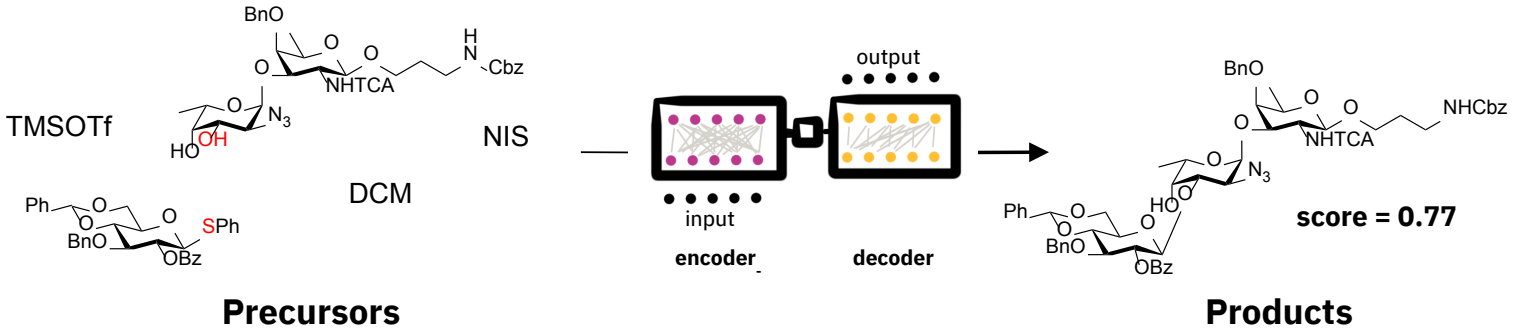
Philipp Seidl, Philipp Renz, Natalia Dyubankova, Paulo Neves, Jonas Verhoeven, Jörg K. Wegner, Marwin Segler, Sepp Hochreiter, and Günter Klambauer*

-> Probably more on that on Wednesday

Stereochemistry & experimental validation

- **14-step synthesis** of a lipid-linked oligosaccharide
- **>40% accuracy** increase with Carbo Transformer
- Similar performance gains on JACS/CARBO test sets





Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates

Pesciullesi*, **Schwaller* et al.,** Nature Communications, 2020

UNIVERSITÄT BERN

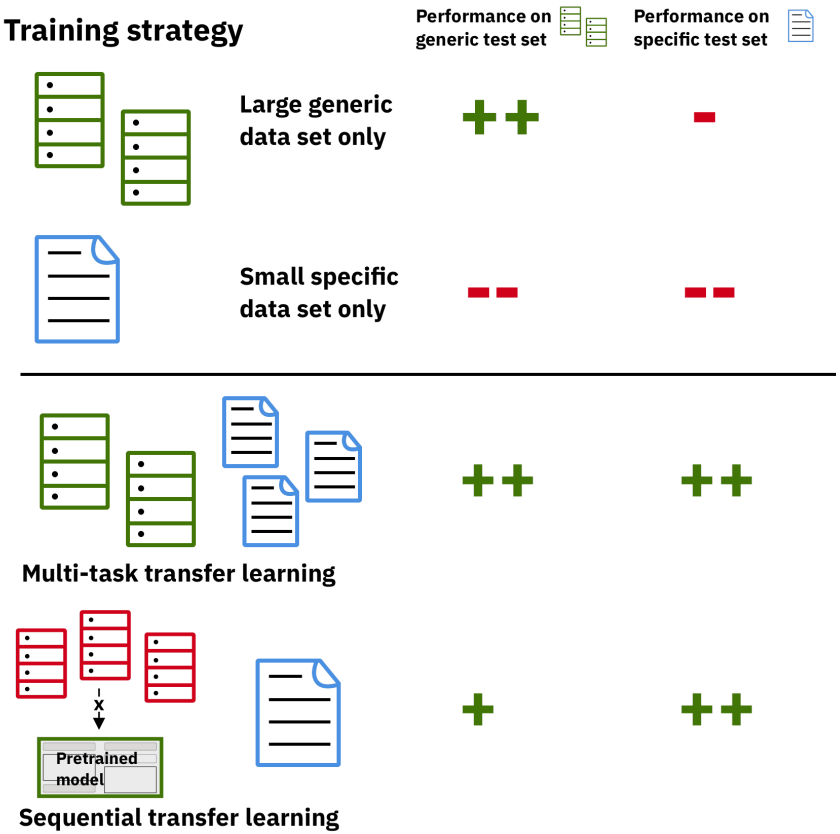
IBM **Research**



What is *transfer learning* ?

- Patent reactions (**1 million**)
- Carbohydrate reactions (**few thousands**)

Carbohydrate Transformer ←

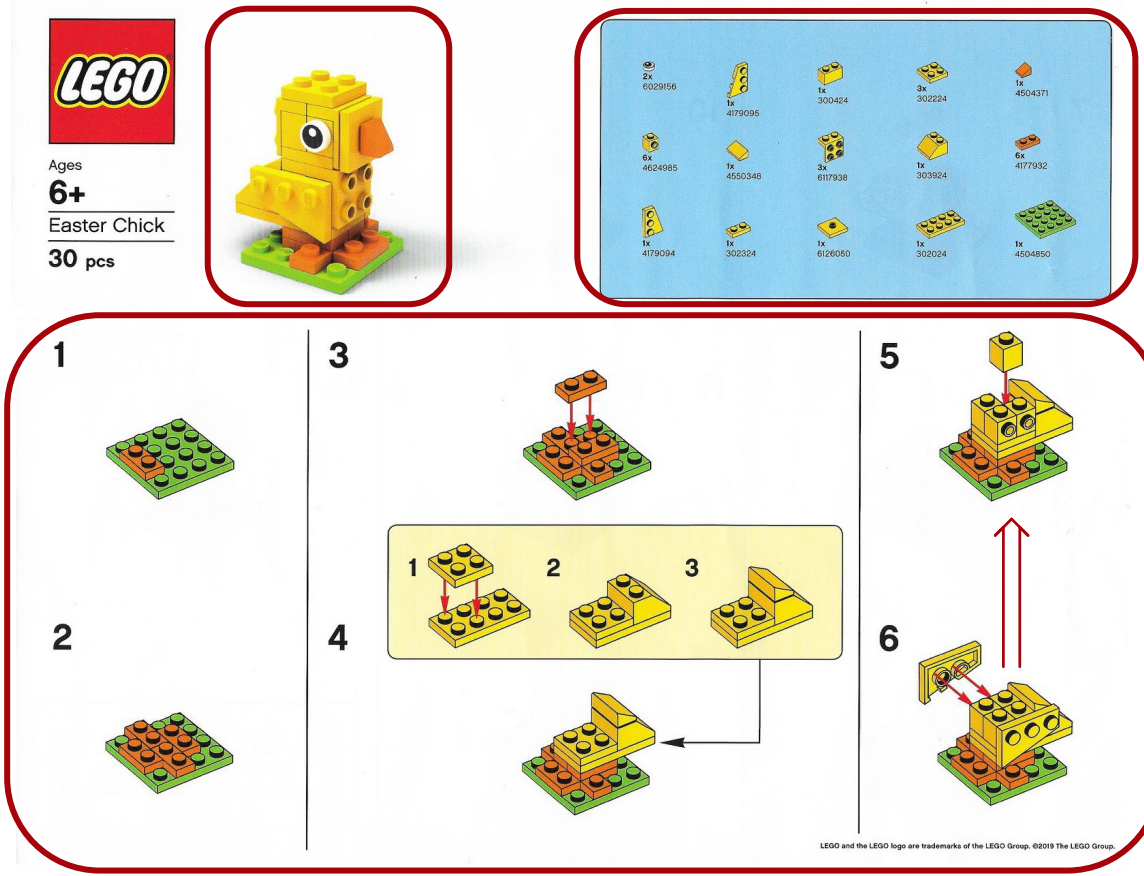


Transfer learning applicable to any reaction subspace of interest!

Retrosynthesis (Corey, Nobel prize, 1990)

Target molecule

Known (commercially available) building blocks



Steps to construct the target

Lego analogy: Amol Thakkar

Single-step retrosynthesis using language models (2017)

Introduction of **USPTO-50k** benchmark dataset

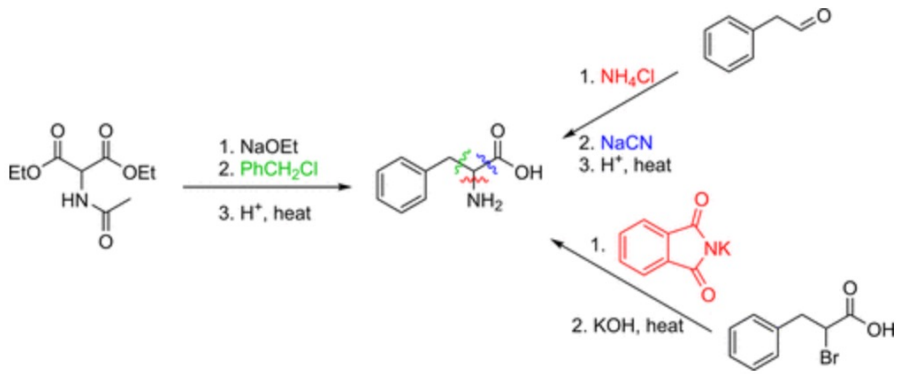


Figure 1. Phenylalanine synthetic scheme.

Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models

Bowen Liu[†], Bharath Ramsundar[‡], Prasad Kawthekar[‡], Jade Shi[†], Joseph Gomes[†], Quang Luu Nguyen[†], Stephen Ho[†], Jack Sloane[†], Paul Wender[†§], and Vijay Pande[*†‡∥]

Single-step retrosynthesis

Work with IBM intern
Riccardo Pisoni
Slide from ACS Boston 2018
Release of IBM RXN 4 chemistry

products → reactants

<RX_1> C O C (= O) C C C (= O) ....

C 1 = C O C C C 1 . ... c 1 c c c (O) c c 1 O

Liu et al. *ACS Cent. Sci.* 3, 10, 1103-1113

USPTO_50k	Top-1 [%]	Top-2 [%]	Top-3 [%]	Top-5 [%]	Top-10 [%]	Top-20 [%]	Top-50 [%]
Rule-based	35.4		52.3	59.1	65.1	68.6	69.5
Liu et. al. Seq2Seq	37.4		52.4	57.0	61.7	65.9	70.7
Molecular Transformer (unpublished)	**54.0**	**65.6**	**69.9**	**74.0**			

Retrosynthesis model

Starting products → Top-1 Reactants (not unique)

Prediction model

Trained on Pistachio

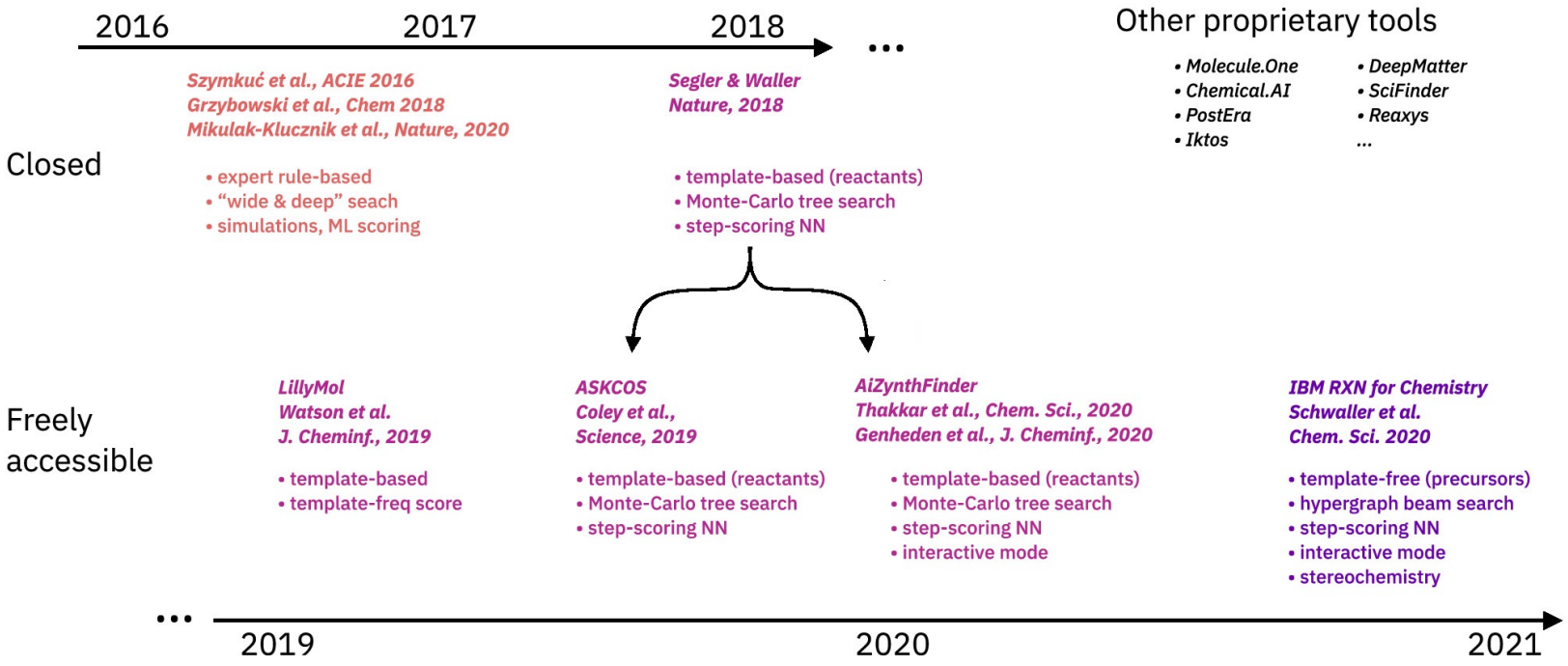
75% predicted products equal starting products

Single-step retrosynthesis (USPTO 50k)

- no reaction class information given

Model	Top-1	Top-5	Top-10
SMILES-based			
SCROP [51]	43.7	65.2	68.7
Two-way transformer [52]	47.1	73.1	76.3
Aug transformer [5]	48.3	**73.4**	**77.4**
Chemformer	53.6	61.1	61.7
Chemformer-Large	**54.3**	62.3	63.0
Graph-based			
MEGAN [53]	48.1	**78.4**	**86.1**
GLN [54]	52.5	75.6	83.7
GraphRetro [55]	**53.7**	72.2	75.5

Multi-step synthesis planning



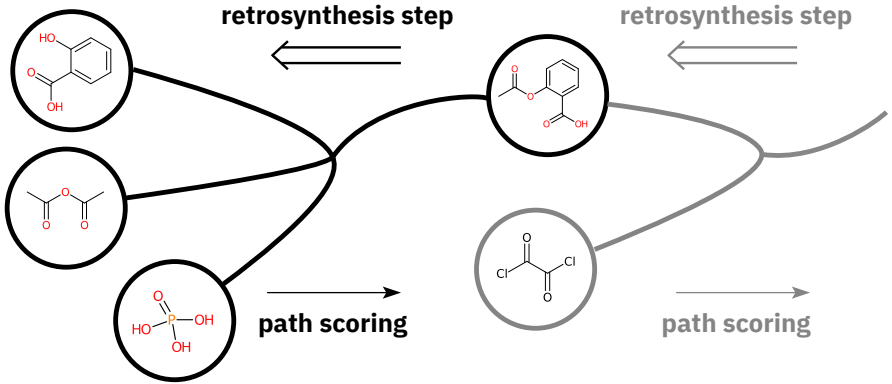
Single-step prediction approach: template-based / graph-based / SMILES-based

Search algorithm: MCTS / RL / A* / beam search

Multi-step synthesis planning

IBM **Research**

Molecular Transformer **Retro**



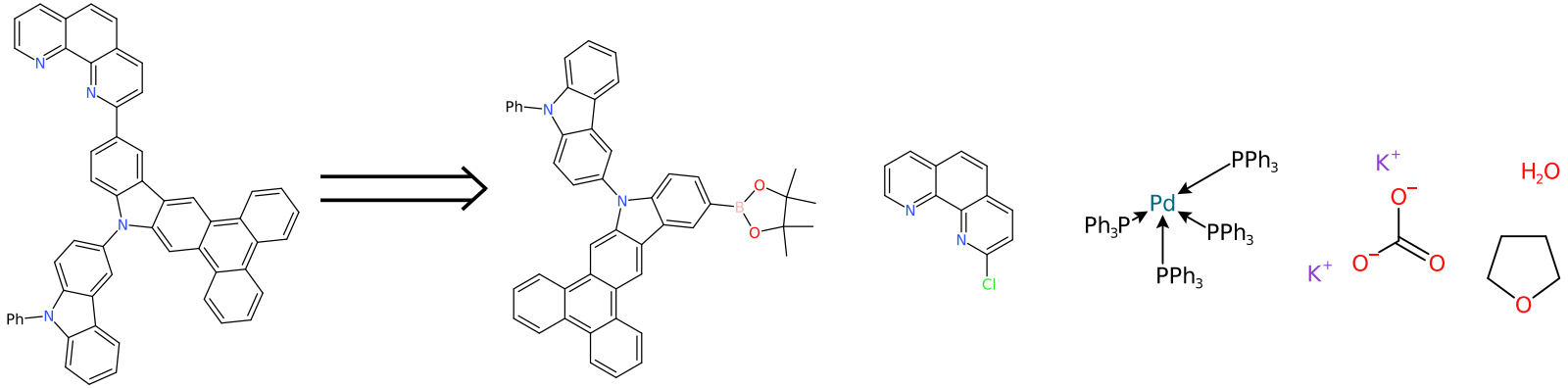
Molecular Transformer **Forward**

Predicting retrosynthetic pathways using a combined linguistic model and hyper-graph exploration strategy

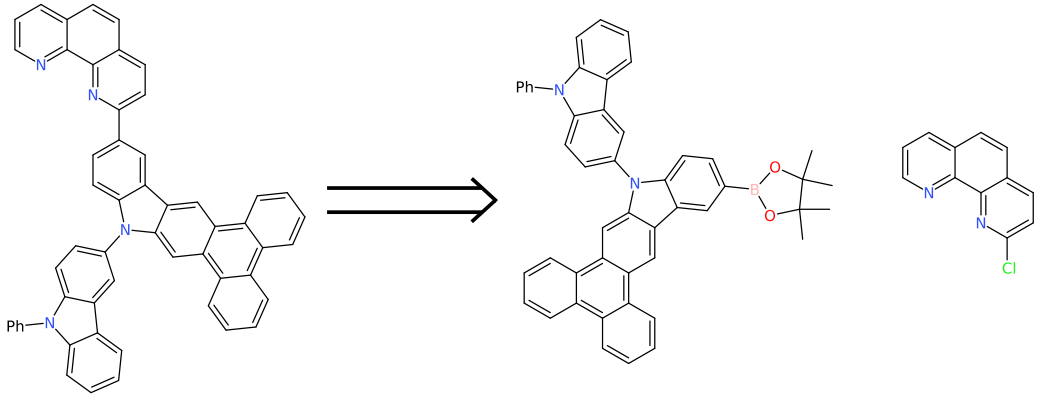
Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, Teodoro Laino

Chem. Sci., 2020, Advance Article
https://doi.org/10.1039/C9SC05704H

Our approach: Precursors prediction (reactants, reagents, catalysts, solvents, ..)



Reactants prediction (atom-mapping dependent)



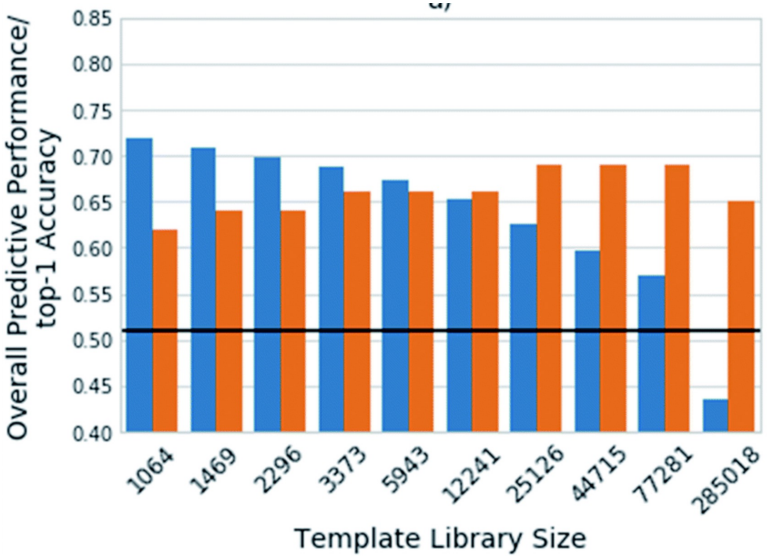
Rule-based approaches:

- Segler et al. Nature, 555, 604–61
- Coley et al. Science 365 (6453)
- Genheden et al. J Cheminf. 12 (1), 1-9
- Thakkar et al. Chem. Sci. 11 (1), 154-168

and other SMILES-based approaches:

- Liu et al. ACS Cent. Sci. 3 (10), 1103-1113
- Tetko et al. Nature Comm. 11 (1), 1-11

Single step -> multi-step performance



Top-1 accuracy (single step)
Multi-step performance

New metrics/benchmarks are required!!
(collaborations?)

Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain†

Amol Thakkar *[ab], Thierry Kogej [a], Jean-Louis Reymond [b], Ola Engkvist [a] and Esben Jannik Bjerrum *[a]

Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy†

Philippe Schwaller *[a], Riccardo Petraglia [a], Valerio Zullo [b], Vishnu H. Nair [a], Rico Andreas Haeuselmann [a], Riccardo Pisoni [a], Costas Bekas [a], Anna Iuliano [b] and Teodoro Laino [a]

Model-based metrics

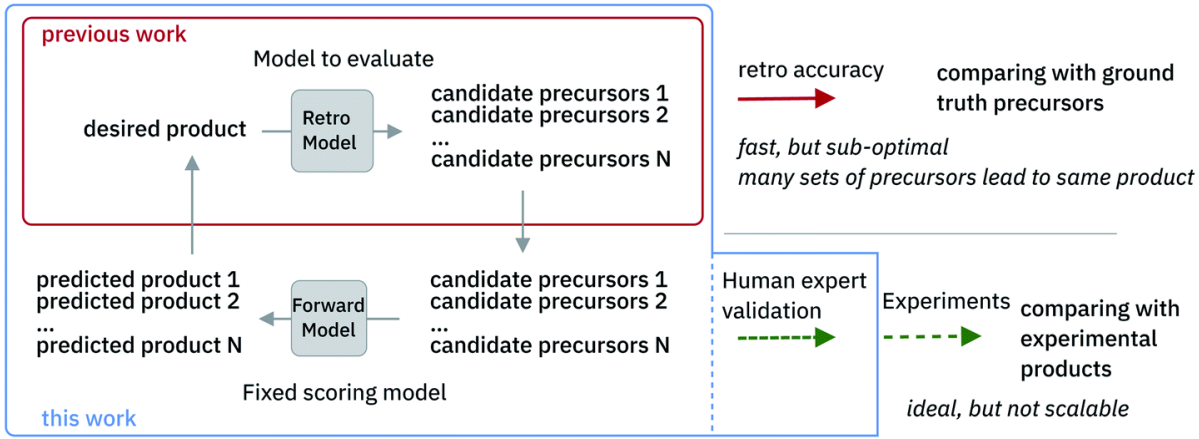
round-trip accuracy
% of predicted product == desired product

coverage
at least 1 predicted product == desired product

class diversity
diversity of suggested reaction classes

Jensen-Shannon divergence
similarity of class probability distributions

scalable single-step retrosynthesis evaluation metrics



Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy†

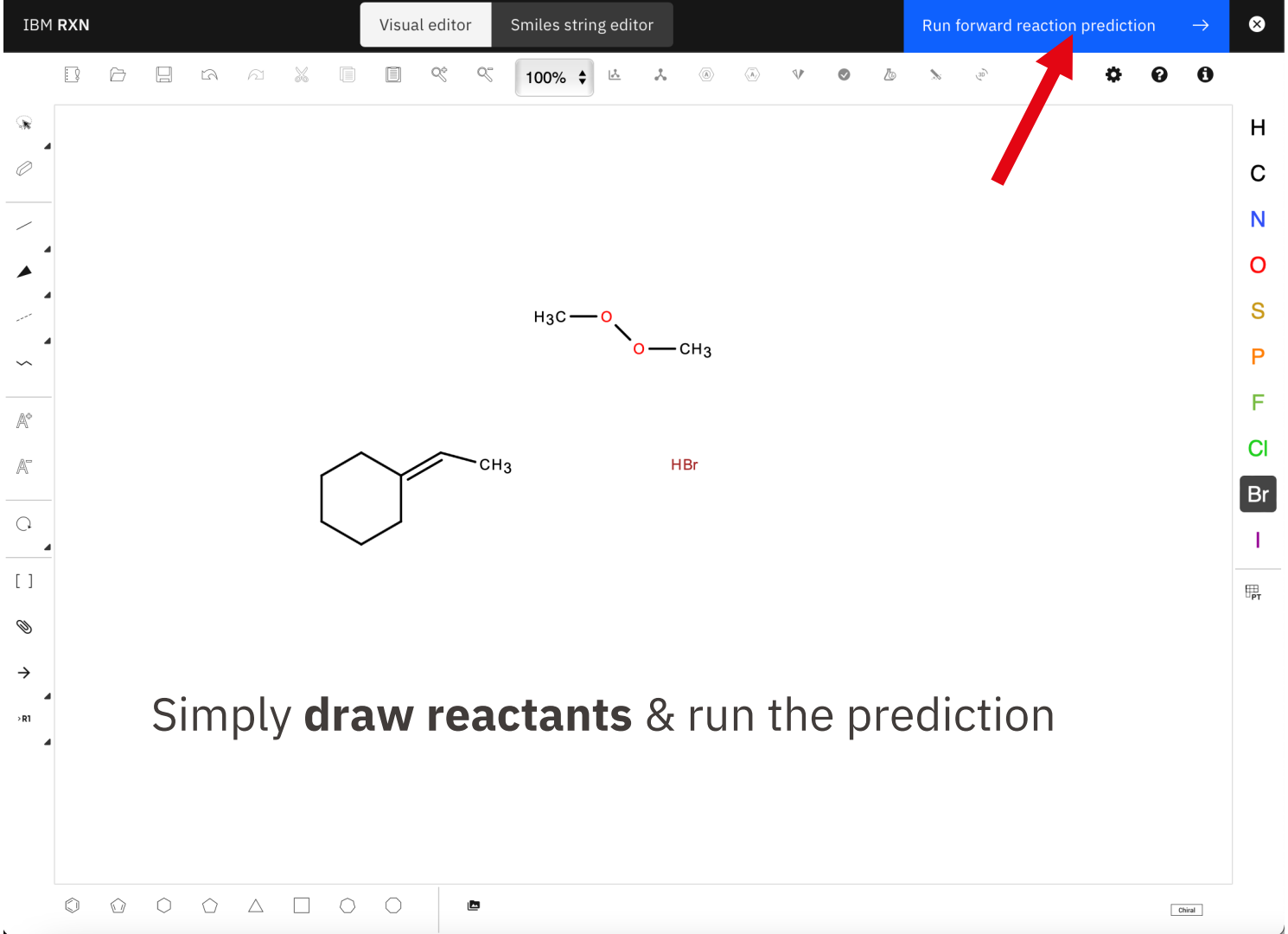
Philippe Schwaller *[a], Riccardo Petraglia [a], Valerio Zullo [b], Vishnu H. Nair [a], Rico Andreas Haeuselmann [a], Riccardo Pisoni [a], Costas Bekas [a], Anna Iuliano [b] and Teodoro Laino [a]

EPFL

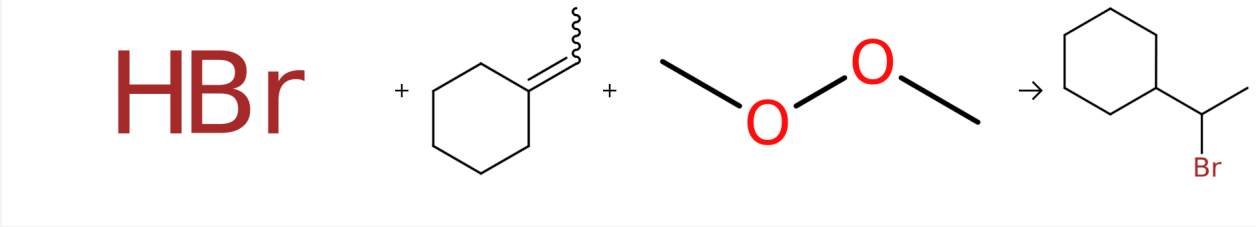
IBM **Research**

How to *facilitate adoption*?

IBM **RXN for Chemistry**
Free platform: rxn.res.ibm.com and API access.


IBM RXN
Visual editor
Smiles string editor
Run forward reaction prediction
100%
H
C
N
O
S
P
F
Cl
Br
I
H3C — O
O — CH3
CH3
HBr
Simply **draw reactants** & run the prediction

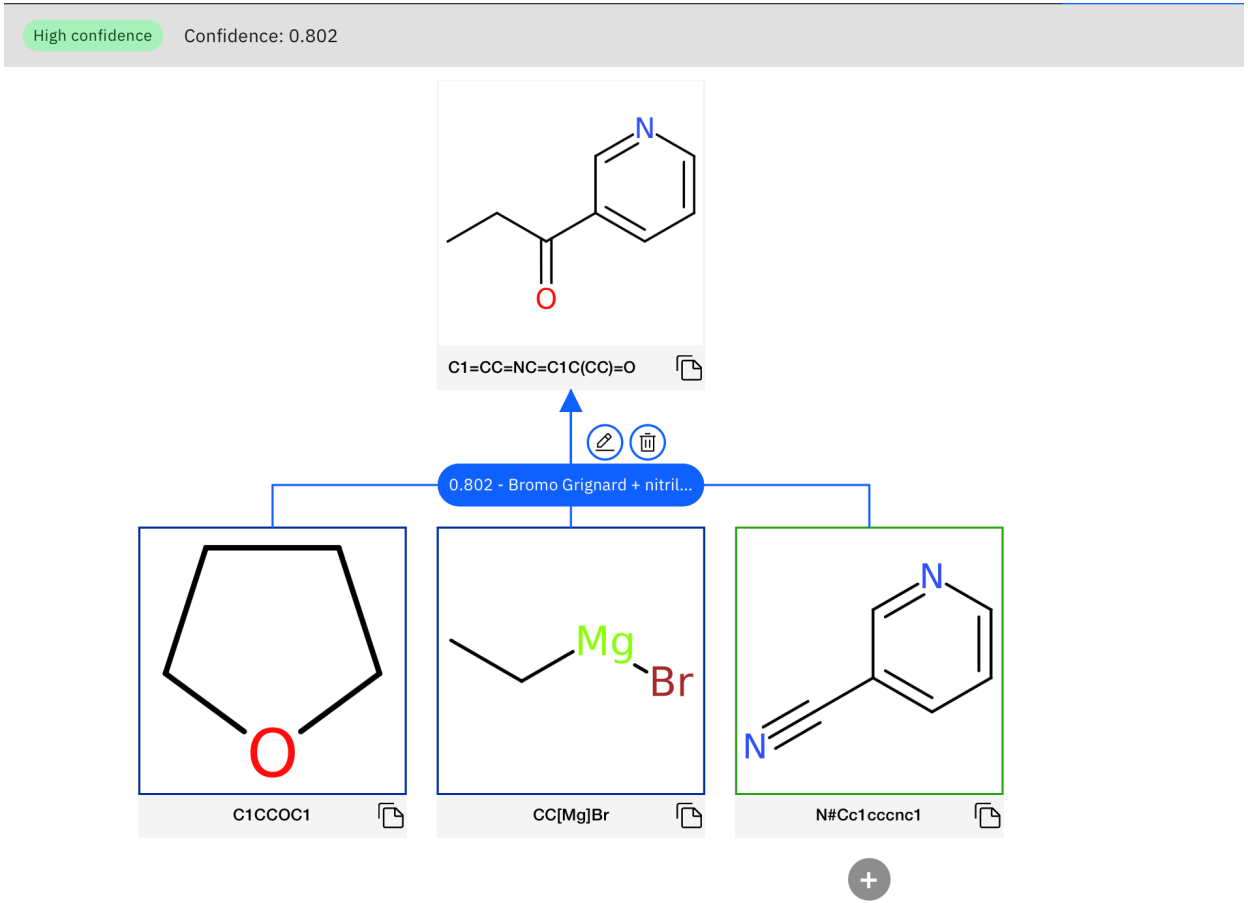
Showcase_20220216_07:18:35.707 Confidence: 0.9999883176160712



Br.CC=C1CCCCC1.COOC>>CC(Br)C1CCCCC1

Get back the **product** and a **confidence score**

Retrosynthesis interface



Retrosynthesis *user feedback*



Alessandro Castrogiovanni
Sparr Research Group
University of Basel, 2019

What’s next?

IBM **Research**

What's next? #RoboRXN

IBM **Research**



Teodoro Laino & RoboRXN Team

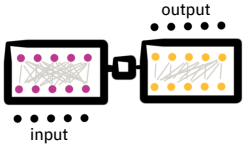
Cloud

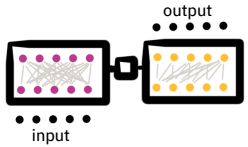
Remote AI-driven synthesis

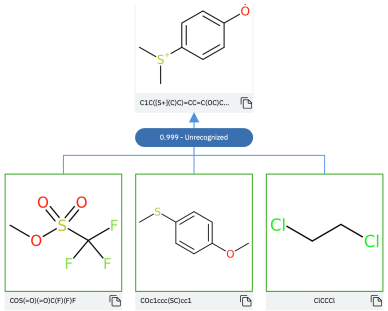
Automation



AI









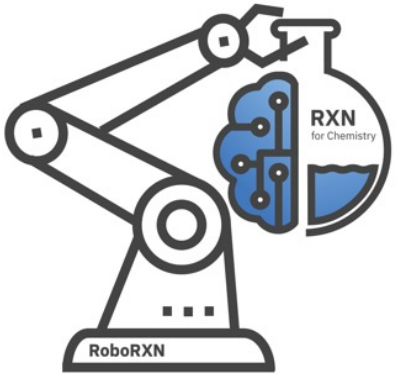
Chemists at home

- design of target molecule
- creative process

Synthesis planning

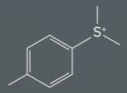
- Add: 1mmol 4-methoxythioanisole
- Add: 10 ml 1,2-dichloroethane
- Add: 1mmol methyl trifluoromethanesulfonate
- Stir: at 25ºC for 1 day
- Filter: keep filtrate
- Concentrate

Recipe formulation



Execution on robot

Synthesizing new molecule



Started: Nov 30 2020, 6:49am PT

Live from IBM RoboRXN

Action 2

Overview

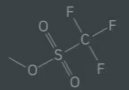
Adding $C_2H_3F_3O_3S$

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

Methyl trifluoromethane

$C_2H_3F_3O_3S$

2D 3D



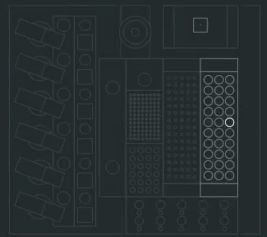
Methyl trifluoromethane sulfonate is a brown liquid. Insoluble in water. This material is a very reactive methylating agent, also known as methyl triflate.

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



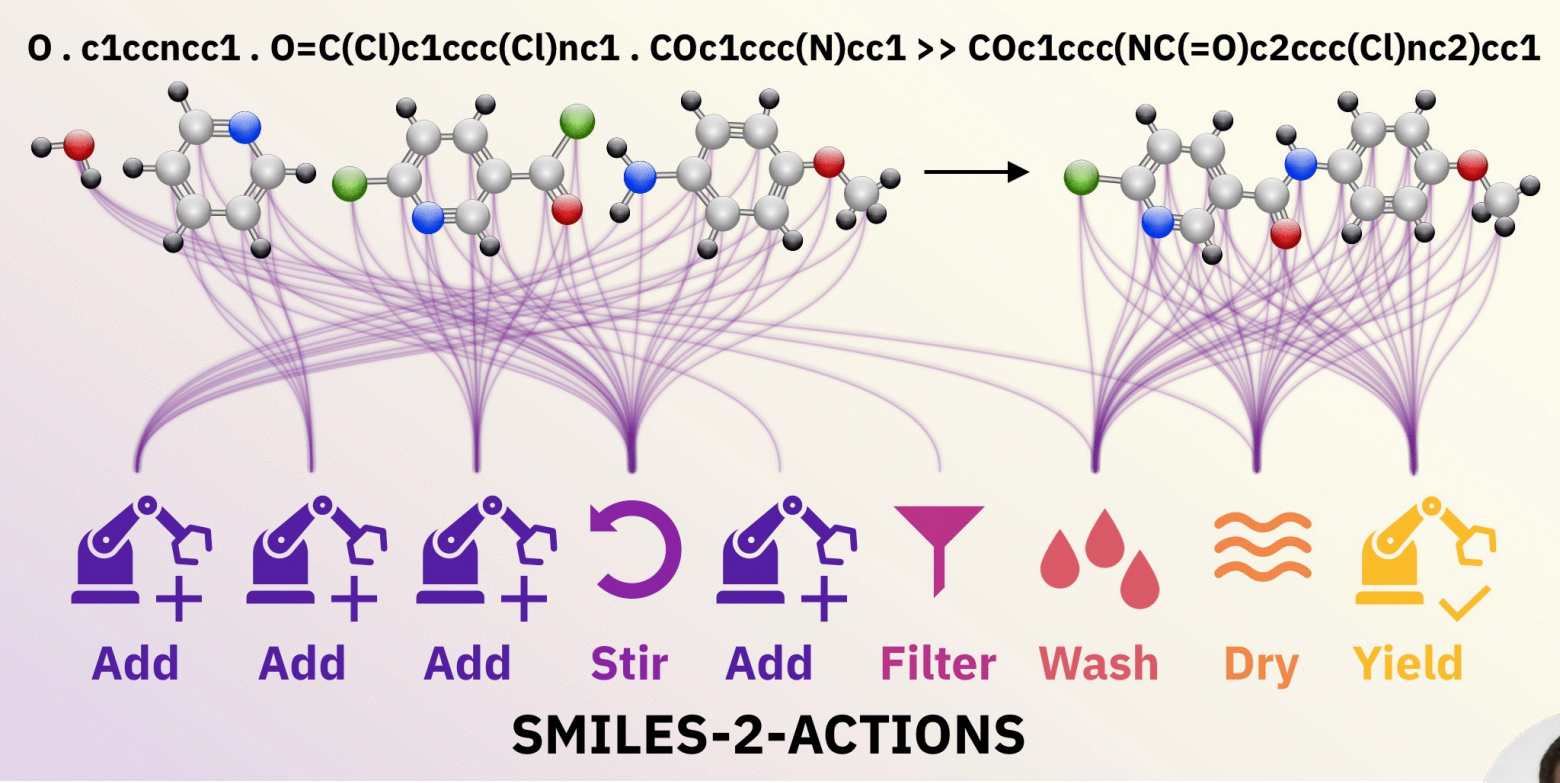
Live view module



2 Adding $C_2H_3F_3O_3S$

00:06:00

LIVE



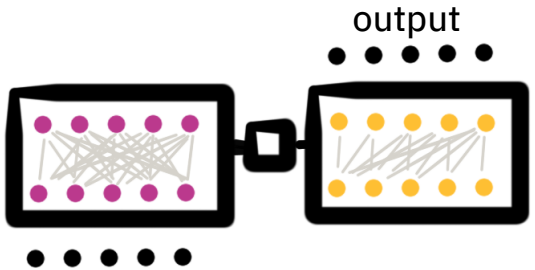
Vaucher, A.C., Schwaller, P., Geluykens, J. *et al.* Inferring experimental procedures from text-based representations of chemical reactions. *Nat Commun* **12,** 2573 (2021).
Vaucher, A.C., Zipoli, F., Geluykens, J. et al. Automated extraction of chemical synthesis actions from experimental procedures. Nat Commun 11, 3601 (2020).



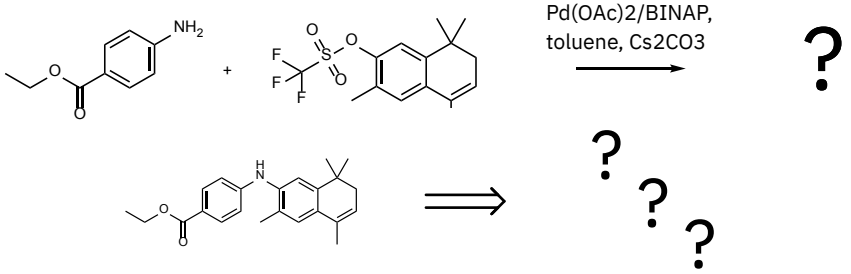
Alain Vaucher

Reaction Transformer models

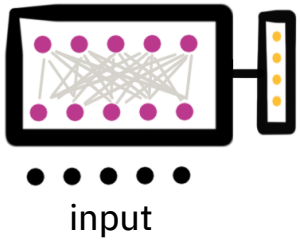
From a sequence to a sequence.



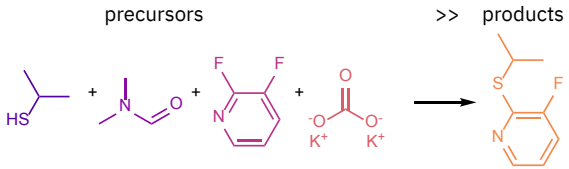
encoder **decoder**



From a sequence to a single value/label.



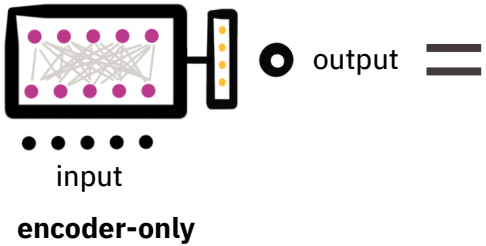
encoder-only

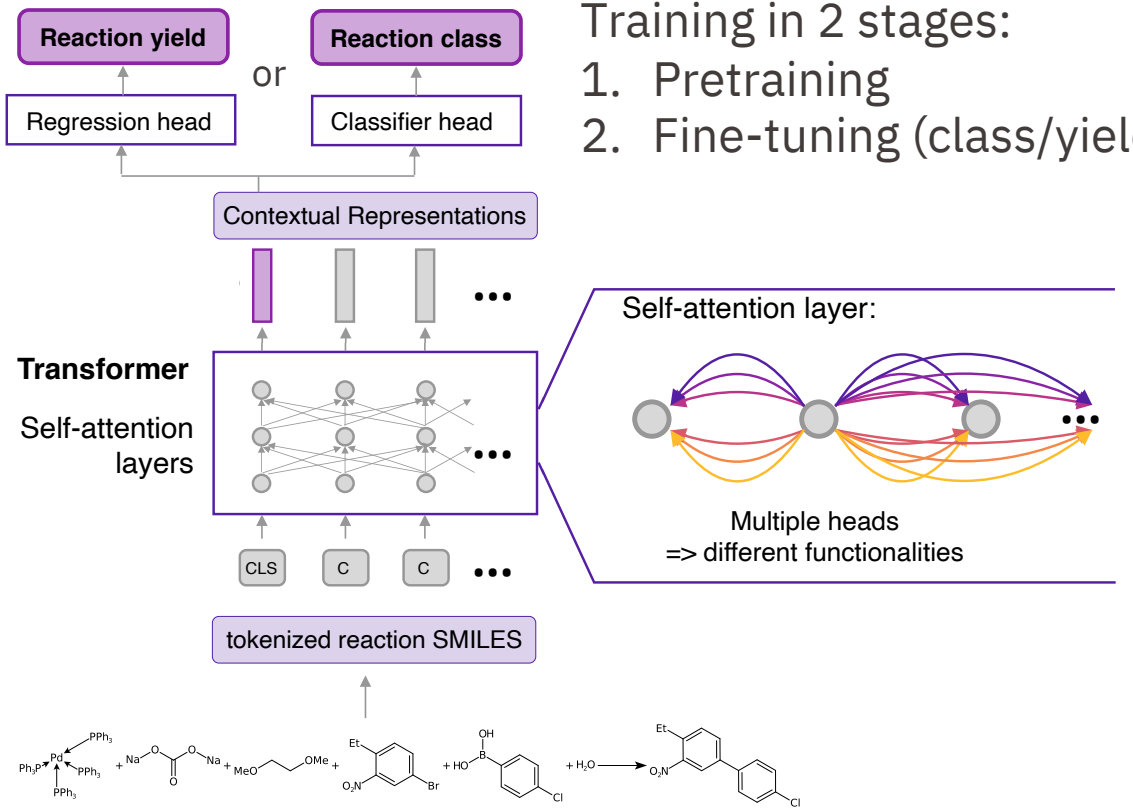


Reaction classification
Schwaller et al., Nature Mach. Int., 2021

Yields predictions
Schwaller et al., MLST, 2021
Schwaller et al., chemrxiv.13286741

Zooming into a reaction Transformer





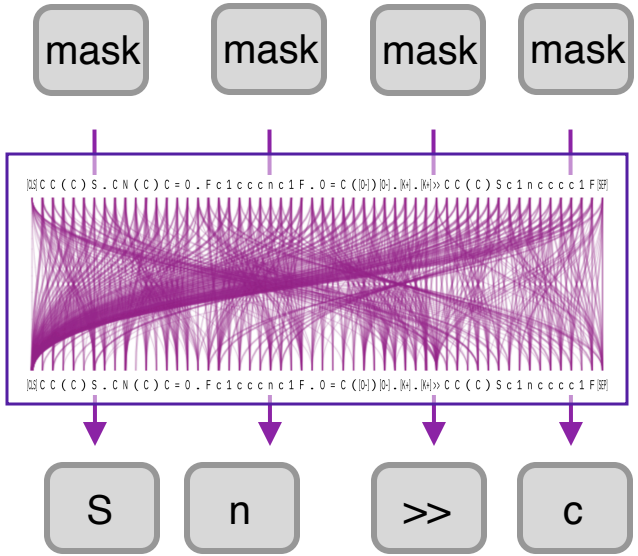
Training in 2 stages:

1. Pretraining
2. Fine-tuning (class/yield)

(BERT, Devlin et al. NeurIPS 2019)

Self-supervised training

Pretraining



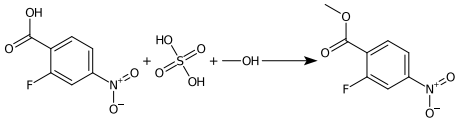
Masked **L**anguage **M**odelling

- Predict mask given context
- Unlimited training data

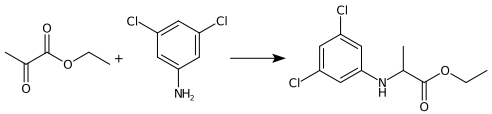
(BERT, Devlin et al. NeurIPS 2019)

Reaction *classification*

Reaction SMILES

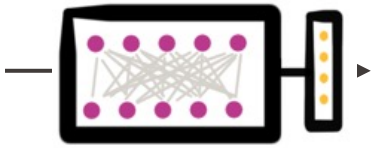


CO.O=C(O)c1ccc([N+](=O)[O-])cc1F.O=S(=O)(O)O>>COC(=O)c1ccc([N+](=O)[O-])cc1F



CCOC(=O)C(C)=O.Nc1cc(Cl)cc(Cl)c1>>CCOC(=O)C(C)Nc1cc(Cl)cc(Cl)c1

Classification model



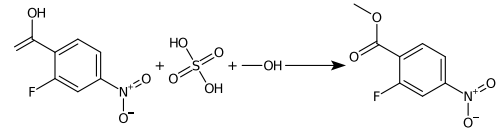
Reaction class

Fischer-Speier Esterification
2.6.3

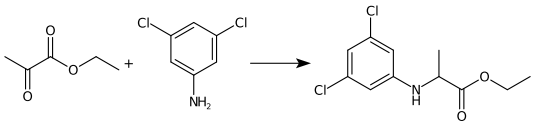
Ketone reductive amination
1.2.5

>98% accurate

NextMove Software

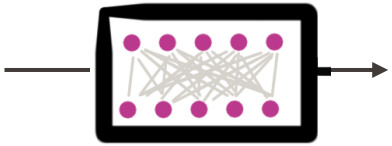


CO.O=C(O)c1ccc([N+](=O)[O-])cc1F.O=S(=O)(O)O>>COC(=O)c1ccc([N+](=O)[O-])cc1F



CCOC(=O)C(C)=O.Nc1cc(Cl)cc(Cl)c1>>CCOC(=O)C(C)Nc1cc(Cl)cc(Cl)c1

Reaction encoder



Fingerprints

= Encoded reaction properties

[0.14, 0.25, ..., 0.9]

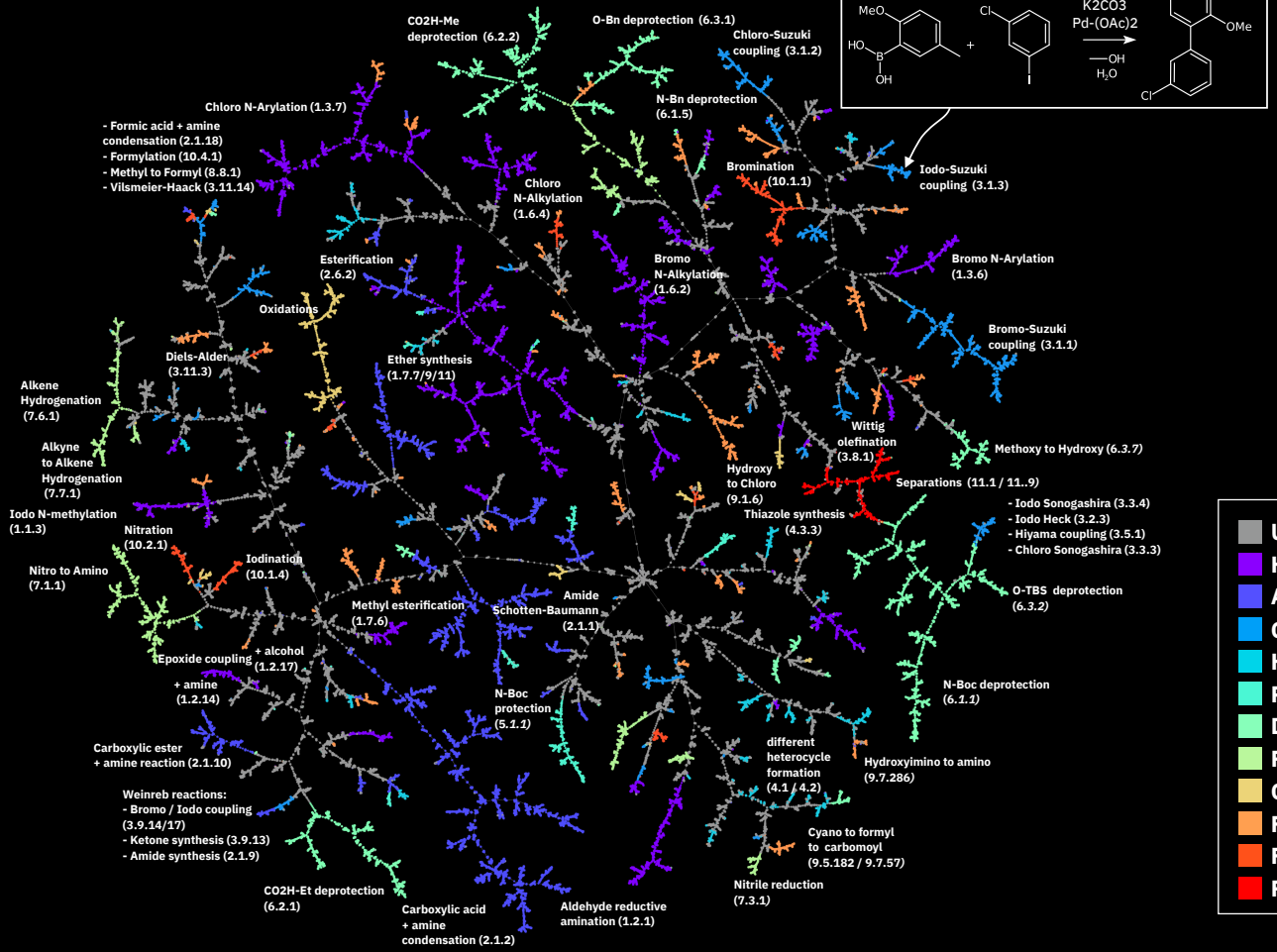
[0.22, 0.83, ..., -0.12]

Mapping the space of chemical reactions using attention-based neural networks

Philippe Schwaller[1,2✉], Daniel Probst[2], Alain C. Vaucher[1], Vishnu H. Nair[1], David Kreutter[2], Teodoro Laino[1] and Jean-Louis Reymond[2]

nature machine intelligence

Chemical Reaction Atlas



Powered by TMAP

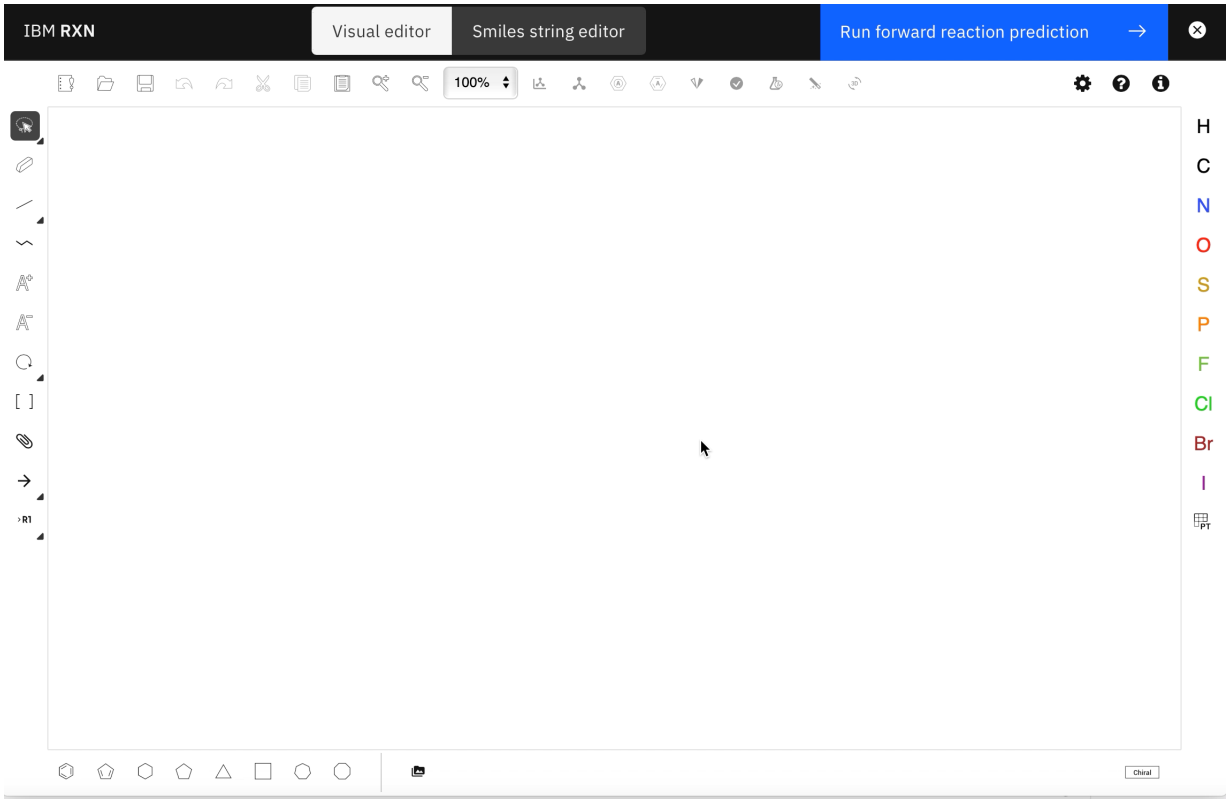


- Unrecognised
- Heteroatom alkylation and arylation
- Acylation and related processes
- C-C bond formation
- Heterocycle formation
- Protections
- Deprotections
- Reductions
- Oxidations
- Functional group interconvertions
- Functional group additions
- Resolutions

Chemical reaction *space exploration*



Search for *similar reactions*



→ Access to reaction conditions, procedures, articles

Open-source code

https://github.com/rxn4chemistry/rxnfp | pip install rxnfp

nature machine intelligence

MATTERS ARISING

https://doi.org/10.1038/s42256-021-00367-2

Check for updates

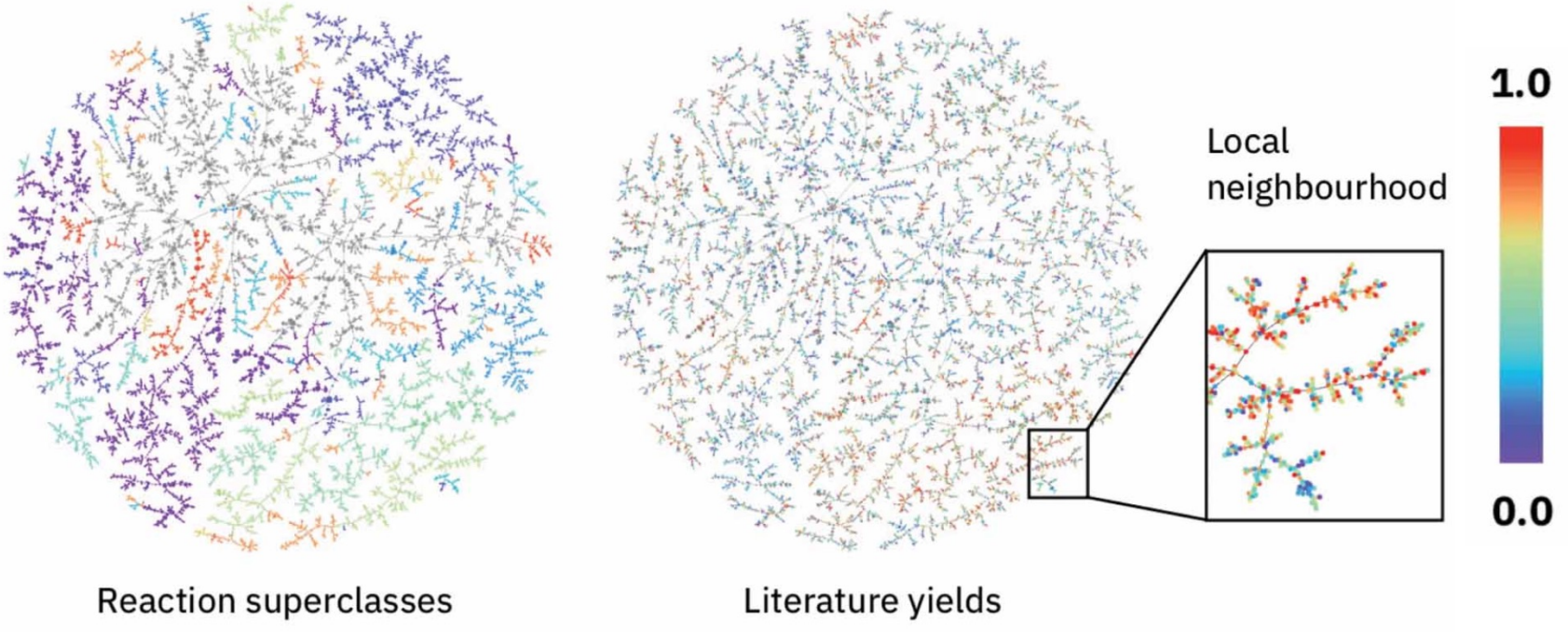
Reusability report: Learning the language of synthetic methods used in medicinal chemistry

Jon Paul Janet, Anna Tomberg and Jonas Boström ✉

ARISING FROM P. Schwaller et al. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-020-00284-w (2020).

Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden. ✉e-mail: jonas.bostrom@astrazeneca.com

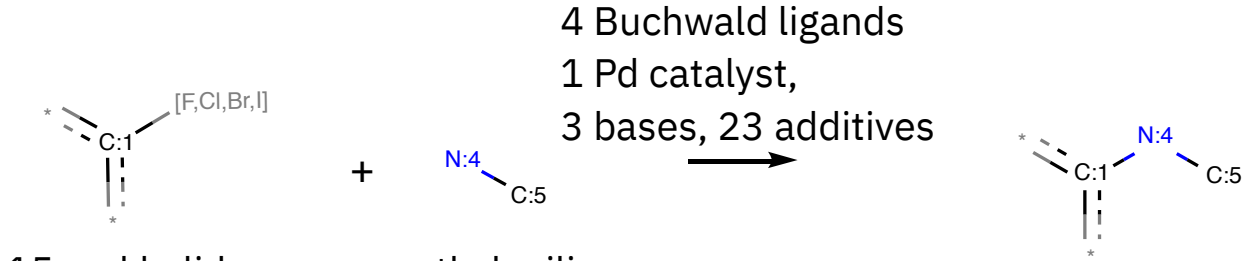
Prediction of *patent reaction yields*



Long story short → I didn't work

Prediction of *HTE reaction yields*

3955 **Buchwald-Hartwig reactions** with measured yield



15 aryl halides methylaniline

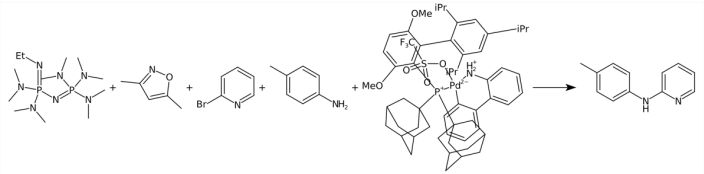
Yield prediction using DFT descriptors + random forest (RF)

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).

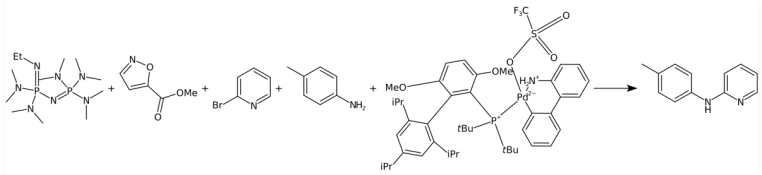
Multi fingerprint features (MFF) + RF

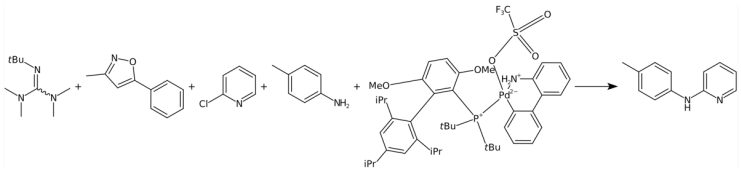
Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* (2020).

Learning to predict 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





Reaction Transformer



encoder-only

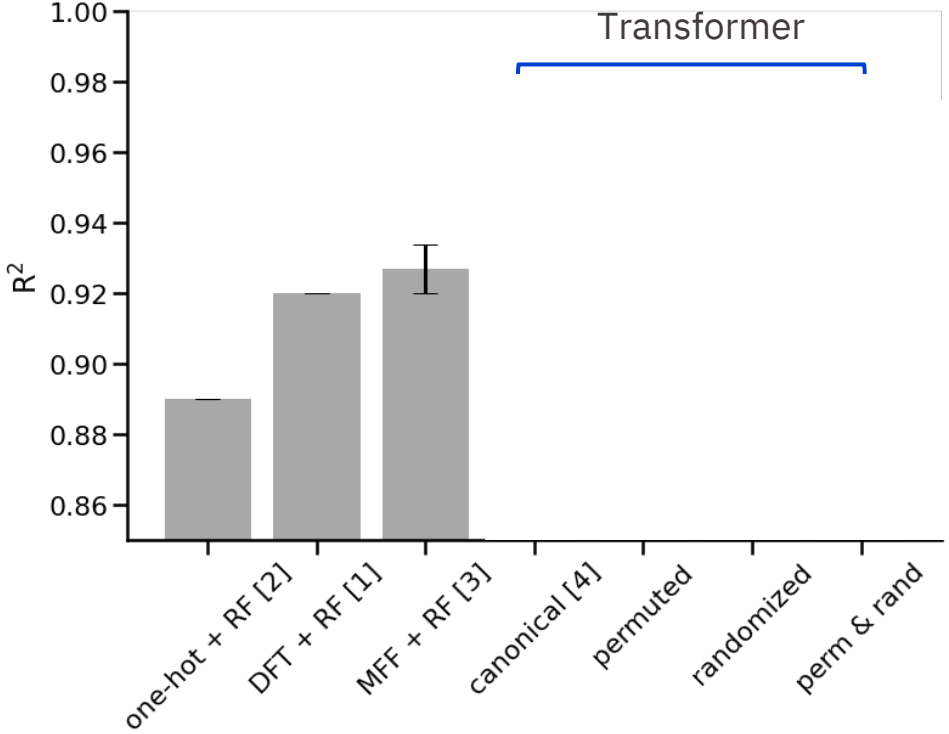
77.6 %

20.0 %

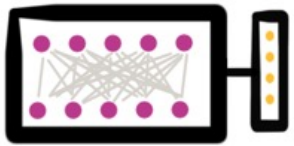
56.3 %

H-L data from: 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).

Results: *70/30* random split



reaction SMILES



70.4%

Yield

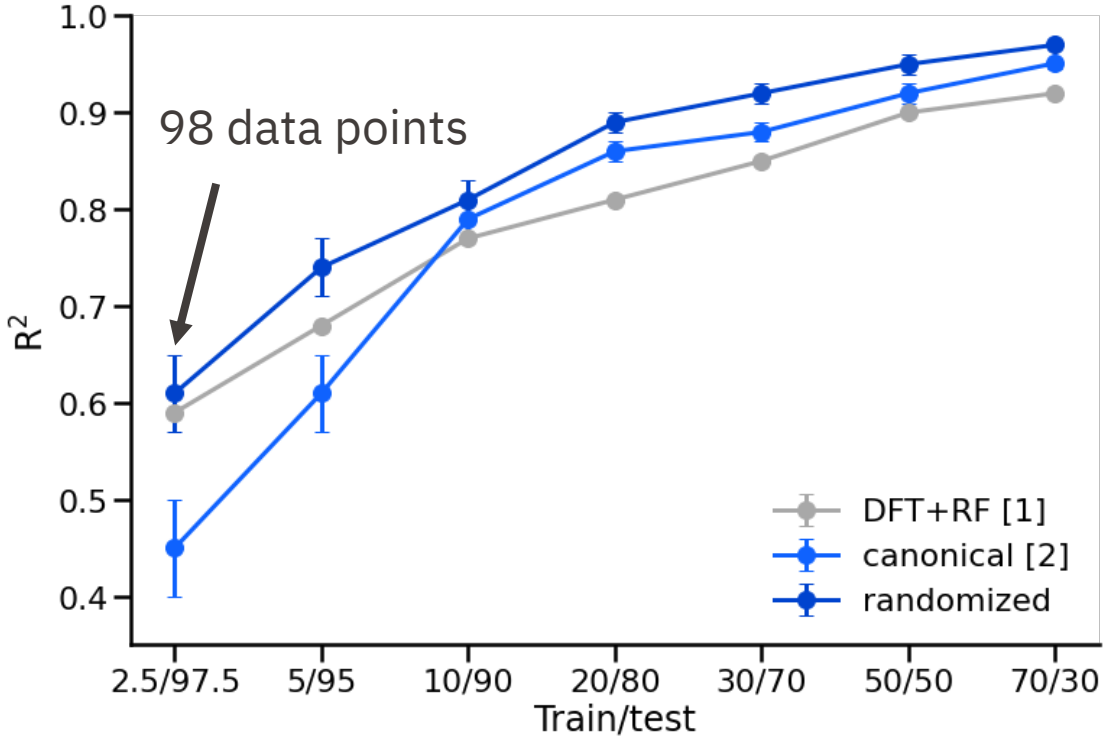
[1] Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).

[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).

Results: *reduced* training data



[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] 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).

But why do *Transformers* work so well?

Work with
Ben Hoover and
Hendrik Strobelt



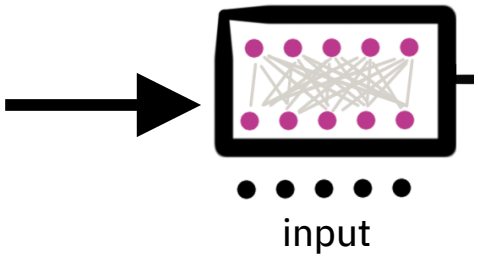


Self-supervised training

millions of **unlabeled** reaction SMILES

```
BrC(Br)(Br)Br.CC...>>...
CO.Nc1cccc([N+]...>>...
...
(OC(C)=O)OC(C)=O..>>...
precursors>>products
```

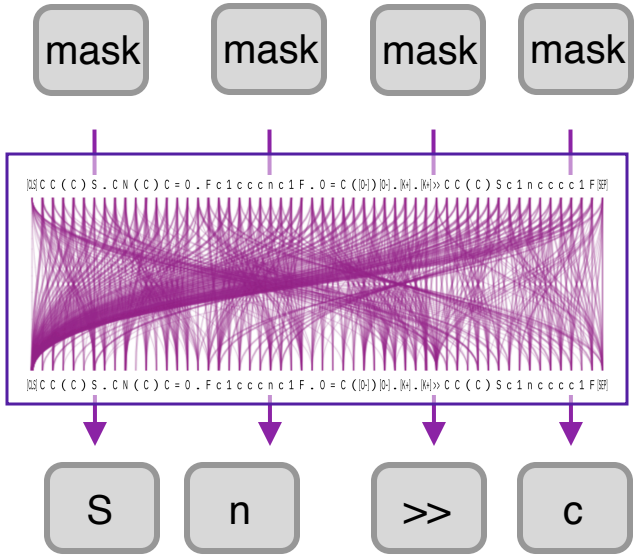
Lowe, 2017



encoder-only

IBM **Research**

Self-supervised training



Masked **L**anguage **M**odelling:

- Reaction corruption task
- Unlimited training data

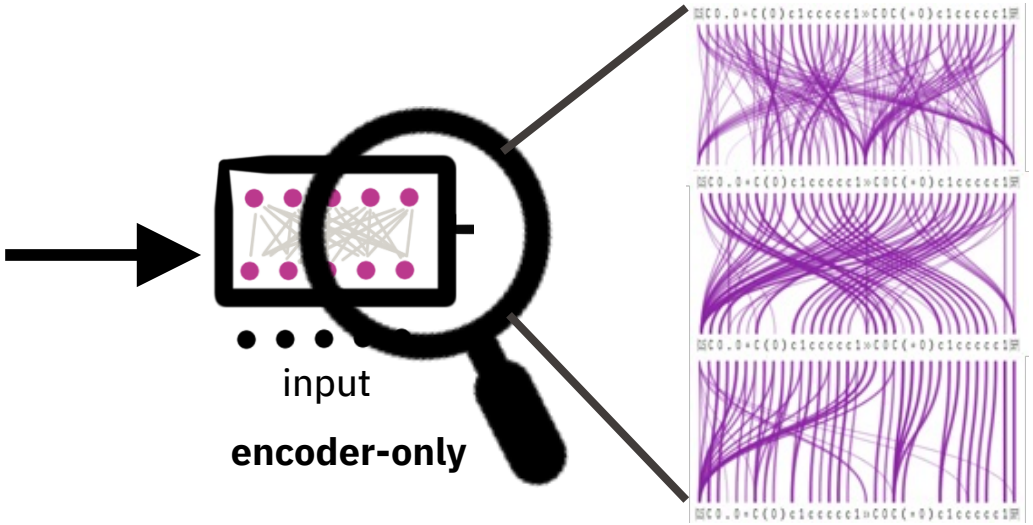
But why do *Transformers* work so well?

Self-supervised training **Visual inspection**

millions of **unlabeled** reaction SMILES

```
BrC(Br)(Br)Br.CC...>>...
CO.Nc1cccc([N+]...>>...
...
(OC(C)=O)OC(C)=O..>>...
precursors>>products
```

Lowe, 2017



Work with
Ben Hoover and
Hendrik Strobelt

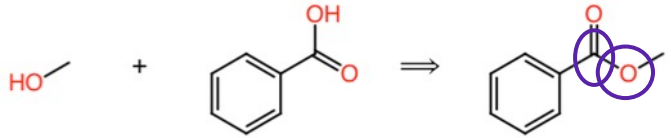




IBM **Research**

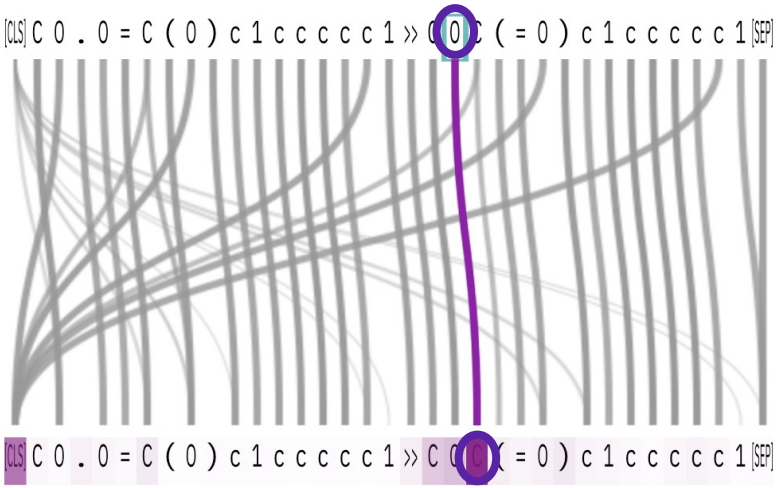
UNIVERSITÄT BERN

Visual inspection of attention weights



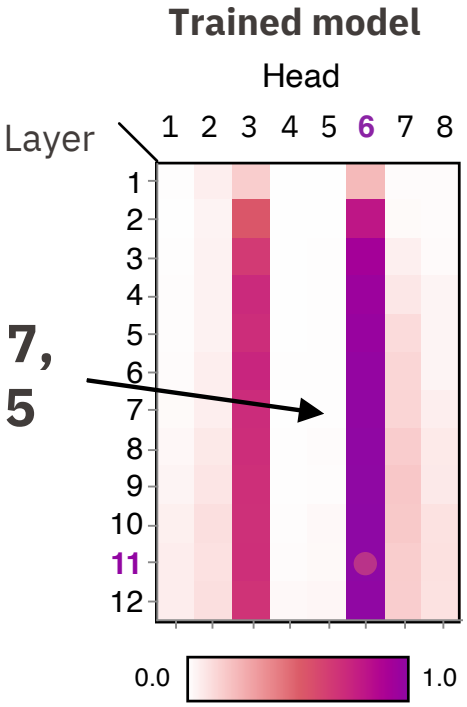
Input: precursors >> product

attention weights



Input: precursors >> product

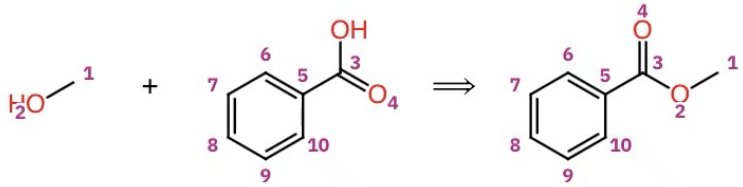
layer 7, head 5

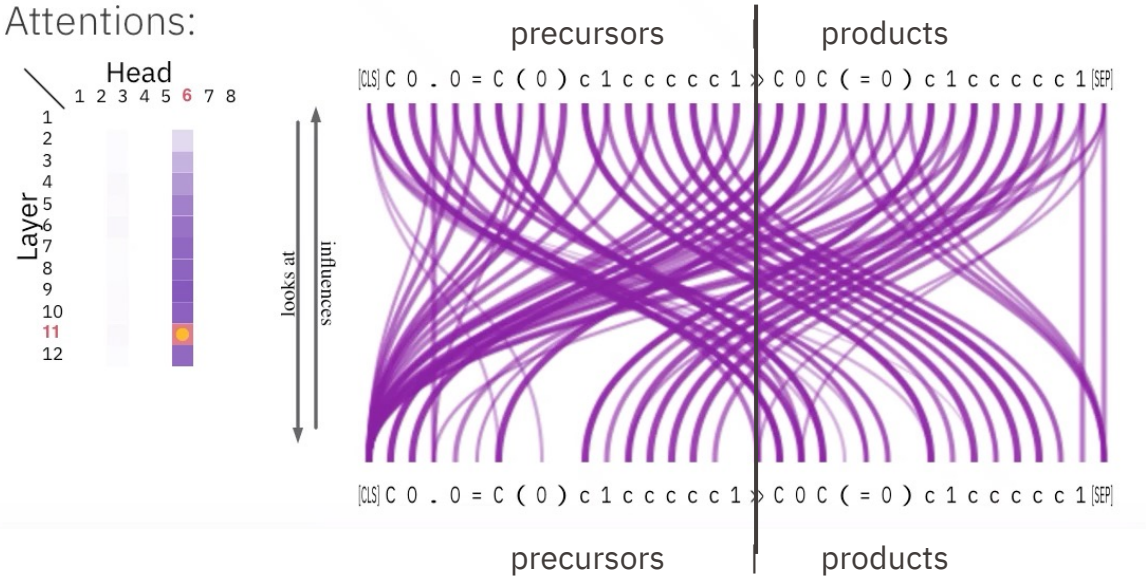


12 self-attention layers with 8 heads each

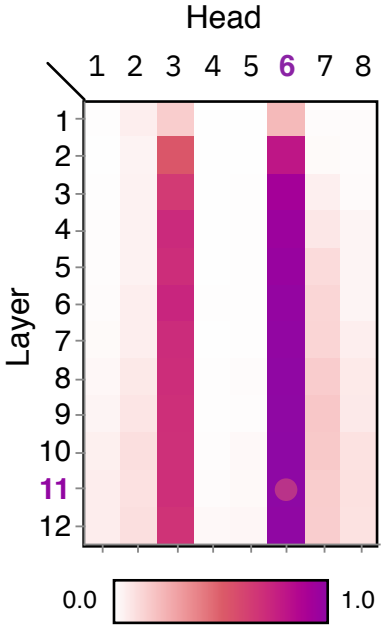
Visual inspection of attention weights

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



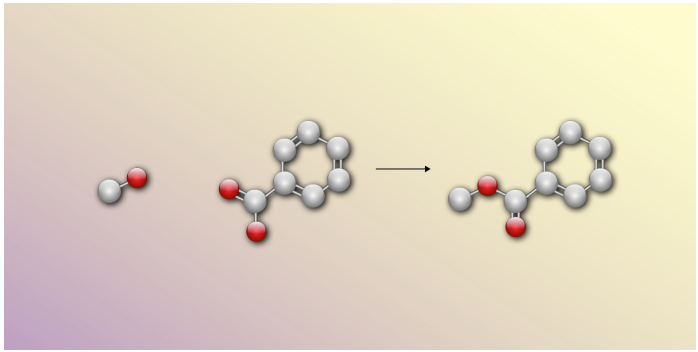


Atom-mapping
→ How atoms rearrange in chemical reactions



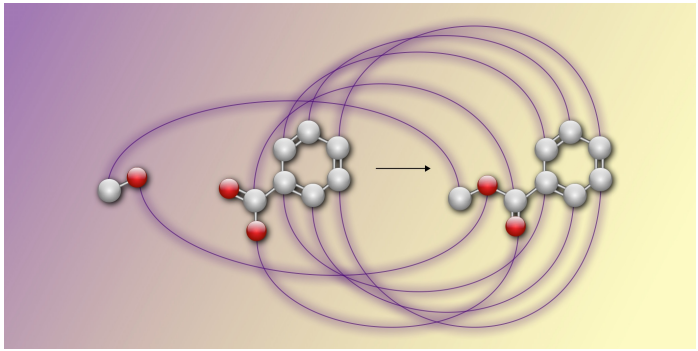
Discovery: Atom-mapping

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



RXNMapper

[CH3:1][OH:2].O[C:3](=[O:4])[c:5]1[cH:6][cH:7][cH:8][cH:9][cH:10]1>>[CH3:1][O:2][C:3](=[O:4])[c:5]1[cH:6][cH:7][cH:8][cH:9][cH:10]1



Independent *benchmark*

- 1851 curated reactions
- 5 popular tools (RXNMapper, ChemAxon, NameRXN, Indigo, and RDTools)

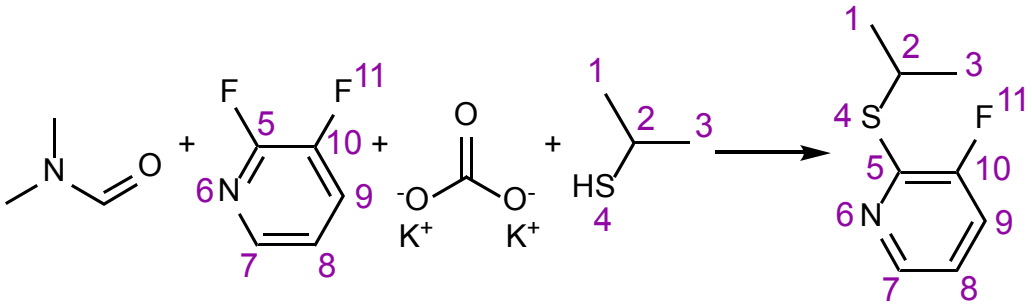
"**RXNMapper was ranked best** with an accuracy of **83.74%** by being the second fastest algorithm (0.05 seconds)."

Compared to **38.47%** accuracy for Indigo (open-source alternative)

Atom-to-atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies

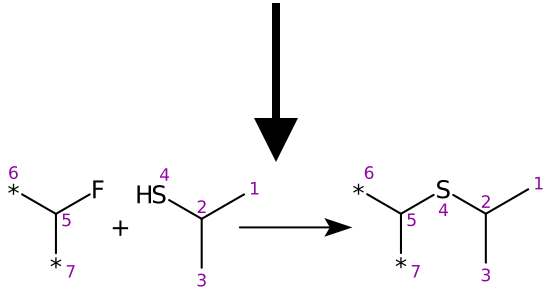
Arkadii Lin, Natalia Dyubankova, Timur I. Madzhidov, Ramil I. Nugmanov, Jonas Verhoeven, Timur R. Gimadiev, Valentina A. Afonina, Zarina Ibragimova, Assima Rakhimbekova ... See all authors

First published: 02 November 2021 | https://doi.org/10.1002/minf.202100138

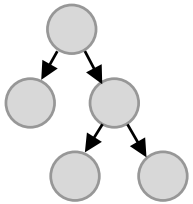


```
> python scripts/run_rxnmapper_on_dataset.py  \
>      --file_path examples/example_rxns.csv \
>      --output_path  out.json \
>      --batch_size 8
```

7sec/1k imbalanced reactions



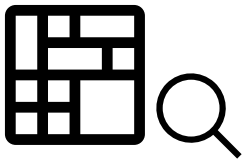
reaction center,
bond changes



Synthesis planning

$$H\psi(x) = E\psi(x)$$

Quantum mechanical simulations



Reaction accessibility and interpretability

Grammar of chemical reactions

...

Atom-mapping in *AI-assisted synthesis planning*

Template-based approaches

Planning chemical syntheses with deep neural networks and symbolic AI

Marwin H. S. Segler[1,2], Mike Preuss[3] & Mark P. Waller[4]

A robotic platform for flow synthesis of organic compounds informed by AI planning

Connor W. Coley[1,*], Dale A. Thomas III[1,2,*,†], Justin A. M. Lummiss[3,*,†], Jonathan N. Jaworski[3,‡], Christopher P....

Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain†

Amol Thakkar, *[ab] Thierry Kogej,[a] Jean-Louis Reymond, [b] Ola Engkvist[a] and Esben Jannik Bjerrum *[a]

Graph NN-based approaches (bond changes / graph edits)

A graph-convolutional neural network model for the prediction of chemical reactivity†

Connor W. Coley [a], Wengong Jin [b], Luke Rogers [a], Timothy F. Jamison [c], Tommi S. Jaakkola [b], William H. Green [a], Regina Barzilay *[b] and Klavs F. Jensen *[a]

Learning Graph Models for Retrosynthesis Prediction

VR Somnath, C Bunne, CW Coley, A Krause, R Barzilay

Neural Information Processing Systems (NeurIPS) 35

SMILES-to-SMILES approaches

Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction

Philippe Schwaller*, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee*

Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy†

Philippe Schwaller *[a], Riccardo Petraglia [a], Valerio Zullo [b], Vishnu H. Nair [a], Rico Andreas Haeuselmann [a], Riccardo Pisoni [a], Costas Bekas [a], Anna Iuliano [b] and Teodoro Laino [a]

Atom-mapping dependent
-> **training data**
atom-maps -> templates
atom-maps -> graph edits

All benefit from better atom-maps!

Atom-mapping independent
-> Atom rearrangement not tracked

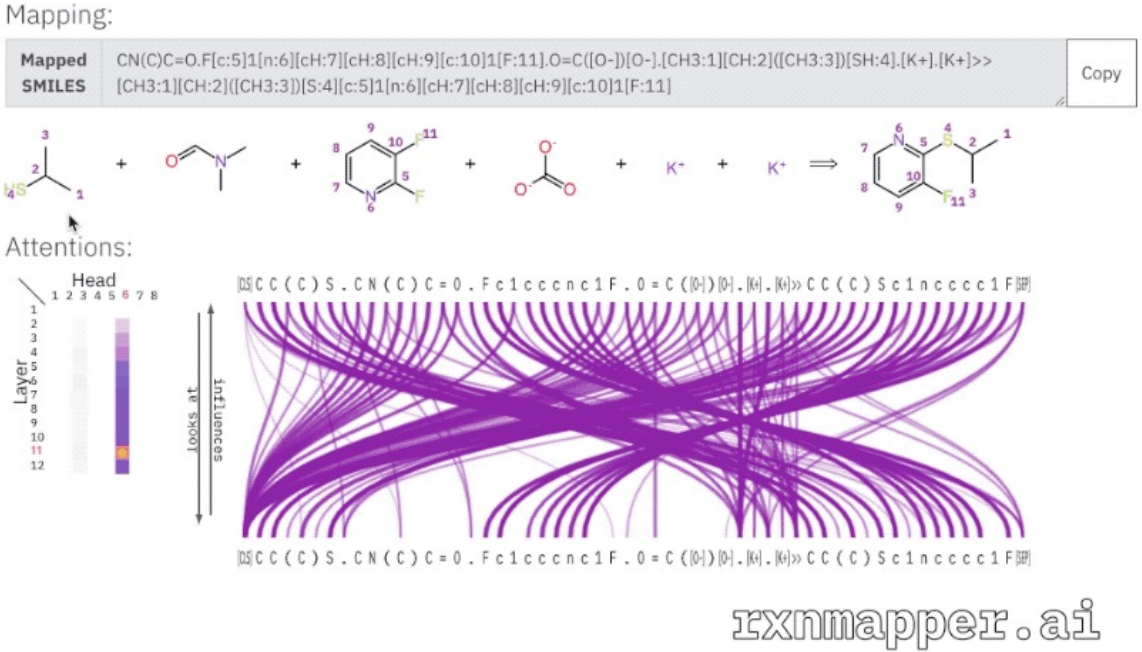
Open-source code

https://github.com/rxn4chemistry/rxnmapper | pip install rxnmapper

```
from rxnmapper import RXNMapper
rxn_mapper = RXNMapper()
rxns = ['CC(C)S.CN(C)C=O.Fc1cccnc1F.O=C([O-])[O-].[K+].[K+]>>CC(C)Sc1ncccc1F', 'C1COCC
results = rxn_mapper.get_attention_guided_atom_maps(rxns)
```

```
[{'mapped_rxn': 'CN(C)C=O.F[c:5]1[n:6][cH:7][cH:8][cH:9][c:10]1[F:11].O=C([O-])[O-].[CH3:1][CH
  'confidence': 0.9565619900376546},
 {'mapped_rxn': 'C1COCCO1.CC(C)(C)[O:3][C:2](=[O:1])[CH2:4][O:5][NH:6][C:7](=[O:8])[NH:9][CH2:
  'confidence': 0.9704424331552834}]
```

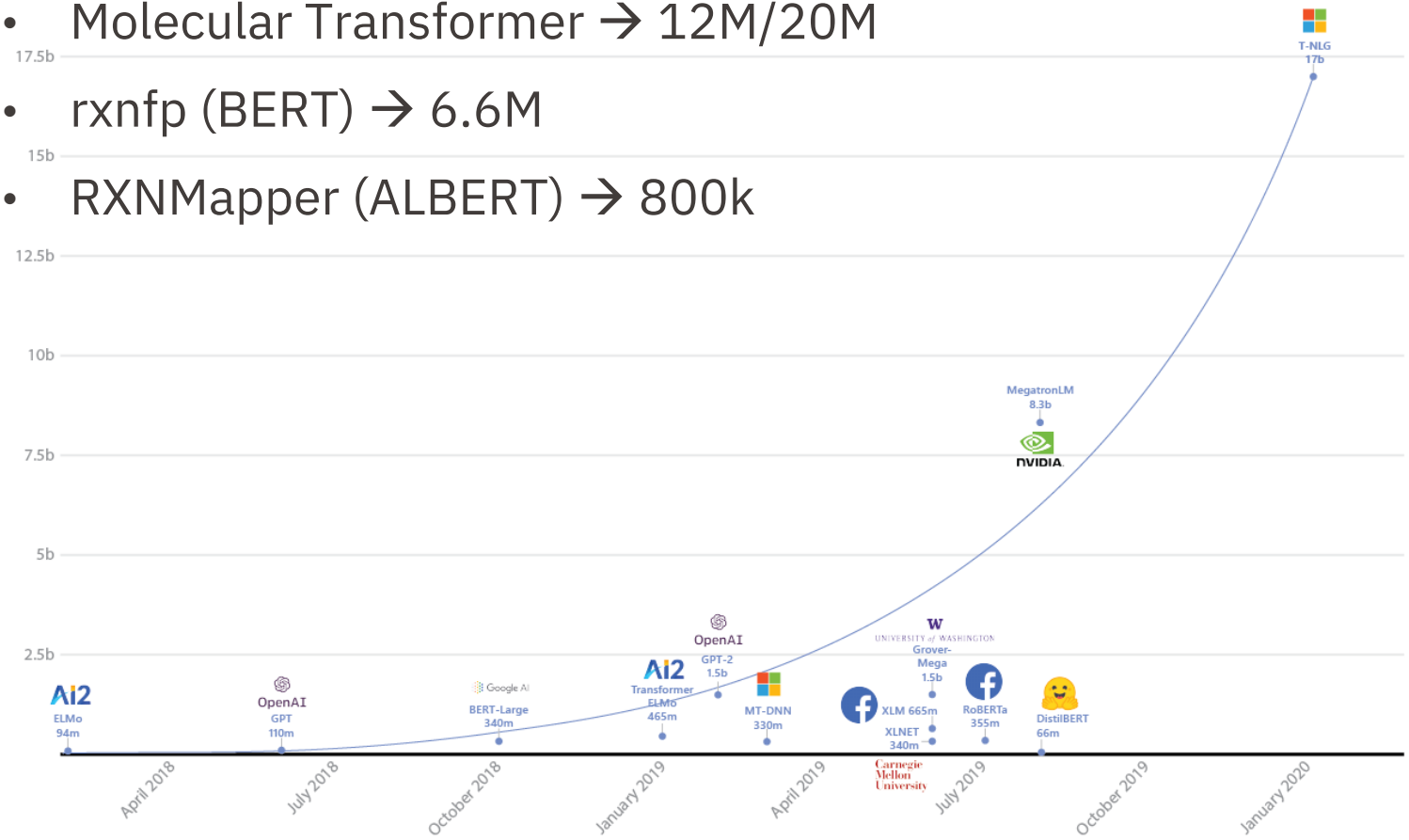
Demo on RXNMapper.ai



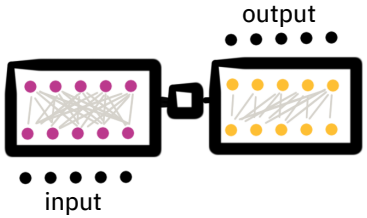
Give it a try!

Model sizes – opposite of trend

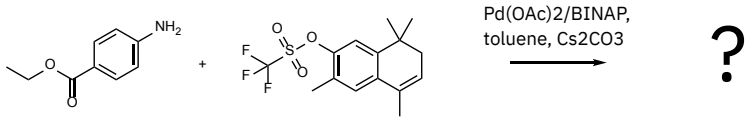
- Molecular Transformer → 12M/20M
- rxnfp (BERT) → 6.6M
- RXNMapper (ALBERT) → 800k



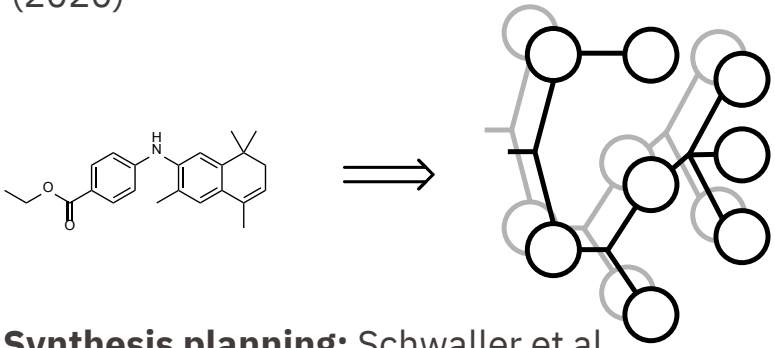
Summary



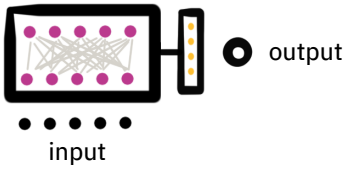
encoder **decoder**



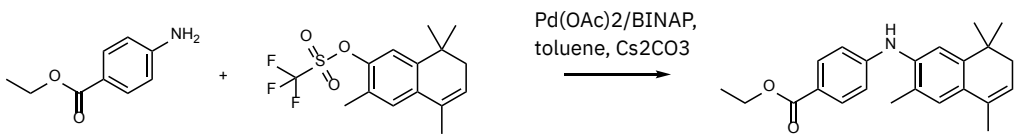
Reaction prediction:
Schwaller et al. *ACS Cent. Sci.* 2019
Pesciullesi et al. *Nature Comm.* 11, 4874 (2020)



Synthesis planning: Schwaller et al. *Chem. Sci., 2020,11, 3316-3325*



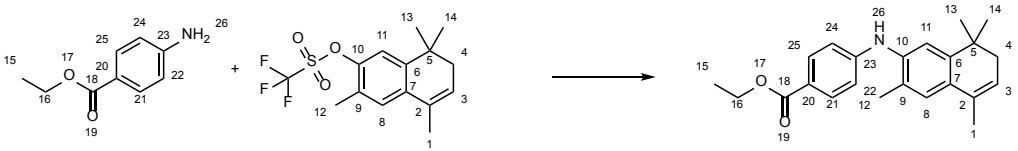
encoder-only



Classification/fingerprints: Schwaller et al. *Nature Mach. Intell., 2021*

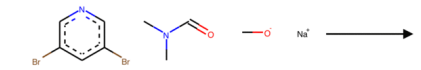
Reaction yields: Schwaller et al. *Mach. Learn.: Sci. Technol.*, 2021; Schwaller et al. NeurIPS 2020 Ml4Mol workshop

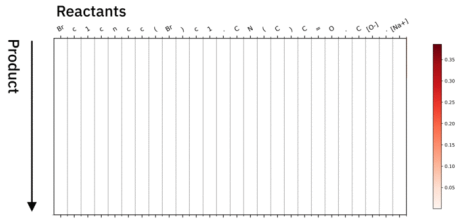
Atom-mapping: Schwaller et al. *Science Advances, 2021*

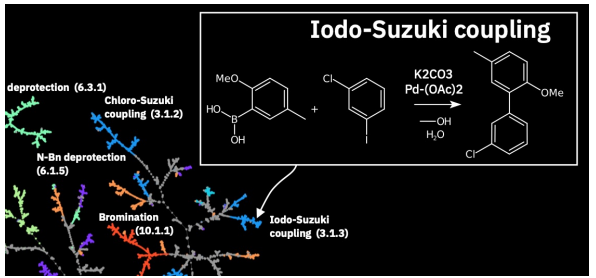


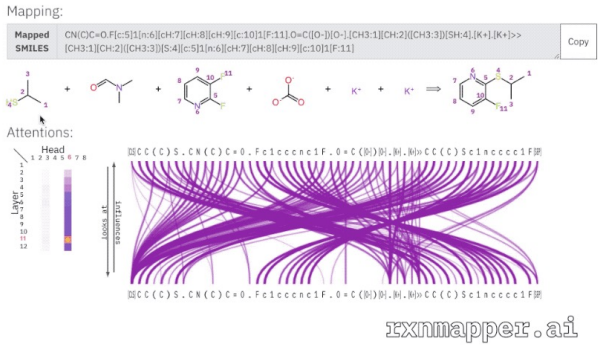
Conclusions

- Apply **ML/NLP methods** to **synthesis**
- **RXN for Chemistry** platform / API access
  - reaction prediction / synthesis planning
  - paragraph-2-actions / RoboRXN simulator
- Data-driven **reaction fingerprints**
  - Reaction atlas / nearest neighbour search
  - Reaction yields
- Transformers capture **the grammar of chemical reactions**
- **Fast, high-quality atom-maps** even on strongly imbalanced reactions.

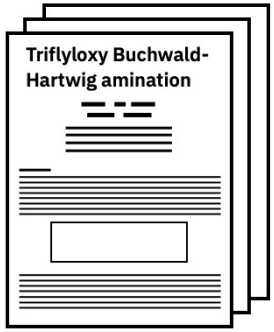


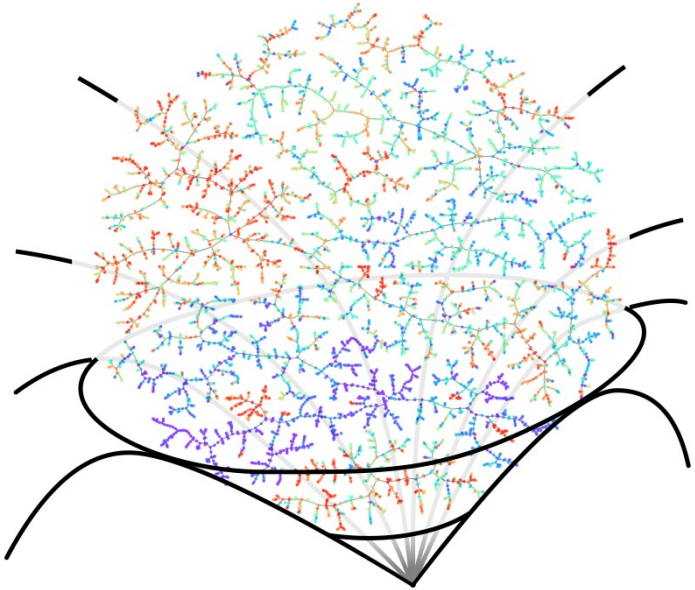




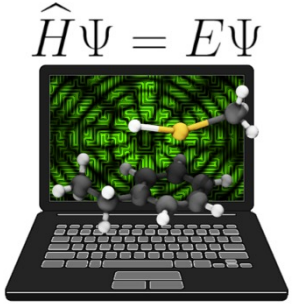


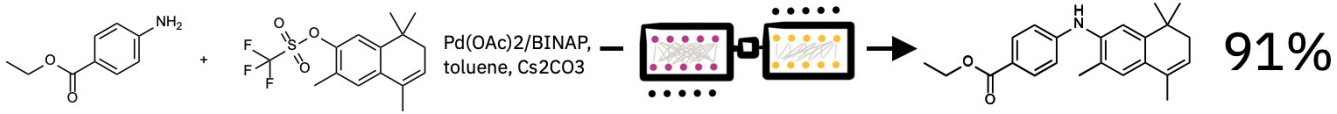
Review & Outlook











https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wcms.1604

Machine intelligence for chemical reaction space

Philippe Schwaller ✉, Alain C. Vaucher, Ruben Laplaza, Charlotte Bunne, Andreas Krause, Clemence Corminboeuf, Teodoro Laino

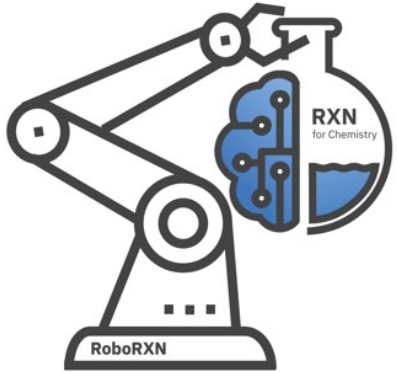
First published: 07 March 2022 | https://doi.org/10.1002/wcms.1604

EPFL LIAC

Bojana Rankovic,
Oliver Schilter (with IBM)

IBM **Research** AI

Teodoro Laino, Alain Vaucher, Alessandra Toniato, Theophile Gaudin, Costas Bekas, Daniel Probst, Matteo Manica and the RoboRXN team, Ben Hoover, Hendrik Strobelt



NextMove Software

docker

RDKit
Open-Source Cheminformatics and Machine Learning

NMT
An open source neural machine translation system.

Jean-Louis Reymond, Giorgio Pesciullesi, David Kreutter, Alice Capecchi, Amol Thakkar, and the Reymond group

UNIVERSITÄT BERN

UNIVERSITY OF CAMBRIDGE

Alpha Lee, Peter Bolgar and Ryan-Rhys Griffiths

PyTorch

Transformers

NCCR Catalysis

Clemence Corminboeuf, Andreas Krause, Ruben Laplaza, Charlotte Bunne

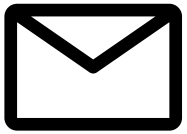
Many thanks for your attention!



@pschwllr @SchwallerGroup



https://github.com/pschwllr
https://github.com/rxn4chemistry



~~phs@zurich.ibm.com~~
philippe.schwaller@epfl.ch

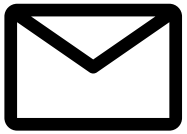
Many thanks for your attention!



@pschwllr @SchwallerGroup



https://github.com/pschwllr
https://github.com/rxn4chemistry



~~phs@zurich.ibm.com~~
philippe.schwaller@epfl.ch

Colab exercise:
https://github.com/schwallergroup/dmds_language_models_for_reactions