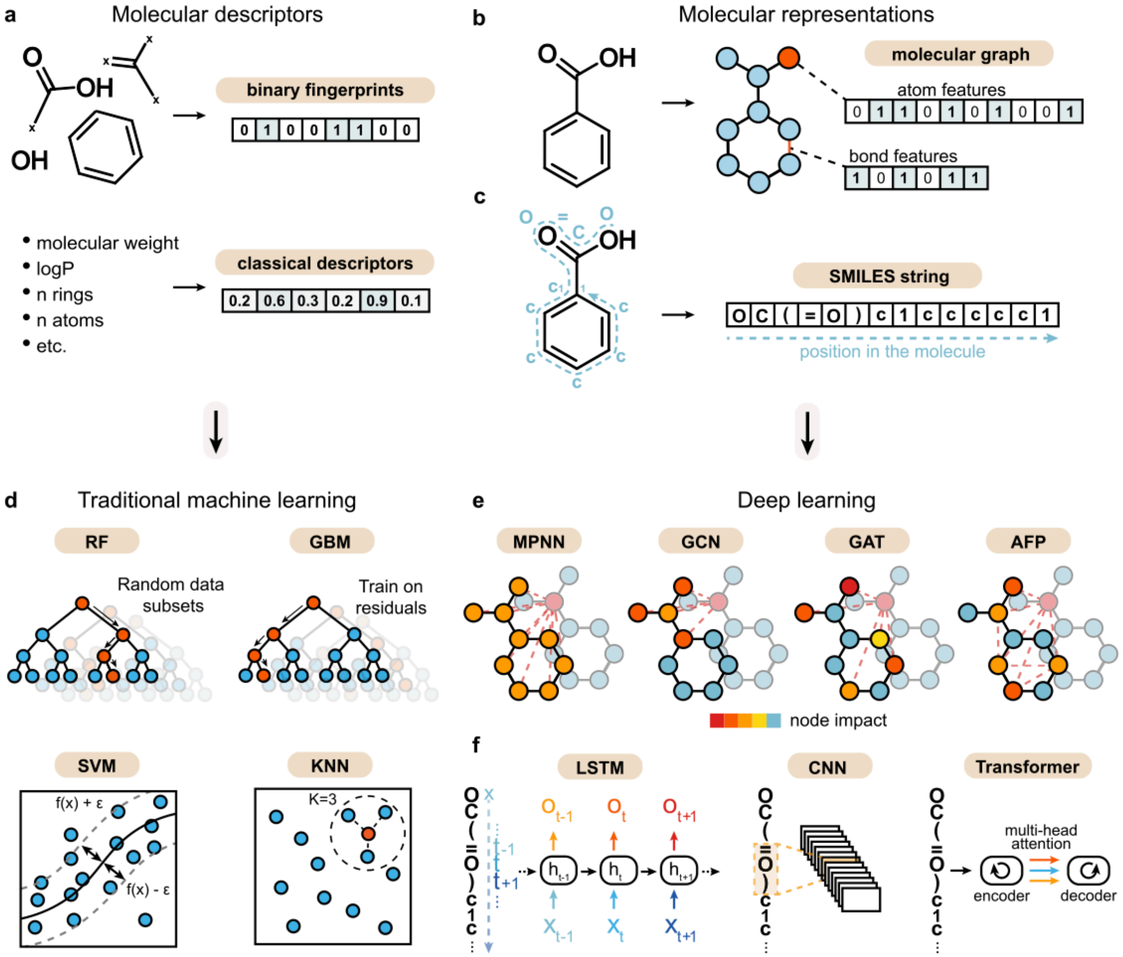
Reaction prediction

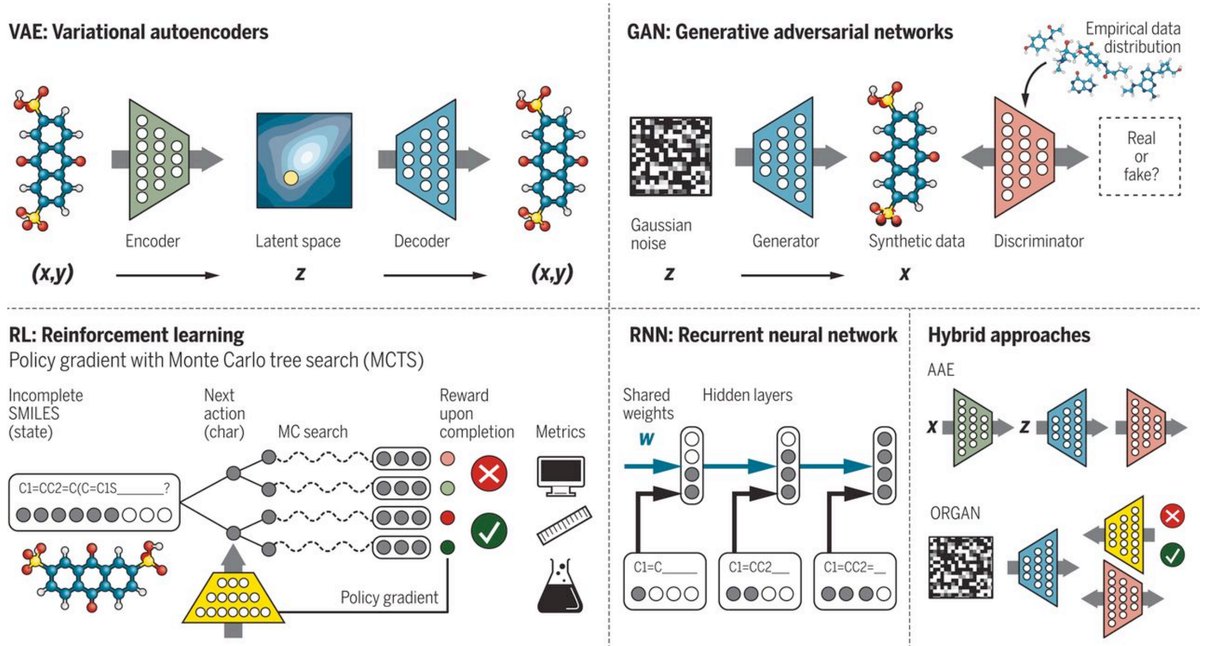
Philippe Schwaller

Laboratory of Artificial Chemical Intelligence (LIAC)

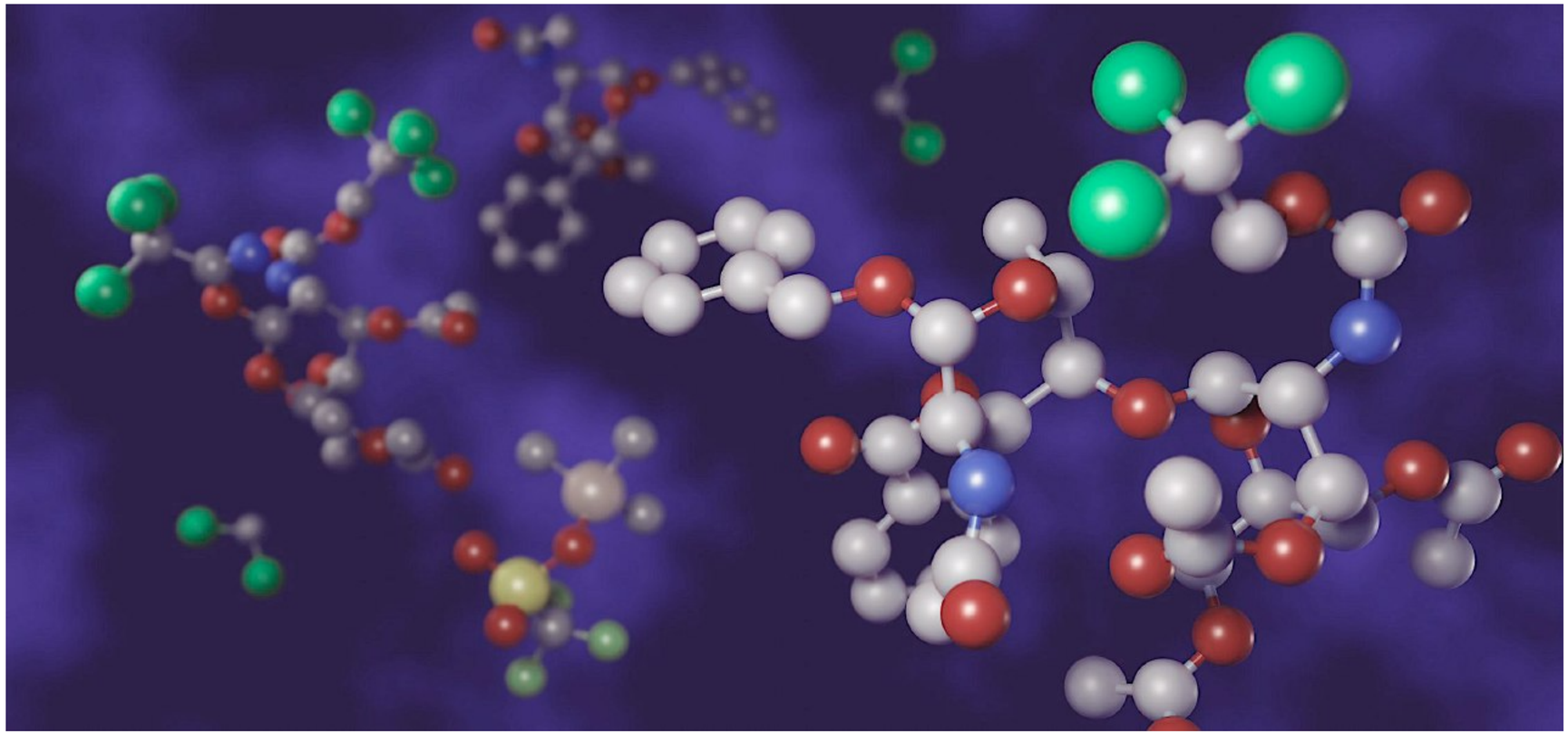
AI for Chemistry

Recap — single molecules

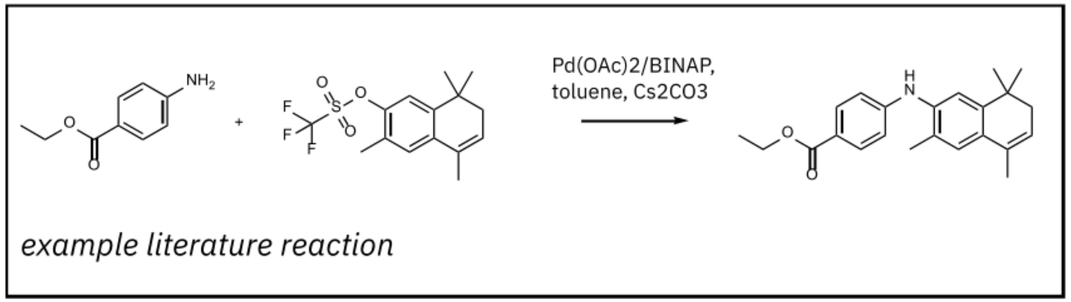




Today: Chemical reactions



A chemical reaction is a process in which one or more substances,
the reactants, are converted to one or more different substances, the products.

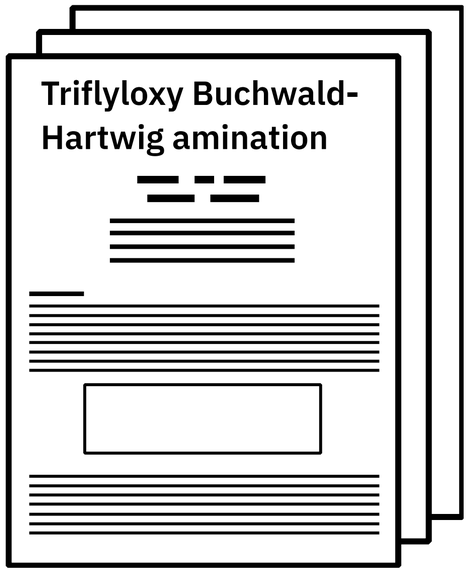


example literature reaction

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

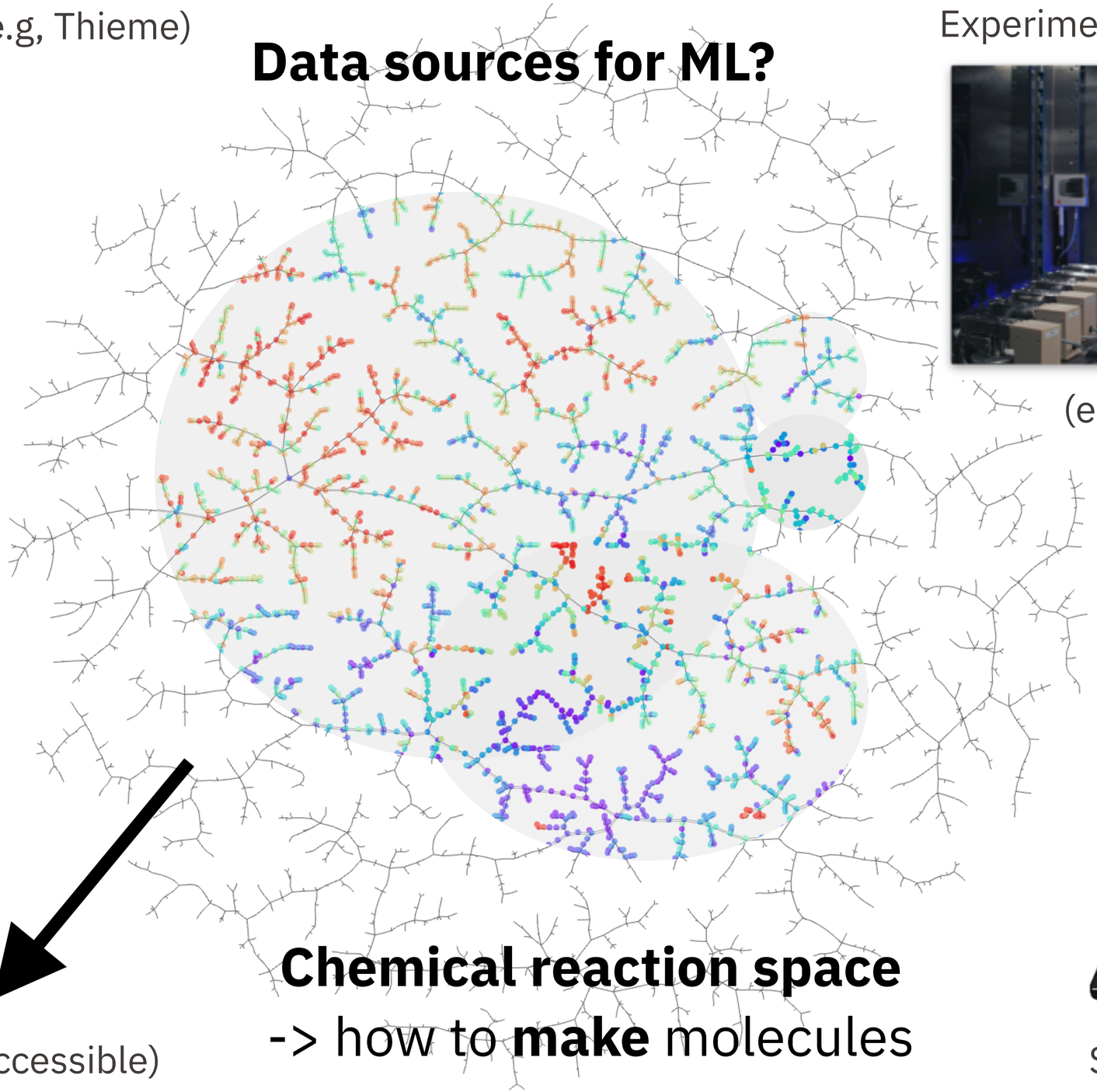
Data sources for ML?

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

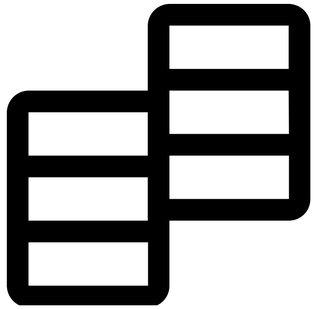
$$\hat{H}\Psi = E\Psi$$

- Patents (broad, accessible)

Simulations (narrow)

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

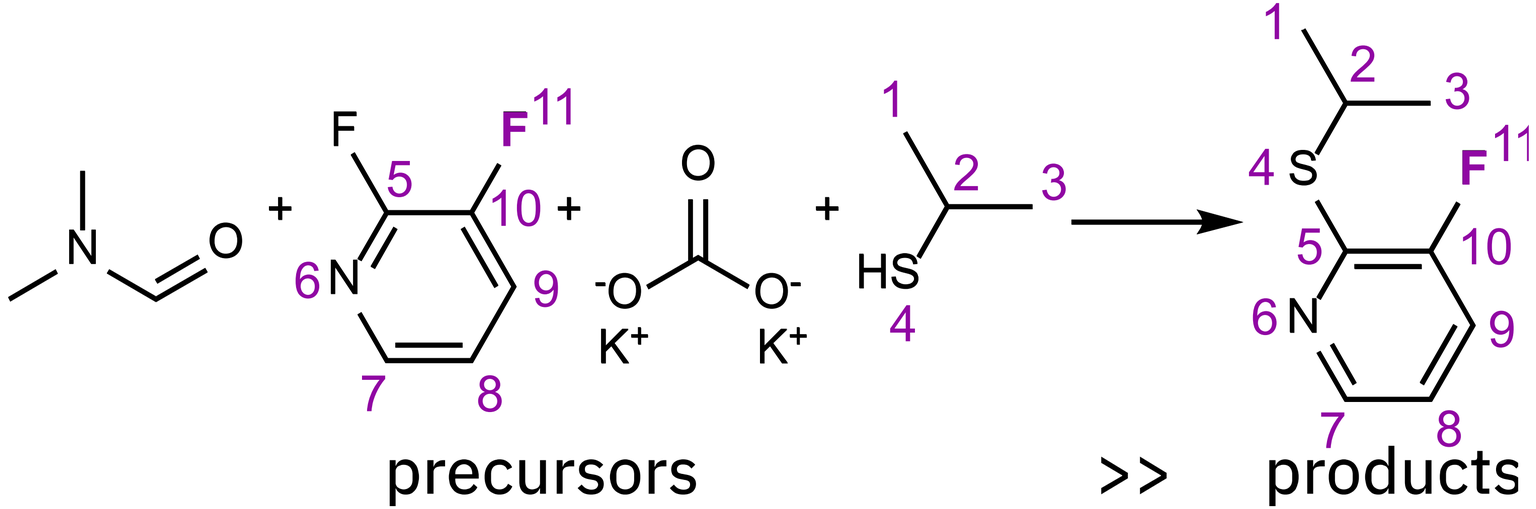
Benchmark sets

USPTO_MIT

USPTO_STEREO

Reaction SMILES

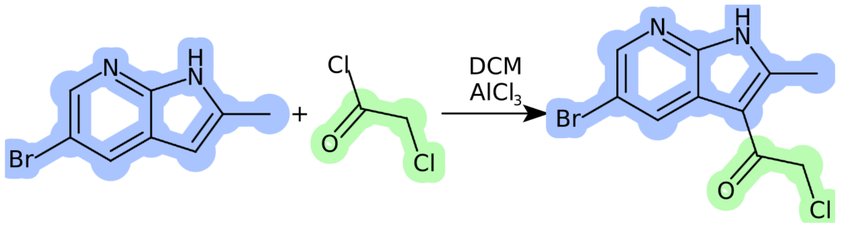
CC(C)S.CN(C)C=O.Fc1cccnc1F.O=C([O-])[O-].[K+].[K+]>>CC(C)Sc1nccccc1F



Synthesis procedures (patents/literature)

To a suspension of AlCl3 (1.57 g, 11.84 mmol) in dichloromethane (50 mL) was added 5-Bromo-2-methyl-1H-pyrrolo[2,3-b]pyridine. After stirring for 30 min, chloroacetyl chloride (1.33 g, 11.84 mmol) was added and the reaction mixture was stirred for 2 hours at room temperature. On completion, solvents were evaporated and quenched with aq. NaHCO3 solution at 0° C. Resulting mixture was extracted with EtOAc. The organic layer was dried over Na2SO4 and filtered through a plug of silica gel. Solvent was evaporated to dryness to give 1-(5-Bromo-2-methyl-1H-pyrrolo[2,3-b]pyridin-3-yl)-2-chloro-ethanone (0.650 g, 95% yield).

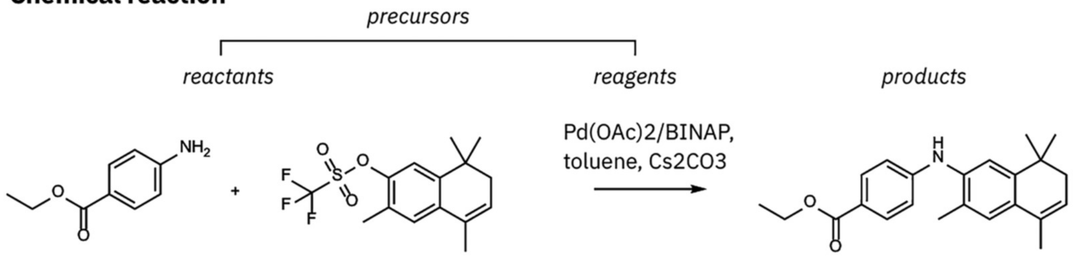
1-(5-Bromo-2-methyl-1H-pyrrolo[2,3-b]pyridin-3-yl)-2-chloro-ethanone	Product	$C_{10}H_8BrClN_2O$	287.54 g/mol	2.261 mmol
5-Bromo-2-methyl-1H-pyrrolo[2,3-b]pyridine	Reactant	$C_8H_7BrN_2$	211.058 g/mol	
chloroacetyl chloride	Reactant	$C_2H_2Cl_2O$	112.943 g/mol	11.84 mmol
AlCl3	Agent	$AlCl_3$	133.34 g/mol	11.84 mmol
dichloromethane	Solvent	CH_2Cl_2	84.932 g/mol	781.209 mmol



Friedel-Crafts acylation (3.10.1)

Reaction *representations*

Chemical reaction



Meta data

reaction class - 1.3.4
Buchwald-Hartwig amination

reaction yield - 80%

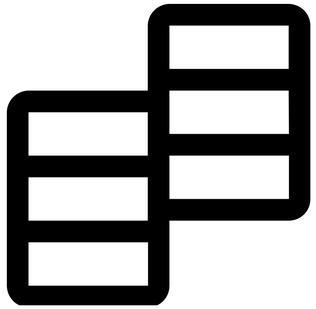
experimental procedures

"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

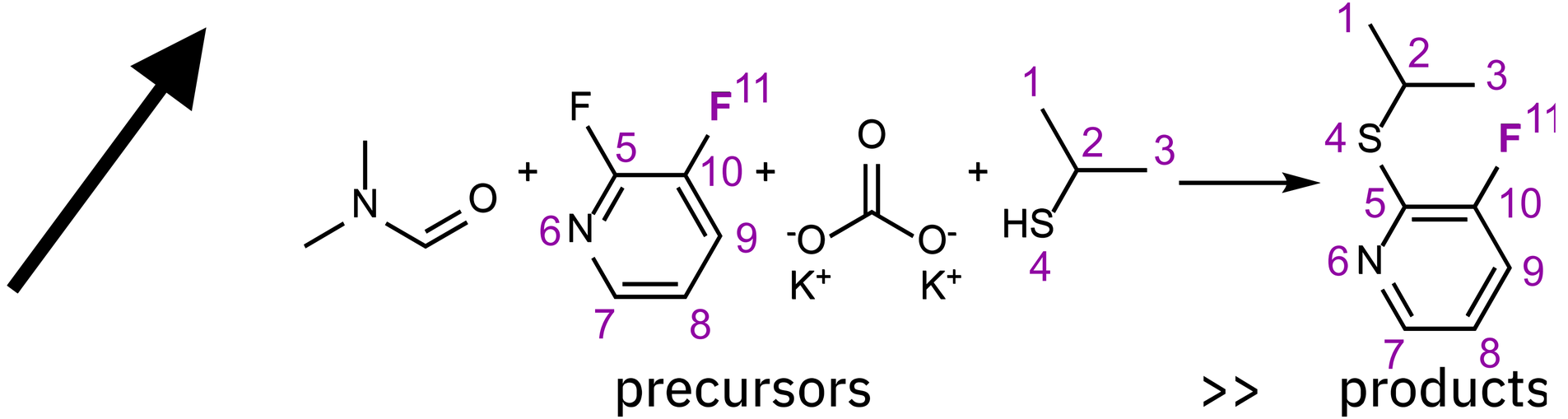
Benchmark sets

USPTO_MIT

USPTO_STEREO

Reaction SMILES

CC(C)S.CN(C)C=O.Fc1cccnc1F.O=C([O-])[O-].[K+].[K+]>>CC(C)Sc1nccc1F



Early work — expert systems (‘60s, ‘70s)

ORGANIZATION OF LHASA

The LHASA program is exceedingly complex - about 400 subroutines, 30,000 lines of FORTRAN code and a data base of over 600 common chemical reactions. To

Empirical Explorations of SYNCHEM

The methods of artificial intelligence are applied to the problem of organic synthesis route discovery.

H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer, J. E. Searleman

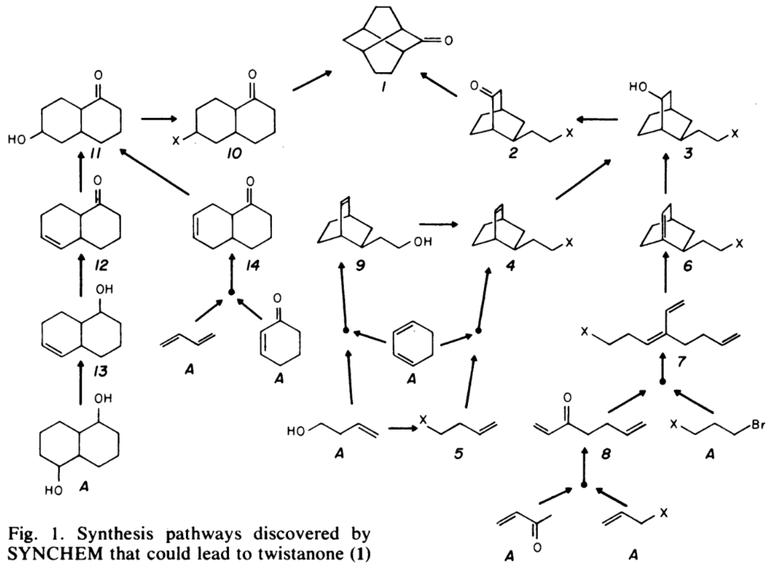


Fig. 1. Synthesis pathways discovered by SYNCHEM that could lead to twistanone (1) from available starting materials. Reaction types used: alkylation alpha to a ketone (1 ← 2, 1 ← 10, 8 ← A + A); oxidation of a secondary alcohol (2 ← 3, 12 ← 13); hydration of an alkene (3 ← 4, 3 ← 6, 11 ← 12, 13 ←A); Diels-Alder reaction (4 ← 5 + A, 6 ← 7, 9 ← A + A, 14 ← A +A); Wittig reaction (7 ← 8 + A); and replacement of an alcohol by a better leaving group (10 ← 11, 4 ← 9, 5 ← A). All compounds labeled A were found by SYNCHEM on its list of available compounds.

- https://pubs.acs.org/doi/10.1021/acs.jcim.0c00448

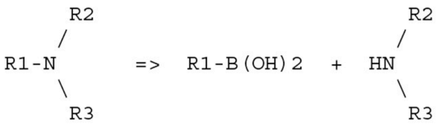
LHASA—Logic and Heuristics Applied to Synthetic Analysis

DAVID A. PENSAK

Central Research and Develop. Dept., E. I. du Pont de Nemours and Co., Wilmington, Del. 19898

E. J. COREY

Dept. of Chemistry, Harvard University, Cambridge, Mass. 02138




```
R1-O,S-R2 => R1-B(OH)2  +  R2-O,S-H

END*REFERENCES
...
TYPICAL*YIELD                    GOOD
RELIABILITY                      GOOD
REPUTATION                       GOOD
HOMOSELECTIVITY                  POOR
HETEROSELECTIVITY                FAIR
ORIENTATIONAL*SELECTIVITY        NOT*APPLICABLE
CONDITION*FLEXIBILITY            POOR
THERMODYNAMICS                   GOOD
```

More expert systems:

- SECS by Wipke
- EROS by Gasteiger
- CAMEO by Jorgensen

An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs

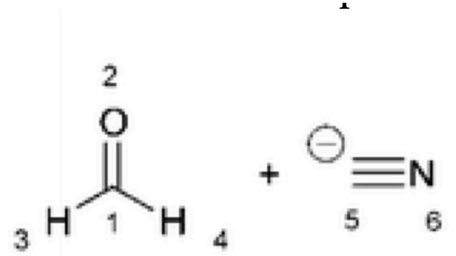
1973

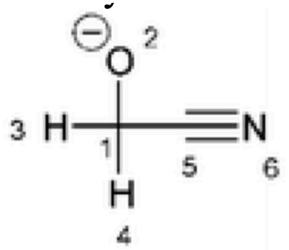
Prof. Dr. James Dugundji

Department of Mathematics, University of Southern California, Los Angeles, California, USA

Prof. Dr. Ivar Ugi

Department of Chemistry, University of Southern California, Los Angeles, California, USA and Laboratorium für Organische Chemie der Technischen Universität München





	C^1	O^2	H^3	H^4	C^5	N^6
C^1	0	2	1	1	0	0
O^2	2	4	0	0	0	0
H^3	1	0	0	0	0	0
H^4	1	0	0	0	0	0
C^5	0	0	0	0	2	3
N^6	0	0	0	0	3	2

be-matrix B

+

	C^1	O^2	H^3	H^4	C^5	N^6
C^1	0	-1	0	0	+1	0
O^2	-1	+2	0	0	0	0
H^3	0	0	0	0	0	0
H^4	0	0	0	0	0	0
C^5	+1	0	0	0	-2	0
N^6	0	0	0	0	0	0

r-matrix R

	C^1	O^2	H^3	H^4	C^5	N^6
C^1	0	1	1	1	1	0
O^2	1	6	0	0	0	0
H^3	1	0	0	0	0	0
H^4	1	0	0	0	0	0
C^5	1	0	0	0	0	3
N^6	0	0	0	0	3	2

be-matrix E

Bond electron matrix
- diagonal: free valence electrons
- off-diagonal: bond order

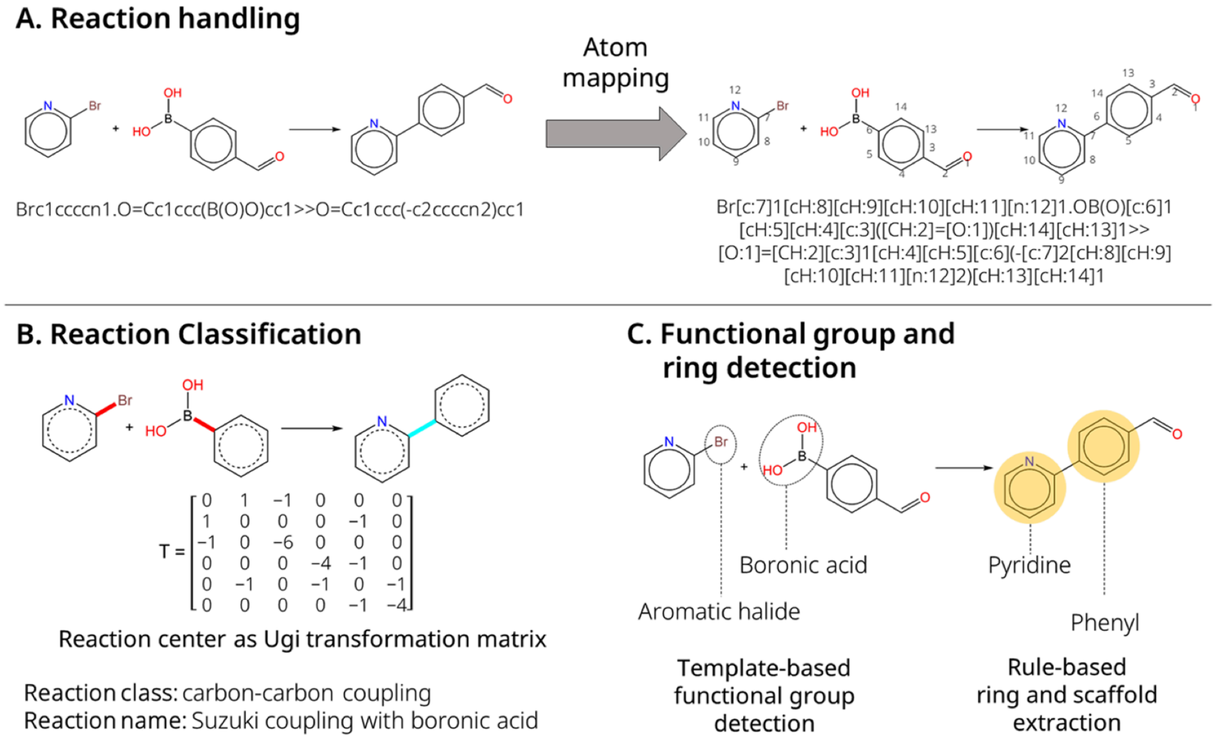
Reaction matrix

Bond electron matrix of the product

RESEARCH Open Access

Rxn-INSIGHT: fast chemical reaction analysis using bond-electron matrices

Maarten R. Dobbelaere[1], István Lengyel[1,2], Christian V. Stevens[3] and Kevin M. Van Geem[1*]



SOPHIA, a Knowledge Base-Guided Reaction Prediction System

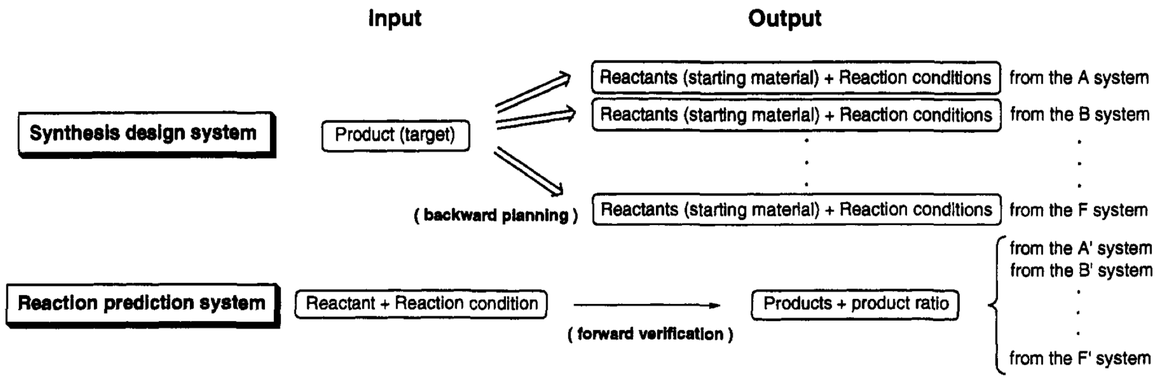


Figure 1. Difference between a synthesis design system and a reaction prediction system.

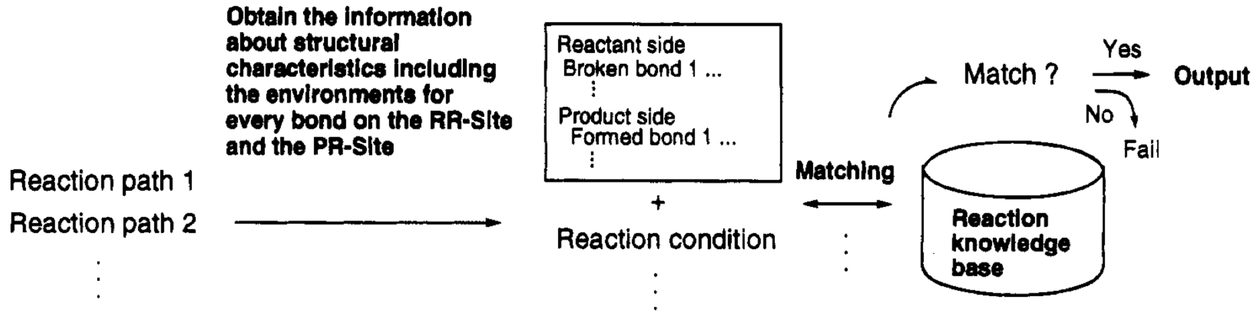


Figure 12. Reaction path evaluation.

Hiroko Satoh and Kimito Funatsu*

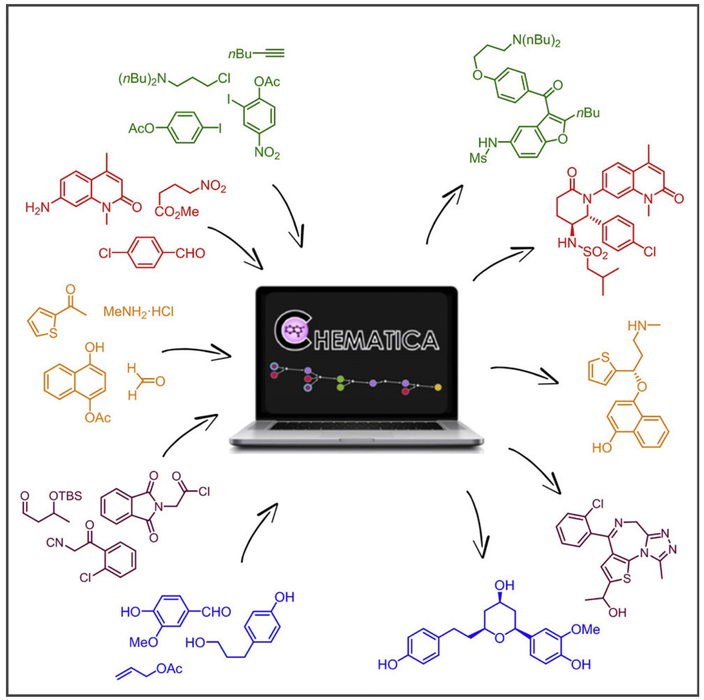
Department of Knowledge-Based Information Engineering, Toyohashi University of Technology, Tempaku, Toyohashi 441, Japan

Received March 28, 1994⊗

Hiroko Satoh
• Associate Prof at ROIS, Japan.
• Researcher at Uni of Zürich.

Chematica (Synthia)

- Largest expert rules system
- More than 100k human expert rules
- Sadly not open-source



Article

Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory

Tomasz Klucznik [1], Barbara Mikulak-Klucznik [1], Michael P. McCormack [2], Heather Lima [2], Sara Szymkuć [1], Manishabrata Bhowmick [2], Karol Molga [1], Yubai Zhou [3], Lindsey Rickershauser [2], Ewa P. Gajewska [1], Alexei Toutchkine [2], Piotr Dittwald [1], Michał P. Startek [4], Gregory J. Kirkovits [2], Rafał Roszak [1], Ariel Adamski [1], Bianka Sieredzińska [1], Milan Mrksich [3], Sarah L.J. Trice [2], Bartosz A. Grzybowski [1 5 6]

Article | Published: 13 October 2020

Computational planning of the synthesis of complex natural products

Barbara Mikulak-Klucznik, Patrycja Gołębiowska, Alison A. Bayly, Oskar Popik, Tomasz Klucznik, Sara Szymkuć, Ewa P. Gajewska, Piotr Dittwald, Olga Staszewska-Krajewska, Wiktor Beker, Tomasz Badowski, Karl A. Scheidt, Karol Molga, Jacek Mlynarski, Milan Mrksich & Bartosz A. Grzybowski

Nature **588**, 83–88 (2020) | Cite this article

26k Accesses | **81** Citations | **188** Altmetric | Metrics

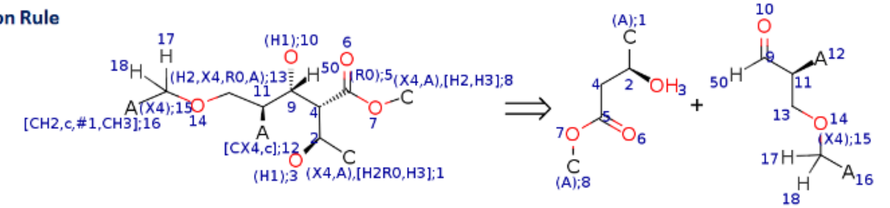
Computer-Assisted Synthetic Planning: The End of the Beginning

Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Dr. Piotr Dittwald, Dr. Michał Startek, Michał Bajczyk, Prof. Dr. Bartosz A. Grzybowski

First published: 08 April 2016 | **https://doi.org/10.1002/anie.201506101** | Citations: 315

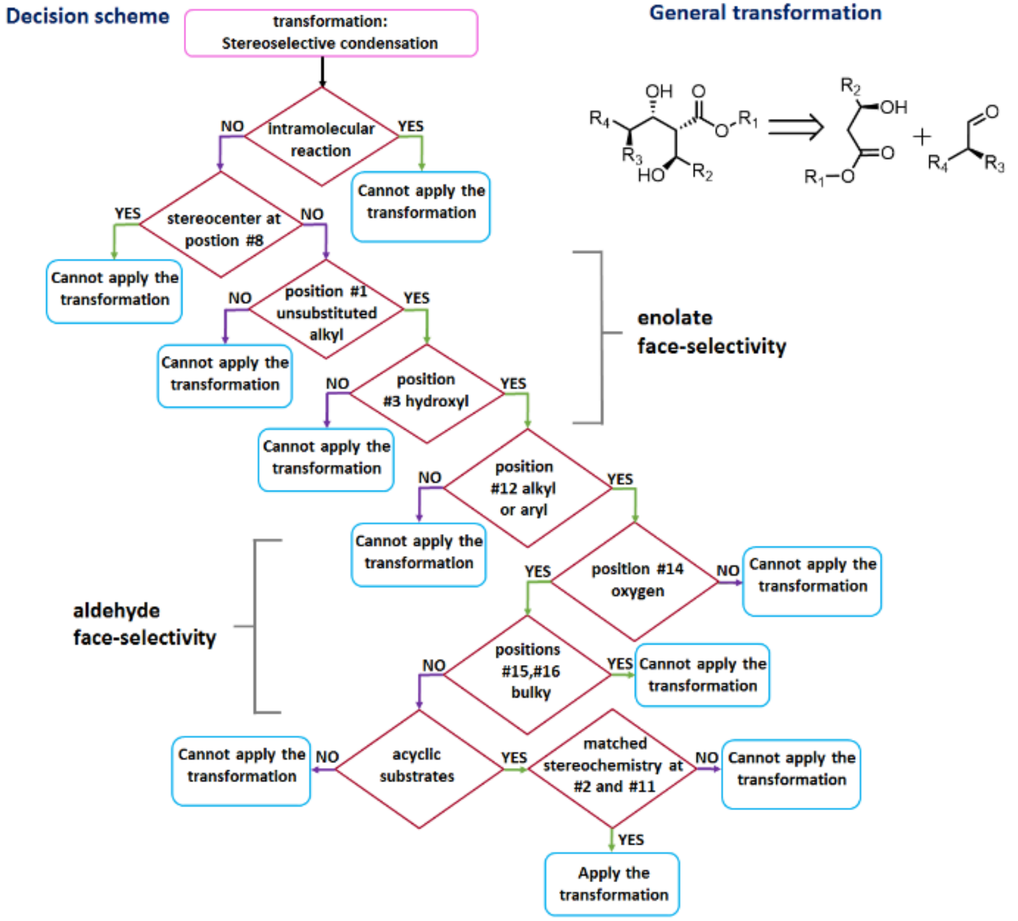
Example

Reaction Rule



name: "Stereoselective condensation of esters with aldehydes"
reaction SMARTS: ["[CX4H2R0,CX4H3:1][C@@H:2]([OH:3])[C@H:4]([C@@:9]([#1:50])([OH:10])[C@@H:11]([CX4,c:12])[CX4H2R0:13][O:14][CX4:15]([#1:17])([#1:18])[CH2,c,#1,CH3:16])[CR0:5](=[O:6])[O:7][CX4H2,CX4H3:8]>>[C:1][C@@H:2]([O:3])[C:4][C:5](=[O:6])[O:7][C:8].[#1:50][C:9](=[O:10])[C@@H:11]([*:12])[C:13][O:14][CX4:15]([#1:17])([#1:18])[*:16]"]
protection_conditions_code: ["SB16","SC88"]
protections: ["[#6][CH2][OH]", "[#6][CH]([#6])[OH]", "[#6][C]([#6])([#6])[OH]", "[c][OH]", "[OH][c][c][OH]", "[#6][CH]=O", "[#6]C([OH])=O", "[CX4,c][SX2H]", "[OH][CX4][CX4][OH]", "[OH][CX4]C[CX4][OH]", "[CX4,c][NH2]", "[CX4,c][NH][CX4,c]", "[nH] "]
incompatibilities: ["[#6][CH]=[SX1]", "[CX3]=[NX3+][O-]", "[CX4][O][S](=O)(=O)[#6]", "[#6]S(=O)[Cl,Br,I]", "[#6]C(=[SX1])[#6]", "[#6][SX3](=O)[OH]", "[CX4]1[SX2][CX4]1", "[#6][S](=O)(=O)[OH]", "[#6][N+]#[C-]", "[#6]N=C=[O,S]", "[#6][SX2,O]C#N", "[#6]C(=O)[Cl,Br,I]", "[#6]C(=[O])OC(=[O])[#6]", "ClC=N", "[#6]O[N+]([O-])=O" , "[#6]O[OH]", "[#6]OO[#6]", "[#6][NX2]=O", "[CX3]=[CX2]=O", "[#6]=[N+]=[N-]", "c[N+]#[N]", "[CX3]=[NX2H]", "[CX3]=[NX2][O]", "[#6][NX3][OH]", "[CX3]=[CX3][OH]", "[OH][CX4][O]", "[#6][Li]", "[#6][BX3]([O,#6])[O,#6]", "[#6][Mg][*]", "[#6][B-](F)(F)F", "[#6][Zn][*]", "[#6][PX3]([#6])[#6]", "N=N", "[#6][SX2][SX2][#6]", "[#6][SX3](=O)[#6]", "[CX4][Cl,Br,I]", "[Cl,Br,I]C#C", "C#[CH]", "[#6][S](=O)(=O)[#6]", "[CX4]1[O,N][CX4]1", "[#6]C(=O)[N]=[N+]=[N-]", "[CX4!H0][N+]([O-])=O", "[CX4!H0]C#N", "[#6]C(=O)[NH2]", "[#6]C([NH][CX4,c])=[O,S]", "[CX4!H0][C](=[O])[OH0]", "[CX4!H0]C(=O)N([#6])[#6]", "[#6][S](=O)(=O)[NH2]", "[CX4,c][NX3][NH2]", "[CX3]([#6,#1])([#6,#1])=[NX2][*!O]", "[CX3!H0]=[CX3]C#N", "[CX3!H0]=[CX3]C(=O)[O,N,S]", "[CX2]#[CX2]C#N", "[CX2]#[CX2]C(=O)[O,N,S]", "[CX3]=[CX2]=[CX3,CX2]", "[n][c;r6]([Cl,F])[n,c]"]
typical reaction conditions: "1.LDA.THF then TMSCl 2.TiCl4.DCM"
references: " 10.1016/S0040-4039(00)82373-4 "
diastereoselective: False

1 of 100k expert-written reaction rules



- https://ars.els-cdn.com/content/image/1-s2.0-S2451929418300639-mmc4.pdf

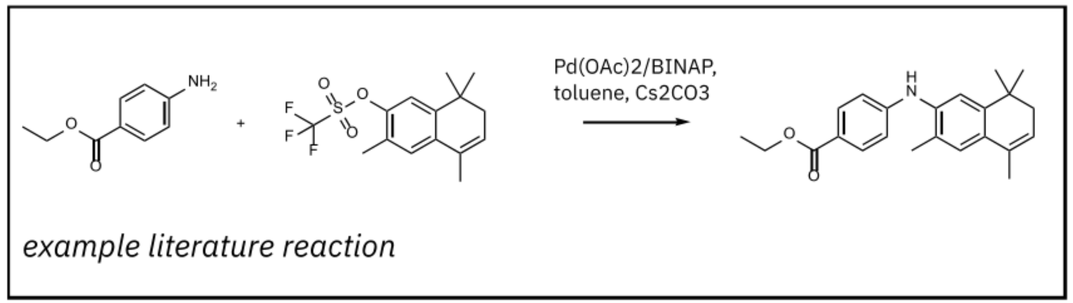
Reaction prediction

reaction classification task

Triflyloxy Buchwald-Hartwig amination

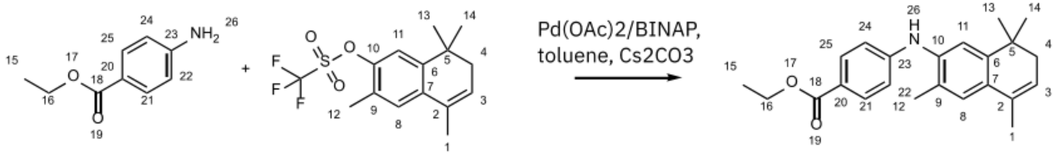
yield prediction task

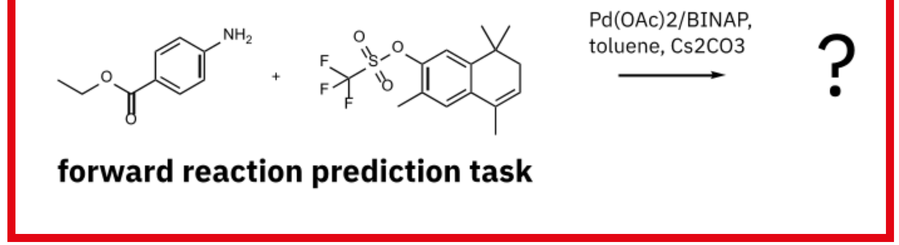
91%



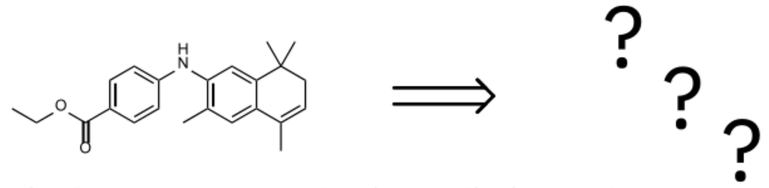
example literature reaction

atom-mapping task

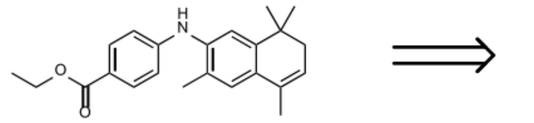




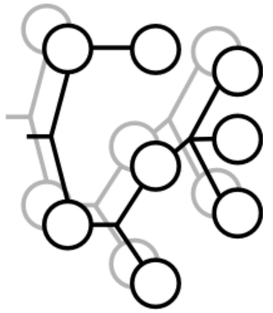
forward reaction prediction task



single-step retrosynthesis prediction task

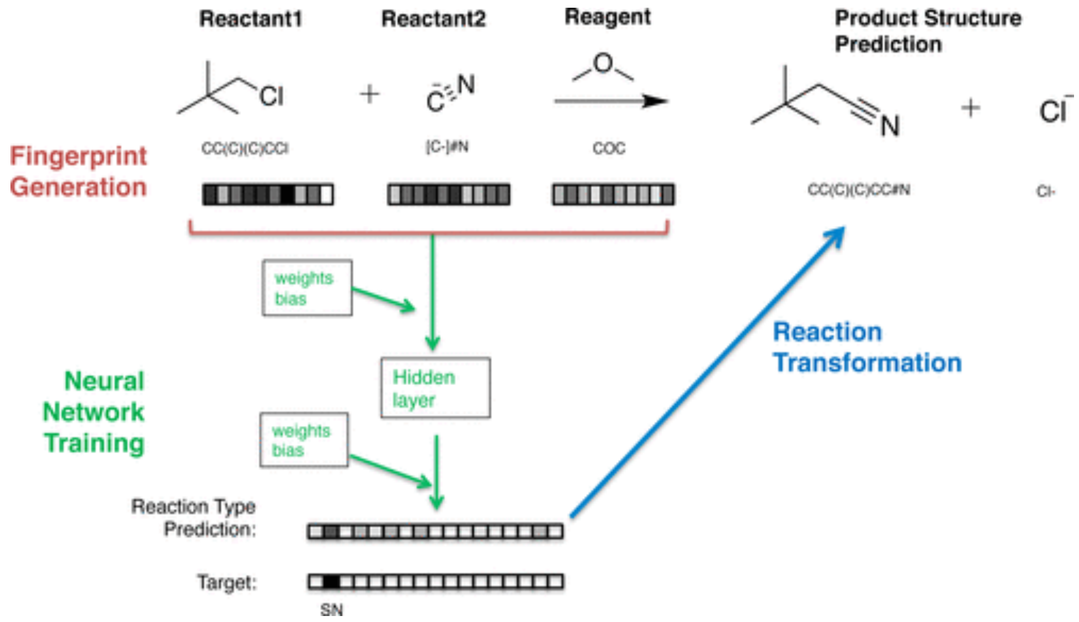


multi-step synthesis planning task



Early deep learning for reaction prediction

- Kayala (2011), Fooshee (2018) & Baldi: neural network to predict mechanistic steps through the identification and ranking of electron sources and sinks
  - Limitation: Hand-crafted rules as training (11k elementary reactions, not open)
- Wei et al. (2016), reaction template prediction (only 17 classes, 2 reactants, 1 reagent)



Neural Networks for the Prediction of Organic Chemistry Reactions

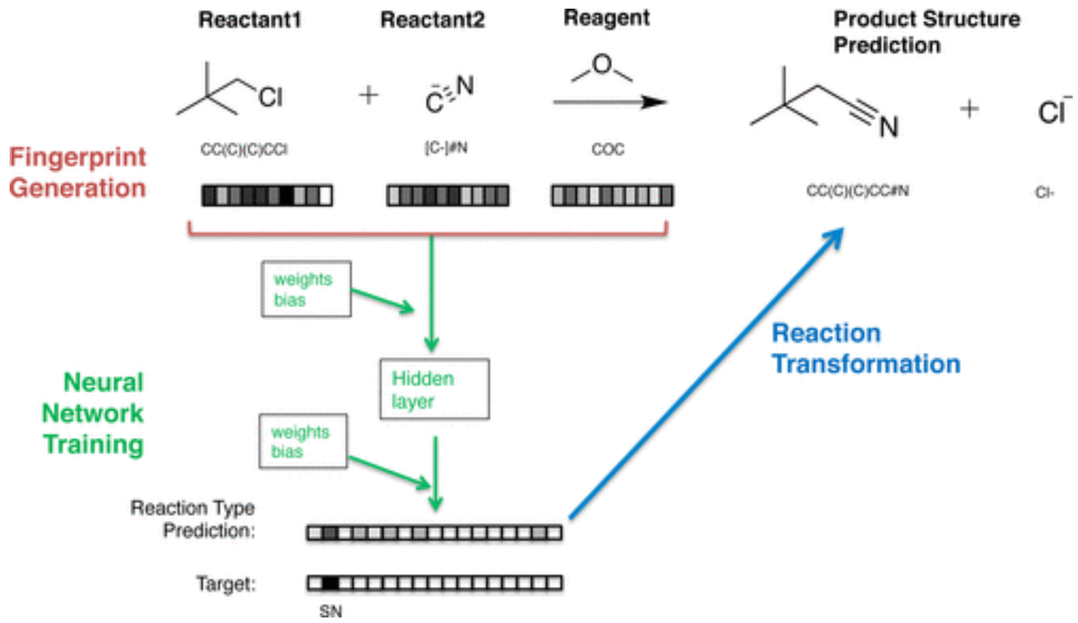
Jennifer N. Wei[†], David Duvenaud[‡], and Alán Aspuru-Guzik[*†]

View Author Information

Cite this: *ACS Cent. Sci.* 2016, 2, 10, 725–732
Publication Date: October 14, 2016
https://doi.org/10.1021/acscentsci.6b00219

Template-based approaches with automated extraction (Segler & Waller)

- Similar to the Wei approach
- But automated extraction of 9k templates from Reaxys
- One ECFP fingerprint to encode reactants



Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction

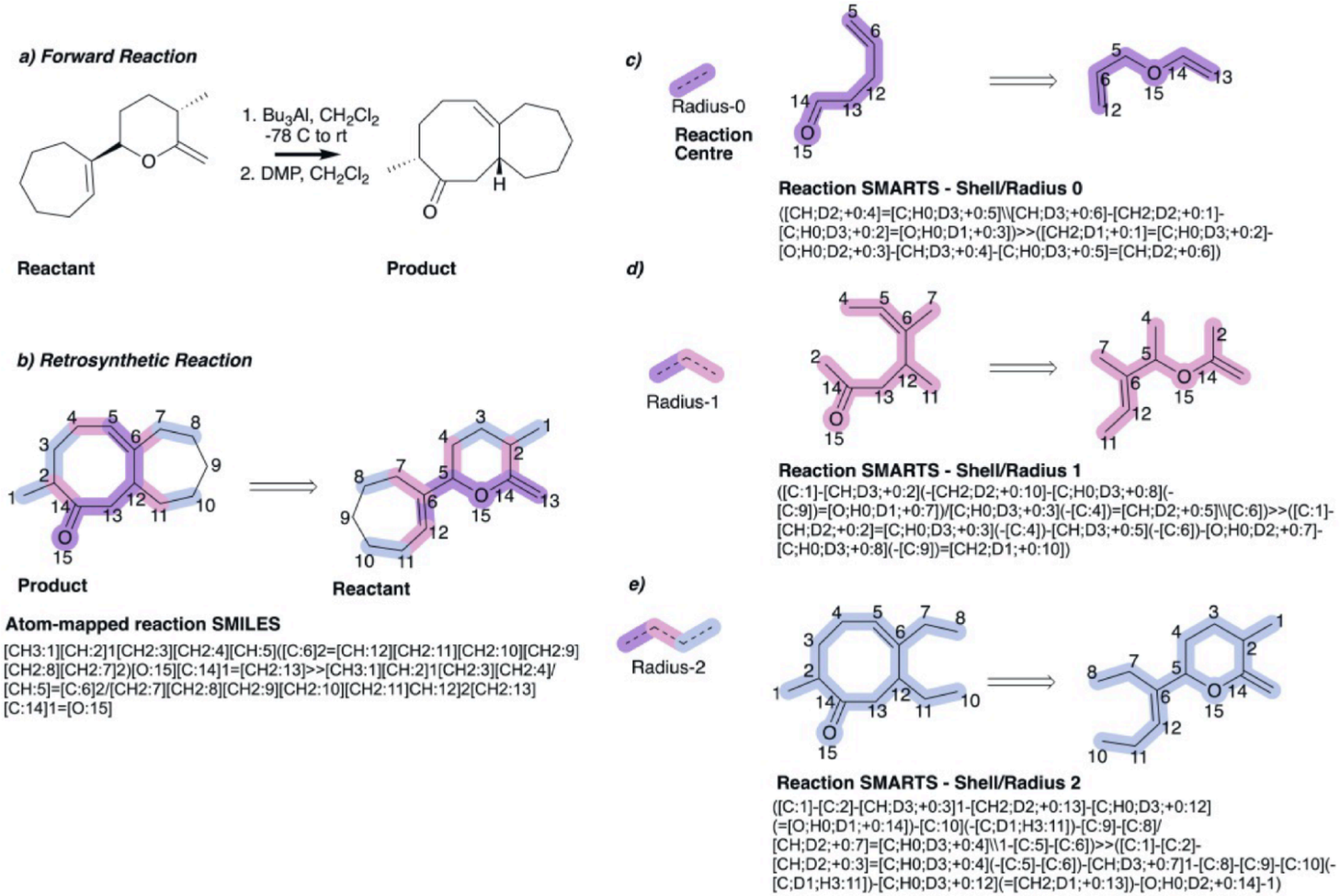
Marwin H. S. Segler, Prof. Mark P. Waller

First published: 30 January 2017 | https://doi.org/10.1002/chem.201605499 | Citations: 244

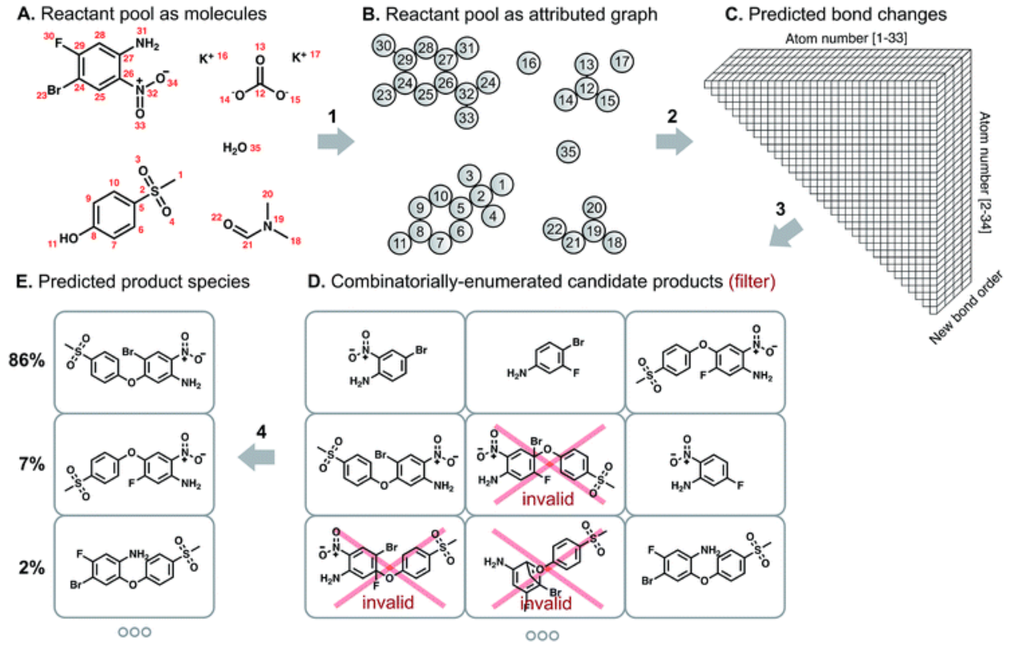
Automatic Extraction of Reaction Templates for Synthesis Prediction

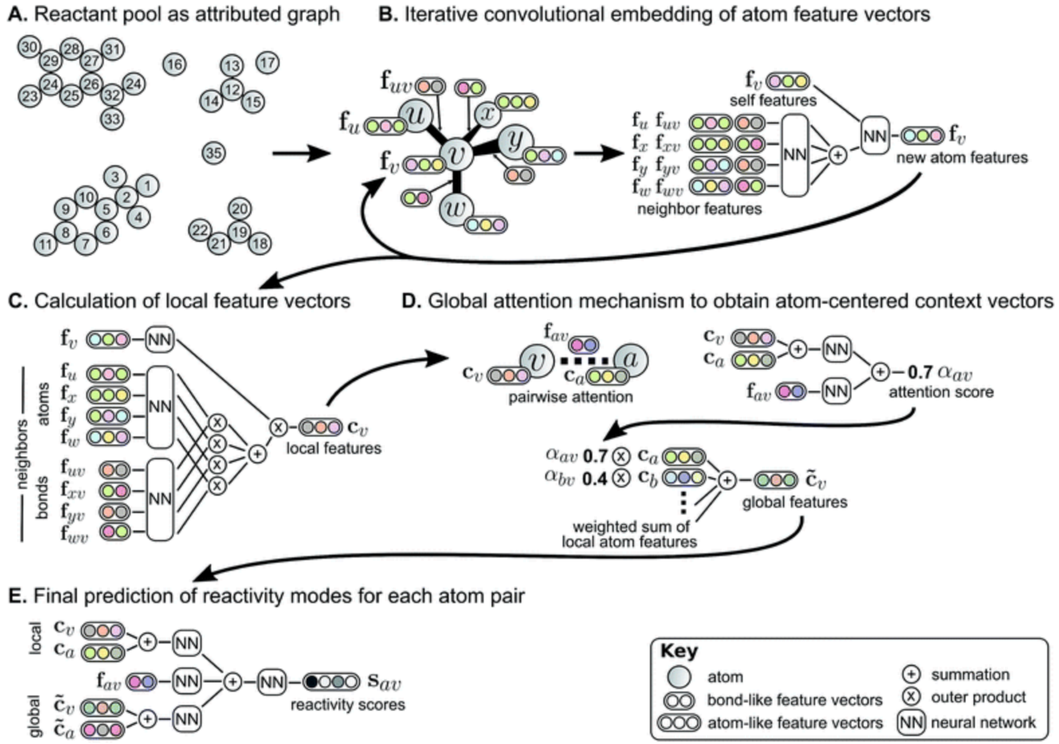
Amol Thakkar§* and Jean-Louis Reymond*

§SCS-Metrohm award for best oral presentation in Computational Chemistry



Bond change prediction / graph edit-based methods





DOI: 10.1039/C8SC04228D (Edge Article) *Chem. Sci.*, 2019, **10**, 370-377

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]

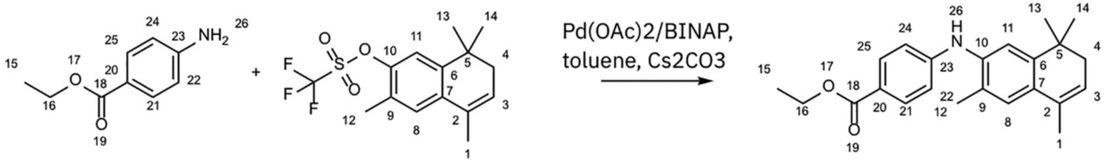
Limitations of atom-mapping dependent approaches

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

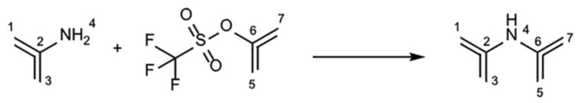
Atom-mapping (e.g. RXNMapper)

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

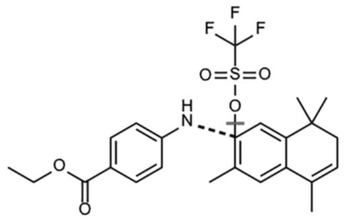


CC(=O)[O-].CC(=O)[O-].Cc1ccccc1.O=C([O-])[O-].O=S(=O)(O[c:11]1[cH:12][c:13]2[c:14]([cH:15][c:16]1[CH3:17])[C:18]([CH3:19])=[CH:20]
[CH2:21][C:22]2([CH3:23])[CH3:24])C(F)(F)F.[CH3:1][CH2:2][O:3][C:4](=[O:5])[c:6]1[cH:7][cH:8][c:9]([NH2:10])[cH:25][cH:26]1.[Cs+].[Cs+].
[Pd+2].c1ccc(P(c2ccccc2)c2ccc3ccccc3c2-c2c(P(c3ccccc3)c3ccccc3)ccc3ccccc23)cc1>>[CH3:1][CH2:2][O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



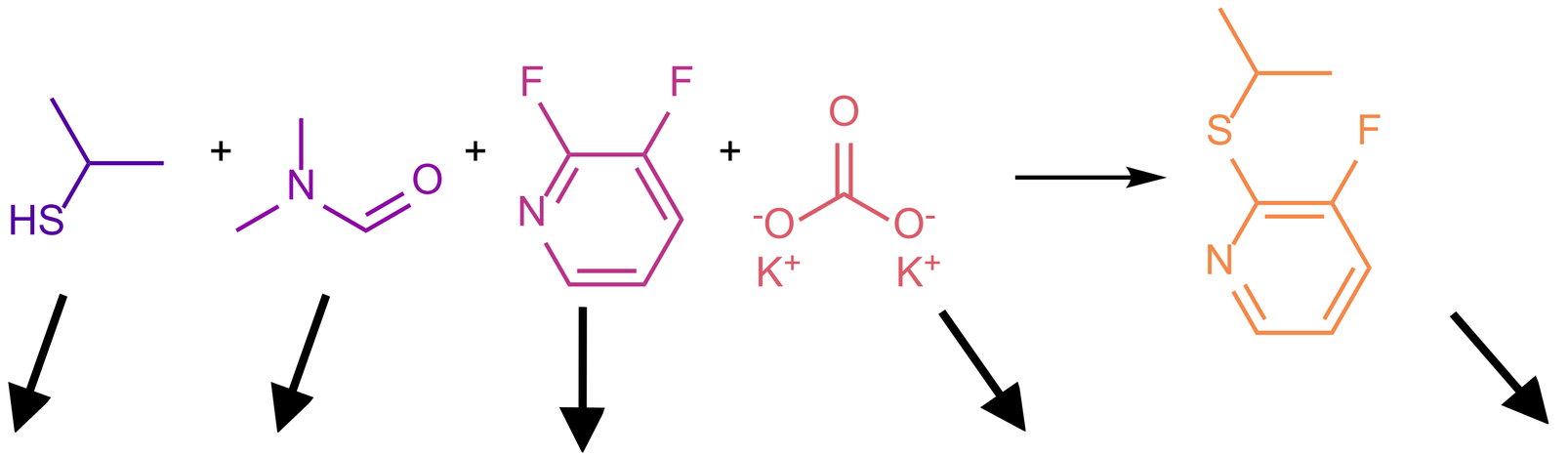
Atom-mapping dependent approaches are only as good as this step.

Wrong atom-mapping
- Wrong graph-edits
- Wrong templates

SMILES-2-SMILES approaches — How to overcome atom-mapping dependence

Atoms as *letters*, molecules as *words*

precursors >> products



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

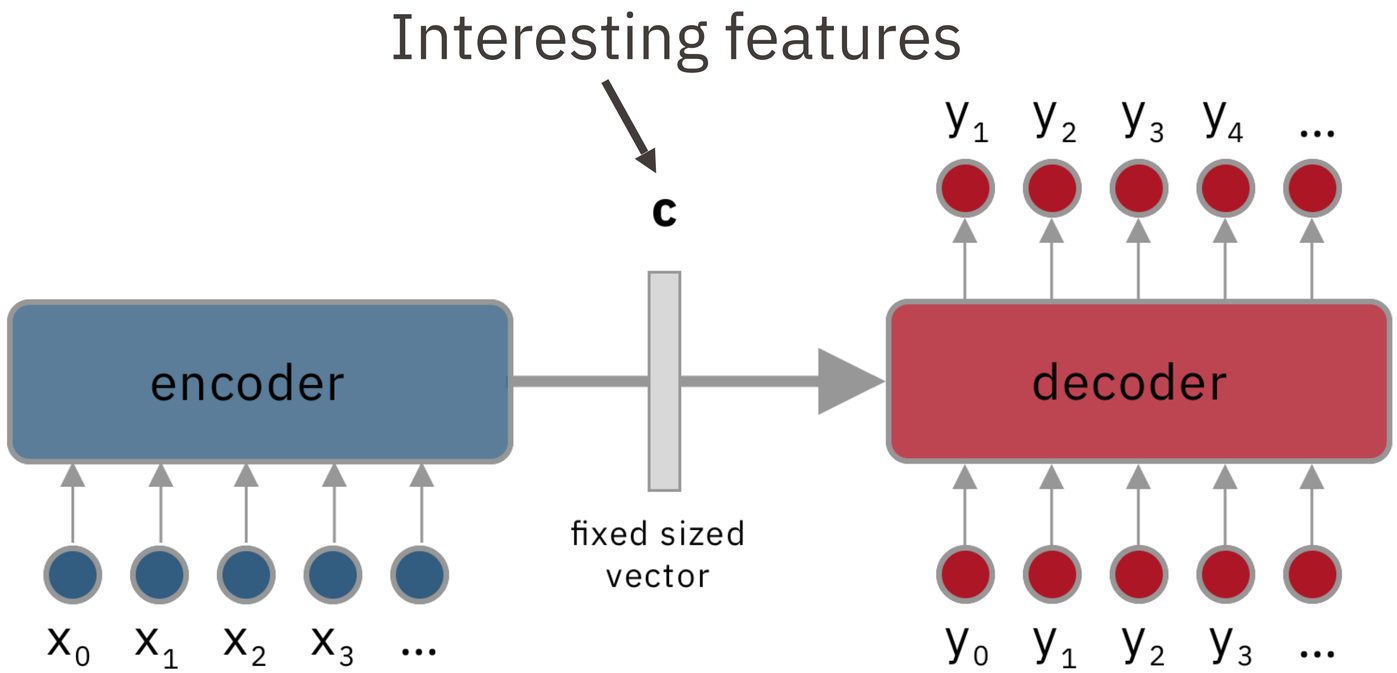
→ Borrow methods developed for human languages

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

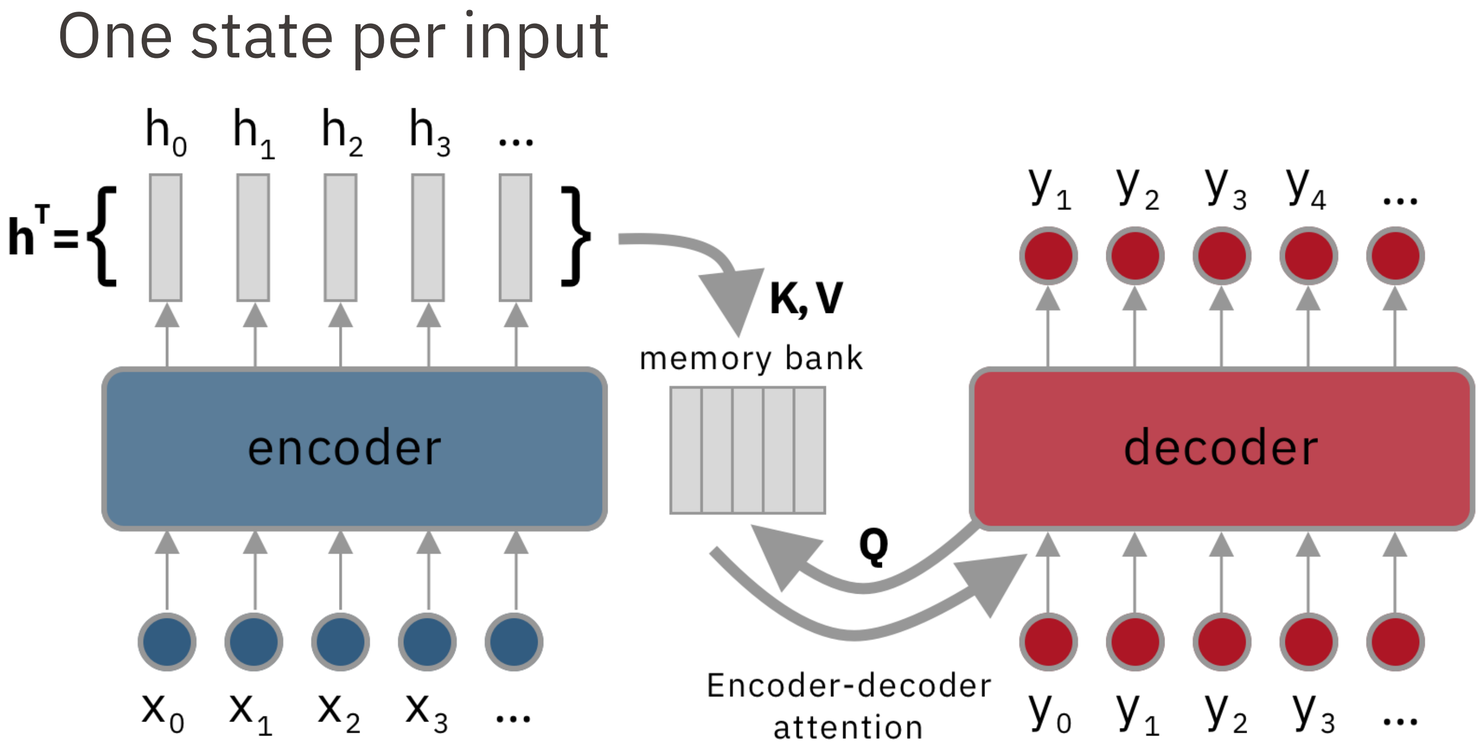
Br c 1 c n c c (Br) c 1 . C N (C) C = O . C [O-] . [Na+]

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.

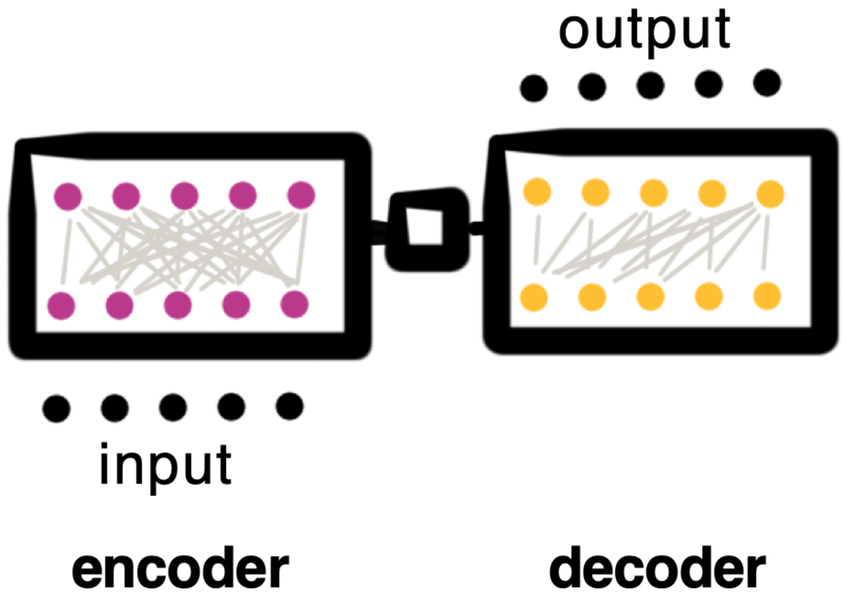
Sequence-2-sequence models *with attention*



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

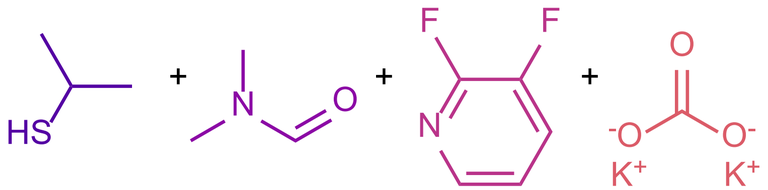


- Stacks of attention layers
- Multi-head attention

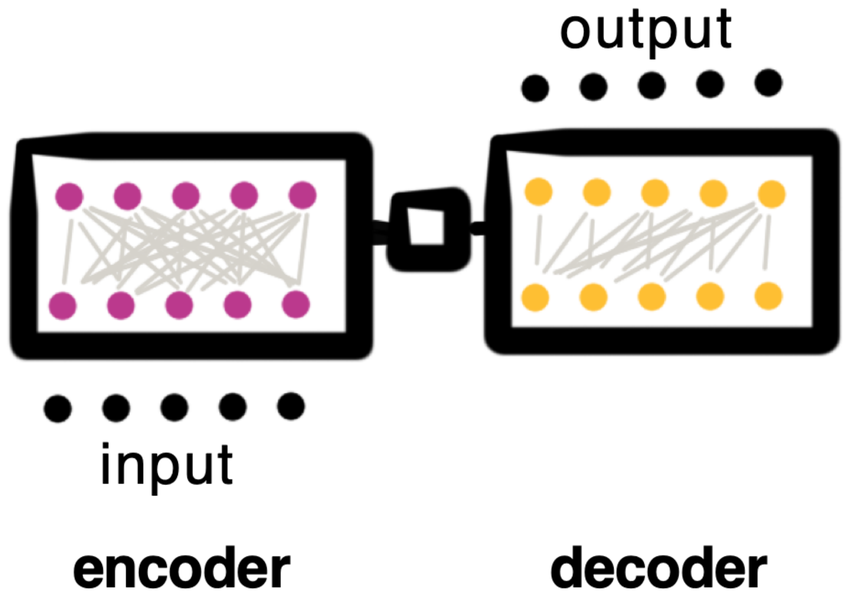
Vaswani et al., Neural Machine Translation by Jointly Learning to Align and Translate. NeurIPS, 2017

Molecular Transformer

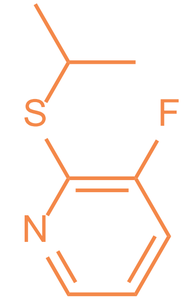
precursors



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

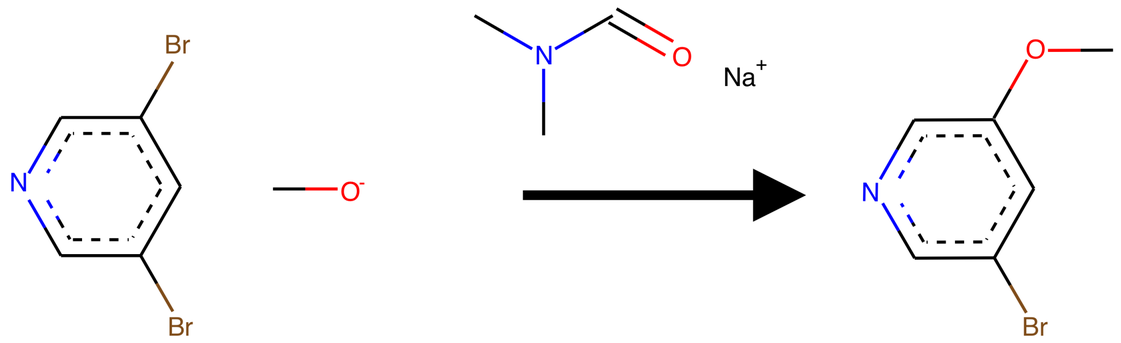
IBM **Research**

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

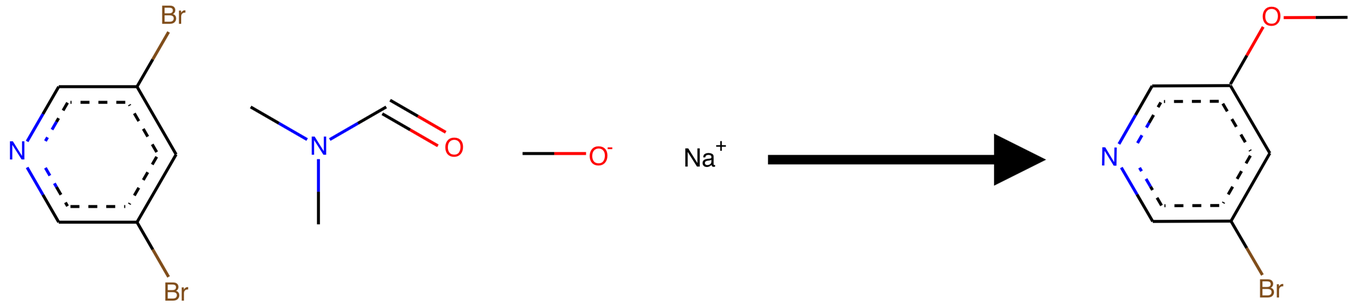
Separated vs mixed setting

Separated



reactants reagents products

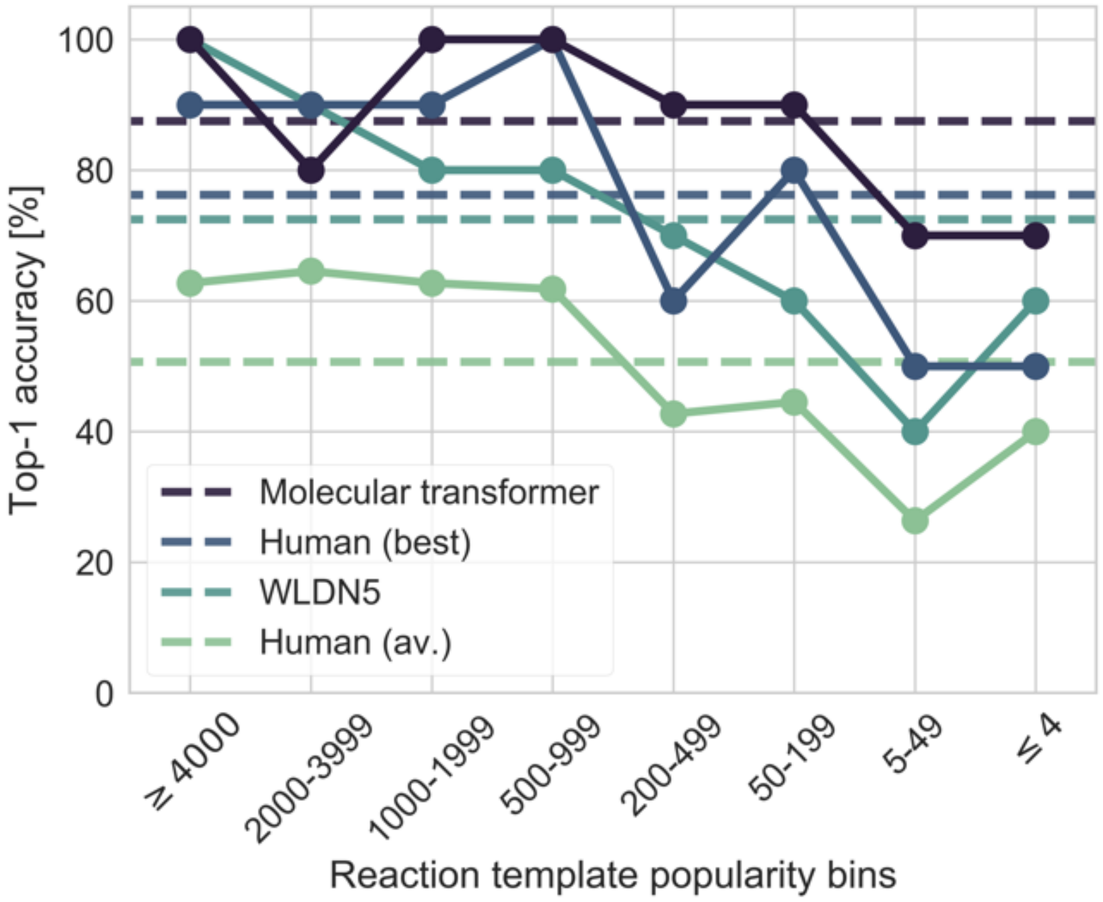
Mixed



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

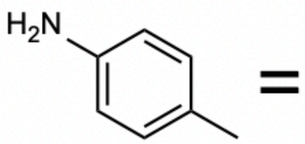
Zhengkai Tu[1,2] and Connor W. Coley[1,3]

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

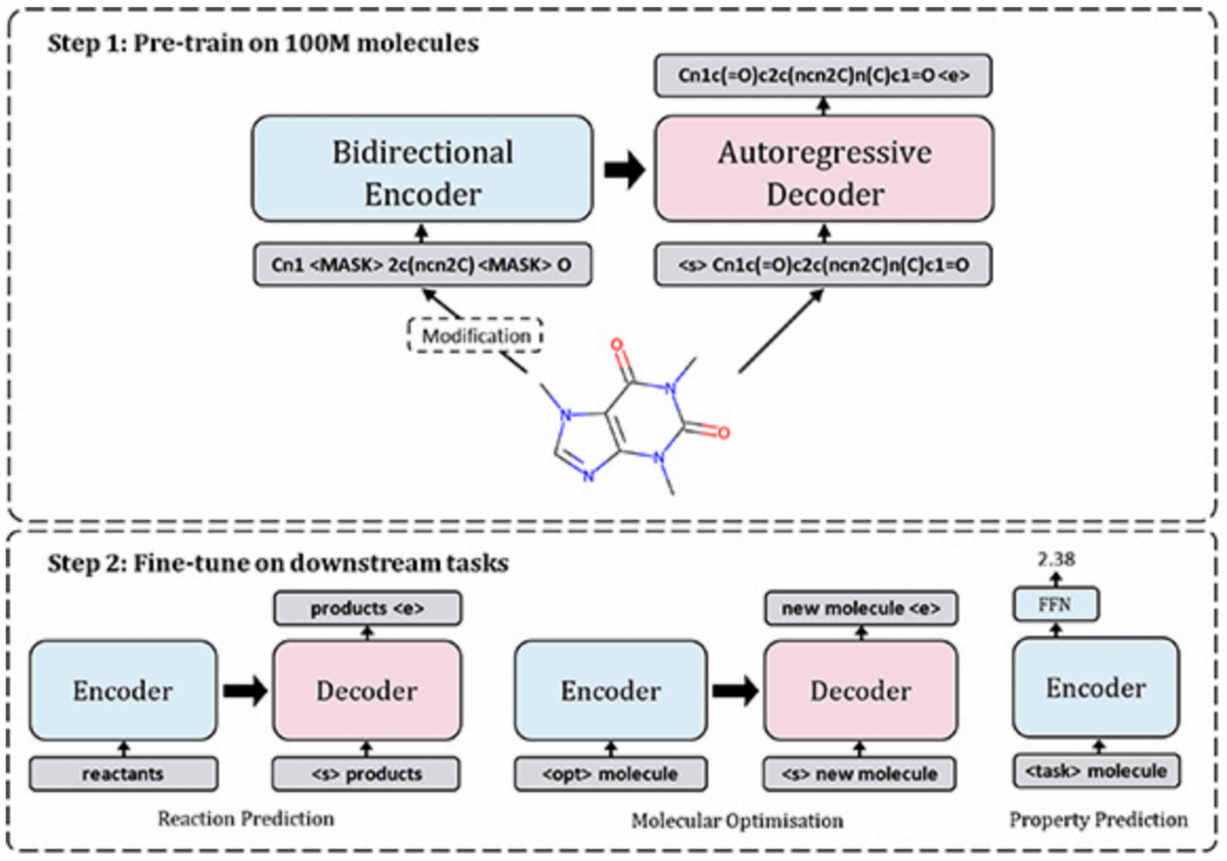
State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis

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

Data augmentation strategies to improve reaction yield predictions and estimate uncertainty

Philippe Schwaller[1,2] phs@zurich.ibm.com
Alain C. Vaucher[1] ava@zurich.ibm.com
Teodoro Laino[1] teo@zurich.ibm.com
Jean-Louis Reymond[2] jean-louis.reymond@dcb.unibe.ch

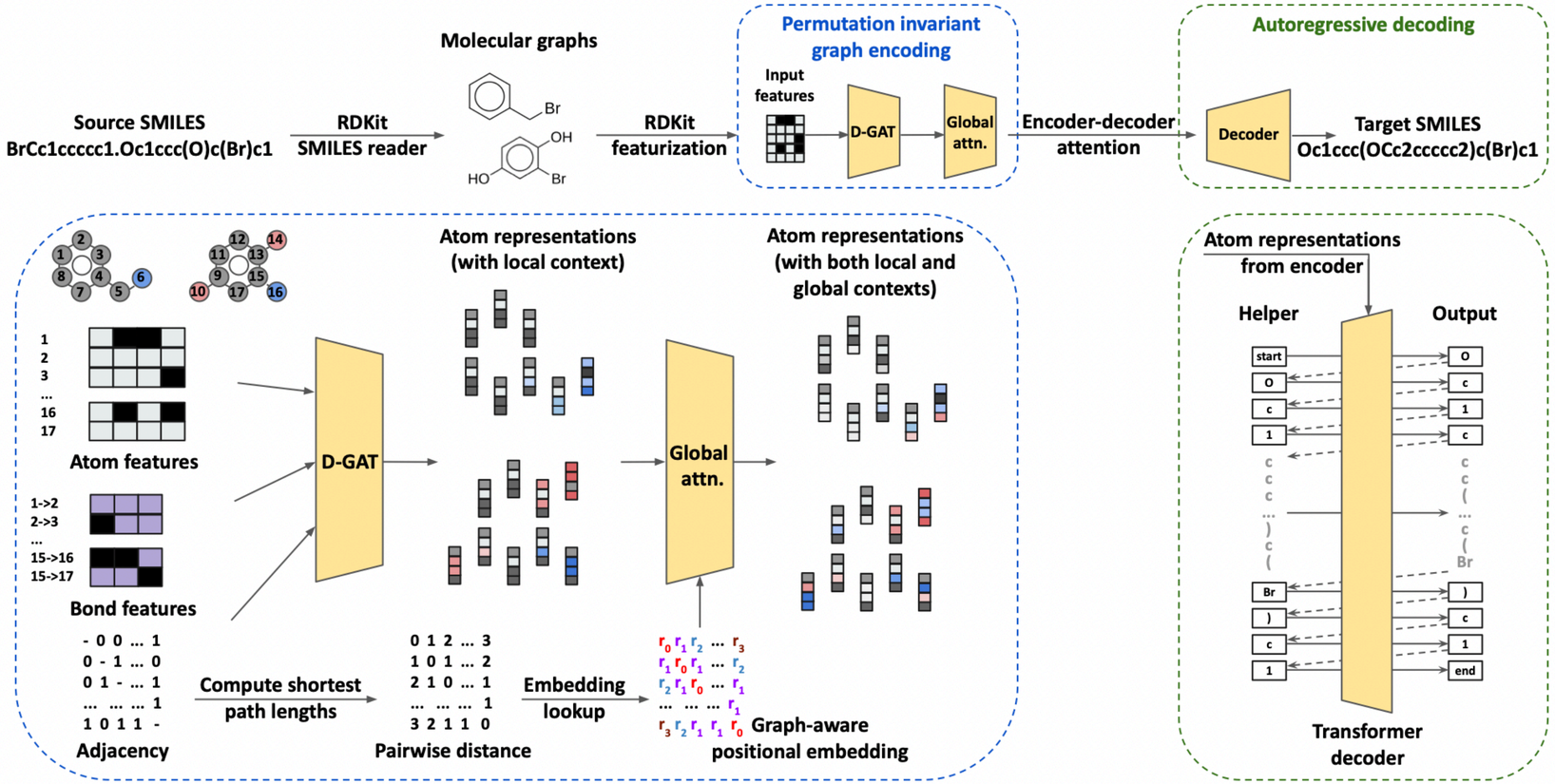
Large-scale pretraining



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

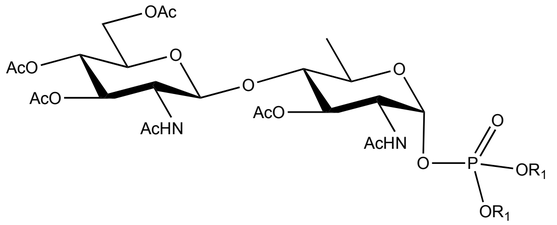


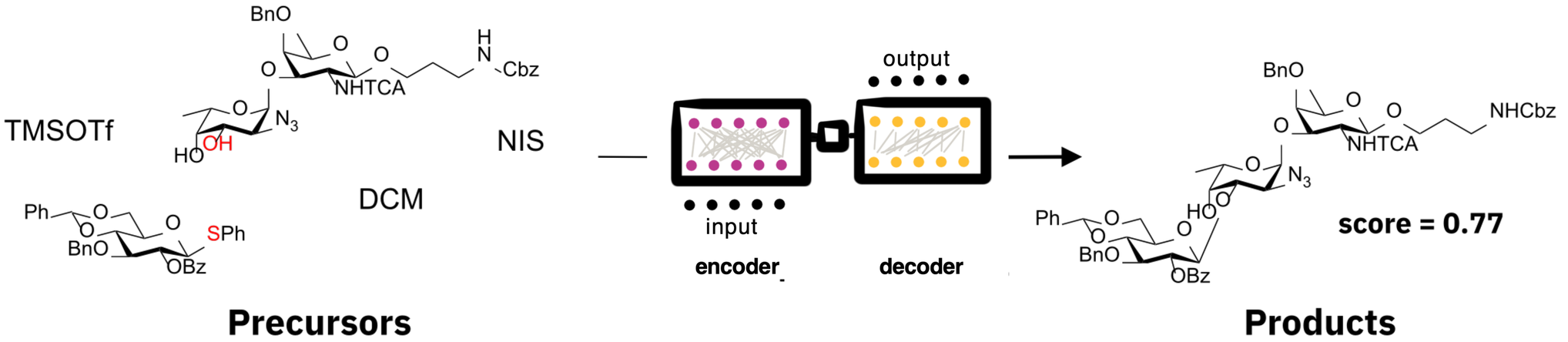
PERMUTATION INVARIANT GRAPH-TO-SEQUENCE MODEL FOR TEMPLATE-FREE RETROSYNTHESIS AND REACTION PREDICTION

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

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





Precursors **Products**

Transfer learning is applicable to any reaction subspace of interest!

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

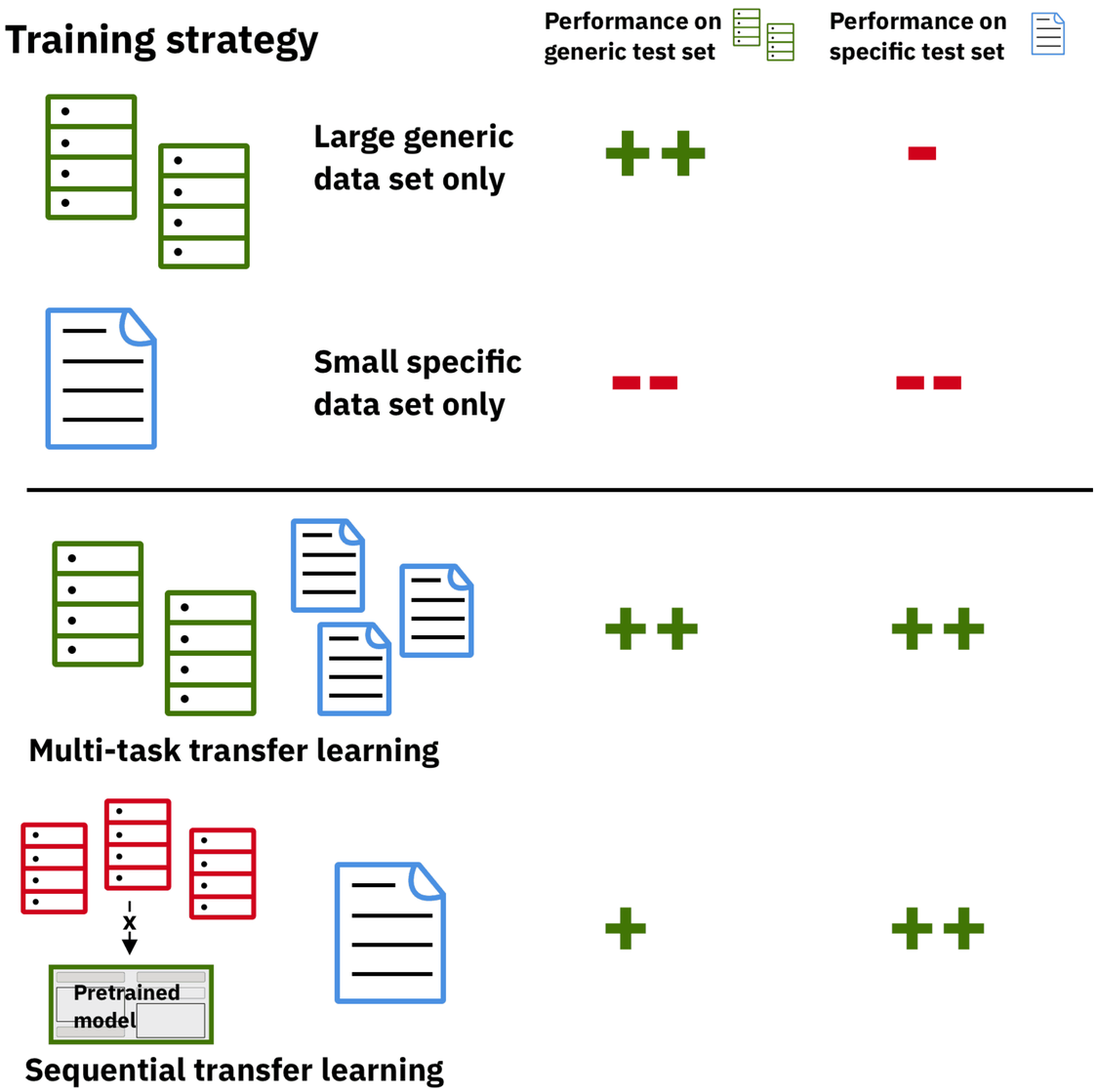
Pesciullesi*, **Schwaller* et al.,** Nature Communications, 2020

IBM **Research**

What is *transfer learning* ?

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

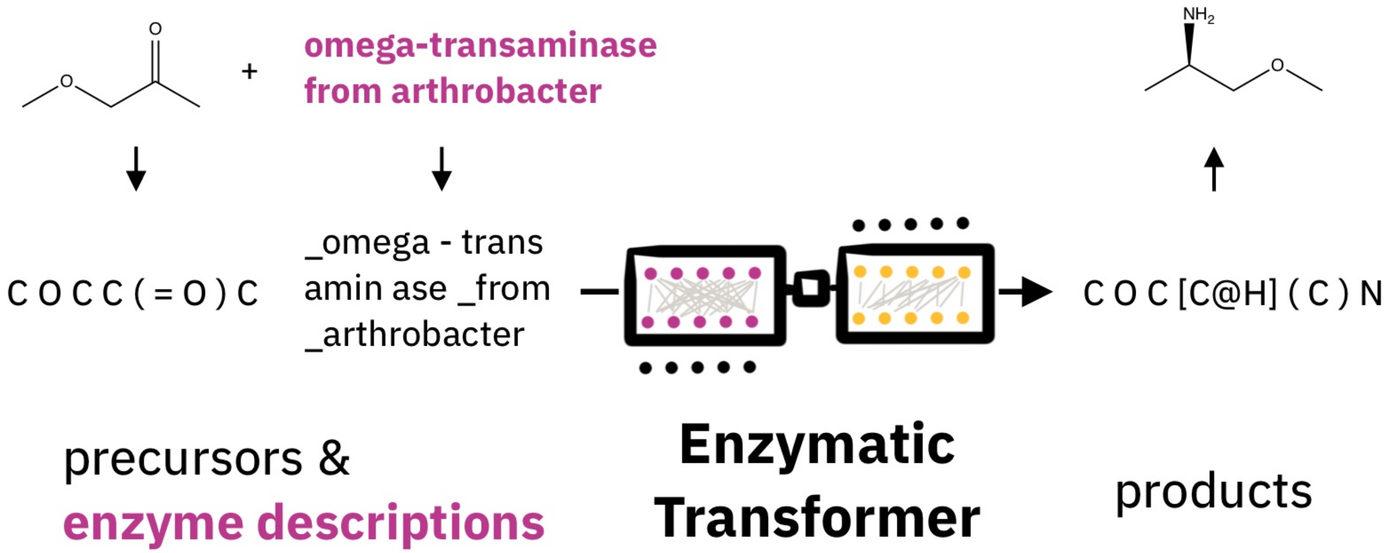
Carbohydrate Transformer ←



- Transfer learning applicable to any reaction subspace of interest!

Molecular Transformer for *enzymatic reactions*

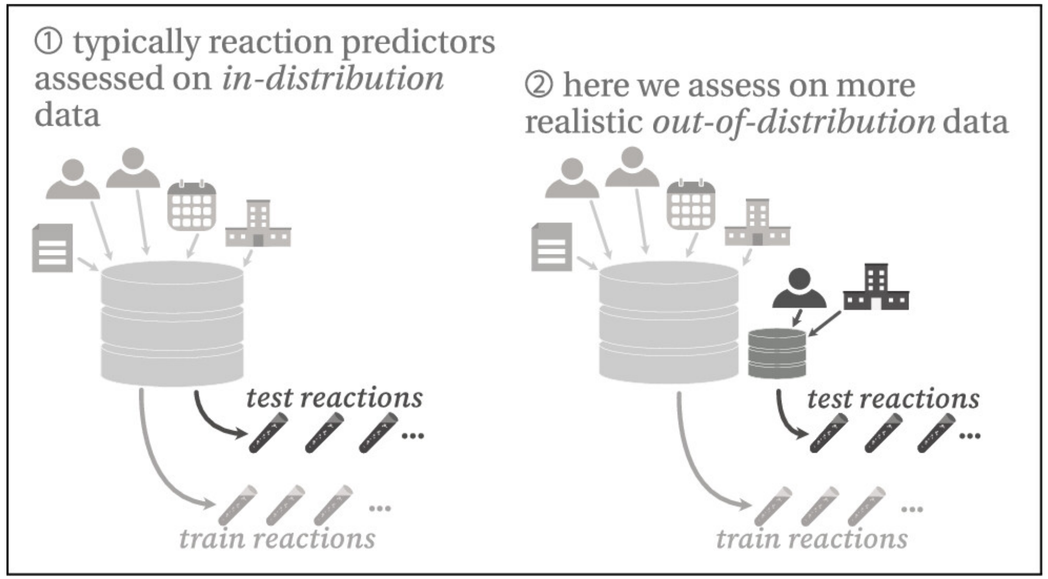
How to represent the enzymes?

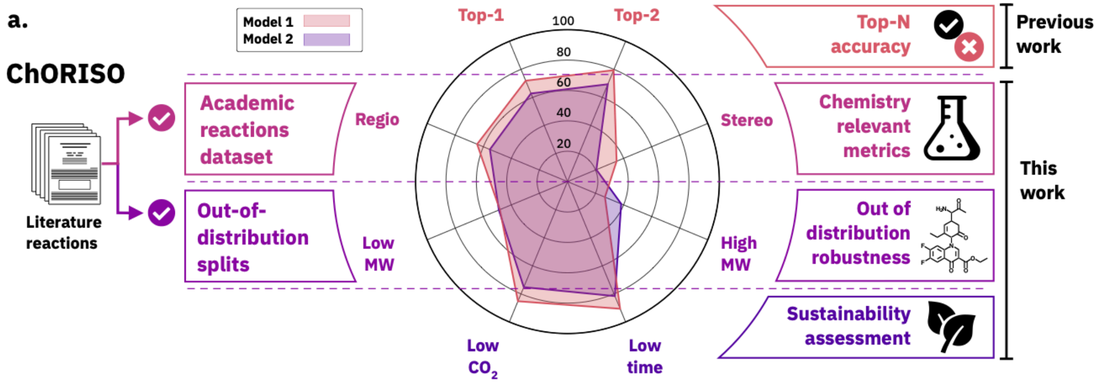


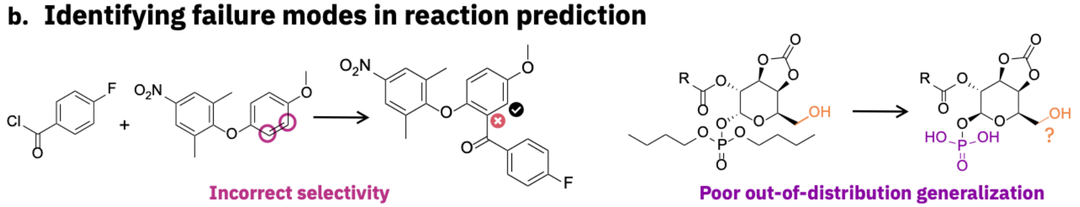
David Kreutter
Reymond group
UNIVERSITÄT BERN

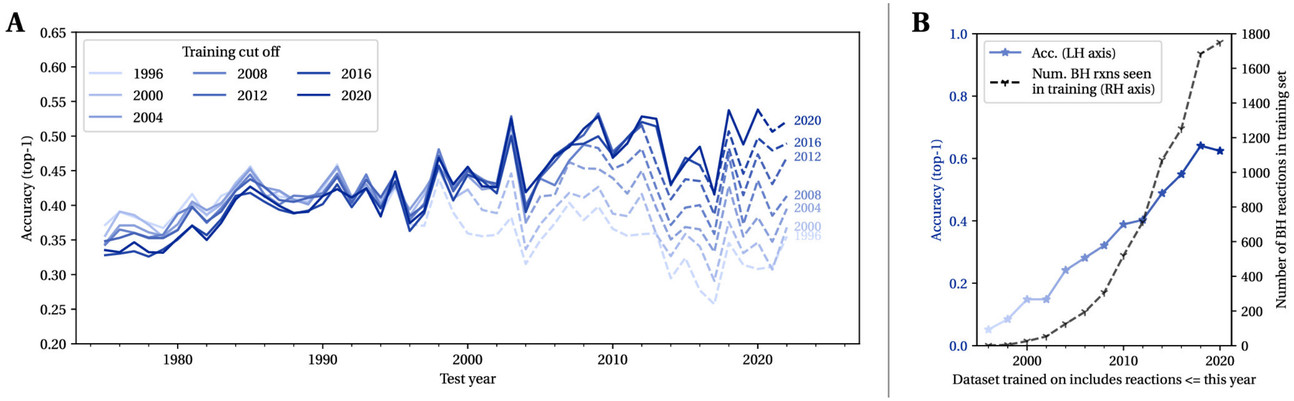
Kreutter et al., Predicting enzymatic reactions with a molecular transformer. Chem. Sci., 2021

Harder benchmarking needed









ARTICLE | March 12, 2025

Challenging Reaction Prediction Models to Generalize to Novel Chemistry

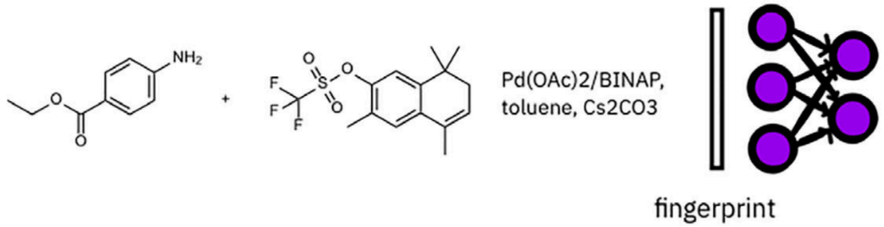
John Bradshaw, Anji Zhang, Babak Mahjour, David E. Graff, Marwin H. S. Segler, and Connor W. Coley*

[Submitted on 14 Dec 2023]

Holistic chemical evaluation reveals pitfalls in reaction prediction models

Victor Sabanza Gil, Andres M. Bran, Malte Franke, Remi Schlama, Jeremy S. Luterbacher, Philippe Schwaller

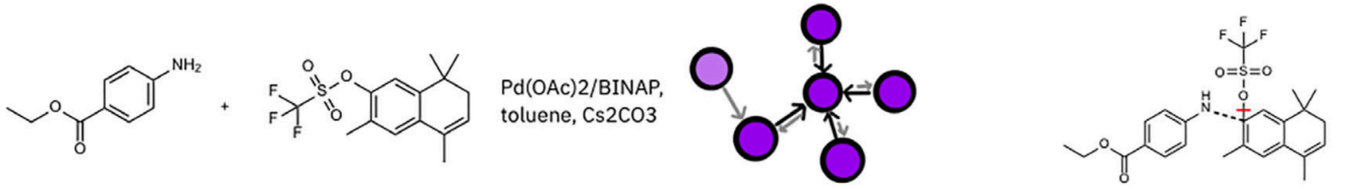
Template-based approaches (atom-mapping dependent)



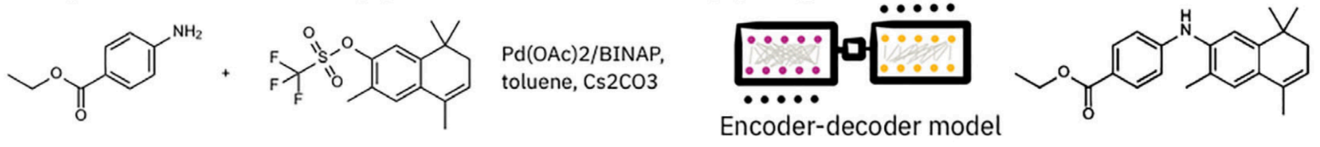
Template classification
- template 772
- template 123
- template 921
...

Product ranking (after template application)
- product 772
- product 521
- product 921
...

Graph edit-based approach (atom-mapping dependent)



Sequence-based approach (atom-mapping independent)



C C (= 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

→

C C O C (= O) c 1 c c c (N c 2 c c 3 c (c c 2 C) C (C) = C C C 3 (C) C) c c 1

atom-wise product prediction

Data is key!