08.05.2023

Performing research with publicly-available data ORD I Lausanne

Philippe Schwaller (he/him/his)

Laboratory of Artificial Chemical Intelligence (LIAC) – SB-ISIC

@SchwallerGroup

philippe.schwaller@epfl.ch

NCCR Catalysis

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

NCCR Catalysis

TT Assistant Professor in Digital Chemistry since Feb 2022

EPFL LIAC

Collaboration with synthetic chemists

Prix Schläfli since 1866

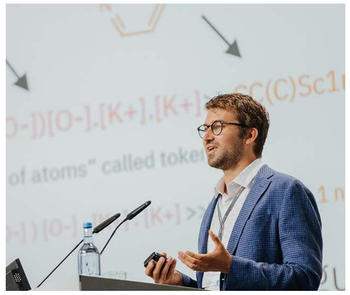
u b

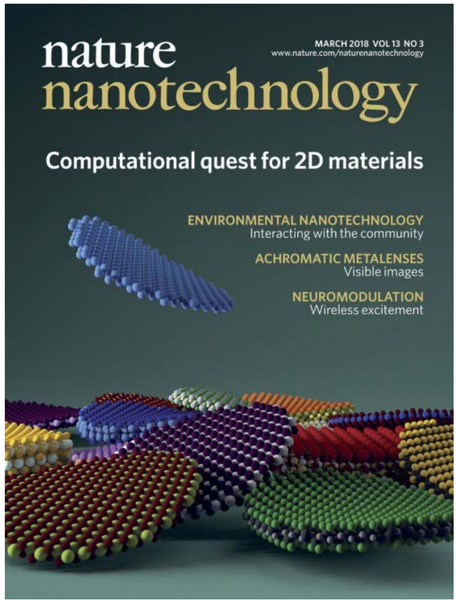
b UNIVERSITÄT BERN

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

EPFL

MANCHESTER 1824
The University of Manchester





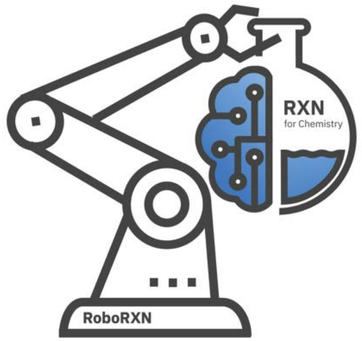
Materials Science & Engineering

Virtual screening & simulation workflows
Prof Nicola Marzari

BSc (‘14)/MSc (‘16)

MPhil in Physics (’19)
Dr Alpha Lee

UNIVERSITY OF CAMBRIDGE



- MaX prize for frontier HPC applications (‘17)
- PRACE HPC Excellence Award (‘22)

Empa
Materials Science and Technology

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

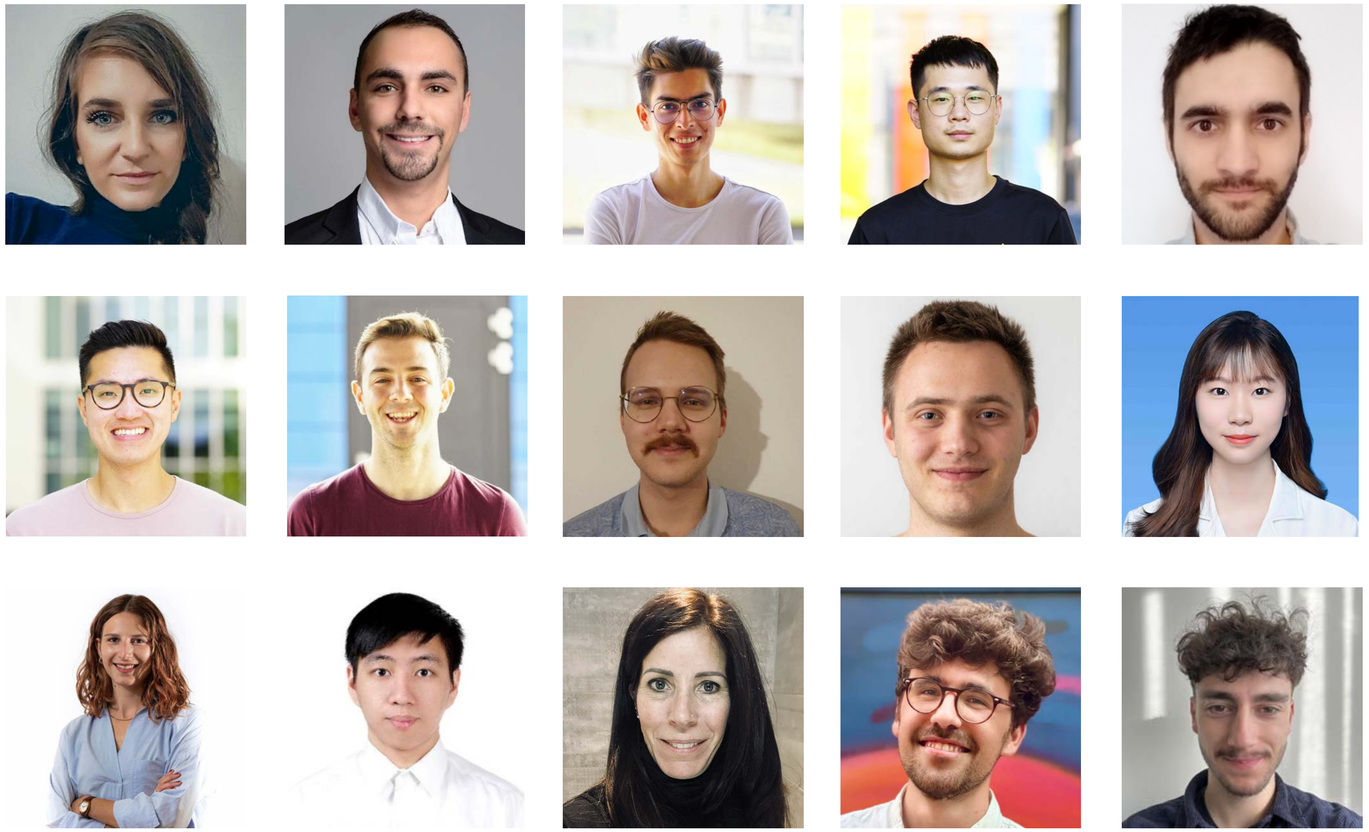
Machine learning for chemical synthesis
Intern/PhD/Postdoc
Dr Teodoro Laino

IBM **Research** Europe

LIAC Team

https://schwallergroup.github.io

PhD students:
Bojana Ranković
Oliver Schilter
Andres CM Bran
Junwu Chen
Jeff Guo
Victor Sabanza Gil
Paulo Neves
Rebecca Neeser



Stéphane D'Ascoli

Cheng-Hua Huang
Artur Stefaniuk
Malte Franke
Xu Huang
Lucien Brey

Annick Delmonaco

NCCR Catalysis

IBM Research

janssen

VANTAI

11 nationalities — one team!

AI4Science at EPFL

Working group:

Goal:

Identify questions that re-appear when applying ML to accelerate various areas of science. Join forces to solve them.

Serendipity is our friend.

Computer Science

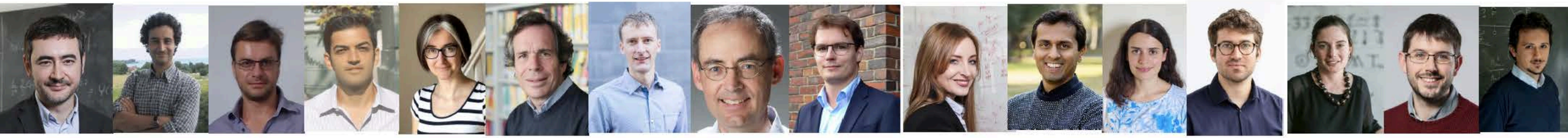
Physics

Life Science

Chemistry

Materials

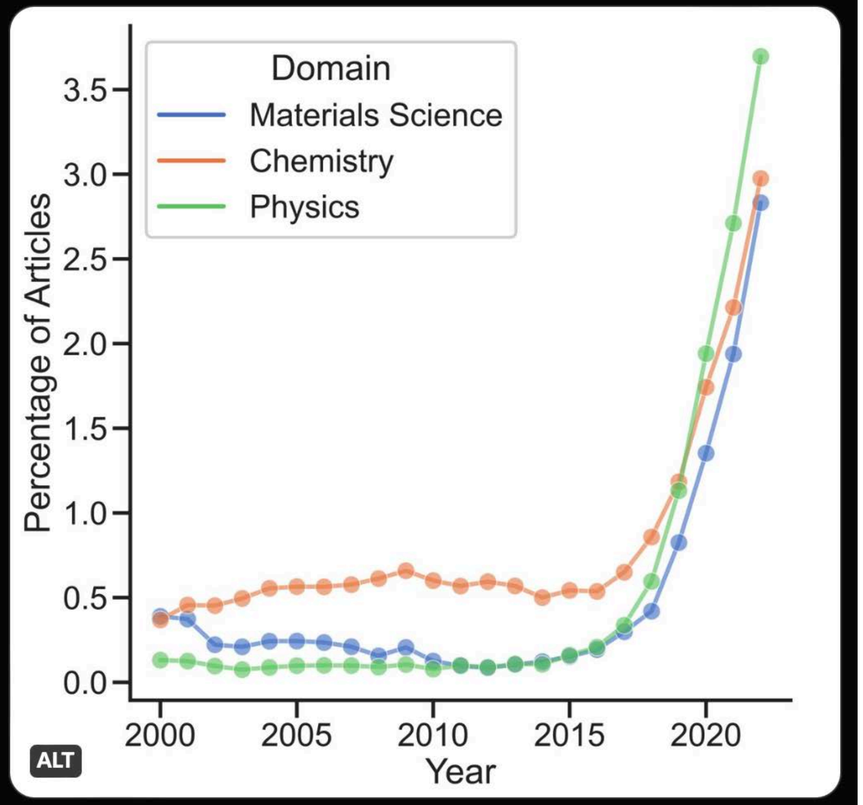
EE



Ben Blaiszik @BenBlaiszik · 45m

Does it feel like you are seeing more impactful #ML and #AI for science publications? This probably explains it. 🚀

📈 We've seen strong continued growth in #AI and #ML for science across a broad set of domains including materials science, chemistry, physics and more.



1 2 9 309

Ben Blaiszik @BenBlaiszik · 45m

Replying to @BenBlaiszik

Computing the YoY growth rates and CAGR (details available in the repo) shows the following.

Percentage Gains in # of matching articles for 2022:

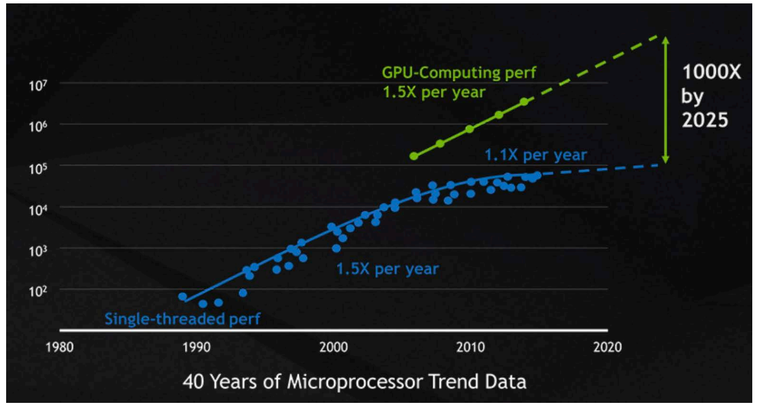
- Materials Science: 39% more articles than 2021
- Chemistry: 27% more articles than 2021
- Physics: 29% more articles than 2021

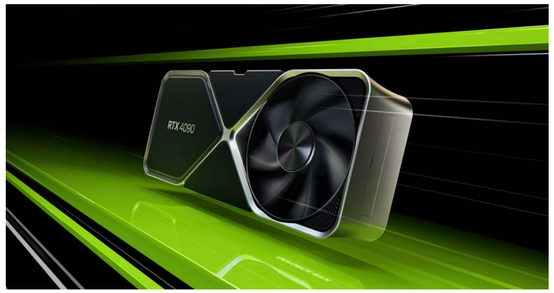
Domain	year	count	CAGR-1 (%)
Materials Science	2022	6180	39.1
Chemistry	2022	8842	27.4
Physics	2022	6829	29.3

ALT

1 58

Why now?





Better performing hardware (GPUs)

TensorFlow

PyTorch

Keras

PyTorch Lightning

Accessible ML frameworks

IMAGENET

RCSB PDB PROTEIN DATA BANK

201,979 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

The Pile *An 800GB Dataset of Diverse Text for Language Modeling*

Open data & code

What is public data?

- openly accessible
- free of charge
- and can be used, reused, and redistributed by anyone
- often with few or no restrictions
- https://creativecommons.org/choose/

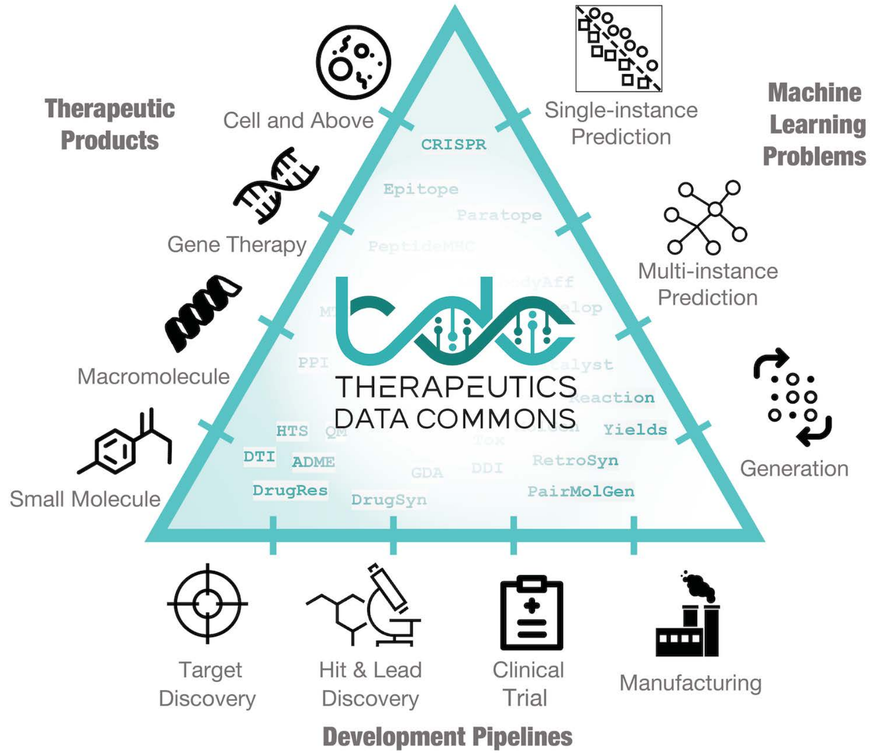
Open research data's significance

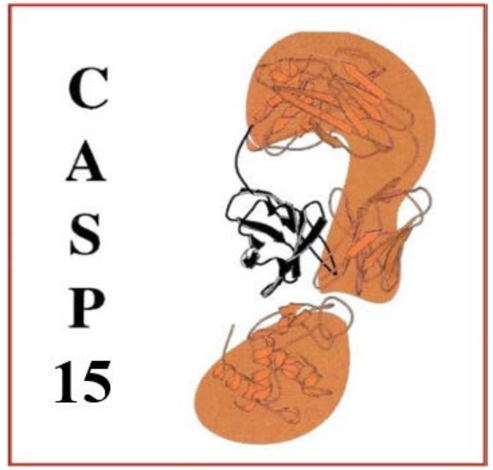
- promotes transparency and trust
- accountability in research
- fosters collaboration among researchers
- enabling researchers to build upon each other's work
- accelerating research & scientific advancements

Standardized benchmarks

- compare the performance of different methods, algorithms, or models
- curated datasets and evaluation metrics
- assess strengths and weaknesses of approaches

- Important to be as close as possible to reality



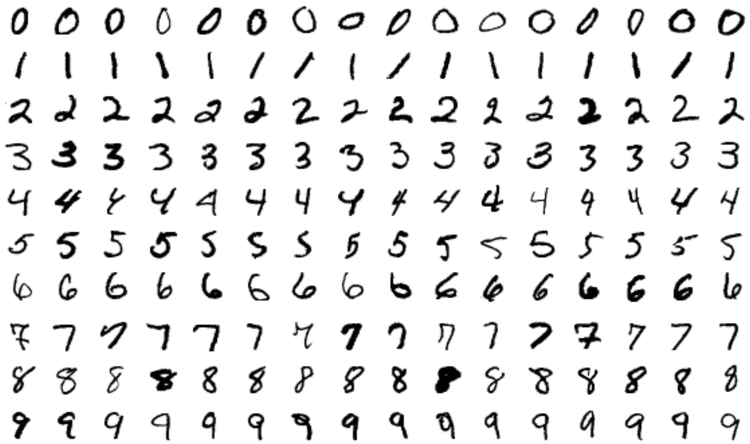


QUANTUM-MACHINE.ORG

Home | Datasets | Publications | Software

QM9

Open data that initiated deep learning wave



THE MNIST DATABASE

of handwritten digits

Yann LeCun, Courant Institute, NYU
Corinna Cortes, Google Labs, New York
Christopher J.C. Burges, Microsoft Research, Redmond

1994, 60k images
6k citations

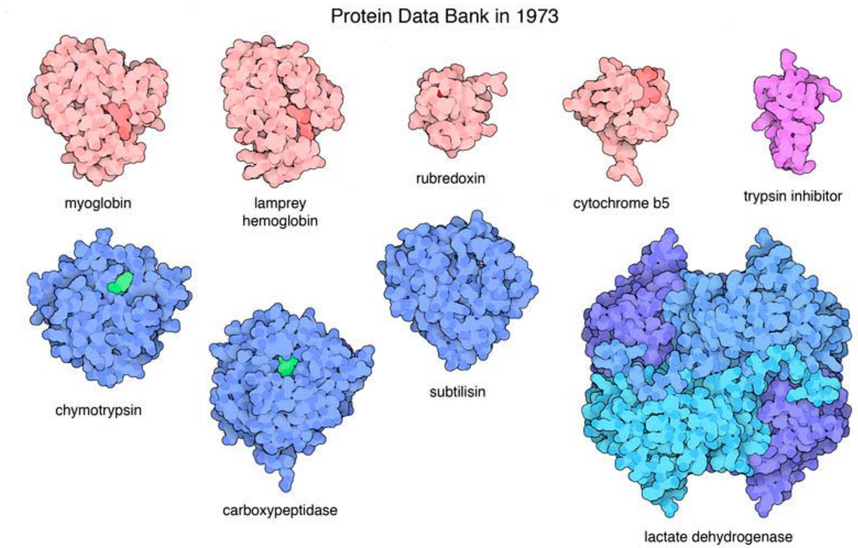
ImageNet: A Large-Scale Hierarchical Image Database

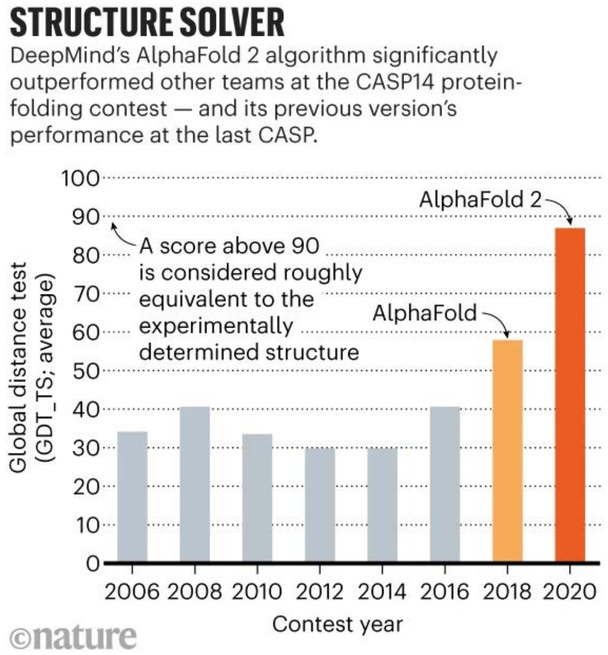
Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei
Dept. of Computer Science, Princeton University, USA
{jiadeng, wdong, rsocher, jial, li, feifeili}@cs.princeton.edu

2009, now 14M images
53k citations

Deepmind's AlphaFold

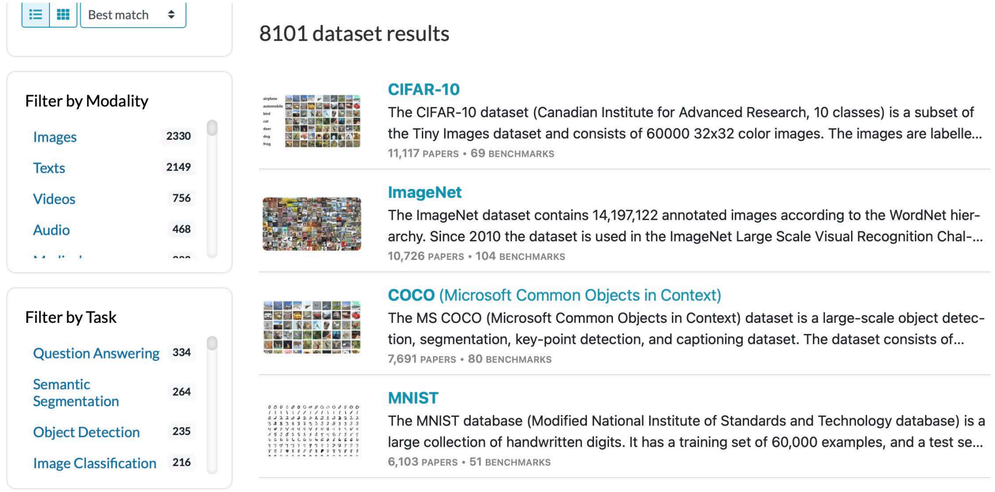
- Won CASP protein-solving contest
- Made possible by the PDB database (established in 1971)

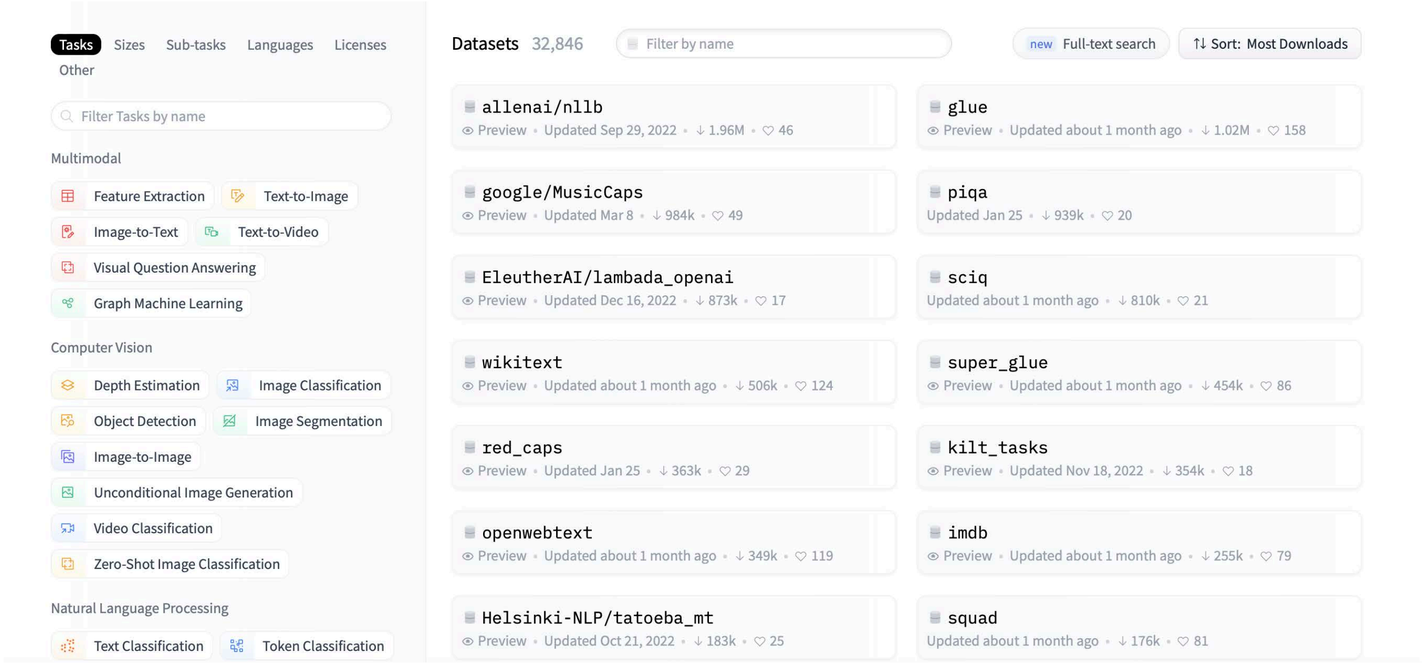




Open datasets

- Paper with Code
  https://paperswithcode.com/datasets
- HuggingFace datasets
  https://huggingface.co/datasets





Without open data I would not stand in front of you today.

Milky-way above the Swiss Alps, Schafarnisch, 2022

What molecule to make?

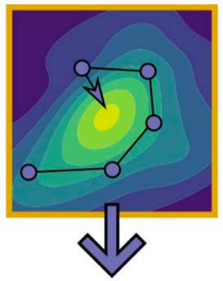
Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

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

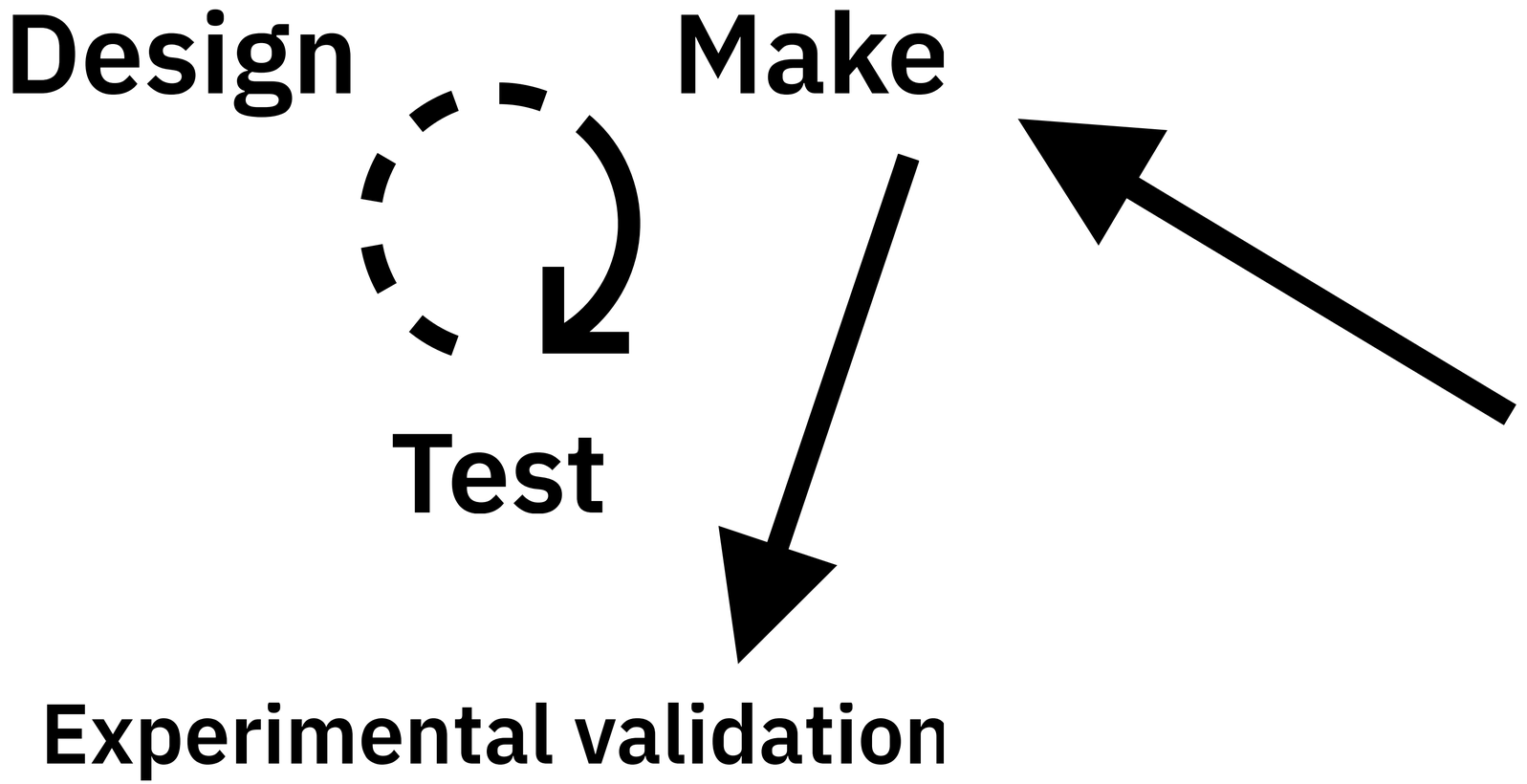


Inverse:
Optimization, evolutionary strategies, generative models

How to make it?

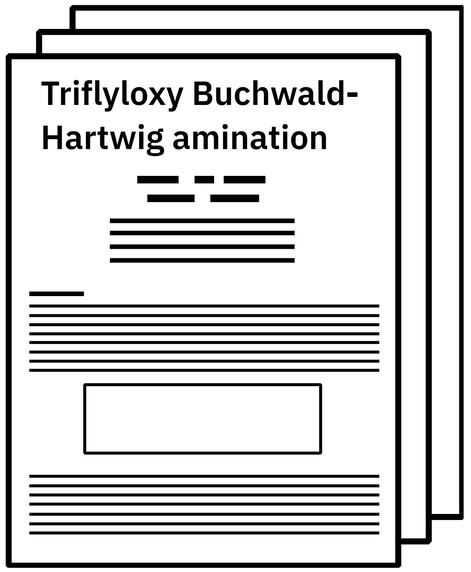
Reaction prediction

Synthesis planning



Data sources for ML?

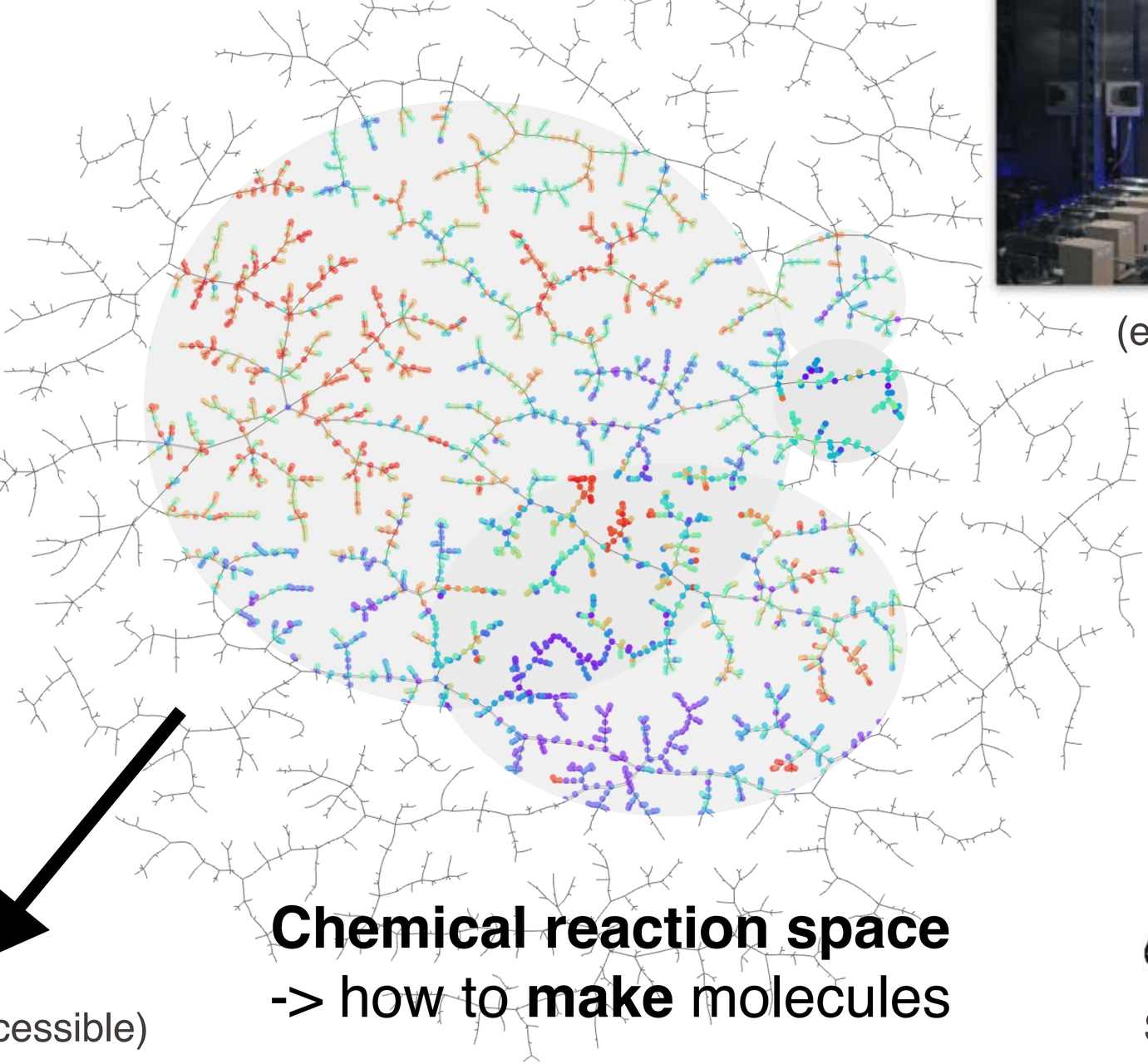
Literature (broad, e.g, Reaxys)



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.

Experiments ELN/HTE (narrow)

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



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

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

Simulations (narrow)

- Patents (broad, accessible)

Open Data for Chemical Reactions

Chemical reactions from US patents (1976-Sep2016)

Cite | Download all (1.39 GB) | Share | Embed | + Collect

Dataset posted on 2017-06-13, 18:49 authored by Daniel Lowe

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.

The reactions were extracted using an enhanced version of the reaction extraction code described in
https://www.repository.cam.ac.uk/handle/1810/244727
with LeadMine (https://www.nextmovesoftware.com/leadmine.html) used for chemical entity recognition.

General tips:
Duplicate reactions are frequent due to the same or highly similar text occurring in multiple patents, this is especially true

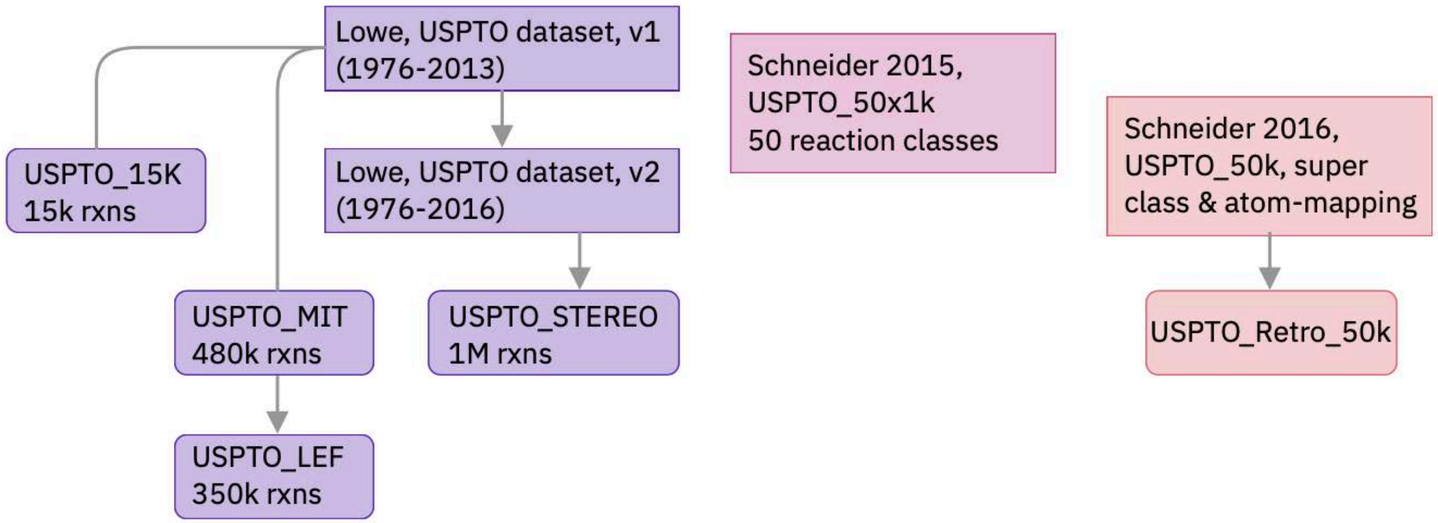
USAGE METRICS

37653	20303	19
views	downloads	citations

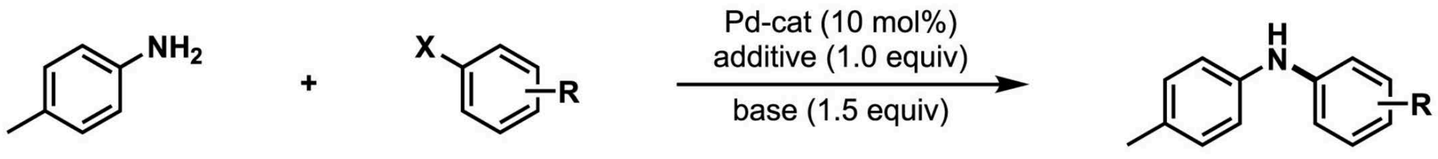
48

CATEGORIES

- Cheminformatics and quantitative structure-activity relationships
- Organic chemical synthesis



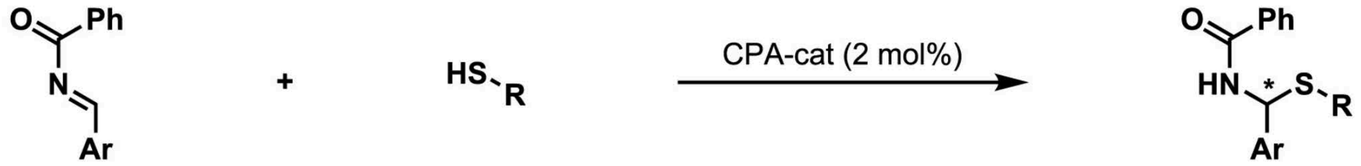
Open Data for Chemical Reactions



- 15 aryl halides
- 3 bases
- 4 Pd-catalysts
- 23 isoxazole additives

total reactions: 4140

C–N cross coupling reactions of 4-methylaniline with various aryl halides by Doyle and co-workers.

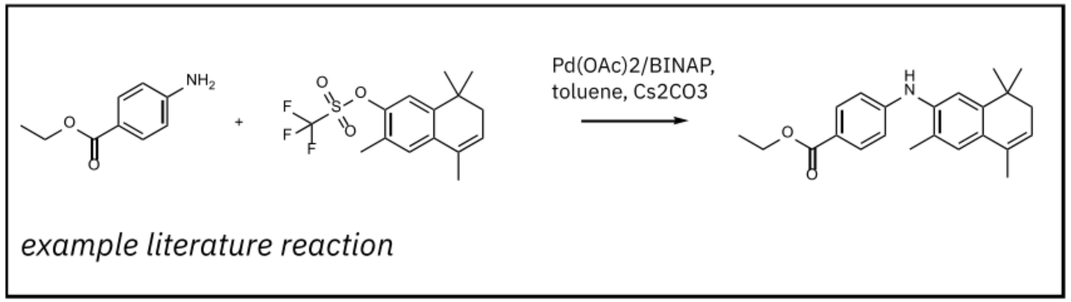


- 5 imines
- 5 thiols
- 43 CPA catalysts

total reactions: 1075

Asymmetric *N,S*-acetal formation using CPA catalysts by Denmark et al.[38]

- https://www.sciencedirect.com/science/article/pii/S2451929420300851



example literature reaction

Data for Chemical Reactions

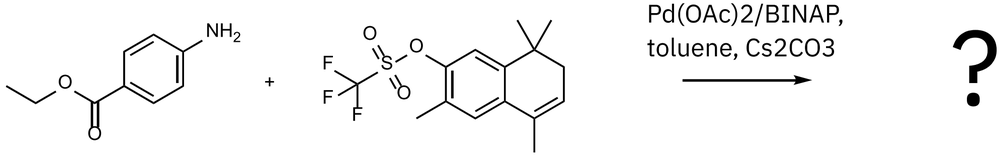
General

- USPTO: https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873
- Train/valid/test splits: https://github.com/coleygroup/Graph2SMILES/blob/main/scripts/download_raw_data.py
- 50k with classes and atom-mapping: https://pubs.acs.org/doi/full/10.1021/acs.jcim.6b00564
- 50k with 50 reaction classes: https://pubs.acs.org/doi/full/10.1021/ci5006614
- 1000 template classes: https://rxn4chemistry.github.io/rxnfp
- Open Reaction Database: https://open-reaction-database.org/client/browse

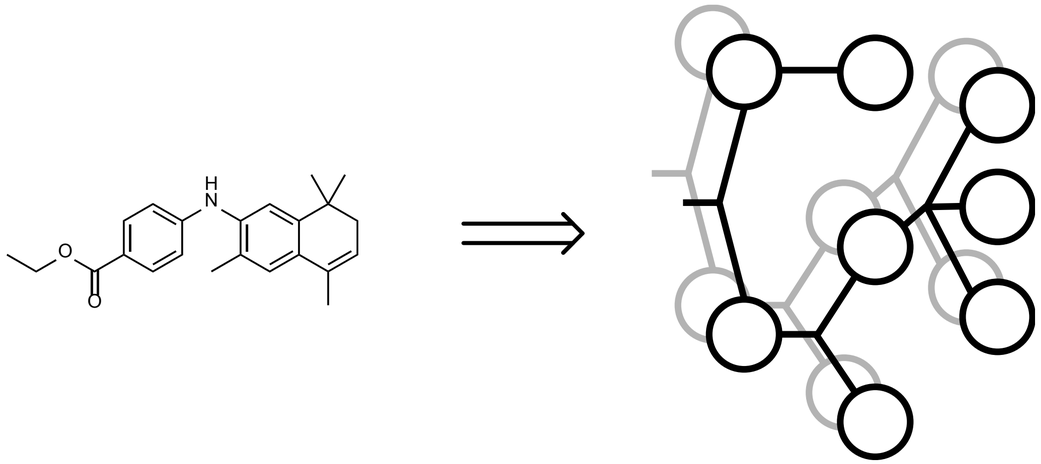
Specific reactions

- 4k Buchwalk-Hartwig reactions (yield): https://www.science.org/doi/10.1126/science.aar5169
- 1k Asymmetric N,S-acetal formation using CPA catalysts (enantioselectivity): https://www.science.org/doi/abs/10.1126/science.aau5631

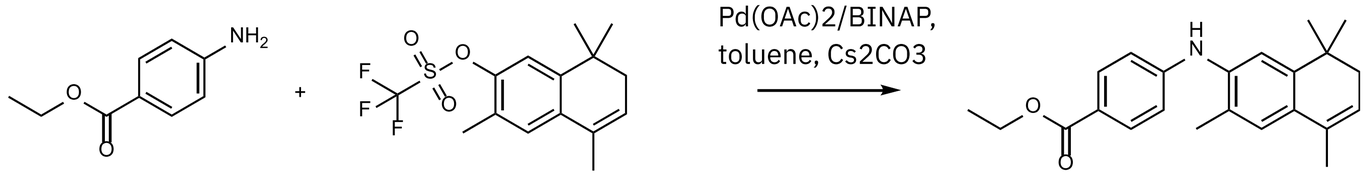
There are many more…



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



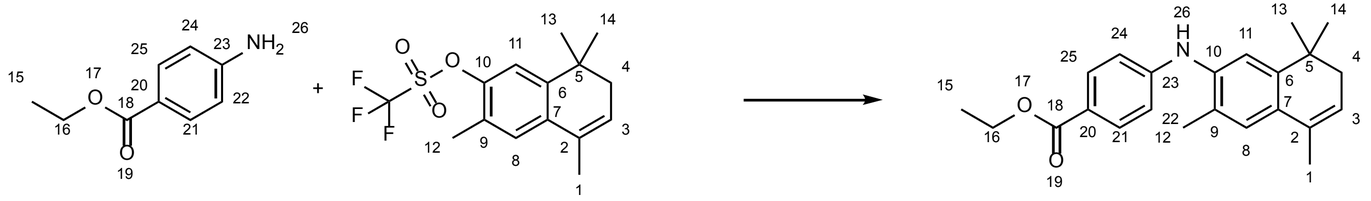
Synthesis planning: Schwaller et al. *Chem. Sci., 2020,11, 3316-3325*

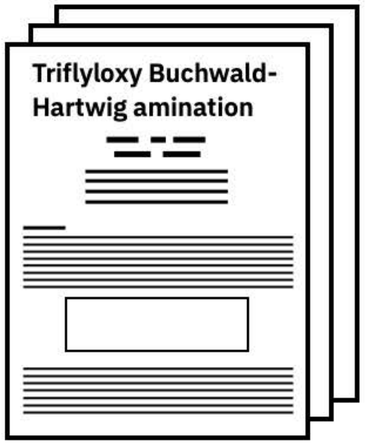


Classification/fingerprints: Schwaller et al. *Nature Mach. Intell., 2021*

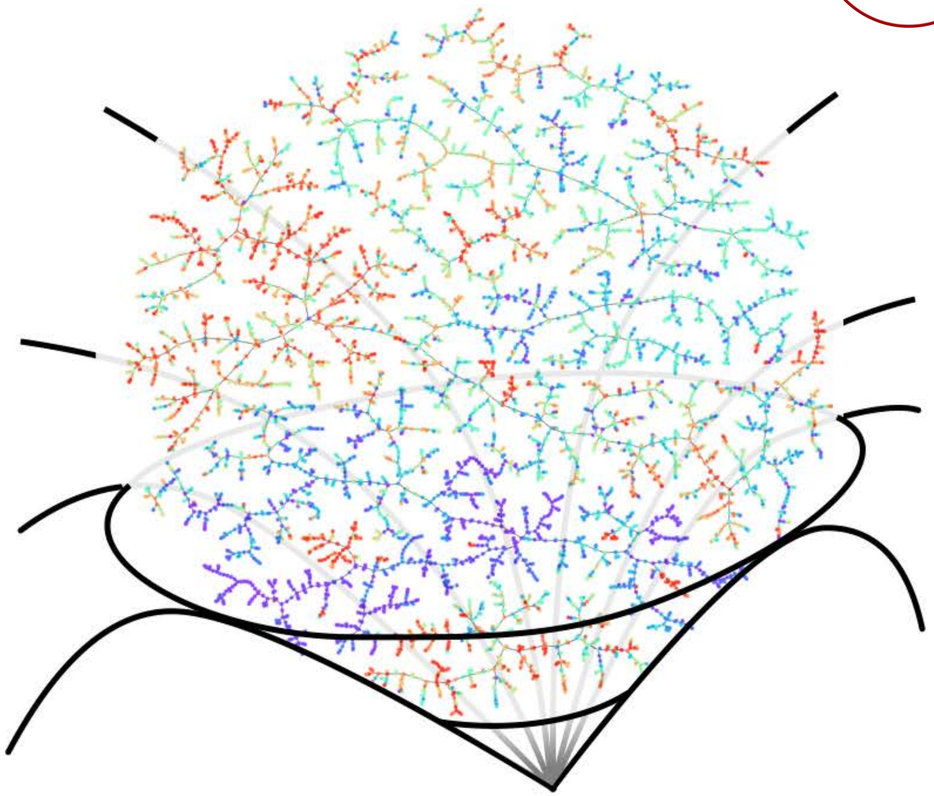
Reaction yields: Schwaller et al. *Mach. Learn.: Sci. Technol.*, 2021; Schwaller et al. NeurIPS 2020 MI4MoI workshop

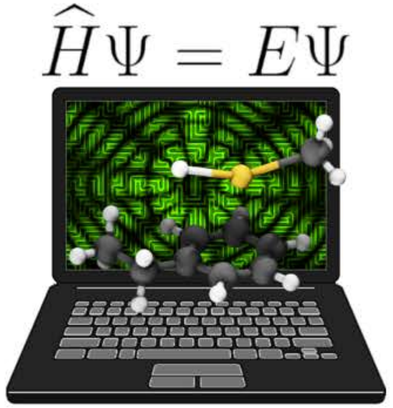
Atom-mapping: Schwaller et al. *Science Advances, 2021*

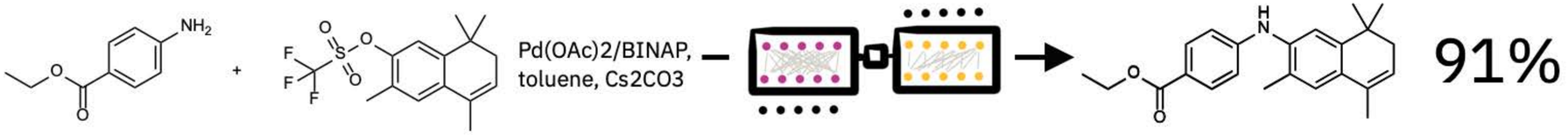




US20030166932A1: General Procedure H A solution of trifluoromethanesulfonic acid 3,5,8,8-tetramethyl-7,8-dihydronaphthalen-2-yl ester (Compound 35, 0.41 g, 1.2 mmol), Pd(OAc)2 (0.027 g, 0.12 mmol), BINAP (0.11 g, 0.18 mmol), Cs2CO3 (0.56 g, 1.72 mmol), ethyl 4-aminobenzoate (0.25 g, 1.5 mmol) and 5 mL of toluene was flushed with argon for 10 min, then stirred at 100° C. in a sealed tube for 48 h. ...







NCCR Catalysis

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

Open Data in the NCCRs

Open Data in NCCR Catalysis

- NCCR Catalysis collects data sets on a Zenodo page: https://zenodo.org/communities/nccr-catalysis/?page=1&size=20

November 30, 2022 (1.0.0) Dataset Open Access

View

Dataset for Reaction-Induced Formation of Stable Mononuclear Cu(I)Cl Species on Carbon for Low-Footprint Vinyl Chloride Production

Faust Akl, Dario; Giannakakis, Georgios; Ruiz-Ferrando, Andrea; Agrachev, Mikhail; Medrano-García, Juan D.; Guillén-Gosalbez, Gonzalo; Jeschke, Gunnar; Clark, Adam H.; Safonova, Olga V.; Mitchell, Sharon; López, Núria; Pérez-Ramírez, Javier;

This dataset complements the publication entitled "Reaction-Induced Formation of Stable Mononuclear Cu(I)Cl Species on Carbon for Low-Footprint Vinyl Chloride Production" by Dario Faust Akl, Georgios Giannakakis, Andrea Ruiz-Ferrando, Mikhail Agrachev, Juan D. Medrano-García, G

Uploaded on April 20, 2023

February 27, 2023 (1.0.0) Dataset Open Access

View

Activation in the rate of oxygen release of Sr0.8Ca0.2FeO3−δ through removal of secondary surface species with thermal treatment in a CO2-free atmosphere

Luongo, Giancarlo; H. Bork, Alexander; M. Abdala, Paula; Wu, Yi-Hsuan; Kountoupi, Evgenia; Donat, Felix; R. Müller, Christoph;

Here we report the data presented in the article titled: "Activation in the rate of oxygen release of Sr0.8Ca0.2FeO3−d through removal of secondary surface species with thermal treatment in a CO2-free atmosphere", Luongo, G., Bork., A., Abdala, P. A., Wu, Y., Kountoupi, E., Donat, F.

Uploaded on April 20, 2023

Romain Graux

Challenges

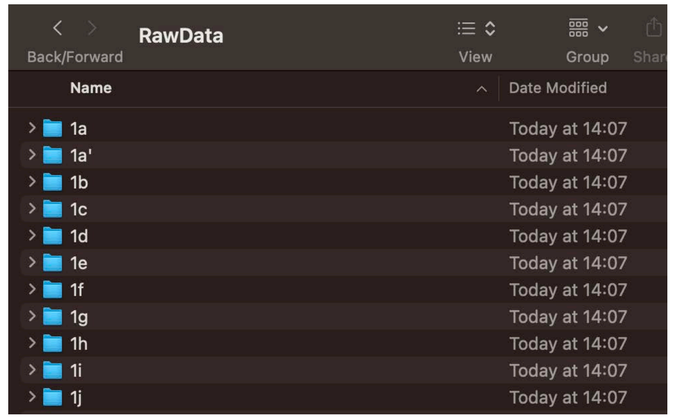
- Raw data is dumped on Zenodo
- No direct link to the compounds
- No machine-actionable reaction data
- Question: How can we do better?

Raw data for the article " Pd(II)-Catalyzed Aminoacetoxylation of Alkenes via Tether Formation"

Raw NMR and MS data for the article "Pd(II)-Catalyzed Aminoacetoxylation of Alkenes via Tether Formation" published in Organic Letters, DOI:

https://doi.org/10.1021/acs.orglett.2c01838

The number of the folders correspond to compounds numbers in the article. All details concerning conditions and equipment for measurements can be found in the supporting information of the article.



Materials Cloud from NCCR Marvel

MATERIALSCLOUD

LEARN WORK DISCOVER EXPLORE ARCHIVE

Built for seamless sharing of resources in computational materials science.

LEARN Lectures and tutorials in computational materials science

WORK Simulation tools and services - in the cloud or on your computer

DISCOVER Curated research data with tailored visualizations

EXPLORE Interactive browser for AiiDA provenance graphs

ARCHIVE An open-access, moderated repository for research data in computational materials science

Please cite L. Talirz et al., Sci Data 7, 299 (2020), if you use Materials Cloud in your research.

Latest records

Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles

DOI 10.24435/materialscloud:55-sd

Aik Rui Tan, Shingo Urata, Samuel Goldman, Johannes C. B. Dietschreit, Rafael Gómez-Bombarelli

Neural networks (NNs) often assign high confidence to their predictions, even for points far out-of-distribution, making uncertainty quantification (UQ) a challenge. When they are employed to model interatomic potentials in materials systems, this problem leads to unphysical structures that disrupt simulations, or to biased statistics and dynamics that do not reflect the true physics. Differentiable UQ techniques can find new informative data and drive active learning loops for robust potentials. However, a variety of UQ techniques, including newly developed ones, exist for atomistic simulations and there are no clear guidelines for which are most effective or suitable for a given case. In this work, we examine multiple UQ schemes for improving the robustness of NN interatomic potentials (NNIPs) through active learning. In particular, we compare incumbent ensemble-based methods against strategies that use single, deterministic NNs: mean-variance estimation, deep evidential ...

Latest version: v1
Publication date: May 04, 2023

Mechanism and prediction of hydrogen embrittlement in fcc stainless steels and high entropy alloys

DOI 10.24435/materialscloud:ct-x8

Xiao Zhou, Ali Tehranchi, W.A. Curtin

The urgent need for clean energy coupled with the exceptional promise of hydrogen (H) as a clean fuel is driving development of new metals resistant to hydrogen embrittlement. Experiments on new fcc high entropy alloys present a paradox: these alloys absorb more H than Ni or SS304 (austenitic 304 stainless steel) while being more resistant to embrittlement. Here, a new theory of embrittlement in fcc metals is presented based on the role of H in driving an intrinsic ductile-to-brittle transition at a crack tip. The theory quantitatively predicts the H concentration at which a transition to embrittlement occurs in good agreement with experiments for SS304, SS316L, CoCrNi, CoNiV, CoCrFeNi, and CoCrFeMnNi. The theory rationalizes why CoNiV is the alloy most resistant to embrittlement and why SS316L is more resistant than the high entropy alloys CoCrFeNi and CoCrFeMnNi, which opens a path for the computationally guided discovery of new embrittlement-resistant alloys.

MARVEL

nature chemistry

Explore content | About the journal | Publish with us

nature > nature chemistry > perspectives > article

Perspective | Published: 04 April 2022

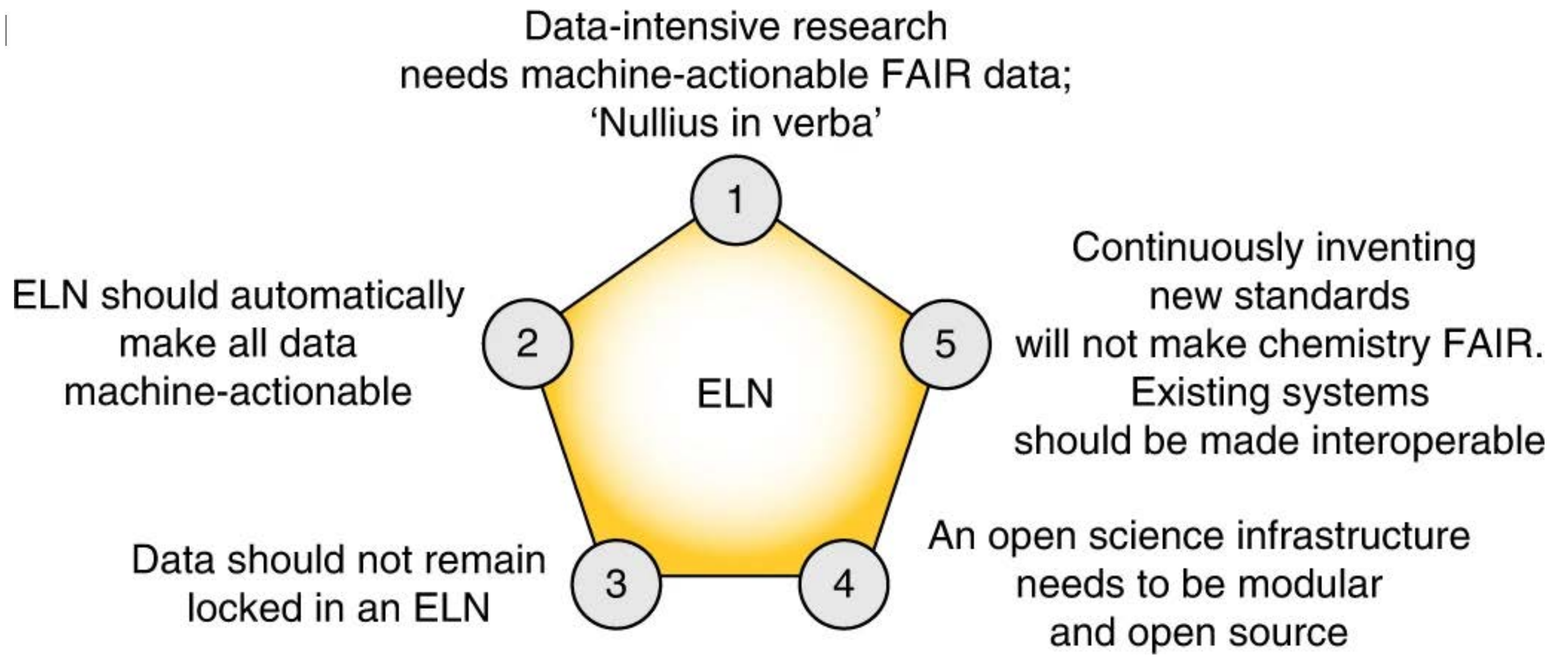
Making the collective knowledge of chemistry open and machine actionable

Kevin Maik Jablonka, Luc Patiny & Berend Smit

Nature Chemistry **14**, 365–376 (2022) | Cite this article

14k Accesses | **13** Citations | **118** Altmetric |

Kevin Jablonka

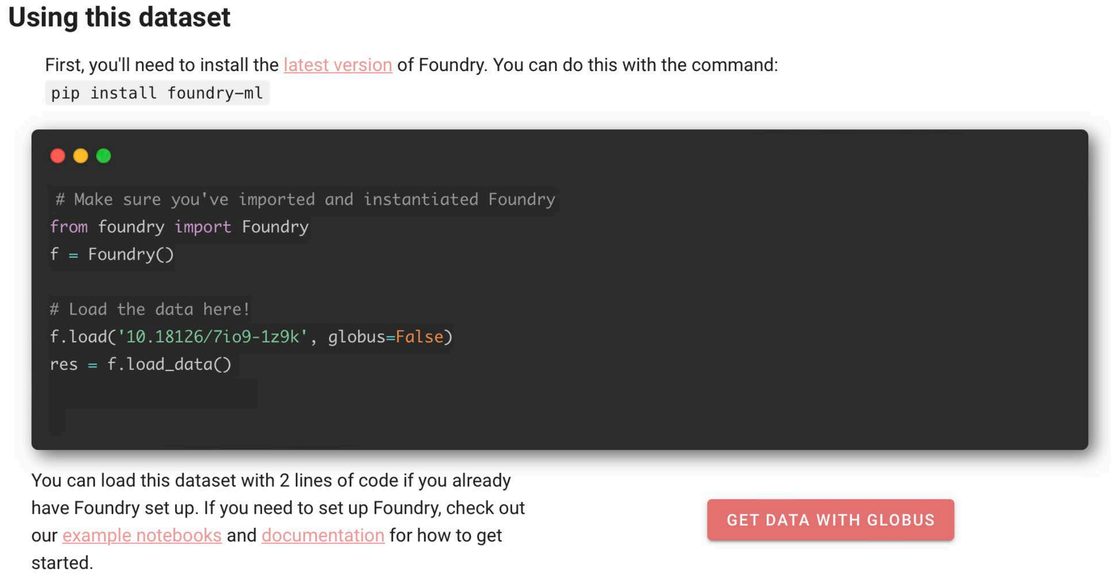


MARVEL

cheminfo

Where can we improve?

- Better metadata / links
- Reusability of datasets
  - Have associated code to load and process the data
  - https://foundry-ml.org/
  - https://github.com/MLMI2-CSSI/foundry/blob/main/examples/bandgap/bandgap_demo.ipynb



Using this dataset

First, you'll need to install the latest version of Foundry. You can do this with the command:

```
pip install foundry-ml
```

```
# Make sure you've imported and instantiated Foundry
from foundry import Foundry
f = Foundry()

# Load the data here!
f.load('10.18126/7io9-1z9k', globus=False)
res = f.load_data()
```

You can load this dataset with 2 lines of code if you already have Foundry set up. If you need to set up Foundry, check out our example notebooks and documentation for how to get started.

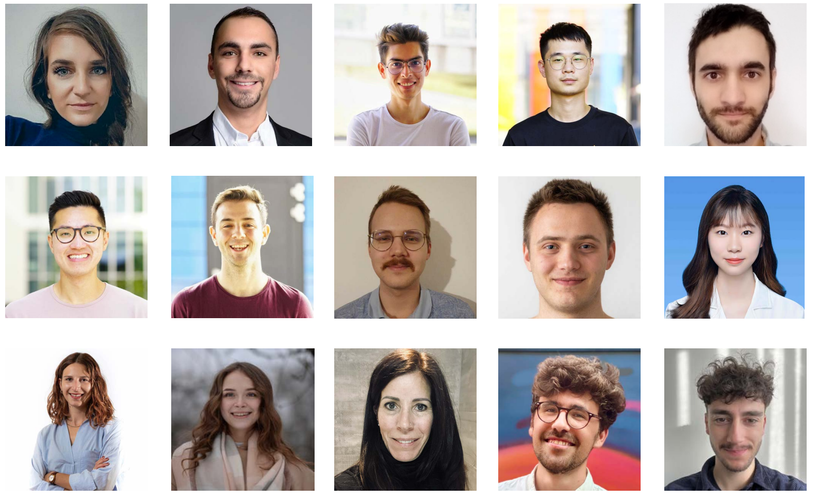
GET DATA WITH GLOBUS

Take home messages

- Open Research Data is crucial
  - Scientific advancement
  - Reliability of results
  - Visibility and citations
  - Greater impact
- There is progress, but we should improve
- Make it accessible through
  - Well-documented linked code
  - Examples
  - Metadata / links

EPFL LIAC

Bojana Rankovic, Andres CM Bran, Junwu Chen, Jeff Guo, Oliver Schilter (with IBM), Victor S Gil (with Luterbacher), Paulo Neves (with Janssen), Stéphane D'Ascoli, Rebecca Neeser (with Correia)



IBM **Research** AI

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

Kevin Jablonka and Berend Smit (EPFL)
Sam Cox and Andrew D White (Universty of Rochester)

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

Alpha Lee, Peter Bolgar and Ryan-Rhys Griffiths

Clemence Corminboeuf, Andreas Krause, Ruben Laplaza, Charlotte Bunne, Jeremy Luterbacher

NEXTMOVE SOFTWARE

PyTorch

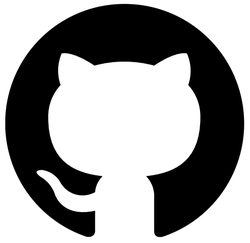
RDKit
Open-Source Cheminformatics and Machine Learning

NMT
An open source neural machine translation system.

Transformers

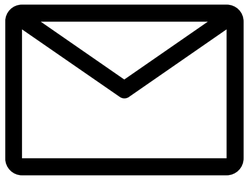
Many thanks for your attention!

@pschwllr @SchwallerGroup



https://github.com/pschwllr

https://github.com/schwallergroup



philippe.schwaller@epfl.ch

https://schwallergroup.github.io

Time for questions and discussion.