Recap and whirlwind intro to ML in Chemistry

Practical Programming in Chemistry

Prof. Philippe Schwaller

Recap of CH-200

The basics
- Using the terminal
- Git and Github
- Environments

Python
- Recap from last year
- Functions & classes
- Pip install and use packages

Cheminformatics
- Molecular representations
- RDKit as main framework

Advanced Practical Programming
- Make reusable code (Python packages)
- APIs/web scraping
- Testing your code
- Writing doc strings and documentation
- Creating applications with streamlit

I hope everyone has learned sth useful for the future!

Exercises gave you practice on different topics

Lecture 1: Terminal

Lecture 2: GitHub & Python recap

Lecture 3: Conda

Lecture 4: Functions, classes, files

Lecture 5: Numpy, Pandas, matplotlib

Lecture 6: RDKit input/output, descriptors and fps

Lecture 7: RDKit substructure matching/conformers

Lecture 8: Python packages introduction

Lecture 9: HTTP requests & web scraping

Lecture 10: Copier templates, tests & coverage

Lecture 11: Visualization & molecular analysis

How many of you have done all exercises?

Lecture 12: Web apps with Streamlit

Where to go from now?

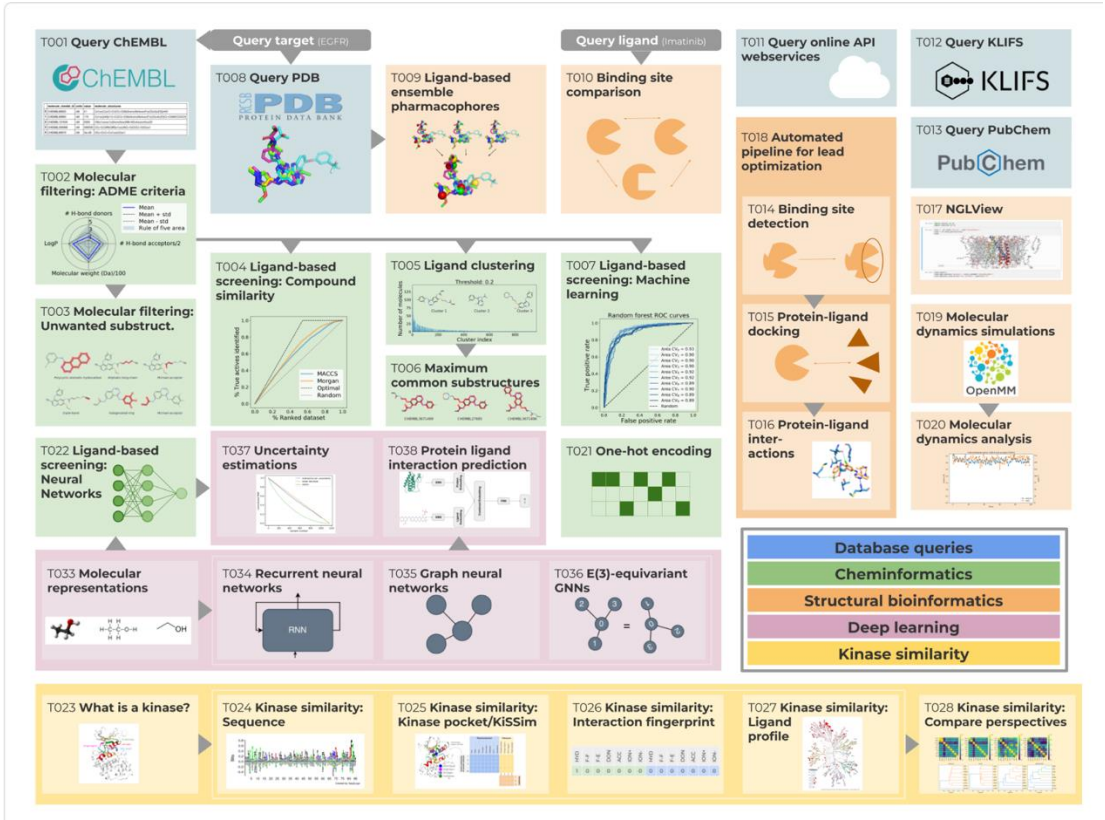
- You have a local setup with environments
  - Remember one environment per project (otherwise, you might have dependency issues!)
- Most open-source Python code is a pip install away from you
  - Get inspired by blogpost, code repositories, and hopefully, don't stop coding ☺
- You all have encountered the most common issues with Github
  - Everyone goes through that at the beginning
  - In the future, you will overcome them more efficiently
  - There is always a solution (→ Google, StackOverflow, and maybe LLMs)

Awesome sources of information

- Pat Walter's – Practical Cheminformatics Tutorials (highly recommended!) (https://github.com/PatWalters/practical_cheminformatics_tutorials)
- Greg Landrum's RDKit blog (https://greglandrum.github.io/rdkit-blog/)
- https://github.com/hsiaoyi0504/awesome-cheminformatics?tab=readme-ov-file#resources

Prof Andrea Volkamers TeachOpenCADD

- https://volkamerlab.org/projects/teachopencadd/



Open course on Computer Assisted Drug Design

Or the AI for Chemistry course (EPFL, Master)

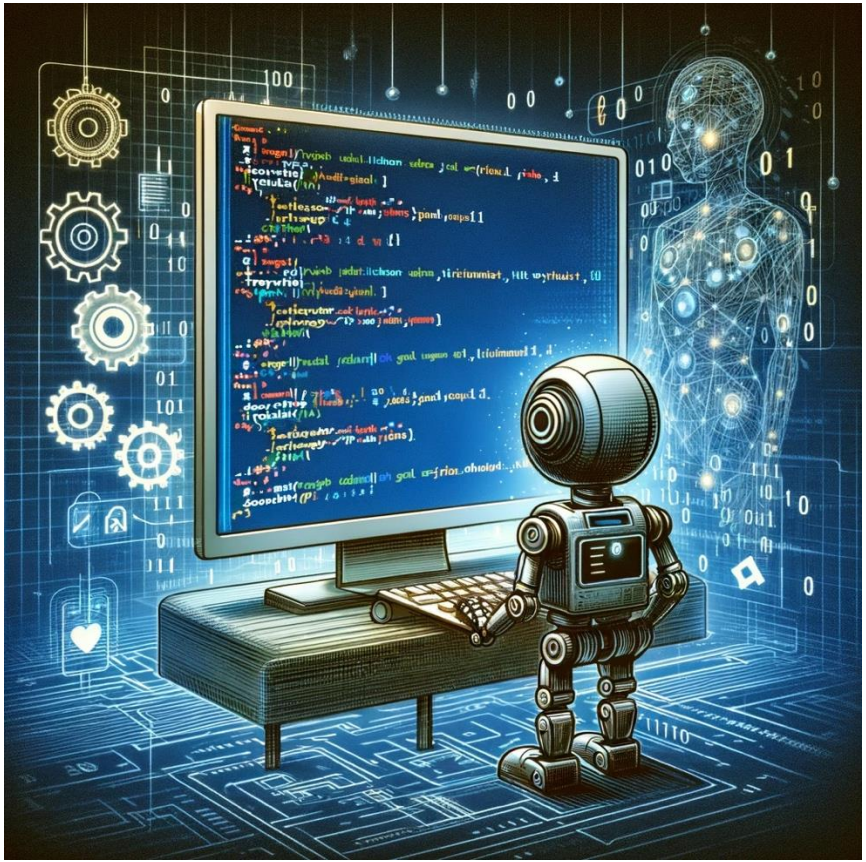
The rest of the lecture will be a whirlwind introduction into machine learning in chemistry.

Made possible through **Machine Learning**...

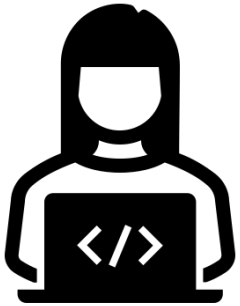
- ChatGPT -> Text
- Midjourney -> Image
- elevenlabs.io
  -> text to speech
- D-ID
  -> image, speech to video

Total time: 15 minutes!

Programming = machines following code instructions



Traditional programming ("Expert system")



Input x

Output y

5.5

3.5

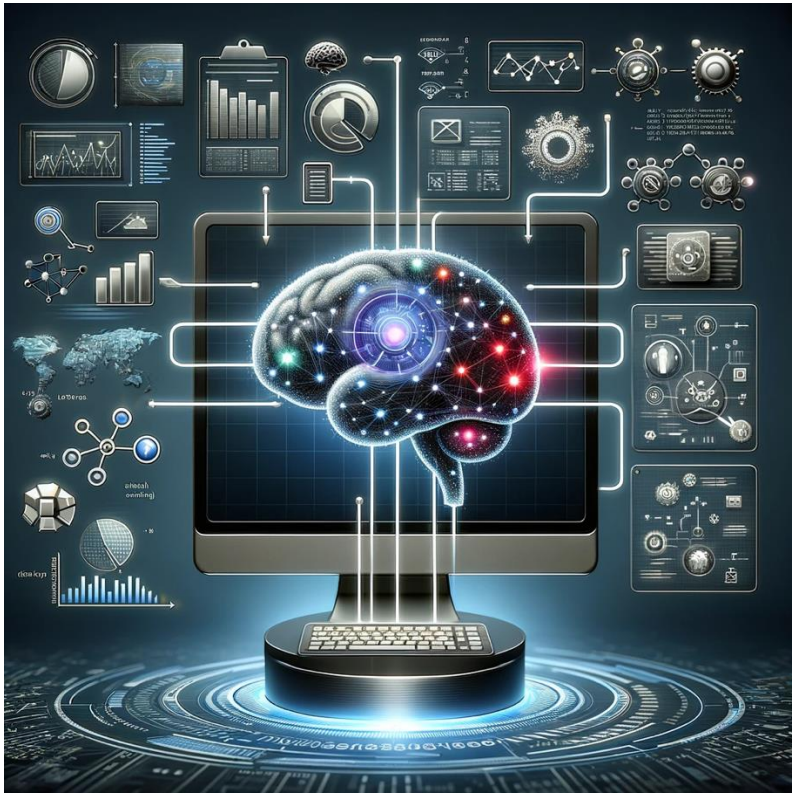
If grade 4 or higher (x >= 4), student passes the course.

Pass

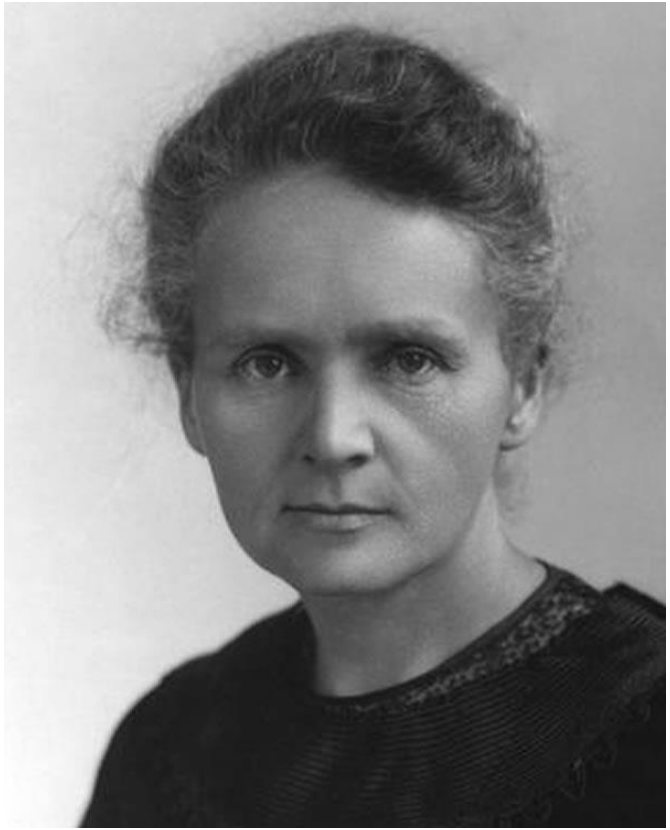
Fail

Predefined human-written rules (knowledge base)

Machine learning – ability for machines to learn without being programmed (from data)

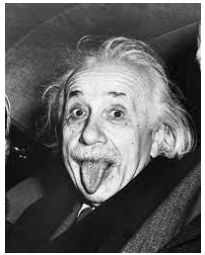


So, let's make an example.



Who is this? **Marie Curie.**

Supervised learning – facial recognition

Input	Output
	Marie Curie
	Marie Curie
	Albert Einstein

Training data

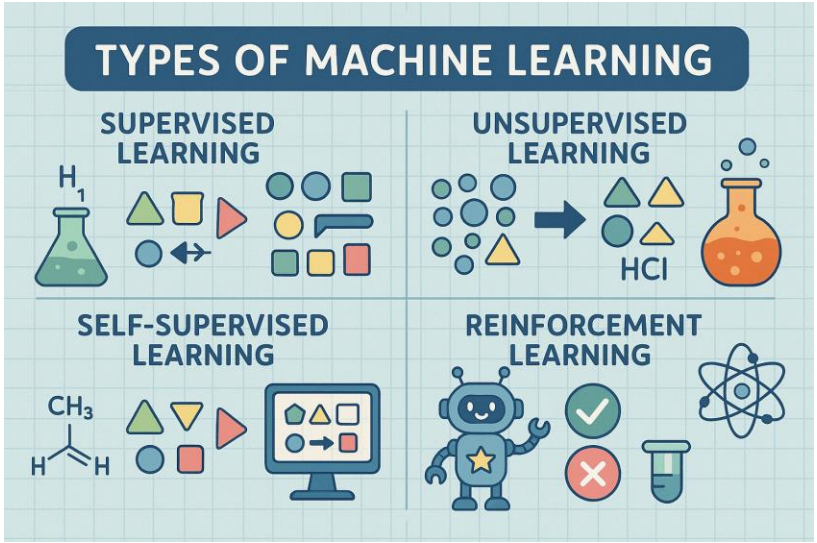
- Goal: learning from **inputs** to **outputs**
- Needs **training data (!)**
- **Model** => neural networks popular
- **Classification** task
- How good is my model? **Accuracy**

Model

?

Several flavours of ML

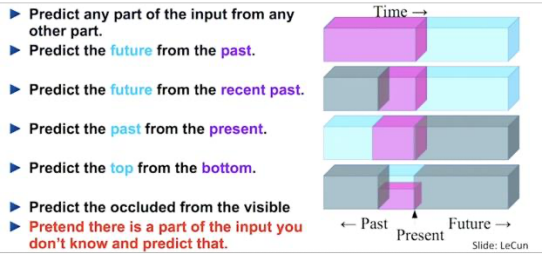
Supervised learning
- learning from labelled data
- classification/regression



Unsupervised learning
- learning from unlabelled data
- patterns and structures

→ K-means clustering and PCA.

Self-supervised learning
- learning by creating labels from the data you have



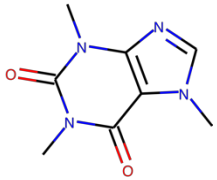
Reinforcement learning
- learning by taking actions in an environment, and getting rewards

Key ingredients for Machine Learning

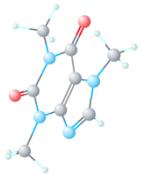
Molecular fingerprints
000010000....0100

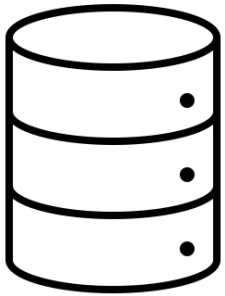
Text-based representations
CN1C=NC2=C1C(=O)N(C(=O)N2C)C

Graph-based representations



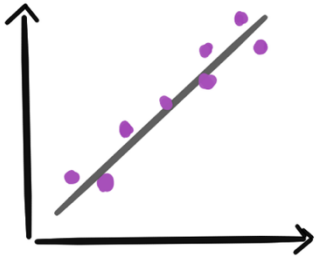
3D coordinates & surface



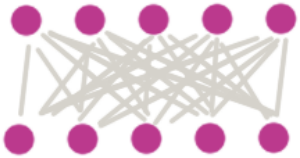


Examples are:
- Molecules & properties
- Chemical reactions
- Synthesis procedures

Linear regression model



Neural networks



And many more..

Representations (machine-readable)

Data (garbage in = garbage out)

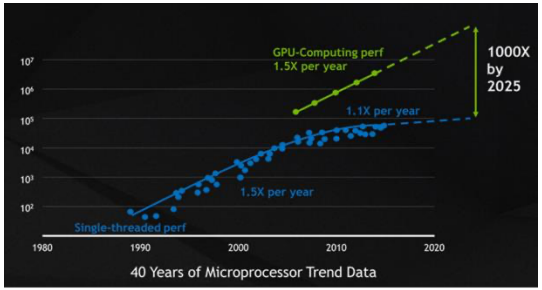
Models/algorithms

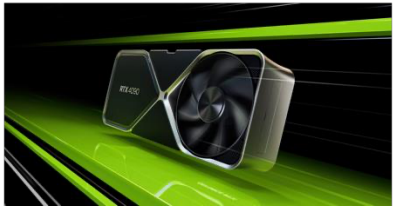
Deep Learning (neural network-based ML) – why did it take off in the last decade?

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*





TensorFlow

PyTorch

Keras

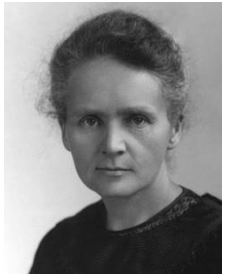
PyTorch Lightning

Immense datasets

Cheap compute

Open research & frameworks

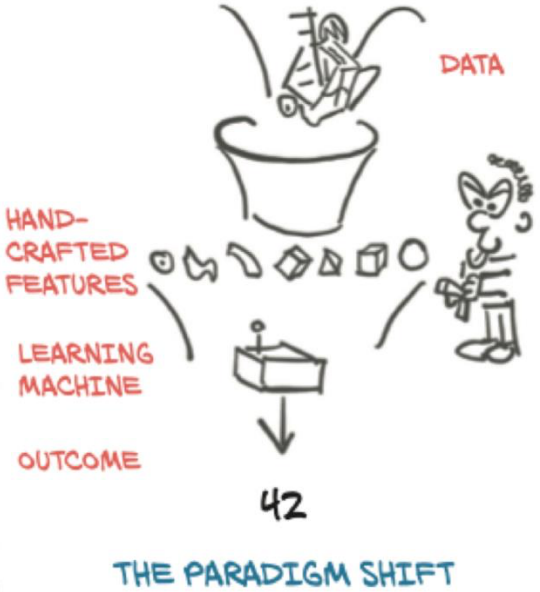
Main difference...



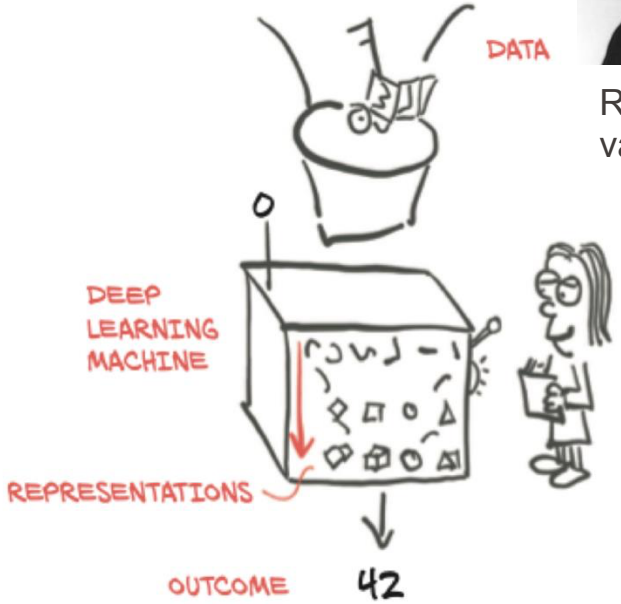
Texture, colour,
… as tabular data

Or for molecules:
Molecular weight,
HOMO-LUMO gap,
number of H-donors

Traditional machine learning
(requires tabular data)



Deep learning
(requires a lot of data)



Raw pixel
values

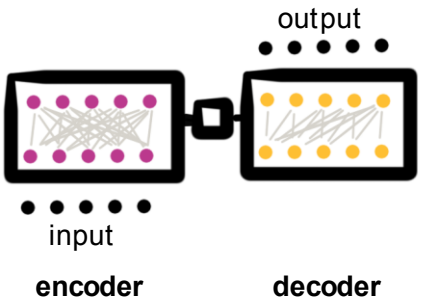
Learning
directly
from structure:

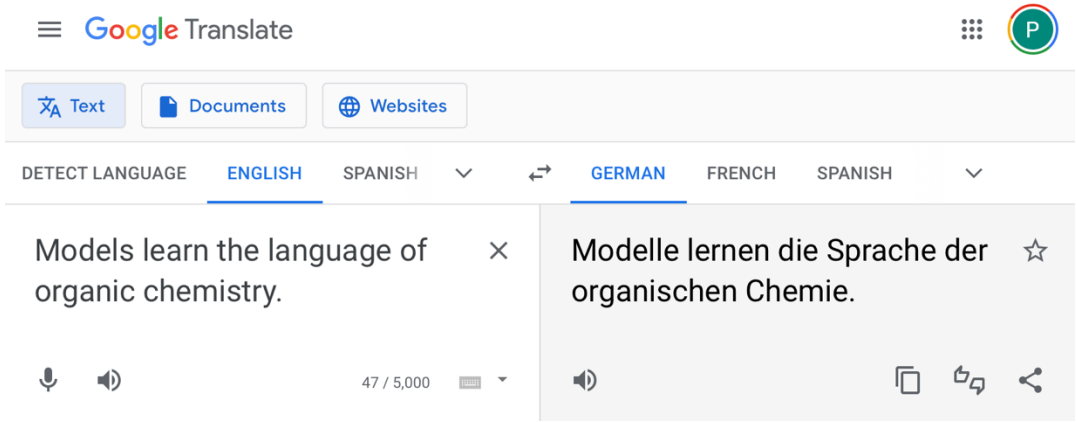
SMILES or
atomic coord.

- Stevens et al., Deep Learning with PyTorch, 2020

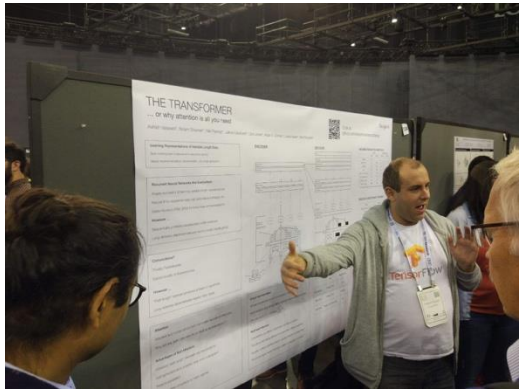
Going beyond regression and classification tasks.

One of the most important neural network architectures is the Transformer.



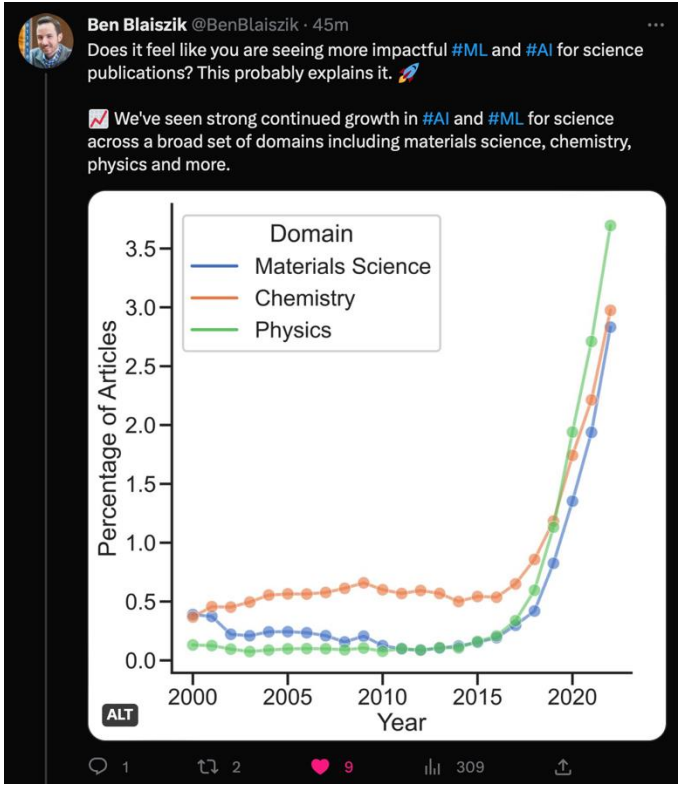


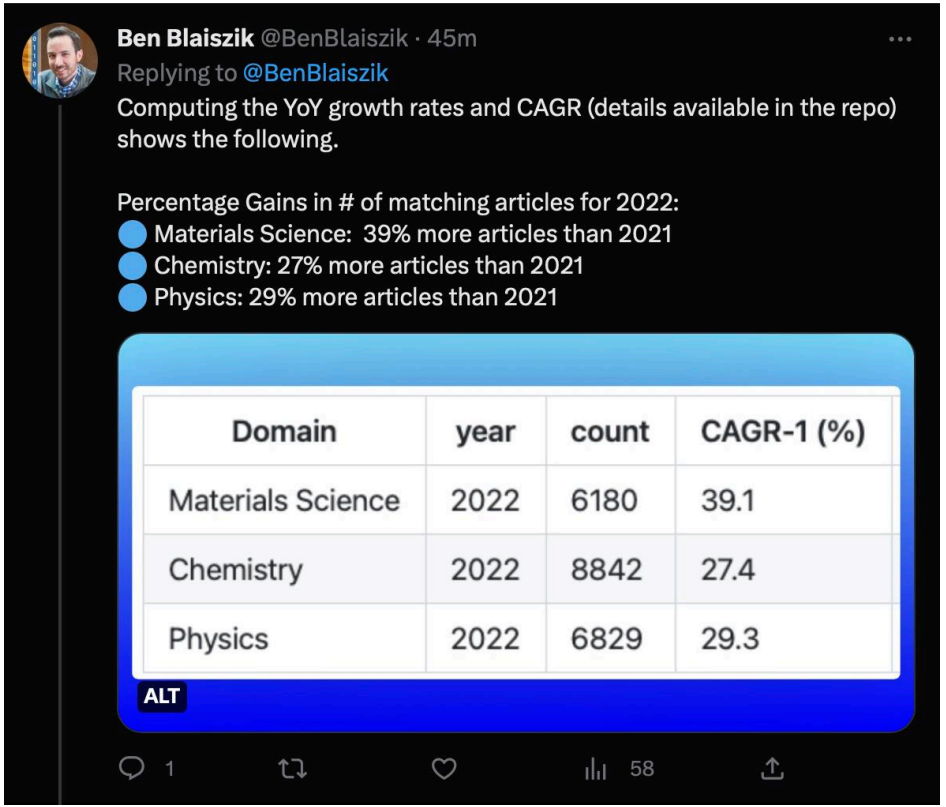
- Learns translation from examples
- GPT → decoder part



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

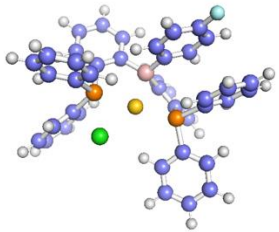
So, why should I care about all that as a chemist?



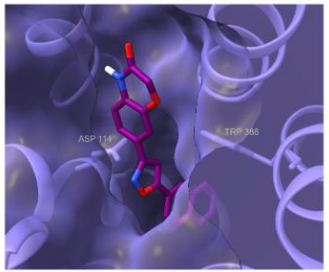


Chemical design cycle

What molecule to make?



Catalysis

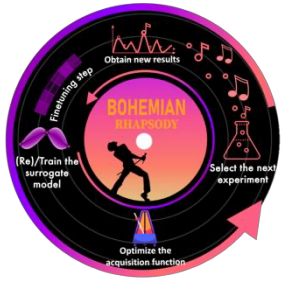


Drug discovery

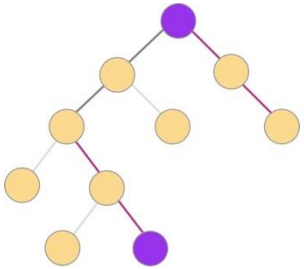
Design

Make

How to make it?



Synthesis optimization

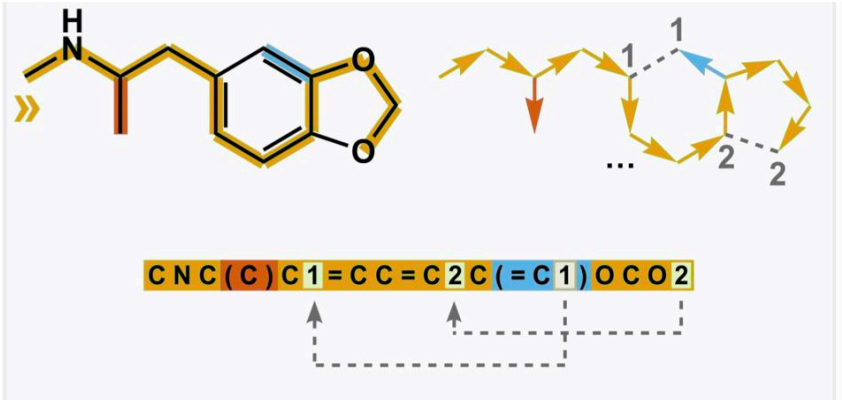


Synthesis planning

Test

Experimental validation

Text-based molecule representation – SMILES



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

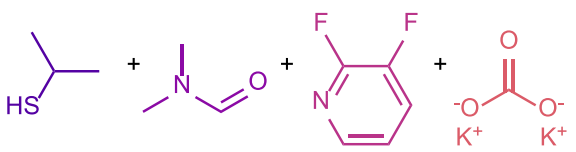
DAVID WEININGER
Medicinal Chemistry Project, Pomona College, Claremont, California 91711

Received June 17, 1987

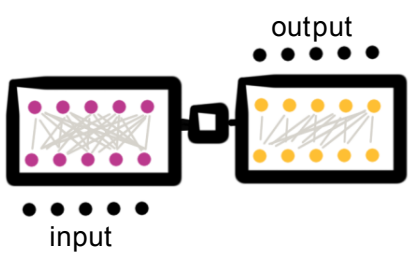
Krenn, Mario, et al. "SELFIES and the future of molecular string representations.", Patterns, 2023.

E.g. reaction prediction task as machine translation task (**Molecular Transformer – first transformer in Chem**)

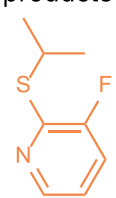
precursors



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+]



products

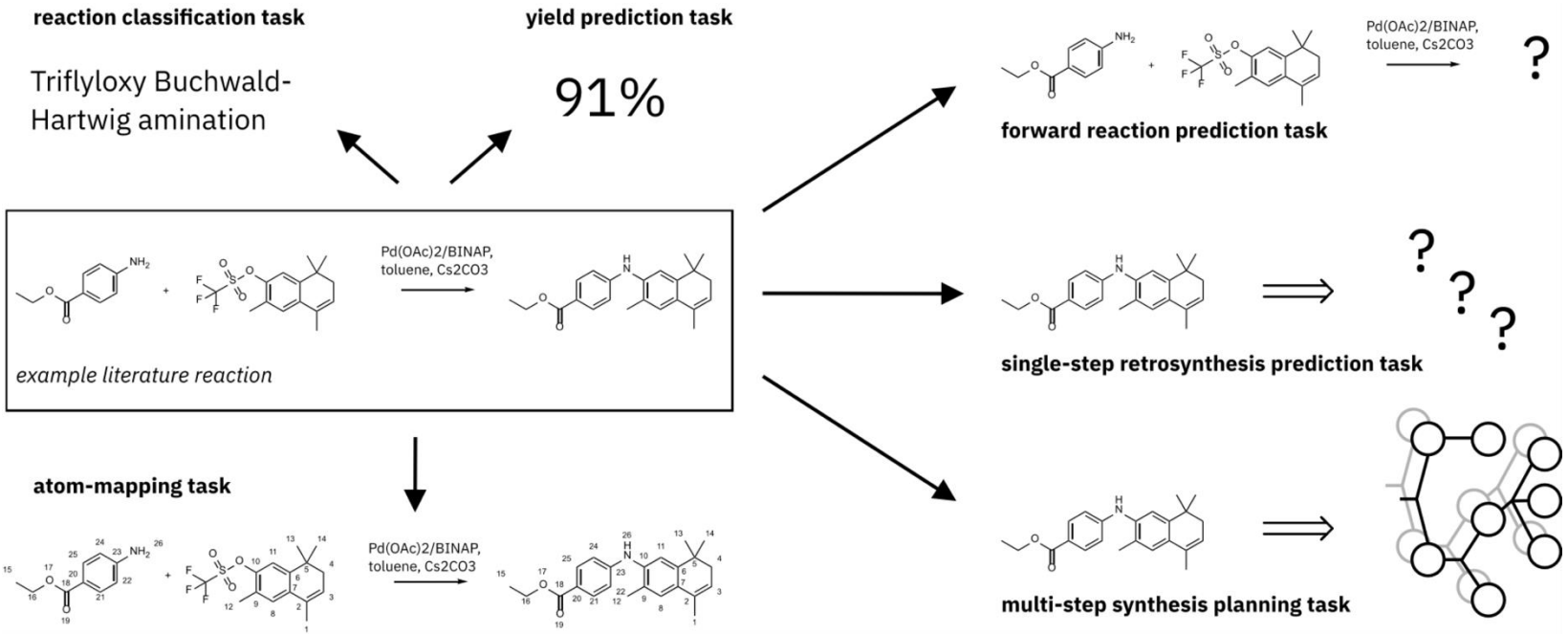


C C (C) S c 1 n c c c c 1 F

Language models for chemistry

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

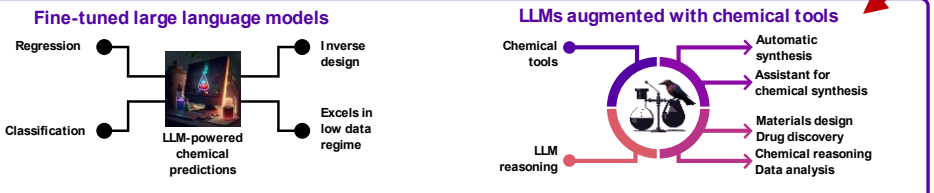
ML for chemical reactions



+ more, such as recipe prediction

Trends in *Language Models* for Chemical Synthesis

General task solvers



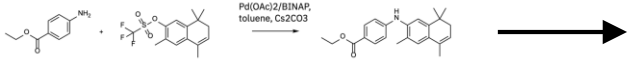
Jablonka et al. NeurIPS ML for Materials, 2022

Bran and Cox et al. ACS Fall Meeting, 2023

Other examples:
- Clairify (Skreta et al.)
- Coscientist (Boiko et al.)

Multiple modalities

Reaction to synthesis procedure

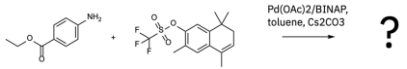


Vaucher et al. Nature Comm. 2020
Vaucher et al. *Nature Comm.* 2021

1. MAKESOLUTION with trifluoromethanesulfonic acid 3,5,8,8-tetramethyl-7,8-dihydronaphthalen-2-yl ester (0.41 g, 1.2 mmol) and Pd(OAc)2 (0.027 g, 0.12 mmol) and BINAP (0.11 g, 0.18 mmol) and Cs2CO3 (0.56 g, 1.72 mmol) and ethyl 4-aminobenzoate (0.25 g, 1.5 mmol) and toluene (5 mL)
2. ADD SLN
3. STIR for 48 hours at 100° C
4. SETTEMPERATURE room temperature
5. CONCENTRATE
6. PURIFY
7. YIELD product

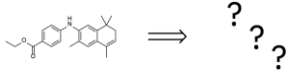
Single modality

Reaction prediction



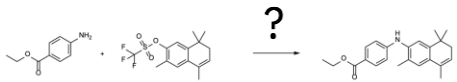
Schwaller et al. *ACS Cent. Sci.* 2019
Pesciullesi et al. *Nature Comm.* 2020

Retrosynthetic planning



Schwaller et al. *Chem. Sci., 2020*
Thakkar et al. *ACS Cent. Sci.* 2023

Reaction condition prediction



Gao et al. *ACS Cent. Sci.* 2018
Schilter et al. *ACS Fall Meeting* 2023

LIAC Highlights

Augmenting language models with *chemistry tools*

- Large language models (GPT4/ChatGPT) are **bad at chemistry**
- There are **excellent chemistry tools** (but in isolation)

So can what do we do? → **ChemCrow**

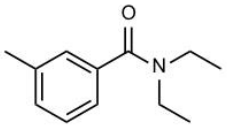
What **tools** for ChemCrow?

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White & Philippe Schwaller

Nature Machine Intelligence **6**, 525–535 (2024) | Cite this article

ChemCrow's toolset

Molecule tools

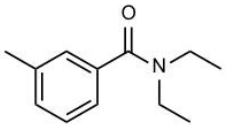


- SMILES to Weight
- SMILES to Price
- SMILES to CAS
- Similarity
- Modify Mol
- Func Groups
- Patent Check
- Name to SMILES

- RDKit: Open-source cheminformatics. https://www.rdkit.org
- PubChem Compound Database https://pubchem.ncbi.nlm.nih.gov/
- Synspace: https://github.com/whitead/synspace
- Molbloom: https://github.com/whitead/molbloom
- Chem-space: https://chem-space.com/

ChemCrow's toolset

Molecule tools



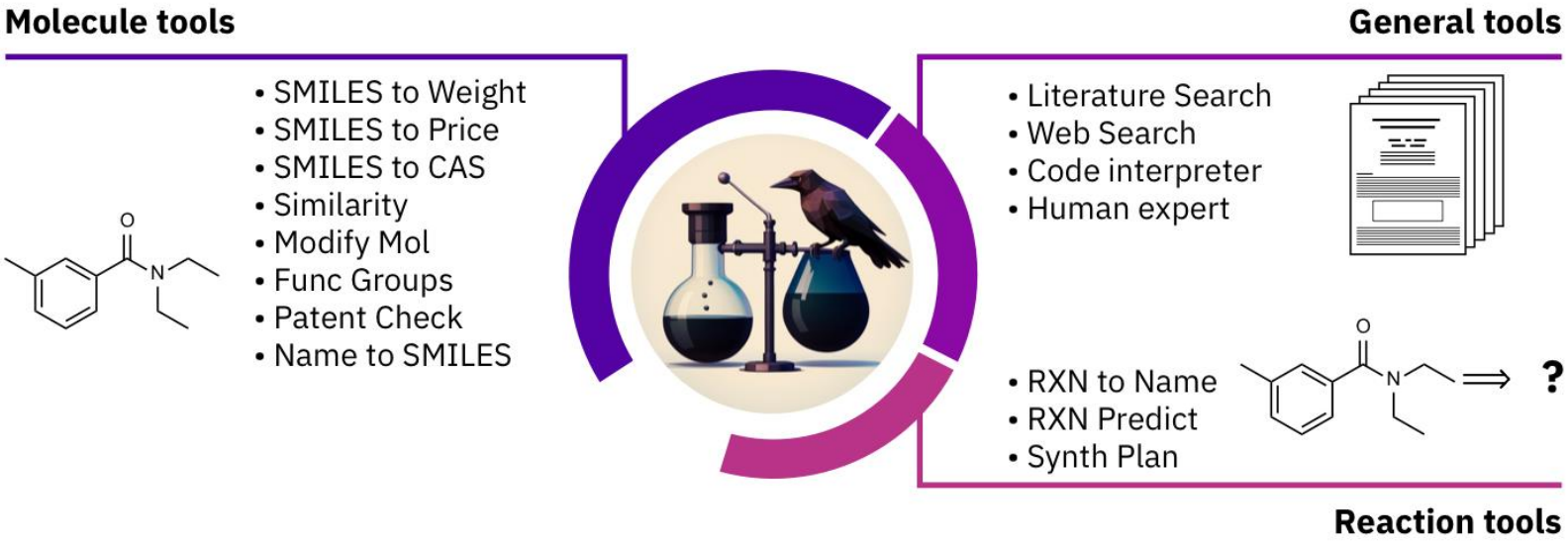
- SMILES to Weight
- SMILES to Price
- SMILES to CAS
- Similarity
- Modify Mol
- Func Groups
- Patent Check
- Name to SMILES

General tools

- Literature Search
- Web Search
- Code interpreter
- Human expert

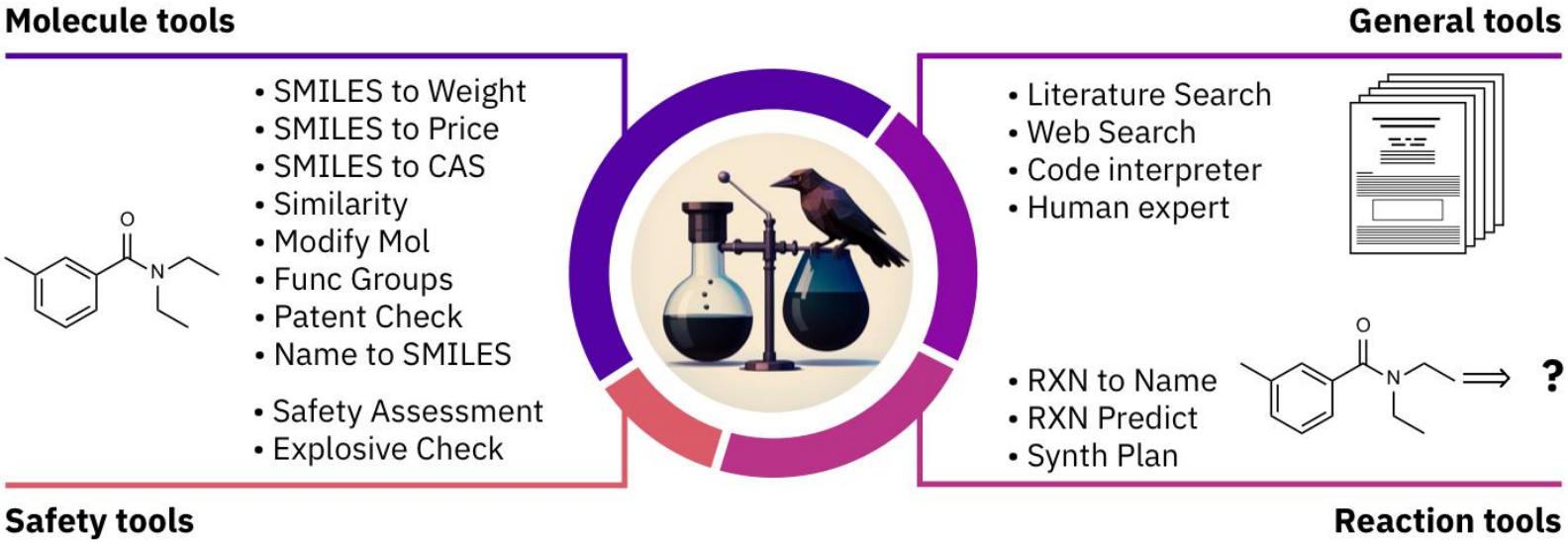
- RDKit: Open-source cheminformatics. https://www.rdkit.org
- PubChem Compound Database https://pubchem.ncbi.nlm.nih.gov/
- Synspace: https://github.com/whitead/synspace
- Molbloom: https://github.com/whitead/molbloom
- Chem-space: https://chem-space.com/
- paper-qa: https://github.com/whitead/paper-qa/tree/main

ChemCrow's toolset



- RDKit: Open-source cheminformatics. https://www.rdkit.org
- PubChem Compound Database https://pubchem.ncbi.nlm.nih.gov/
- Synspace: https://github.com/whitead/synspace
- Molbloom: https://github.com/whitead/molbloom
- Chem-space: https://chem-space.com/
- paper-qa: https://github.com/whitead/paper-qa/tree/main
- IBM RXN for Chemistry: https://github.com/rxn4chemistry/rxn4chemistry/

ChemCrow's toolset



- RDKit: Open-source cheminformatics. https://www.rdkit.org
- PubChem Compound Database https://pubchem.ncbi.nlm.nih.gov/
- Synspace: https://github.com/whitead/synspace
- Molbloom: https://github.com/whitead/molbloom
- Chem-space: https://chem-space.com/
- paper-qa: https://github.com/whitead/paper-qa/tree/main
- IBM RXN for Chemistry: https://github.com/rxn4chemistry/rxn4chemistry/
- ClinTox: Substance Toxicity Dataset: https://www.clintox.org/
- GHS Classification: https://pubchem.ncbi.nlm.nih.gov/ghs/

ChemCrow

Expert-designed chemistry tools

User-defined scientific tasks

ChemCrow

Expert-designed chemistry tools

Example prompt: Plan and execute the synthesis of an insect repellent.

User-defined scientific tasks

ChemCrow

Expert-designed chemistry tools

Example prompt: Plan and execute the synthesis of an insect repellent.

User-defined scientific tasks

1. thought
reason, plan

ChemCrow

ChemCrow

Expert-designed chemistry tools

Example prompt: Plan and execute the synthesis of an insect repellent.

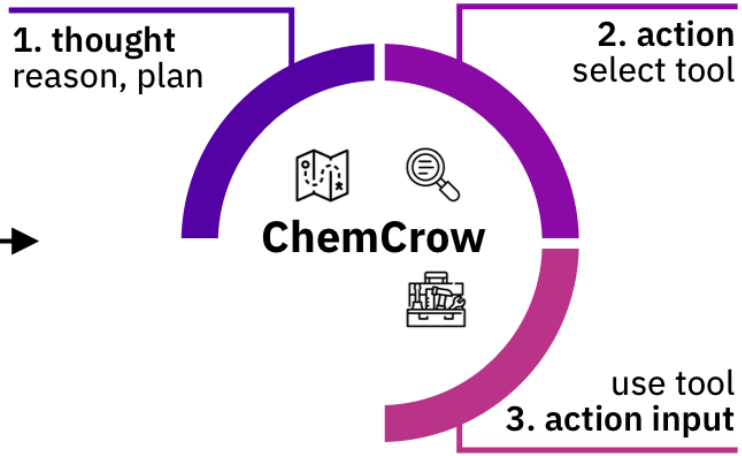
User-defined scientific tasks

ChemCrow

Expert-designed chemistry tools

Example prompt: Plan and execute the synthesis of an insect repellent.

User-defined scientific tasks

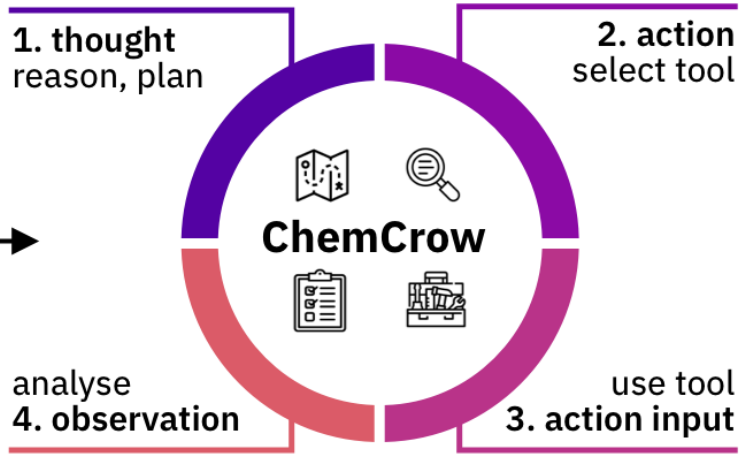


ChemCrow

Expert-designed chemistry tools

Example prompt: Plan and execute the synthesis of an insect repellent.

User-defined scientific tasks

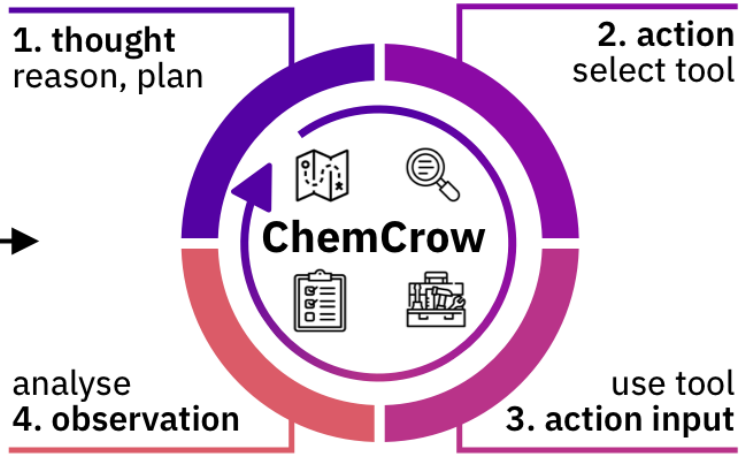


ChemCrow

Expert-designed chemistry tools

Example prompt: Plan and execute the synthesis of an insect repellent.

User-defined scientific tasks



- Yao, S.; React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* **2022**,
- Karpas,E. et al. MRKL Systems: A modular, neuro-symbolic archi- tecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445* **2022**,

Use case: *Automatic synthesis*

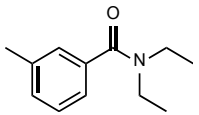
Task: Plan and execute the synthesis of an insect repellent

Thought: I need to know which insect repellent to synthesize. I will do a literature search to find a common insect repellent.
Action: LitSearch
Action input: common insect repellent

Observation: **DEET, picaridin, …**

Thought: I will choose DEET as the insect repellent to synthesize. First I need to get the SMILES of DEET.
Action: Name2SMILES
Action input: DEET

Observation: **CCN(CC)C(=O)c1cccc(C)c1**



Use case: *Automatic synthesis*

Thought: Now that I have the SMILES of DEET, I will plan the synthesis.
Action: RXNPlanner
Action input: CCN(CC)C(=O)c1cccc(C)c1
Observation:
To make CCN(CC)C(=O)c1cccc(C)c1, follow these steps:

1. Add 0.3 mL of diethylamine to the reaction vessel.
2. Add 15 mL of dichloromethane to the reaction vessel.
3. Add 3.2 mmol of 3-methylbenzoyl chloride to the reaction vessel at 0°C.
4. Stir the mixture at 25°C for 1 hour […]

RXN4Chemistry:
Recipe prediction

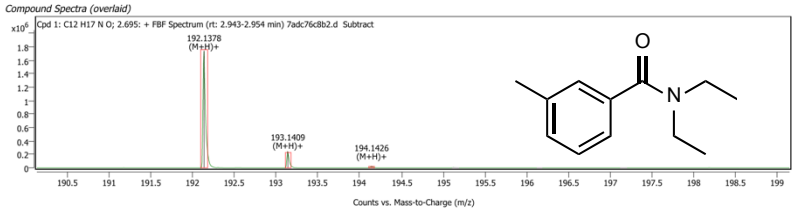
Thought: I have a synthesis plan for DEET. Now I will execute the synthesis.
Action: RXNExecute
Action input: CCN(CC)C(=O)c1cccc(C)c1

Observation: **Synthesis procedure completed!**

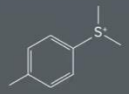
RoboRXN

Compound Summary

Cpd	Name	Formula	CAS	RT	Mass	Mass (Tgt)
1		C12 H17 N O		2.695	191.1305	191.1310
2		C4 H11 N		3.405	73.0896	73.0891



Synthesizing new molecule



Started: Nov 30 2020, 6:49am PT

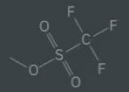
Live from IBM RoboRXN

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

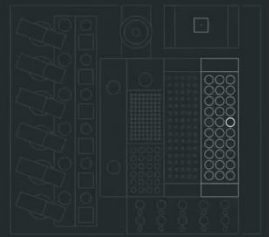


Methyl trifluoromethane sulfonate is a brown liquid. Insoluble in water. This material is a very reactive methylating agent, also known as methyl triflate.

NOW

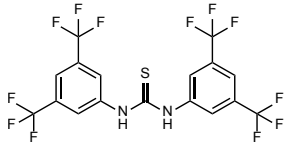
10 ml of reagent containing methyl trifluoromethane sulfonate is being moved from Vial 61 and added to Reactor 2.

Position of the robot arm
Moving to Vial 61

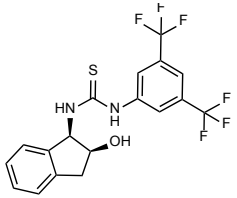


Use case: *Automatic synthesis*

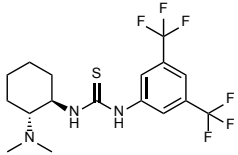
- Find and synthesize a thiourea organocatalyst which accelerates a Diels-Alder reaction.



Schreiner's



Ricci's



Takemoto's

Organocatalysts

ETH zürich

EPFL

CAT+

ChemCrow

Available tools: 18

Tool	Description
✅ Name2SMILES	Input molecule name, returns SMILES.
✅ SMILES2Price	Input SMILES, returns price of compound.
✅ Similarity	Input two SMILES (sep by .), returns Tanimoto similarity.
✅ ModifyMol	Input SMILES, returns list of modified molecules. Determinist
✅ PatentCheck	Input SMILES, returns if molecule is patented.
✅ FuncGroups	Input SMILES, returns list of functional groups.
✅ SMILES2Weight	Input SMILES, returns molecular weight.
✅ MOL2CAS	Input SMILES or molecule name, returns CAS number

RUNNING... Stop

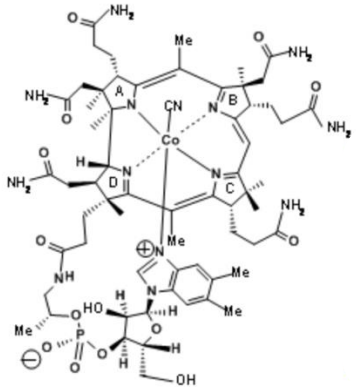
Find 3 psychedelic substances. what happens when they react with acetic chloride?

Your message

Is automated synthesis a solved problem? No...

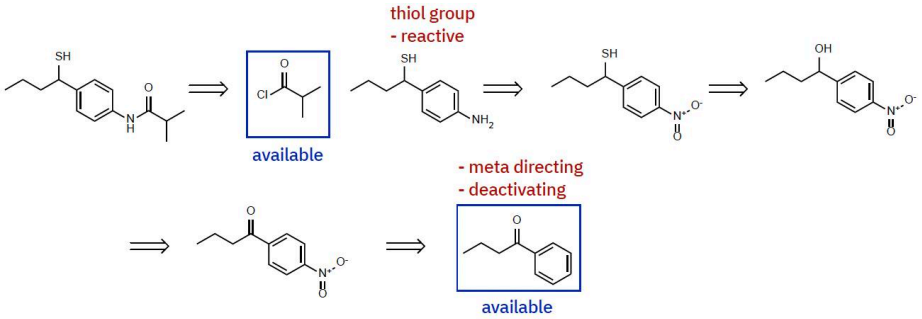
- Supply chain/robotics challenges
- Weak synthesis planning models

Daniel Armstrong & Zlatko Jončev

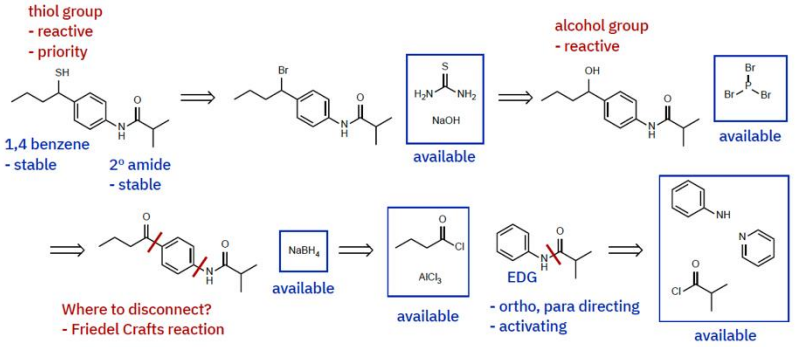


Complexity that an chemist would like to target: Vitamin B12

- 12 year
- ~90 postdocs, 12 PhD students
- Eschenmoser, Woodward



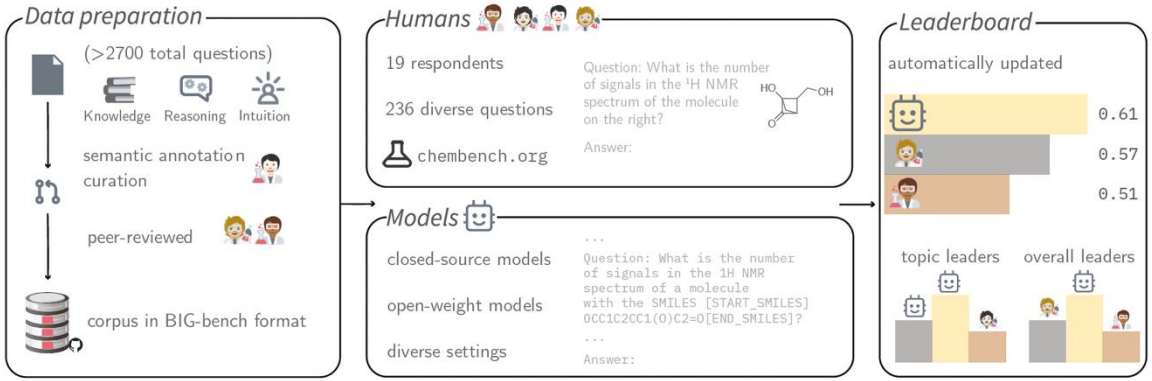
Typical output of unnamed system

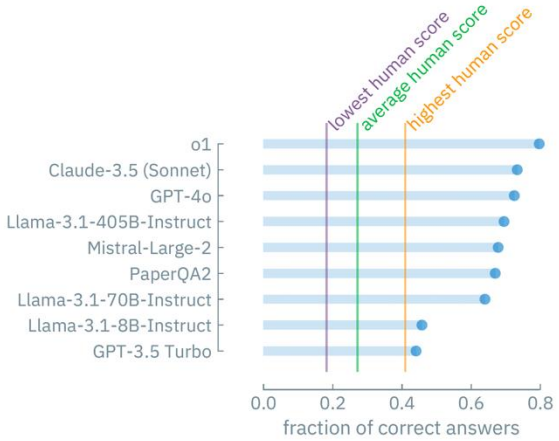


More strategic synthesis plan (including reagents)

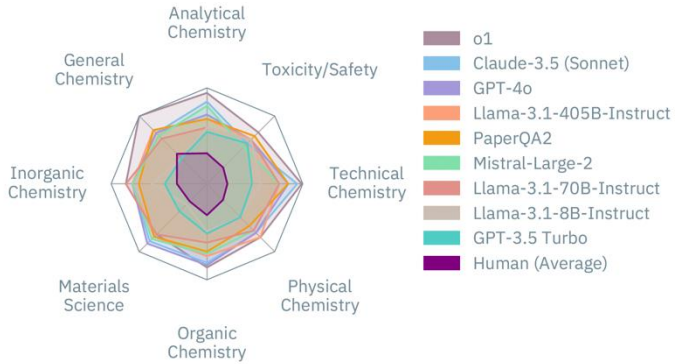
https://www.slideshare.net/EngelbertZass1/of-a-landmark-total-synthesis-yet-unpublished-in-full-experimental-detail-vitamin-b12

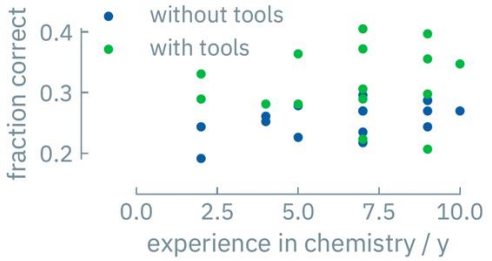
Benchmarking LLMs in Chemistry – ChemBench





Collaboration led by Kevin Jablonka





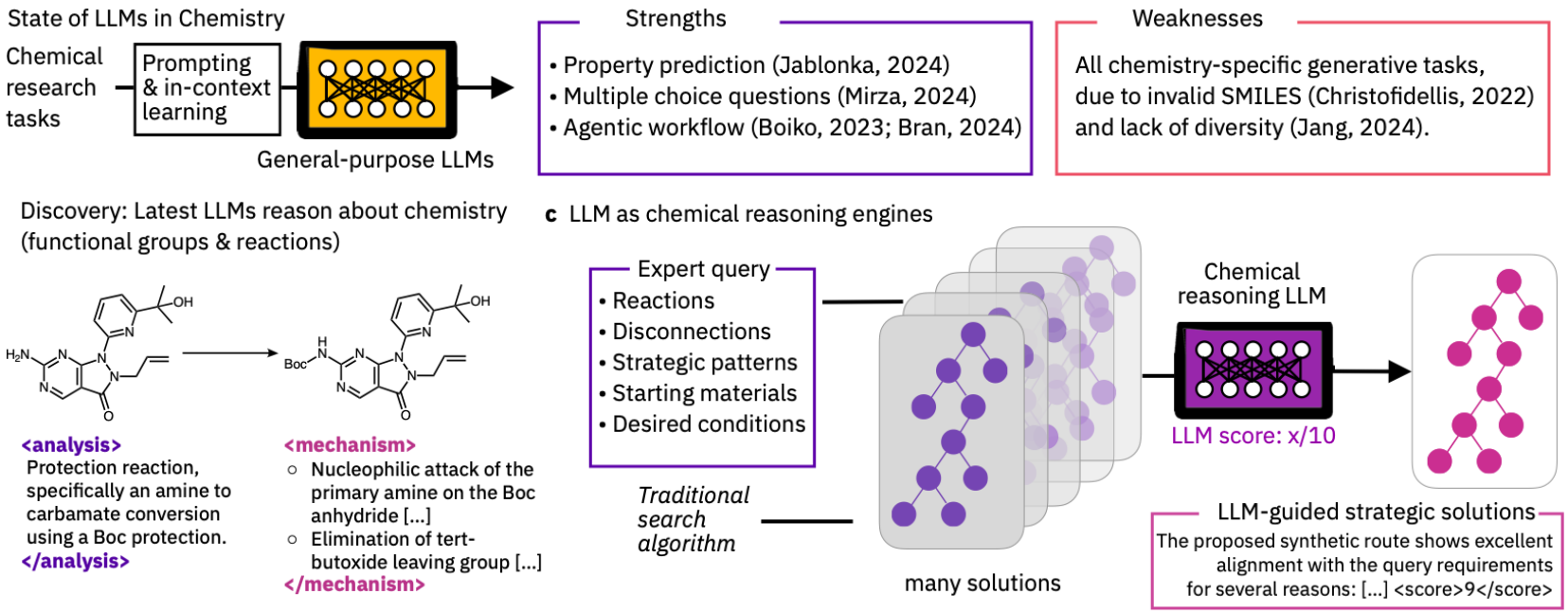
19 human experts

https://www.chembench.org

- https://arxiv.org/abs/2404.01475, Are large language models superhuman chemists?

Benchmarking beyond multiple choice questions.

LLMs as chemical reasoning engines



Chemical reasoning in LLMs unlocks steerable synthesis planning and reaction mechanism elucidation

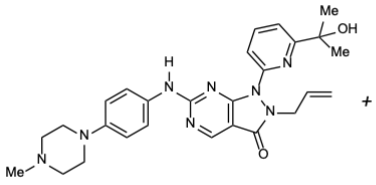
AM Bran, TA Neukomm, DP Armstrong, Z Jončev, P Schwaller

arXiv preprint arXiv:2503.08537




Steerable synthesis planning

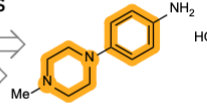
Top-ranked synthetic route



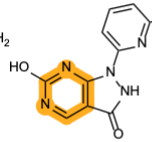
Wee1 kinase inhibitor by Merck

Expert query:
Break **pyrimidine** *in the* **early stage** *but get all other rings from commercially available materials.*

15 steps

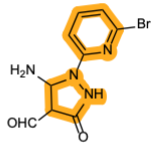


available



Query satisfied

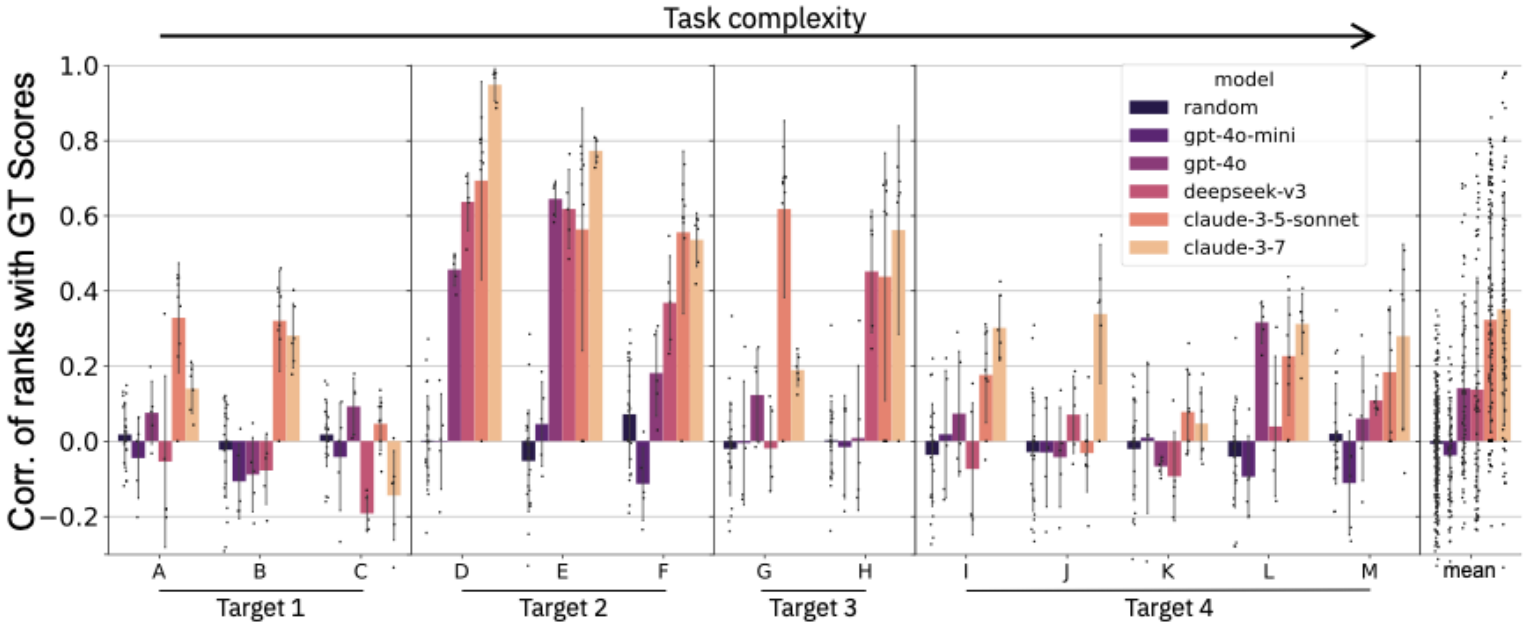
last step



available

- Natural language
- Full route analysis
- Route selection

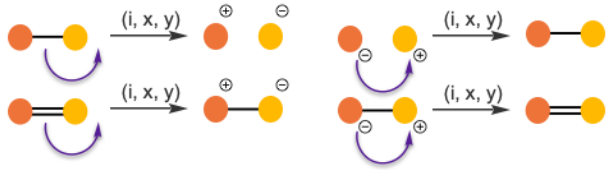
LLM score: 9/10

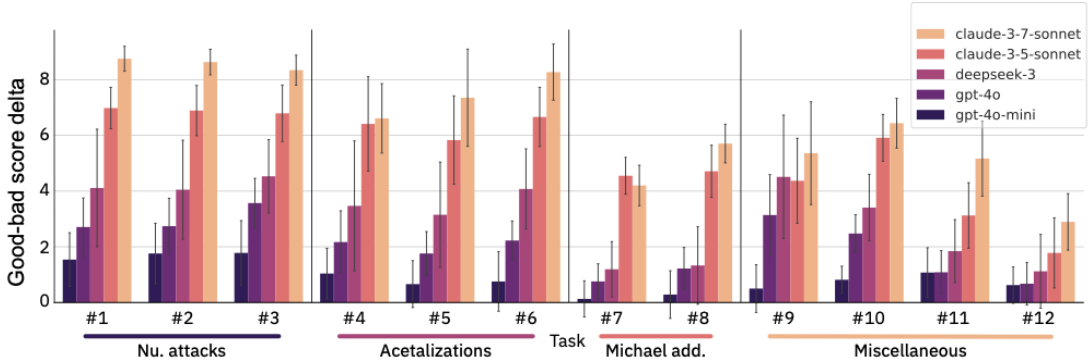


Reaction mechanism elucidation

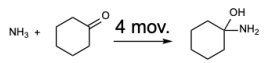
Actions: elementary steps

- Ionization moves
- Attack moves

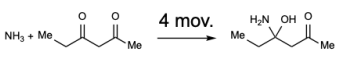




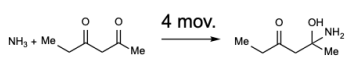
Task #1:
Nu attack of NH_3 on cyclohexanone:



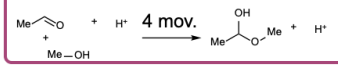
Task #2:
Selective Nu attack of NH_3 on dione:



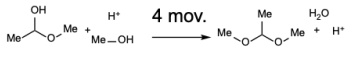
Task #3:
Selective Nu attack of NH_3 on dione:



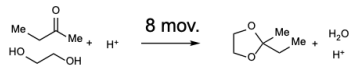
Task #4:
Hemiacetal formation:



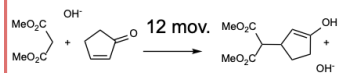
Task #5:
Hemiacetal to Acetal:



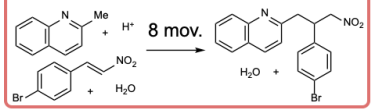
Task #6:
Intramolecular acetal formation:



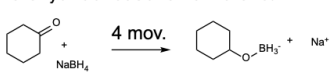
Task #7:
Enolate Formation + Michael Additon:



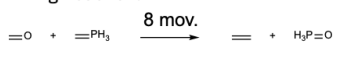
Task #8:
Tautomerisation + Michael Addition:



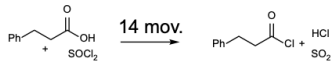
Task #9:
Borohydride reduction of ketone:



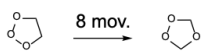
Task #11:
Wittig Reaction:



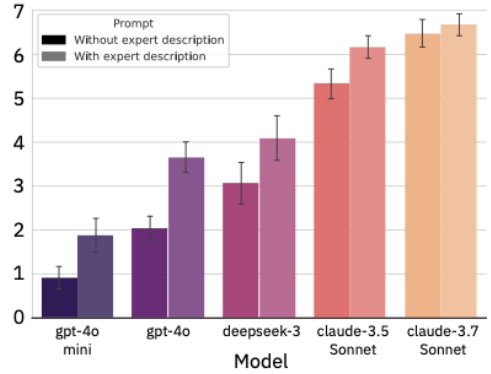
Task #10:
Acyl chloride formation with $SOCl_2$:



Task #12:
Transformation of molozonide to ozonide:



Nucleophilic additions | Acetalizations | Michael additions | Miscellaneous reactions



Expert reaction description in prompt helps weaker models.

Gold standard would be experimental validation.

But it takes time…

I have 5 different parameters to adjust for my reaction – how should I tune them?

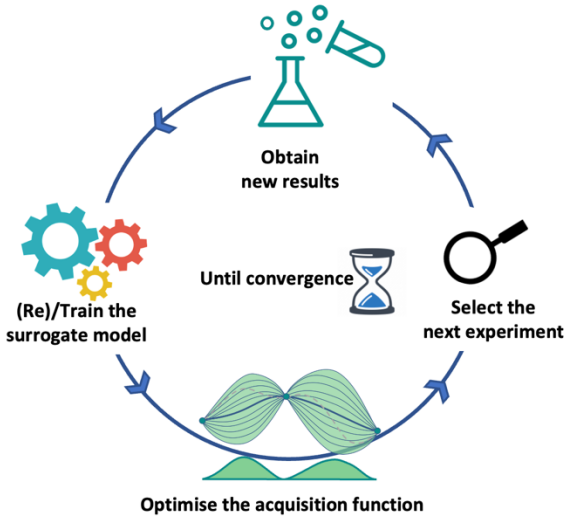
- Bayesian Optimization (using ML model to guide experiments) is currently popular for optimizing chemical reaction conditions/procedures

Article | Published: 03 February 2021

Bayesian reaction optimization as a tool for chemical synthesis

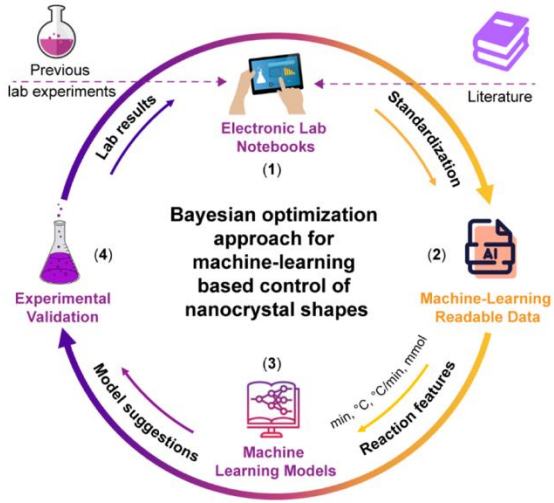
Benjamin J. Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I. Martinez Alvarado, Jacob M. Janey, Ryan P. Adams & Abigail G. Doyle

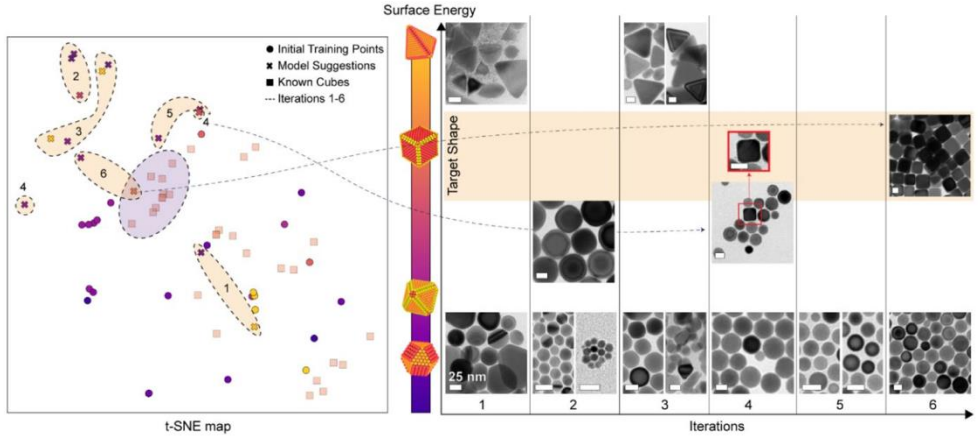
Nature **590**, 89–96 (2021) | Cite this article



ML for linking synthesis to outcome

Can we design Cu nanocrystals with a particular shape (=> reactivity)?

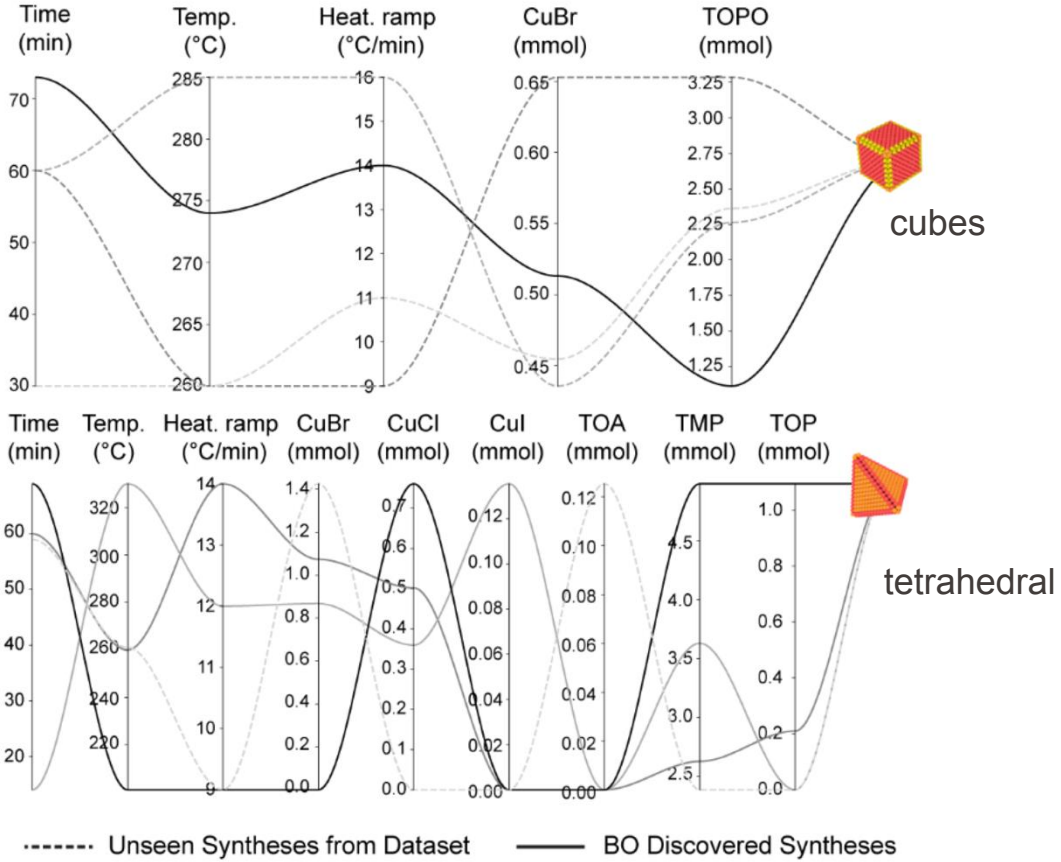


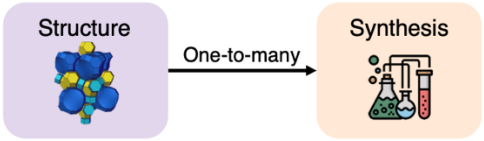


Bojana Ranković

A holistic data-driven approach to synthesis predictions of colloidal nanocrystal shapes
LEF Zaza, B Rankovic, **P Schwaller, R Buonsanti** (JACS)

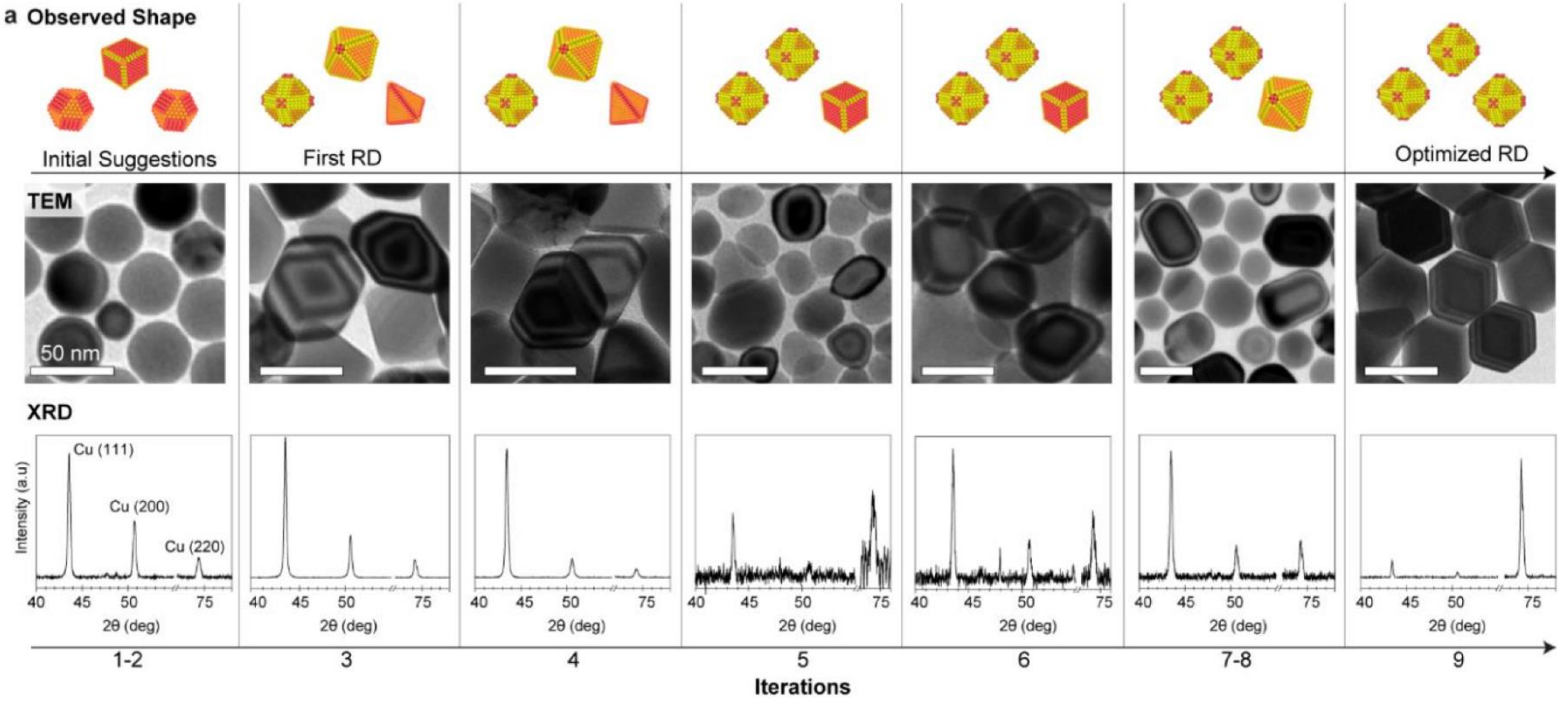
Multiple syntheses lead to the same outcome (ground truth not unique)





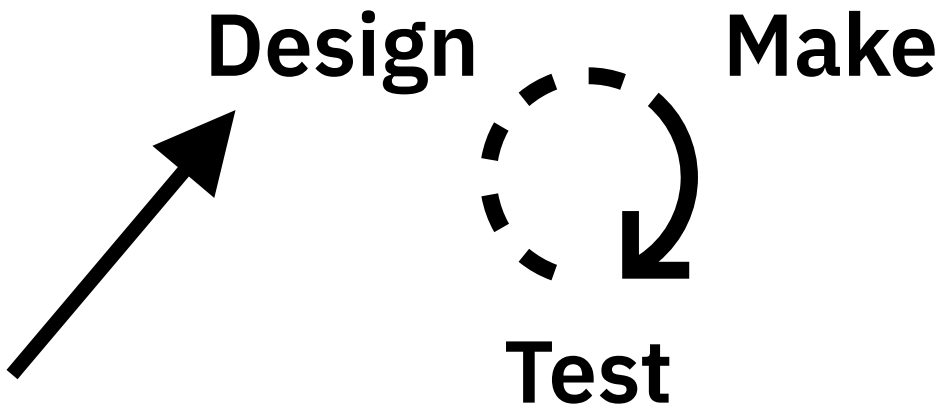
Elton Pan (NeurIPS 2024, AI4Mat workshop)

Discovering a new shape (rhombic dodecahedral)



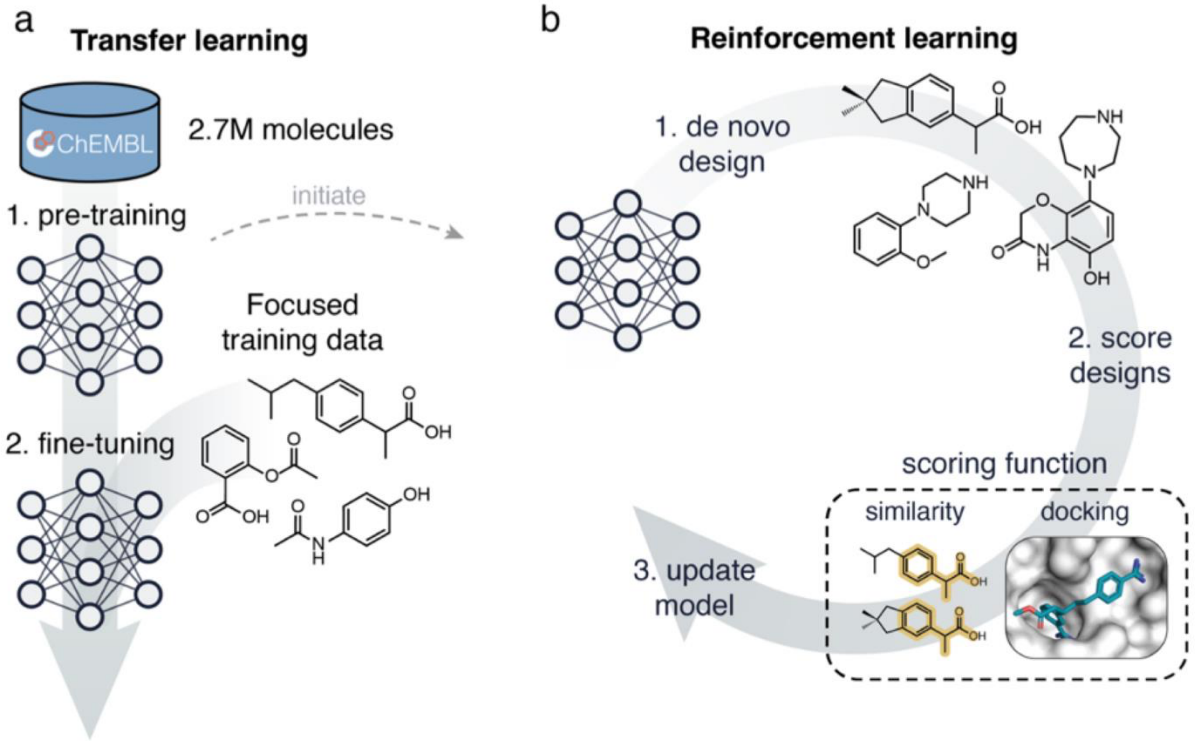
What molecule to make?

How to make it?



Experimental validation

Generating novel molecules with desired properties



Distribution learning

Goal-directed learning

Key Consideration

How correlated is your *in silico* predictor with the desired end-point?

- Grisoni, 2023

Towards Direct Optimization of High-fidelity Oracles

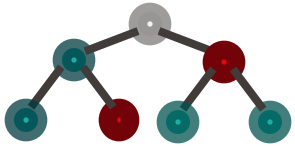
Oracle: Computational prediction or simulation

Sample Efficiency: How few oracle evaluations are required to optimize the objective?

Increasing predictive accuracy *but* also computational cost

Drug Discovery

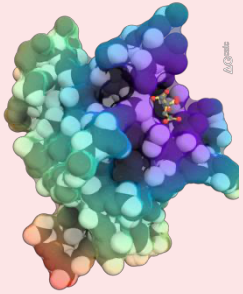
Predictive Models



*Can be accurate but may have narrow domain of applicability

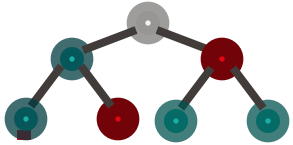
Molecular Docking

MMPB(GB)SA



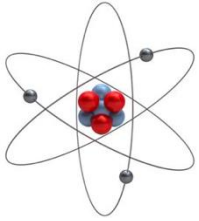
Free Energy Calculations

Materials and Catalyst Design

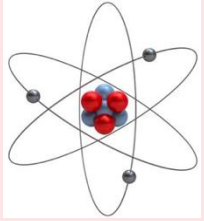


Semi-empirical QM

Single-point DFT



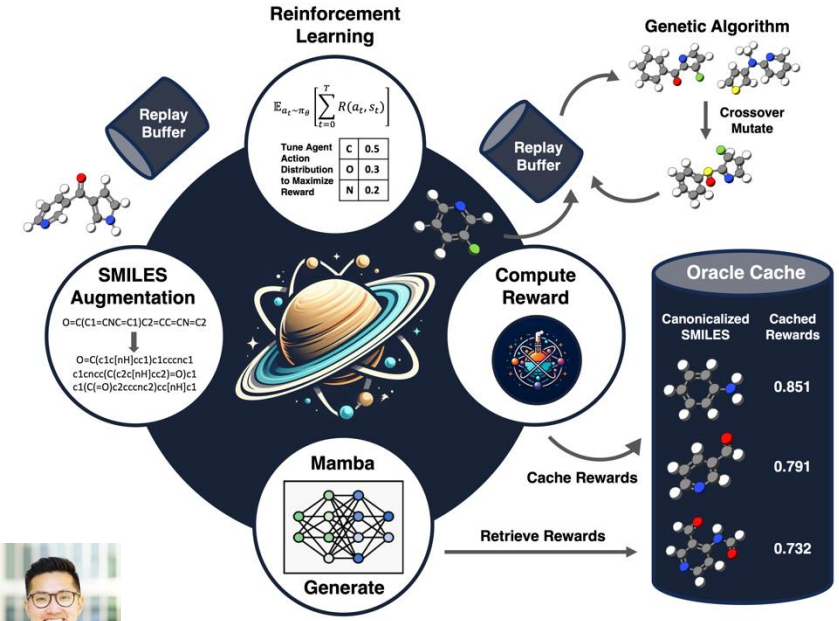
DFT – Varying Functionals / MD



Tackling sample efficiency for high-fidelity feedback

Augmented Memory & Saturn

- State of the art in sample efficiency



Sample efficiency benchmark (PMO, NeurIPS `22)

Model	Rank (/30)	Score
Augmented Memory	1	15.002
REINVENT [4]	2	14.016
SynNet [8] [ICLR `22]	12	11.498
DoG-Gen [9] [NeurIPS `20]	13	11.456
DST [10] [ICLR `22]	15	10.989
MARS [11] [ICLR `21]	16	10.989
MIMOSA [12] [AAAI `21]	17	10.651
DoG-AE [9] [NeurIPS `20]	20	9.790
GFlowNet [13] [NeurIPS `21]	21	9.131
GA+D [14] [ICLR `20]	22	8.964
GFlowNet-AL [13] [NeurIPS `21]	27	8.406
JT-VAE [15] [ICML `18]	28	8.358

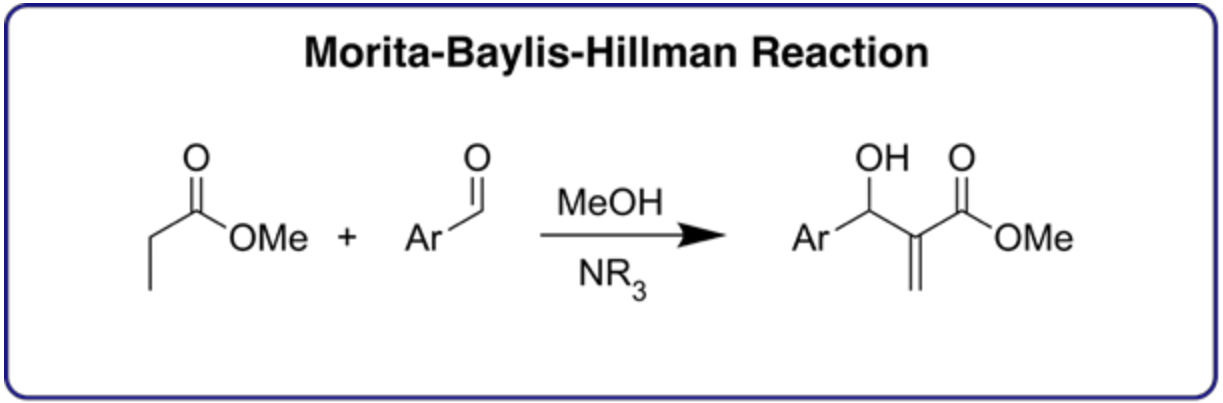
RNN-based

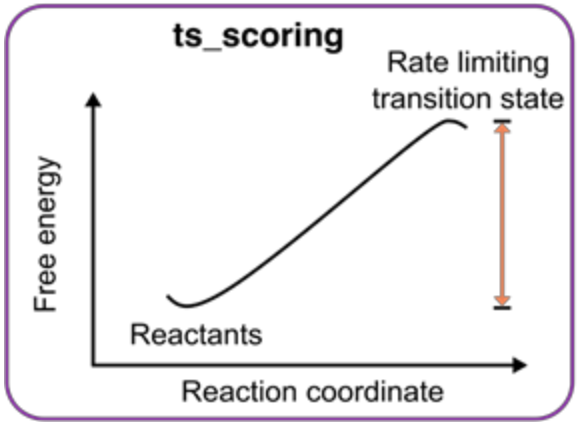
Approaches from big machine learning conferences.

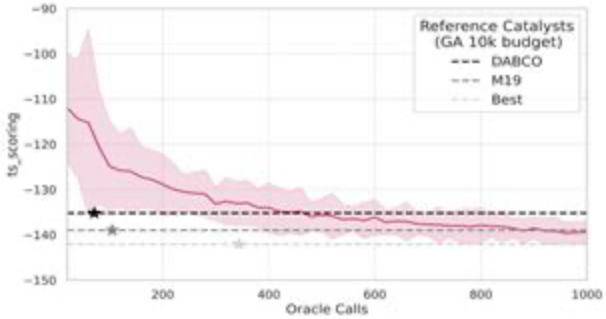
10k oracle calls, 23 tasks

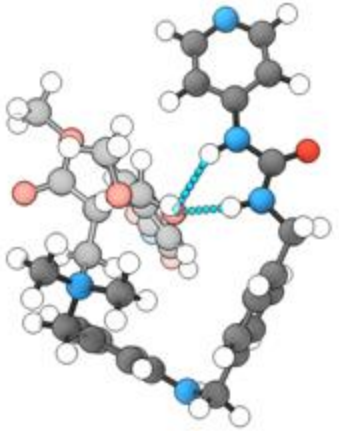
Guo, J. & Schwaller, P. *JACS Au*, **2024**.
Guo, J. & Schwaller, P. *ArXiv*, **2024**.

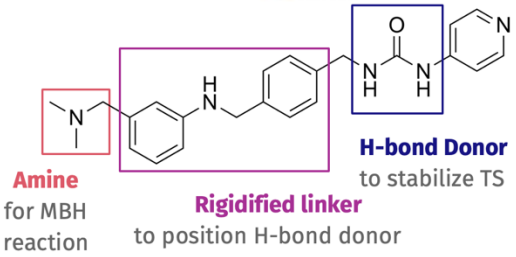
Designing a catalyst for the Morita-Baylis-Hillman Reaction











- 2D SMILES generator → 3D function

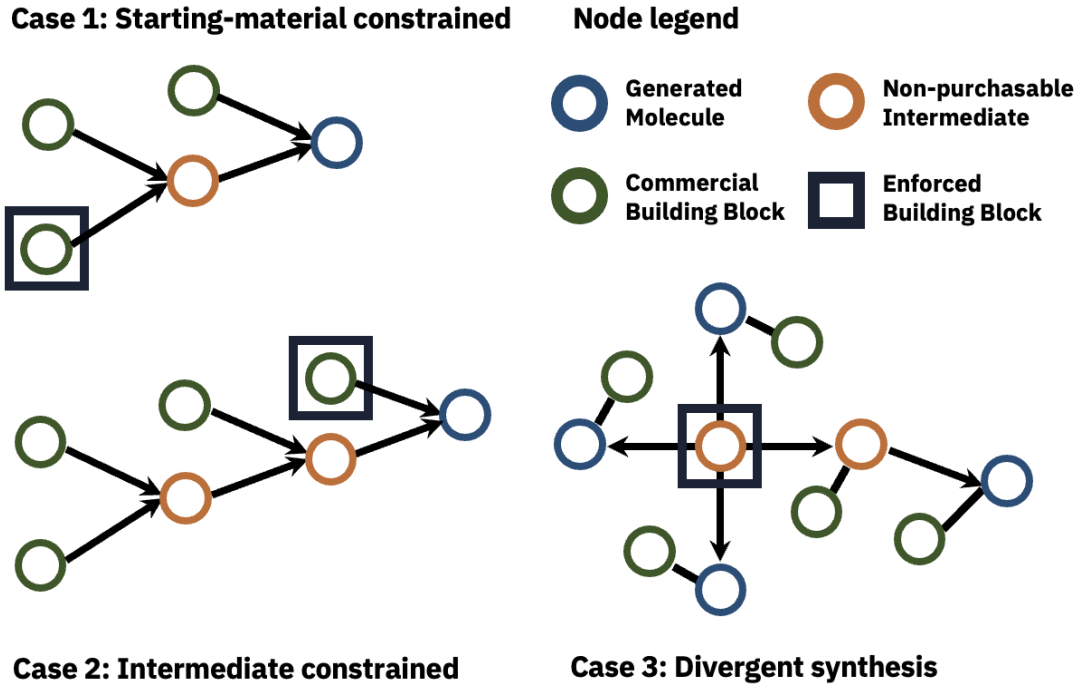
- Better score than best-known catalyst.
- 1k oracle budget, compared to 10k
- Using **Saturn** (Jeff Guo)

Sarina Kopf
(with Nevado group)

Seumer et al., *Angew. Chem.* **2023**

Towards de novo molecular design with experimental validation

Synthesisability & experimental validation is the bottleneck
Saturn's sample efficiency enables **directly optimizing for synthesizability** using retrosynthesis models (https://arxiv.org/abs/2407.12186)

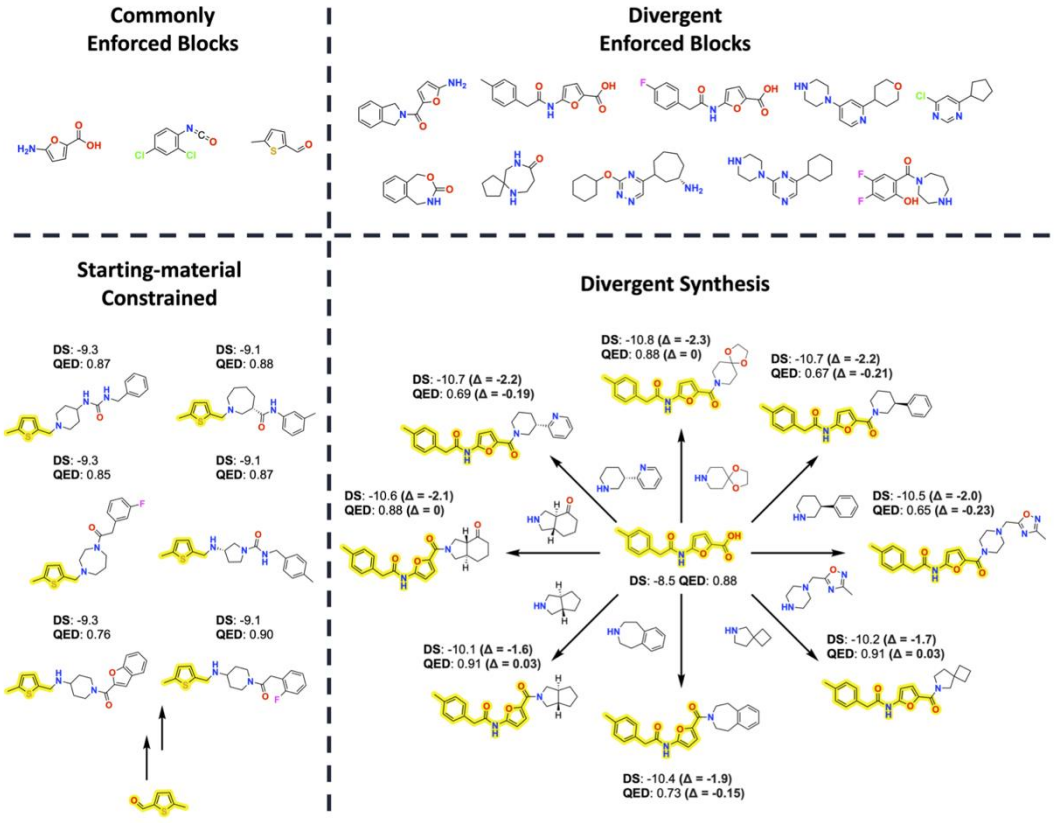


- Upgrading **bio-based building blocks**
- **Improving hits**
- Starting from **available building blocks** in lab

Jeff Guo

Guo, J. & Schwaller, P. *ArXiv*, **2024**.

It takes two to TANGO – enforcing building blocks in synthesis routes

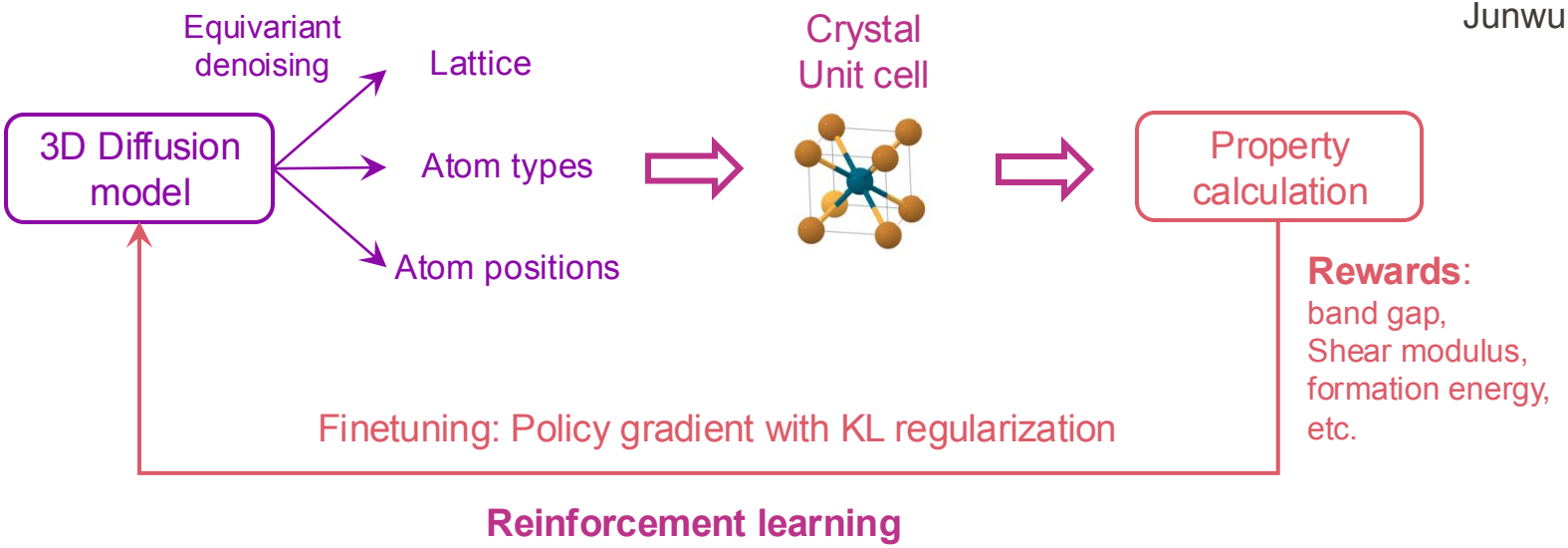


Guo, J. & Schwaller, P. *ArXiv*, **2024**.

Goal-directed learning for de novo crystal generation

Learning to make materials with targeted property profiles

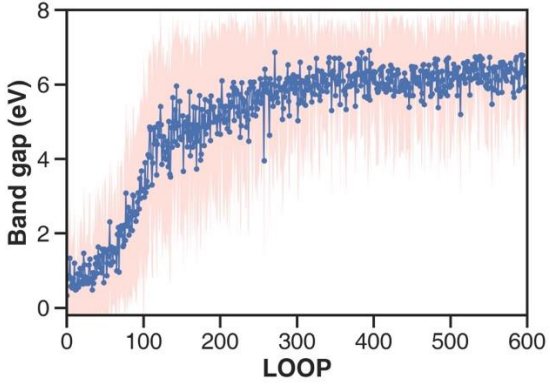
Junwu Chen



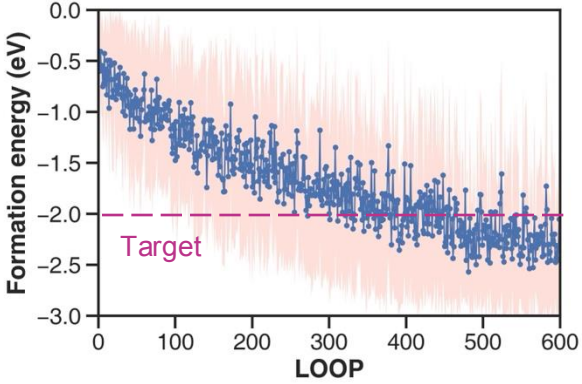
- (unpublished, preliminary work)

Crystal Property Optimization

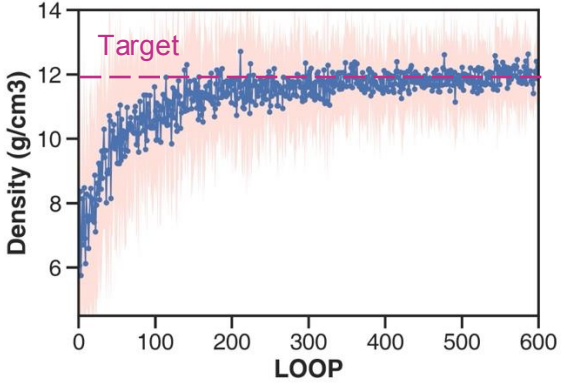
Target: higher band gap



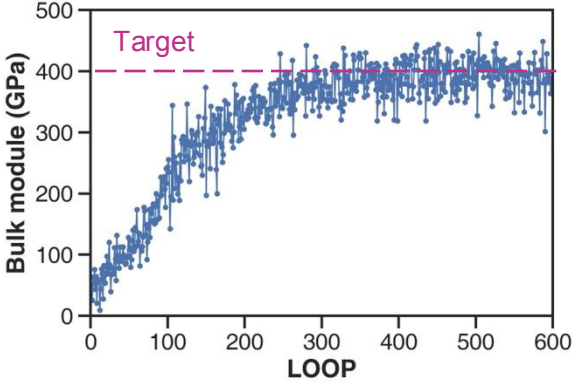
Target: FE lower than -2.0 eV



Target: density = 12g/cm^3

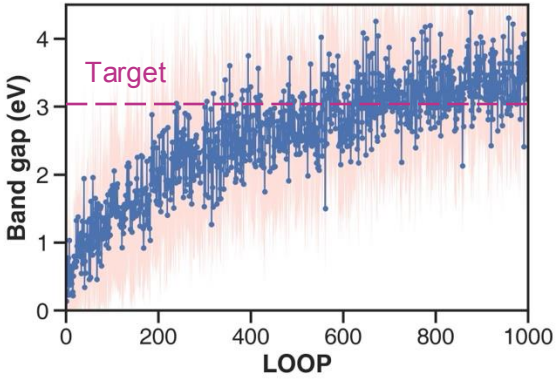


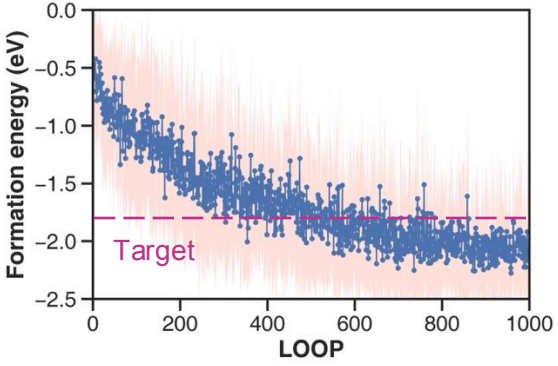
Target: Bulk module = 400GPa



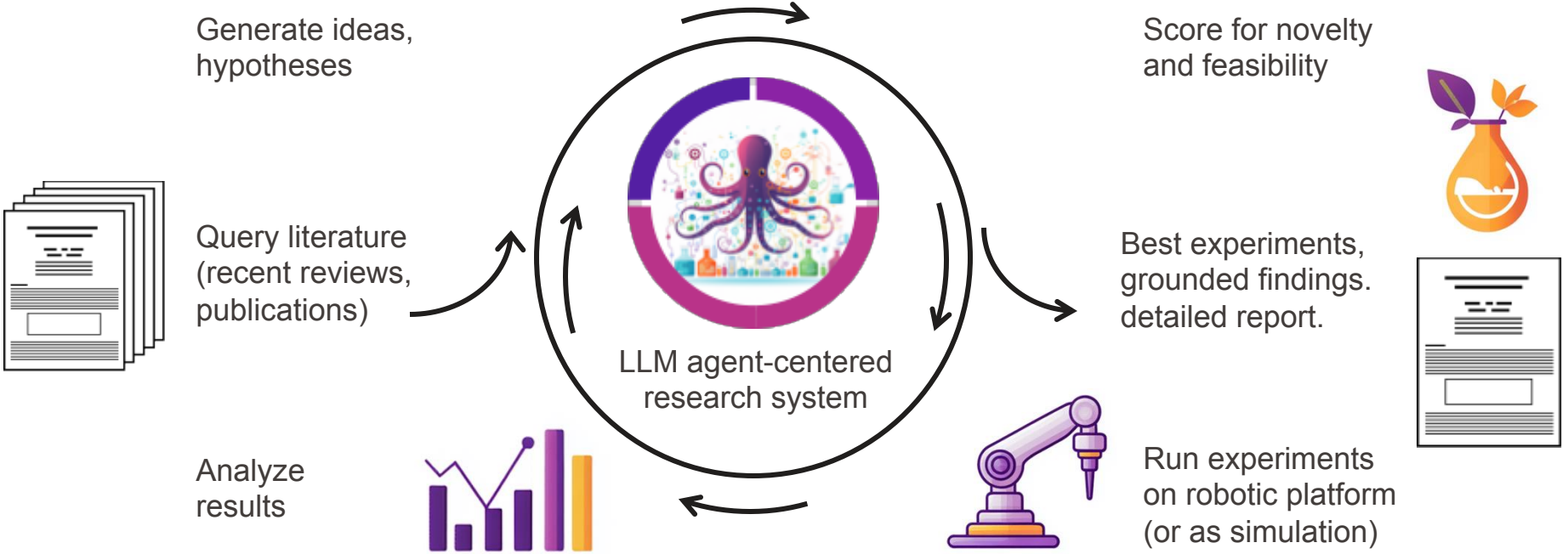
Multiple properties

Target: 1) gap = 3.0 eV
2) FE < -1.8 eV





Open scientific agentic systems with experimental feedback



Summary of whirlwind intro to ML

- Many flavours: supervised (examples with labels), unsupervised (example without labels), self-supervised (examples with artificial labels), reinforcement learning (reward from environment)
- Traditional ML (human expert features) → deep learning (features learned from data)
- You always need training data!
- Recent work goes beyond simple regression and classification task. ML enables you to generate ideas for novel molecules/materials, synthesis routes to never synthesised molecules, etc…
- Programming is needed to do ML in Chemistry.
- If you are excited about this direction, this course was only the beginning of your journey.

LIAC Team

PhD students
Bojana Ranković
Oliver Schilter (IBM Research)
Andres CM Bran
Junwu Chen
Jeff Guo
Victor Sabanza Gil (Luterbacher)
Paulo Neves (Janssen)
Rebecca Neeser (Correia, VantAI)
Sarina Kopf (Nevado)
Daniel Armstrong
Joshua Sin (Roche)
Sacha Raffaud
Sandro Agostini (IBM Research)
Théo Neukomm (Intel/Merck)
Salomé Guilbert (Röthlisberger)
Matt Hart (Trospha)

Admin
Annick Delmonaco

Postdocs/Engineers
Zlatko Jončev
Edvin Fako
Jeremy Goumaz

Project students
Shai Pranesh
Octavian Susanu
David Segura
Vu Nguyen

>15 nationalities — one team!
https://schwallergroup.github.io

Funding:
NCCR Catalysis
Swiss National Science Foundation
IBM Research
janssen
VANTAI
Roche
intel
MERCK

Presentations on **26.05.25 (CM 1 4, 11.15-13h)** and **28.05.25 (here, 11.15-13h).**

- **16 May 2025 (end of day, CET):** Complete project information -> one entry per team in a Google form. We will share the form closer to that date. First come, first serve for date preference.
- **23 May 2025 (end of day, CET):** Code repository including Jupyter notebook-based report. Changes after 23 May 2024 will not be considered for grading. The repository will have to **be public**.
- **26 and 28 May 2025:** Presentations during 2 lectures → more information on exact timing will follow, but roughly 4.5 minutes per team.
- Office hours: **Thursdays 13h-14h**, please write us beforehand
- Any email with project questions should contain **CH-200 in the subject,** and be sent to **me and all TAs.**