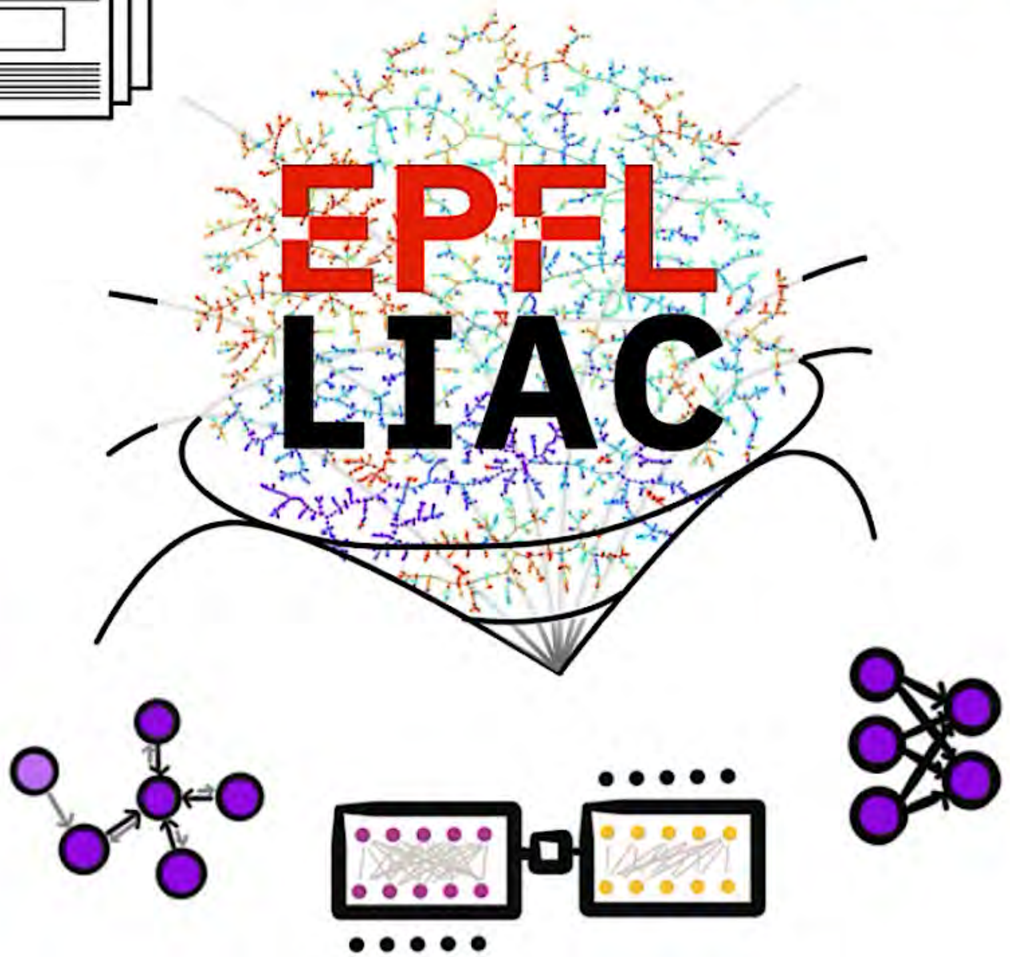
US20030166932A1: General Procedure H A solution of trifluoromethanesulfonic acid 3,5,8,8-tetramethyl-7,8-dihydronaphthalen-2-yl ester (Compound 35, 0.41 g, 1.2 mmol), Pd(OAc)2 (0.027 g, 0.12 mmol), BINAP (0.11 g, 0.18 mmol), Cs2CO3 (0.56 g, 1.72 mmol), ethyl 4-aminobenzoate (0.25 g, 1.5 mmol) and 5 mL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



P. Schwaller (@SchwallerGroup)

Language models for accelerated discovery and synthesis

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

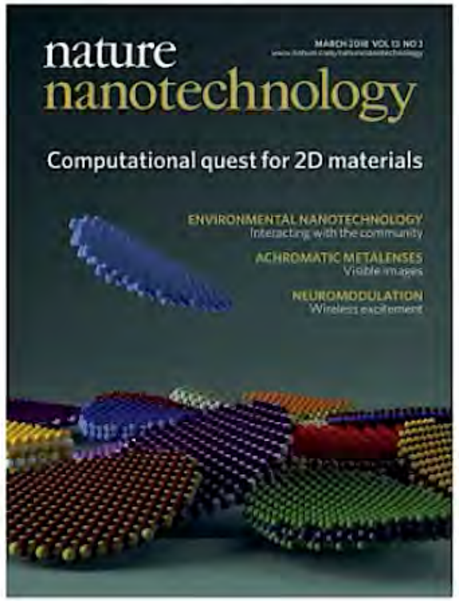
EPFL

MANCHESTER 1824
The University of Manchester

b UNIVERSITÄT BERN

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

Closing the loop...



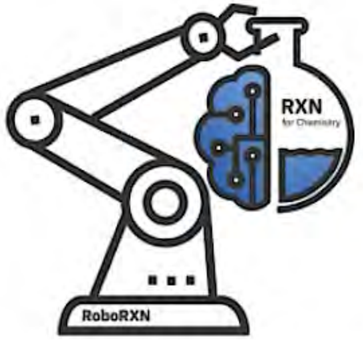
Materials Science & Engineering

Virtual screening & simulation workflows
Prof Nicola Marzari

BSc ('14)/MSc ('16)

MPhil in Physics ('19)
Dr Alpha Lee

UNIVERSITY OF CAMBRIDGE



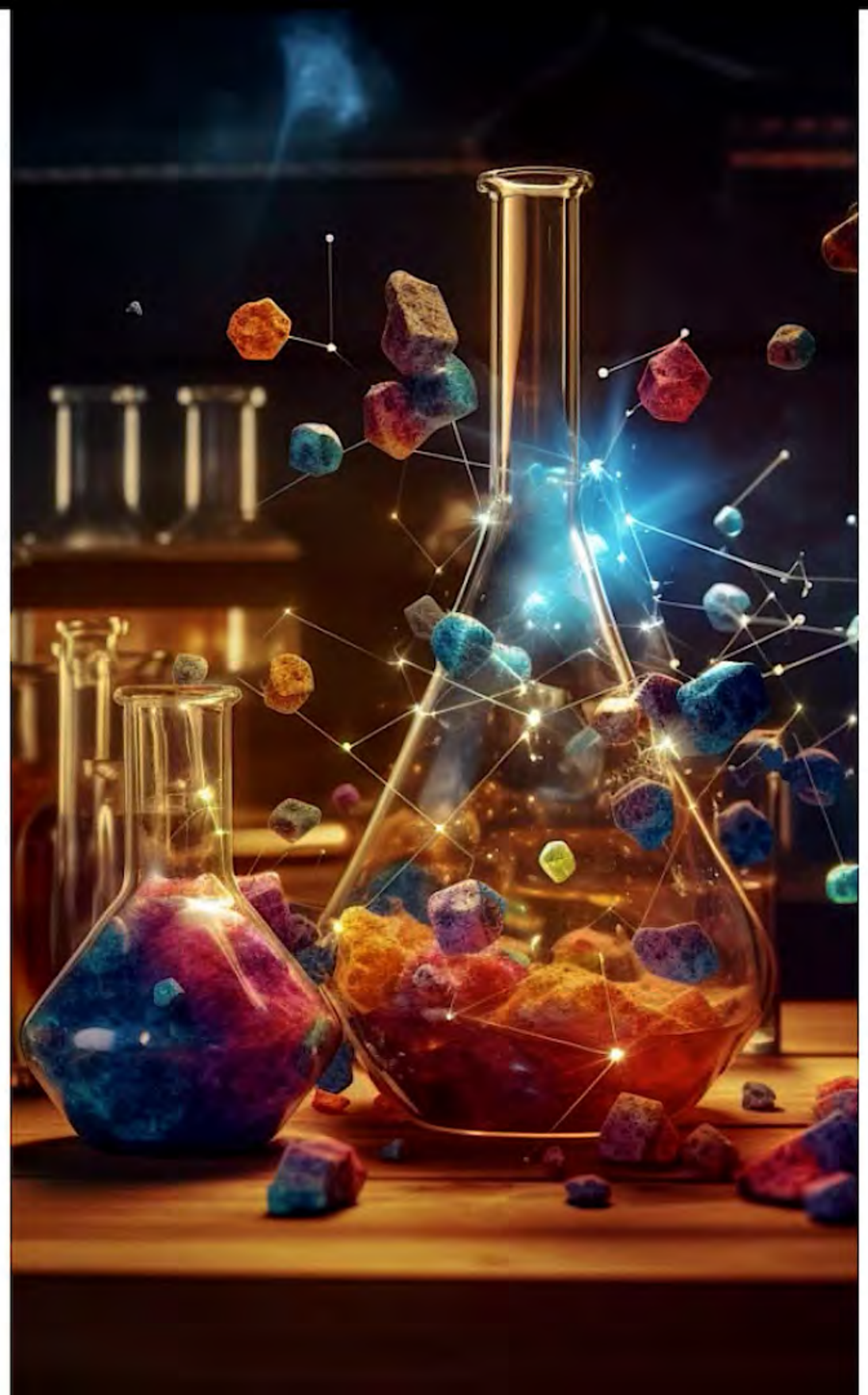
Machine learning for chemical synthesis
Intern/PhD/Postdoc
Dr Teodoro Laino

Empa
Materials Science and Technology

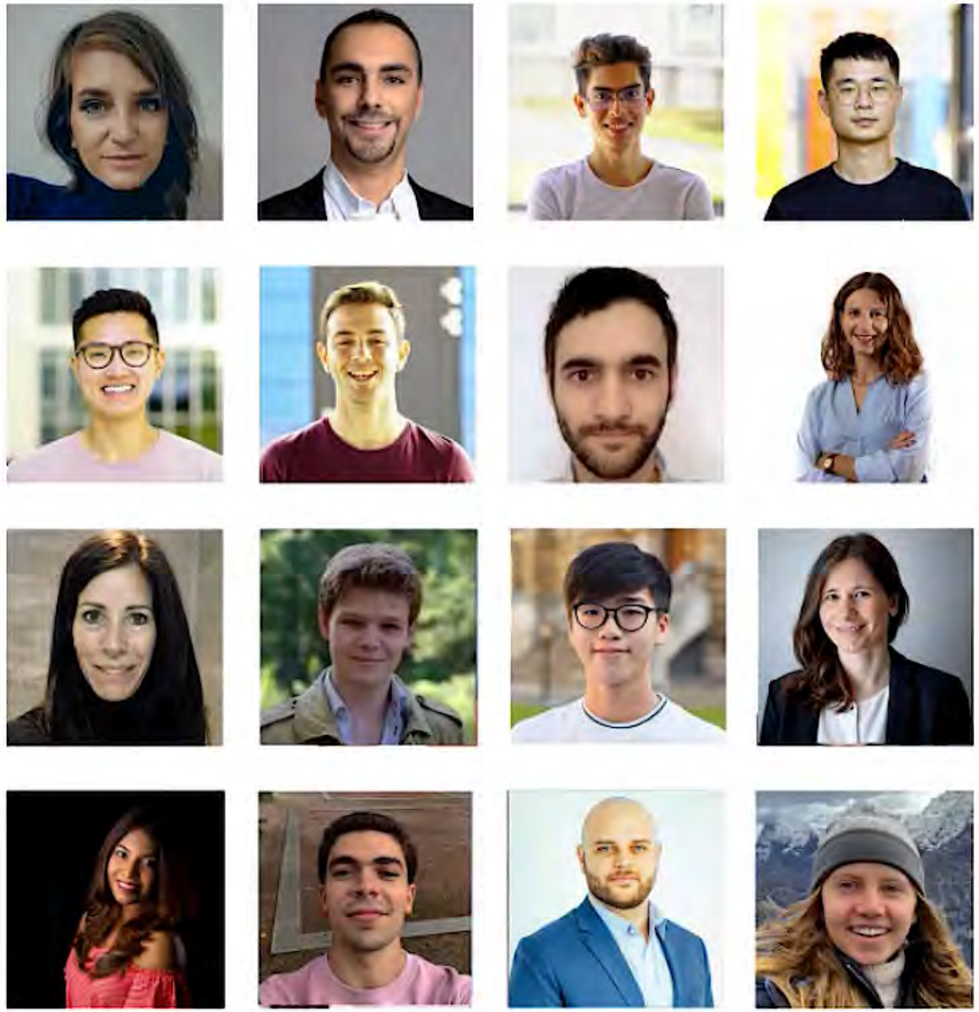
Lab work on ternary polymer blends for organic solar cells
Prof Frank Nüesch

IBM **Research** Europe

Mission:
Accelerating Discovery with Digital Chemistry



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)

Admin
Annick Delmonaco

Postdocs
Zlatko Jončev
Marta Brucka

Project students
Jenestin Anthonipillai
Remi Schlama
Anna Borisova
Magdalena Lederbauer
Tugdual Kerjan

>10 nationalities — one team!
https://schwallergroup.github.io

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

Exploring the nearly *endless chemical space*

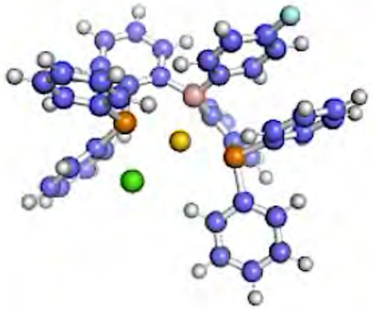
10^{60} drug-like molecules

Milky-way above the Swiss Alps, Schafarnisch, 2022

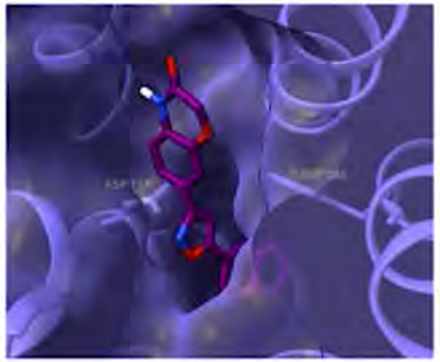
Molecule/Materials Design Cycle

What molecule to make?

How to make it?



Catalysis

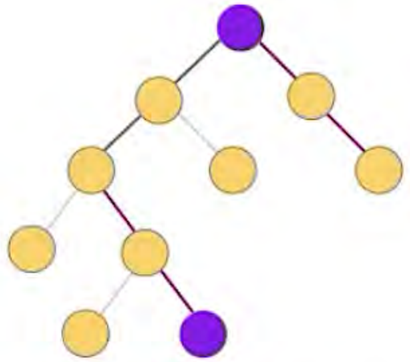


Drug discovery

Design

Make

Test



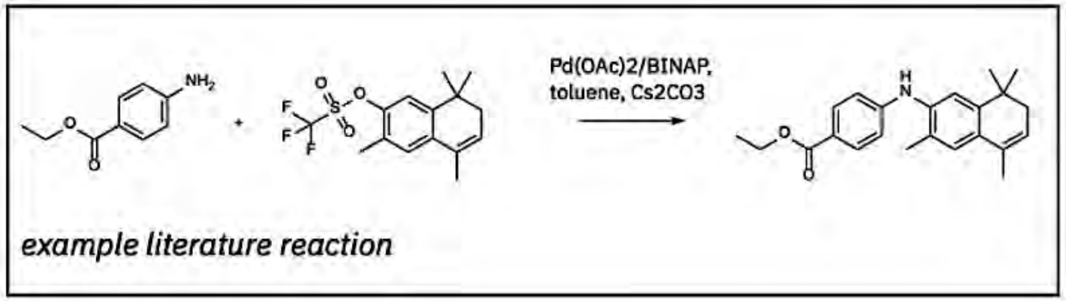
Synthesis planning

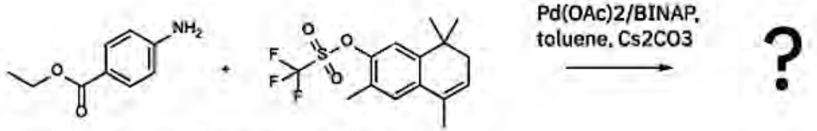
Synthesis optimization

Saturn (Sample-efficient de novo molecular design)

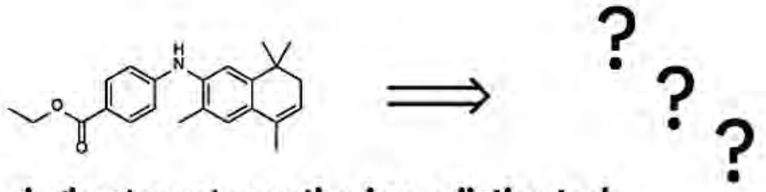
Experimental validation

ChemCrow (LLM agents for chemical research)

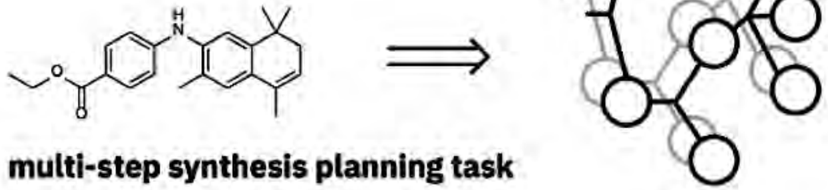




forward reaction prediction task



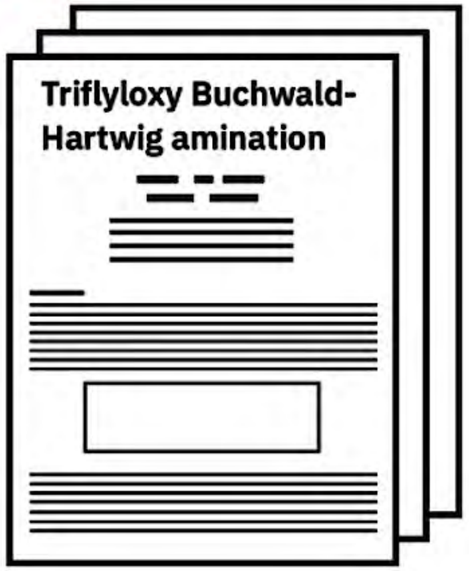
single-step retrosynthesis prediction task



multi-step synthesis planning task

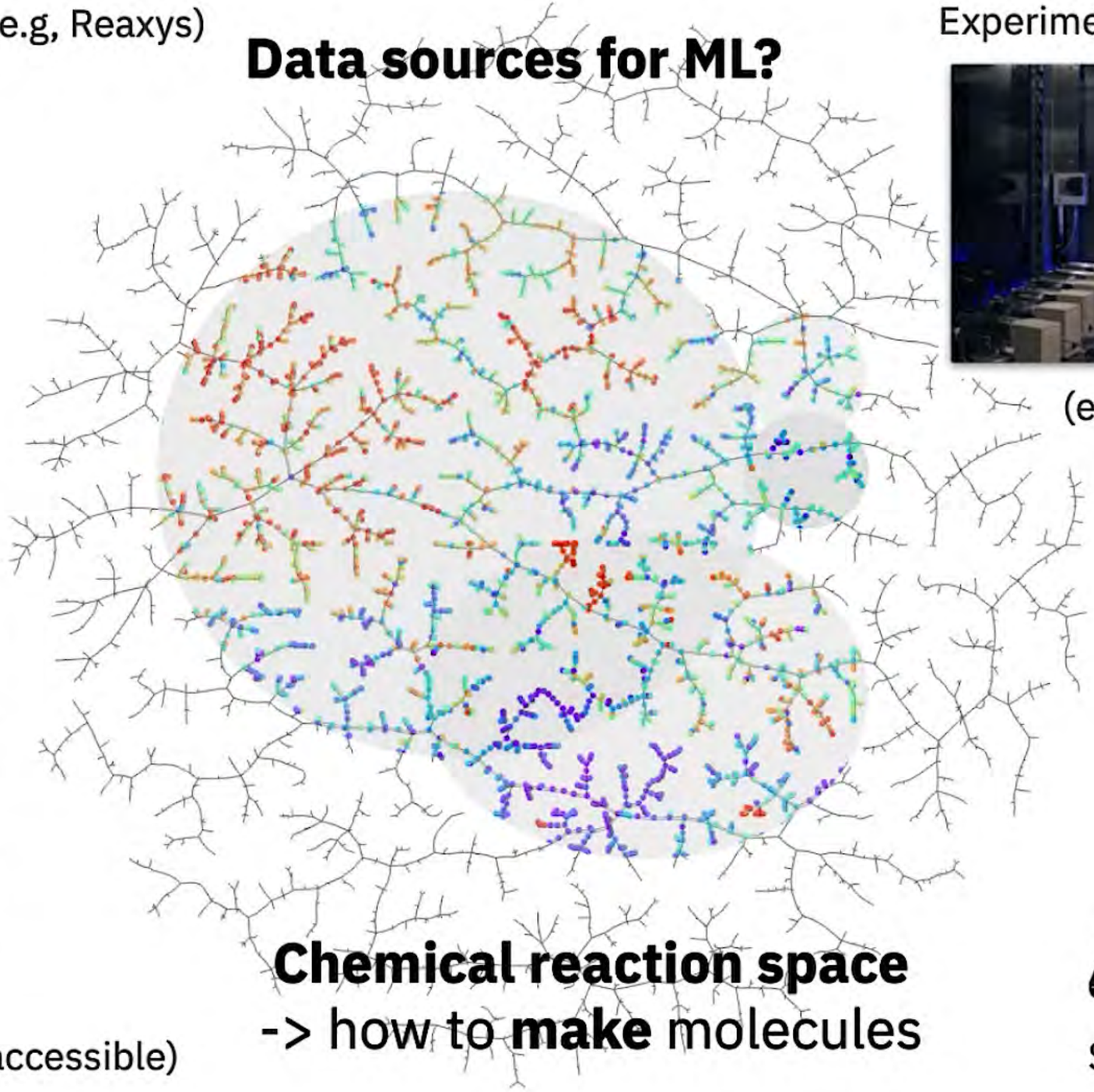
Data sources for ML?

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.

Patents (broad, accessible)



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

Experiments ELN/HTE (narrow)

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

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

Simulations (narrow)

Patent *reaction data*

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. ...

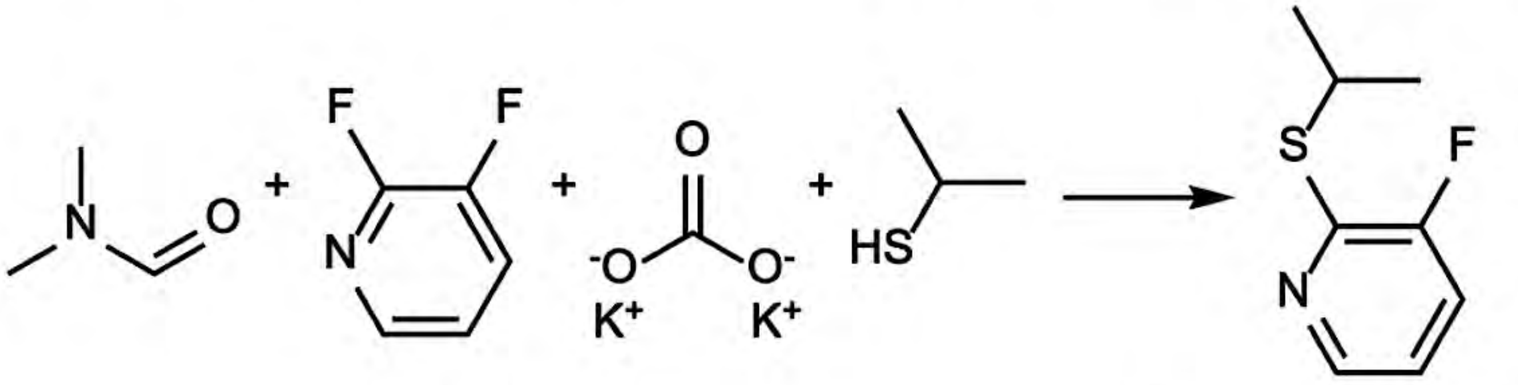
Daniel Lowe & Roger Sayle

Text-mining

Millions of reaction equations and procedures.

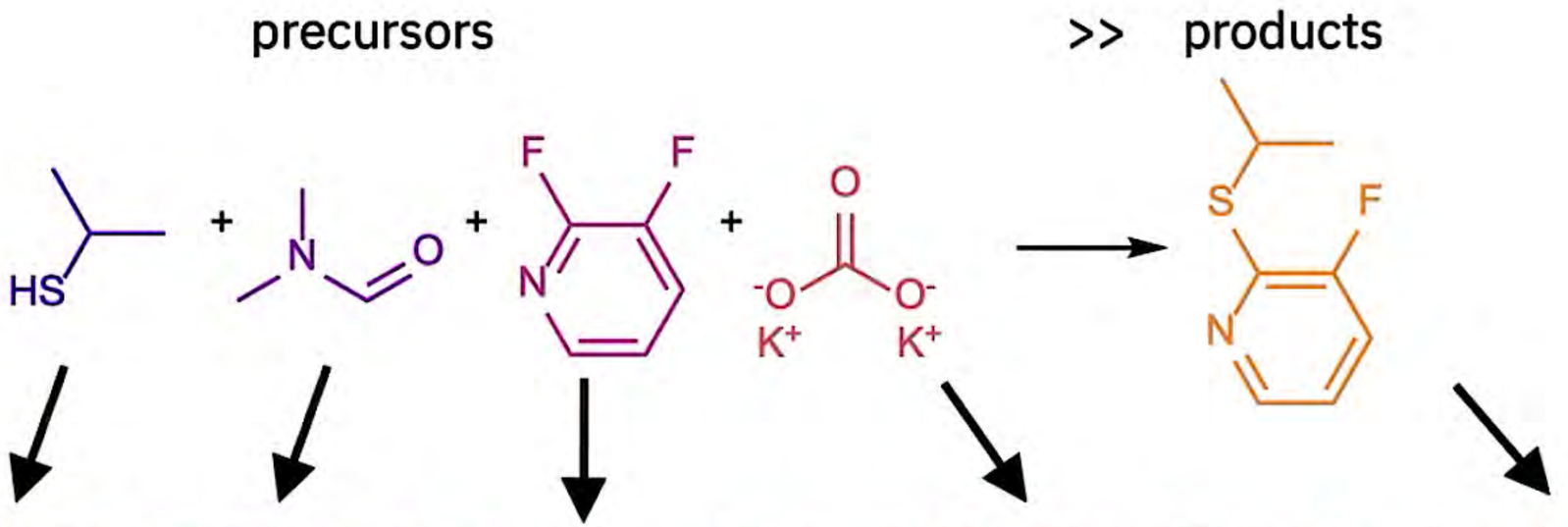
Reaction SMILES

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



precursors >> products

Atoms as *letters*, molecules as *words*



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

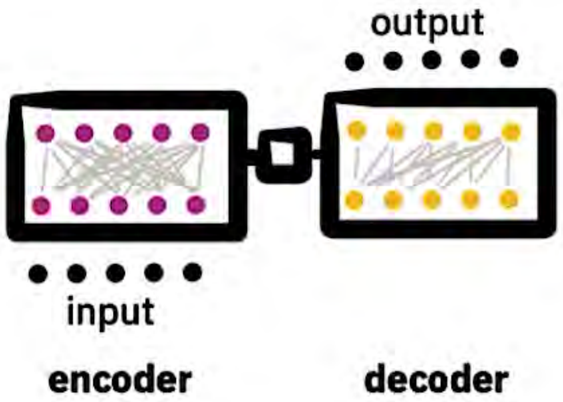
Split -> “sequences of atoms” called tokens

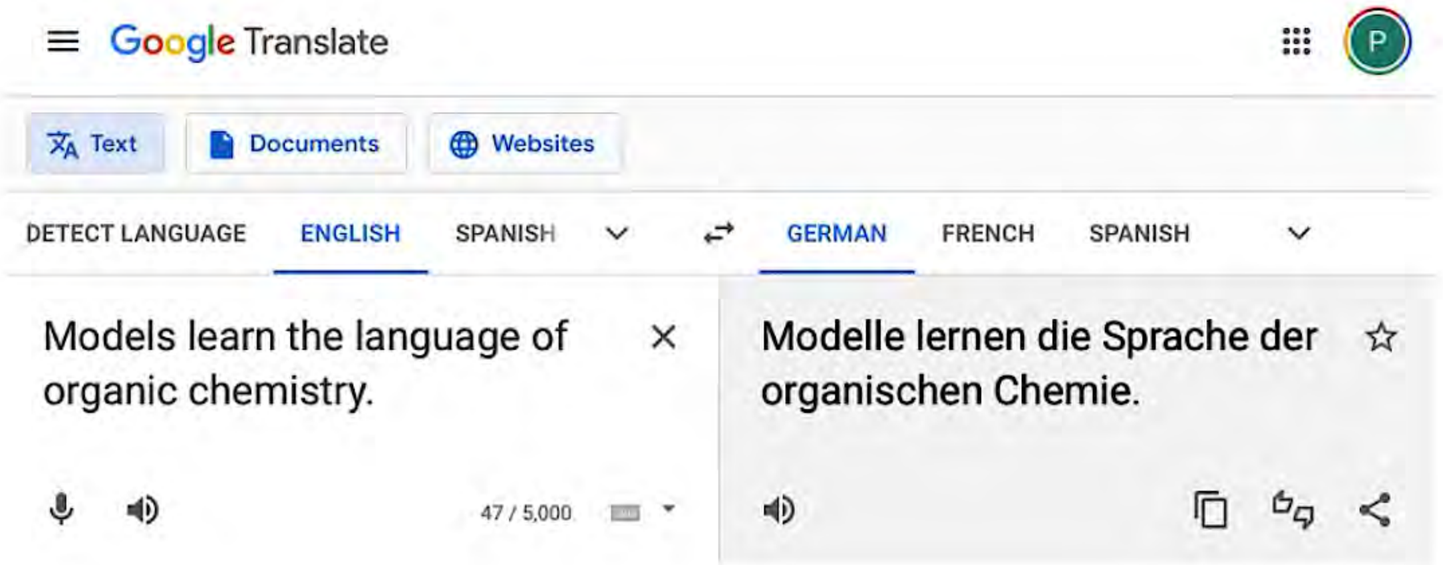
C C (C) S . C N (C) C = O . F c 1 c c c n c 1 F . O = C ([O-]) [O-] . [K+] . [K+] >> C C (C) S c 1 n c c c c 1 F

→ Borrow methods developed for human languages

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

Transformer architecture





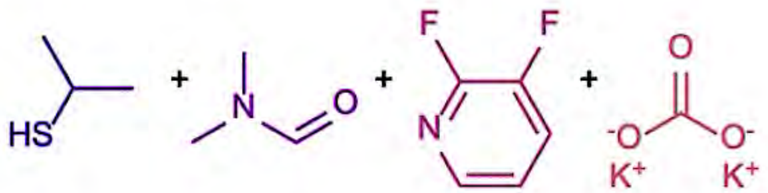
- Learns translation from examples

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

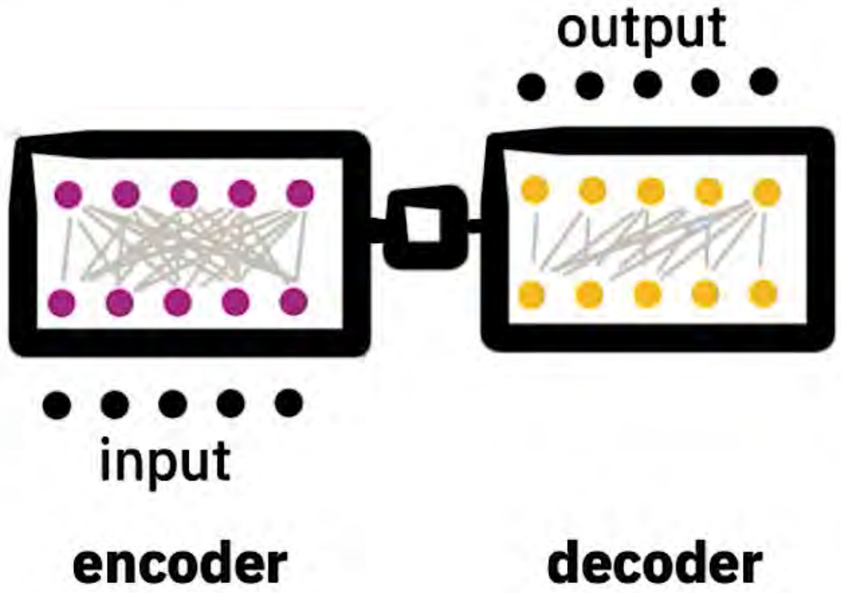
EPFL

Molecular Transformer

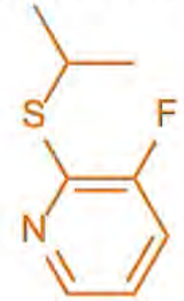
precursors



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 (top on benchmarks, back then)
- Better than rule and graph-based approaches

P. Schwaller (@SchwallerGroup)

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

UNIVERSITY OF CAMBRIDGE

IBM **Research**

Methods	Top-n accuracy (%)	
	1	
USPTO_480k_mixed		
MEGAN (Sacha et al., 2021)	86.3	
Molecular Transformer (Schwaller et al., 2019)	88.6	
Graph2SMILES (D-GCN) (*ours*)	90.3	Transformer-decoder
Graph2SMILES (D-GAT) (*ours*)	90.3	Transformer-decoder
Augmented Transformer (Tetko et al., 2020)	90.6	Same architecture as Molecular Transformer
Chemformer (Irwin et al., 2021)	**91.3**	Same architecture as Molecular 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]

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

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

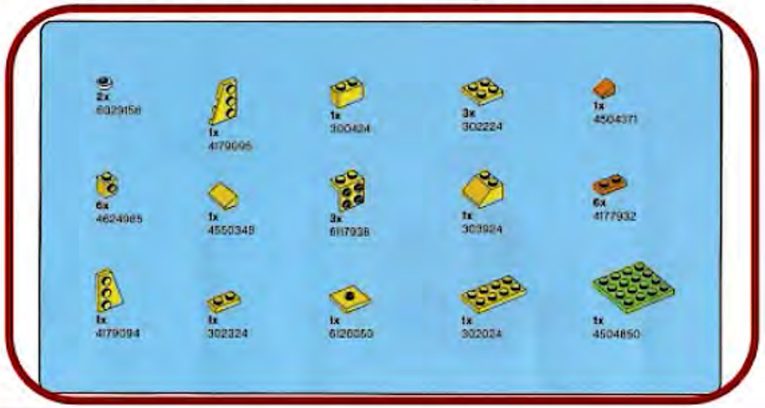
Mikołaj Sacha, Mikołaj Błaż, Piotr Byrski, Paweł Dąbrowski-Tumański, Mikołaj Chromiński, Rafał Loska, Paweł Włodarczyk-Pruszyński, and Stanisław Jastrzębski*

MEGAN

Retrosynthesis (Corey, Nobel prize, 1990)

Target molecule

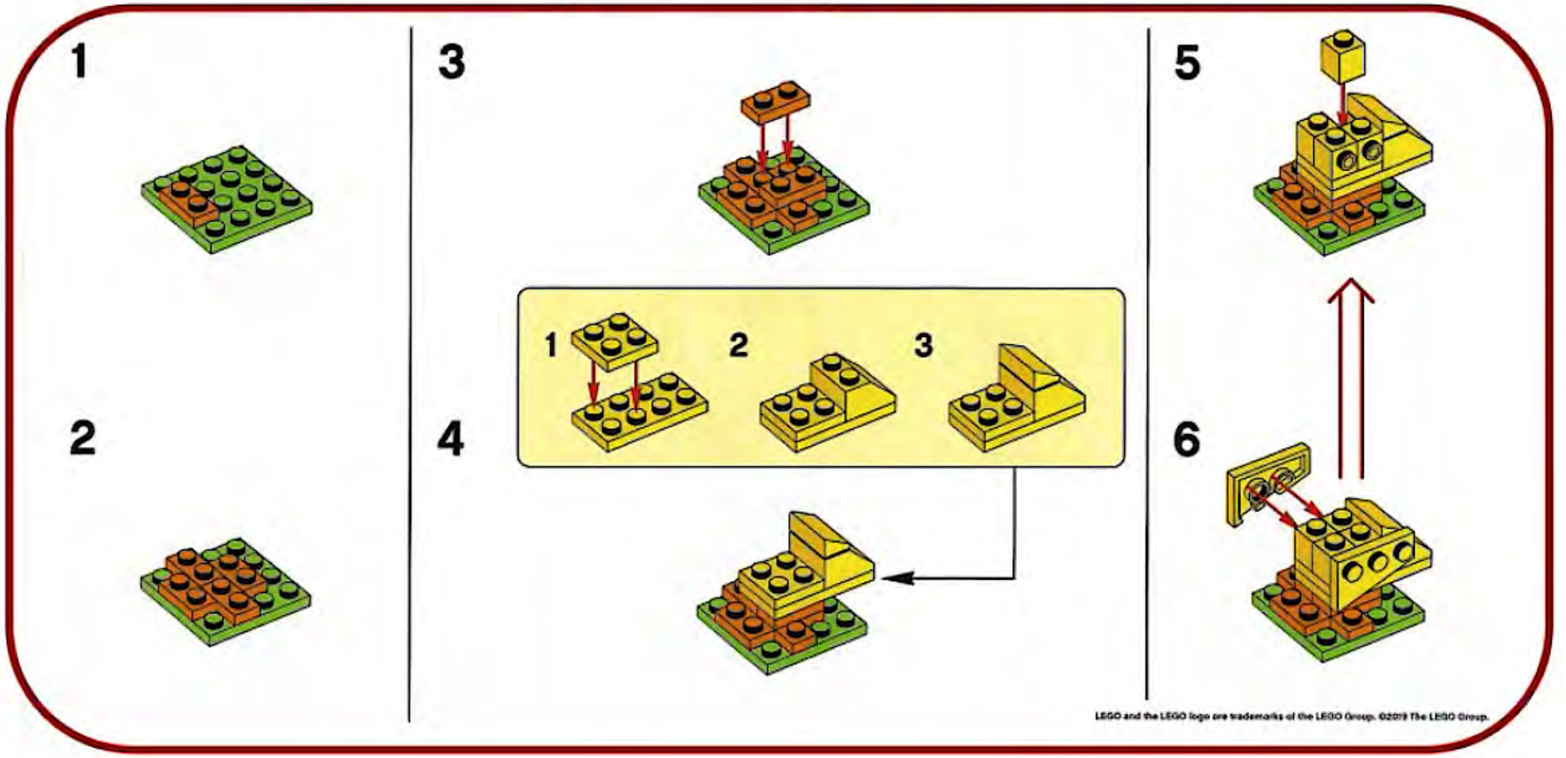
Known (commercially available) building blocks



Szymkuć et al., ACIE 2016
Grzybowski et al., Chem 2018
Mikulak-Klucznik et al., Nature, 2020

- expert rule-based
- "wide & deep" seach
- simulations, ML scoring

Steps to construct the target



Segler & Waller
Nature, 2018

- template-based (reactants)
- Monte-Carlo tree search
- step-scoring NN

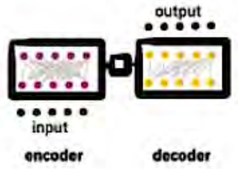
AiZynthFinder
Thakkar et al., Chem. Sci., 2020
Genheden et al., J. Cheminf., 2020

- template-based (reactants)
- Monte-Carlo tree search
- step-scoring NN
- interactive mode

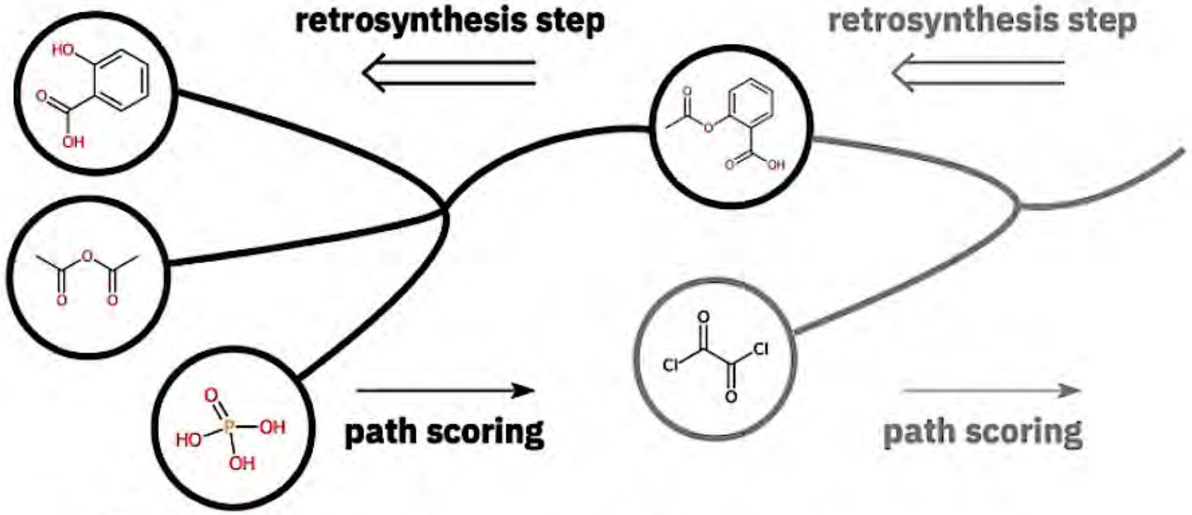
Lego analogy:
Amol Thakkar

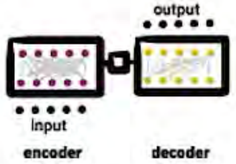
And more…

Multi-step synthesis planning



Molecular Transformer **Retro**





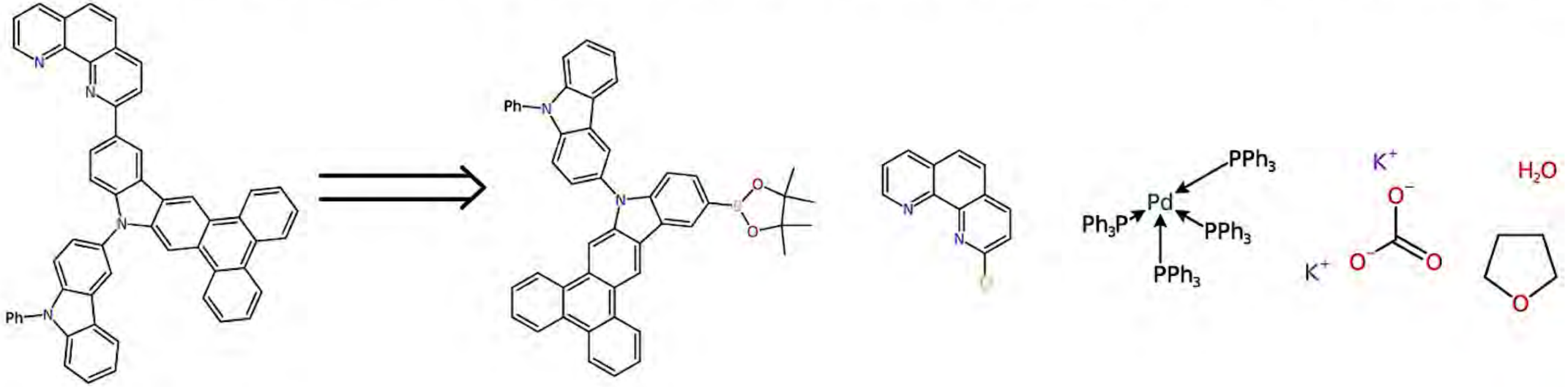
Molecular Transformer **Forward**

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

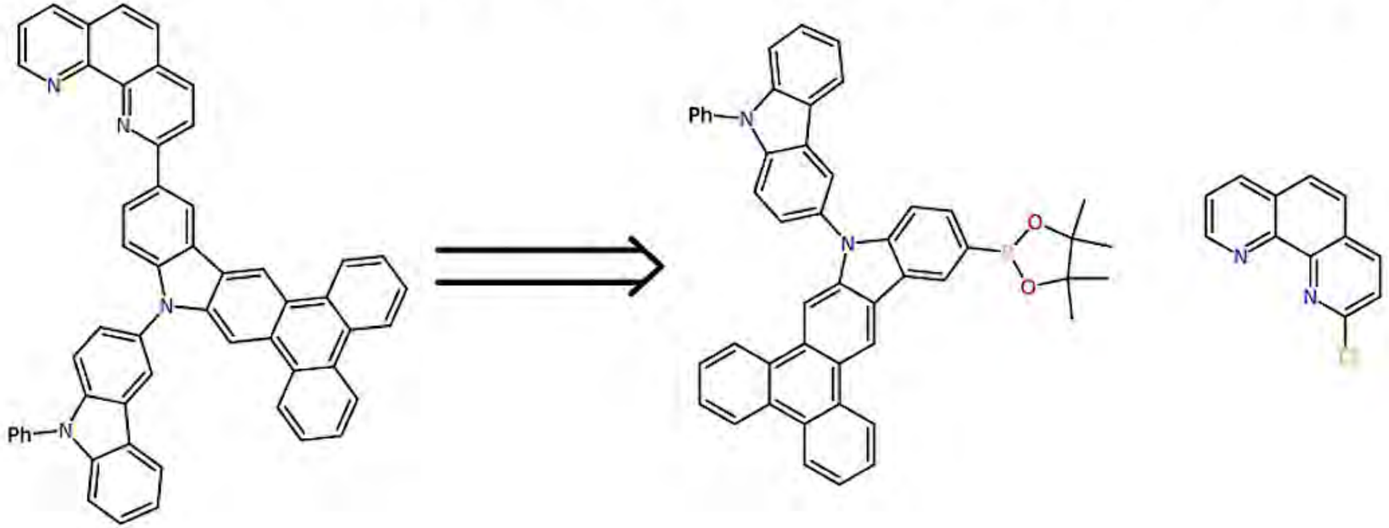
Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, Teodoro Laino

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

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



Reactants prediction (atom-mapping dependent)



Rule-based approaches:

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

and other SMILES-based approaches:

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

RoboRXN

Teodoro Laino & RoboRXN Team

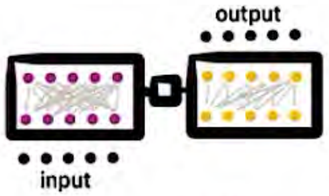
Chemists at home
- design of target molecule
- creative process

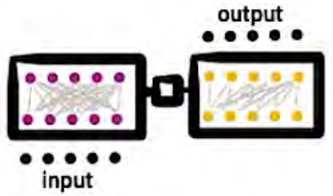
Cloud

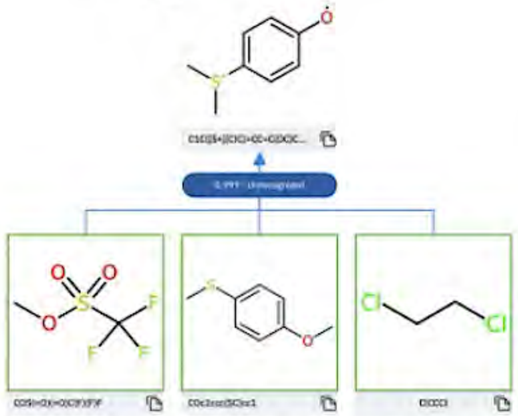
Remote AI-driven synthesis

Automation

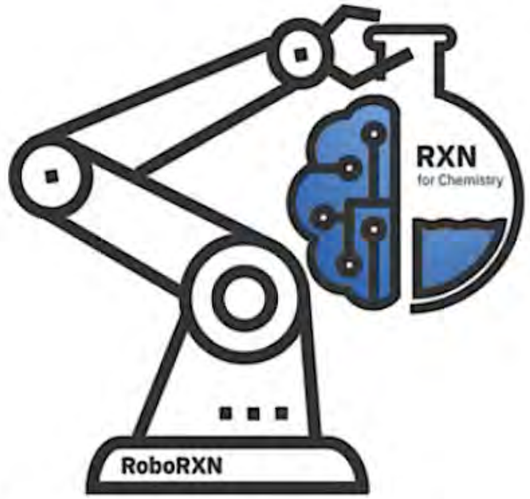
AI







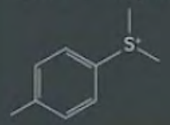
- Add: 1mmol 4-methoxythioanisole
- Add: 10 ml 1,2-dichloroethane
- Add: 1mmol methyl trifluoromethanesulfonate
- Stir: at 25°C for 1 day
- Filter: keep filtrate
- Concentrate



Synthesis planning → Recipe formulation → Execution on robot


Synthesizing new molecule
Started: Nov 30 2020, 6:49am PT
Live from IBM RoboRXN
Action 2
Overview
Adding $C_2H_3F_3O_3S$
In this action, the molecule methyl trifluoromethane sulfonate is added to Reactor 2.
Methyl trifluoromethane
$C_2H_3F_3O_3S$
2D
3D
Methyl trifluoromethane sulfonate is a brown liquid. Insoluble in water. This material is a very reactive methylating agent, also known as methyl triflate.
NOW
10 ml of reagent containing methyl trifluoromethane sulfonate is being moved from Vial 61 and added to Reactor 2.
Position of the robot arm
Moving to Vial 61
Live view module
00:06:02
LIVE

Synthesizing new molecule



Started: Nov 30 2020, 6:49am PT

Live from IBM RoboRXN

Action 2

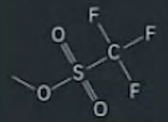
Overview

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



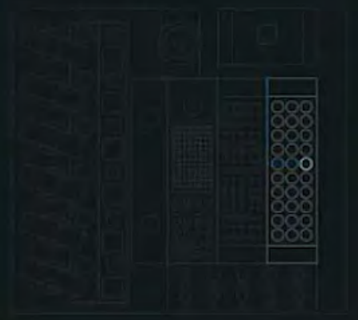
Methyl trifluoromethane sulfonate is a brown liquid. Insoluble in water. This material is a very reactive methylating agent, also known as methyl triflate.

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

Extracting from vial 61



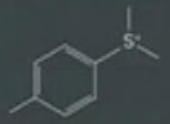
Live view module

2 Adding $C_2H_3F_3O_3S$

00:06:07

LIVE

Synthesizing new molecule



Started: Nov 30 2020, 6:49am PT

Live from IBM RoboRXN

Action 2

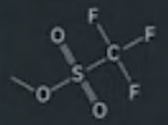
Overview

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

Inserting into reactor 2

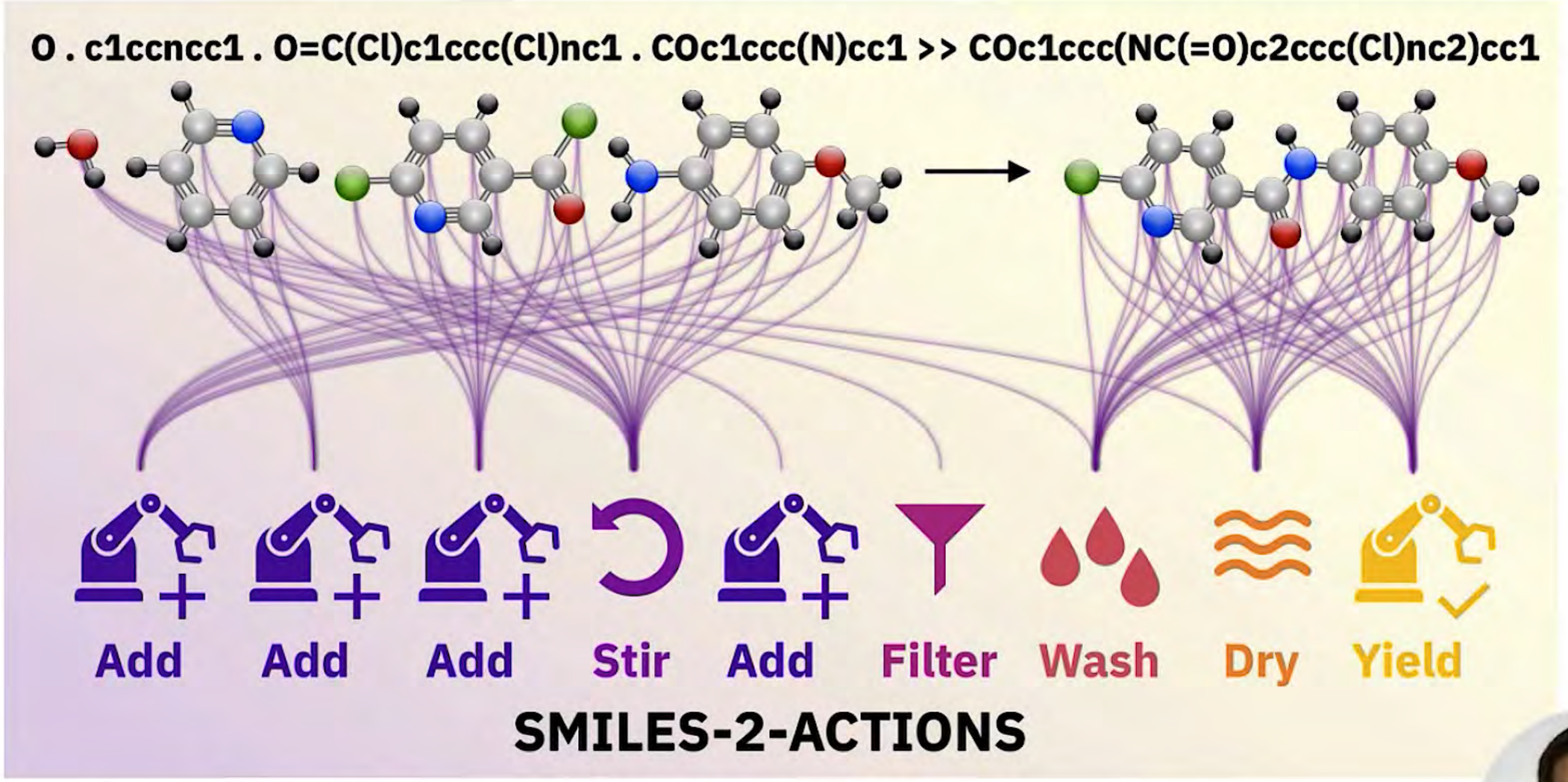
Live view module

Adding $C_2H_3F_3O_3S$

00:06:20

LIVE


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.
2D
3D
Methyl trifluoromethane sulfonate is a brown liquid. Insoluble in water. This material is a very reactive methylating agent, also known as methyl triflate.
NOW
10 ml of reagent containing methyl trifluoromethane sulfonate is being moved from Vial 61 and added to Reactor 2.
Position of the robot arm
Moving to Vial 61
Live view module
00:06:00
LIVE



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

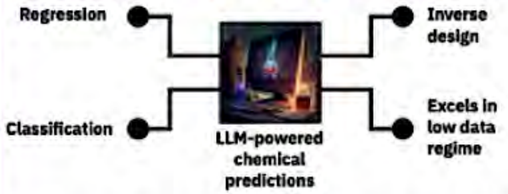
Alain Vaucher

Trends in *Language Models* for Chemistry

General task solvers

Other examples:
- Darwin (open models)

Fine-tuned large language models



Jablonka et al. NeurIPS ML for Materials, 2022

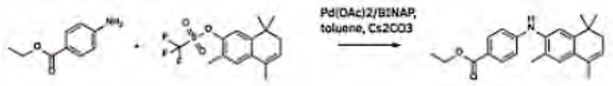
LLMs augmented with chemical tools

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

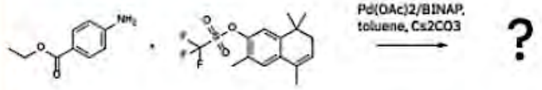


Vaucher et al. Nature Comm. 2020
Vaucher et al. *Nature Comm.* 2021

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

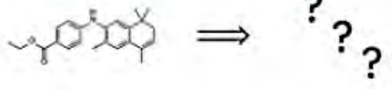
Single modality

Reaction prediction



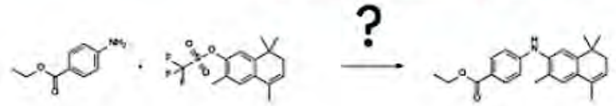
Schwaller et al. *ACS Cent. Sci.* 2019
Pesciullesi et al. *Nature Comm.* 2020

Retrosynthetic planning



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

Reaction condition prediction

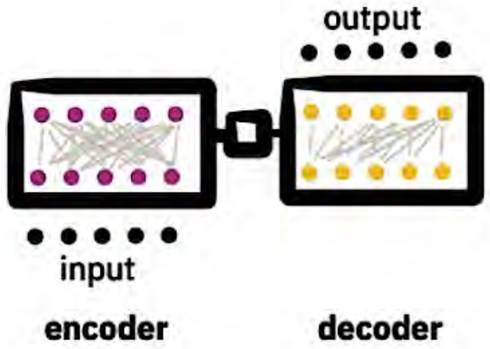


Gao et al. *ACS Cent. Sci.* 2018
Schilter et al. *ACS Fall Meeting* 2023

- So far, all I showed you had max ~20M parameters, trained on 1 consumer GPU in ~2 days.

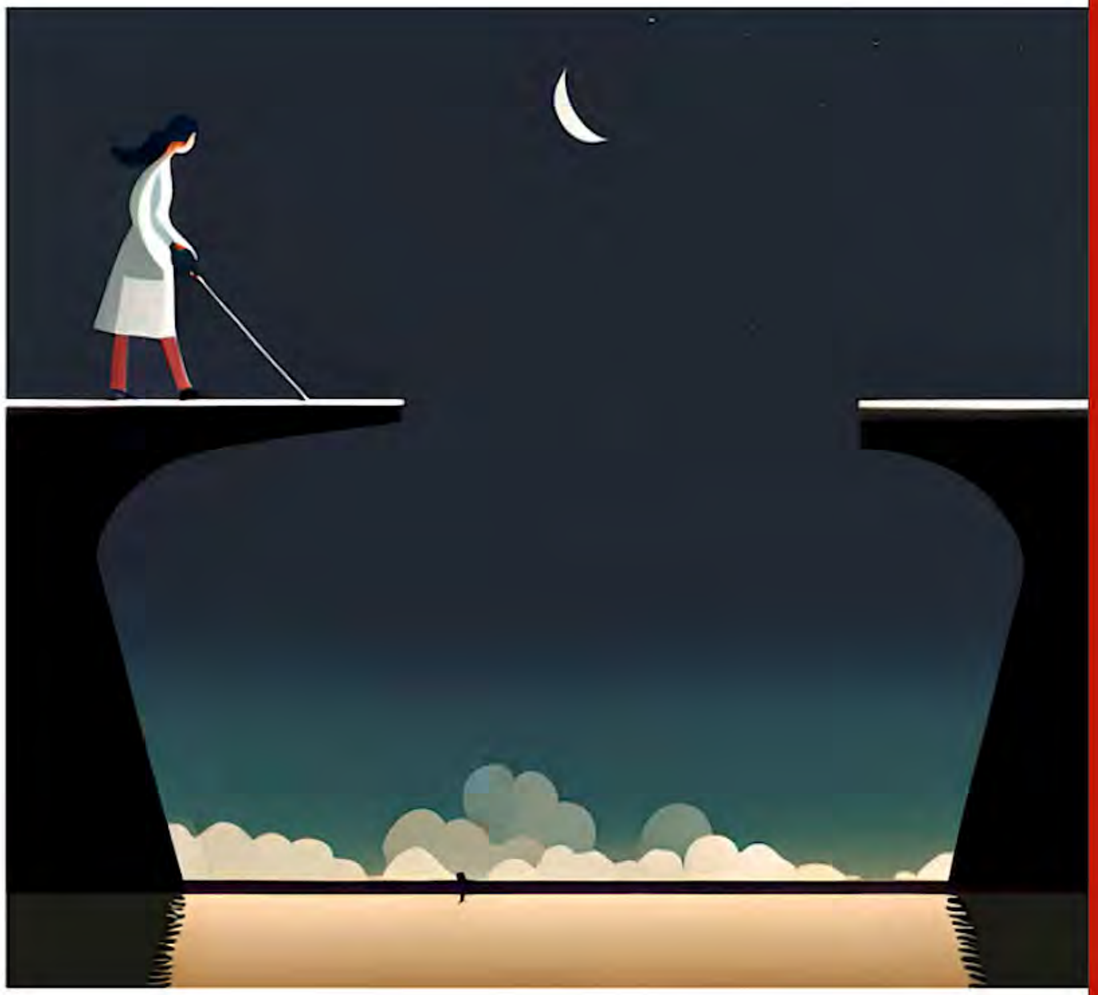
Decoder-only transformer.

Attention is all you need
(Vaswani et al., 2017)



GPT
(Radford et al., 2018)

- Decoder-only models are trained to **predict the next "word"** given all previous words. → GPT3, ChatGPT, GPT4
- Philippe should
- Philippe should prepare
- Philippe should prepare slides
- …
- Philippe should prepare slides for the SPREE seminar.



Not many tools are used

- Too many of them (100+)
- They live in isolation
- Hard to set-up
- Non-intuitive to use

There's a gap between computational and experimental fields!

Aim: Bridge the gap

- ✓ Continue making useful tools
- ✓ Make them usable
  -> also for non-experts

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?



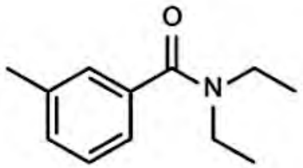


Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White ✉ & Philippe Schwaller ✉

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

ChemCrow's toolset

Molecule tools



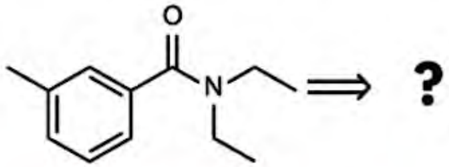
- SMILES to Weight
- SMILES to Price
- SMILES to CAS
- Similarity
- Modify Mol
- Func Groups
- Patent Check
- Name to SMILES

- Safety Assessment
- Explosive Check

General tools

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

- RXN to Name
- RXN Predict
- Synth Plan



Safety tools

Reaction tools

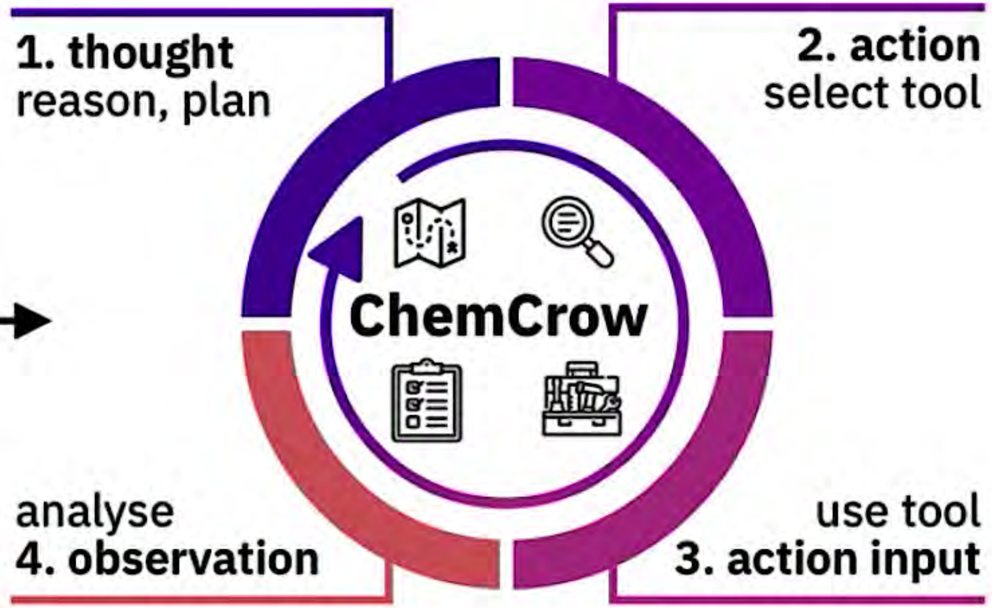
- RDKit: Open-source cheminformatics. https://www.rdkit.org
- PubChem Compound Database https://pubchem.ncbi.nlm.nih.gov/
- Synspace: https://github.com/whitead/synspace
- Molbloom: https://github.com/whitead/molbloom
- Chem-space: https://chem-space.com/
- paper-qa: https://github.com/whitead/paper-qa/tree/main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

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*

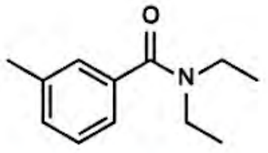
Task: Plan and execute the synthesis of an insect repellent

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**



Philippe Schwaller (@SchwallerGroup)

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

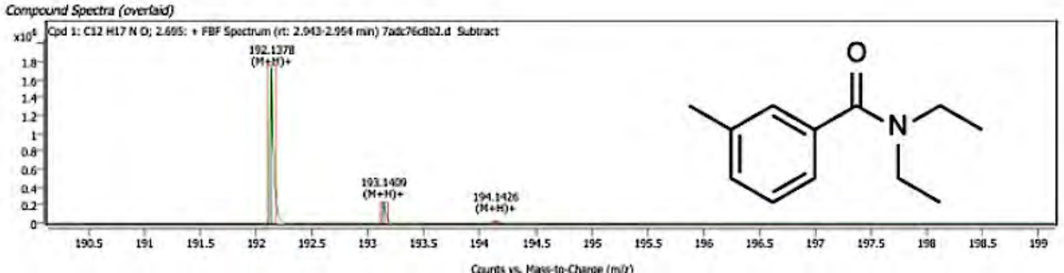
Thought: I have a synthesis plan for DEET. Now I will execute the synthesis.
Action: RXNExecute
Action input: CCN(CC)C(=O)c1cccc(C)c1

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

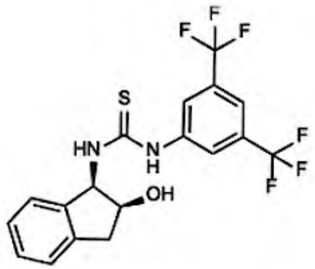


Philippe Schwaller (@SchwallerGroup)

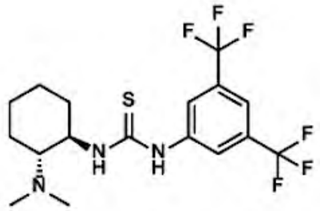
Use case: *Automatic synthesis*

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

Schreiner's



Ricci's



Takemoto's

Organocatalysts

ETH zürich

EPFL

CAT

ChemCrow

liacpc11.epfl.ch:8019

RUNNING... Stop

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

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

Name2SMILES: LSD

First, I need to find the SMILES representation of three psychedelic substances. I will use the Name2SMILES tool for this.
Action: Name2SMILES
Action Input: LSD

ChemCrow

liacpc11.epfl.ch:8019

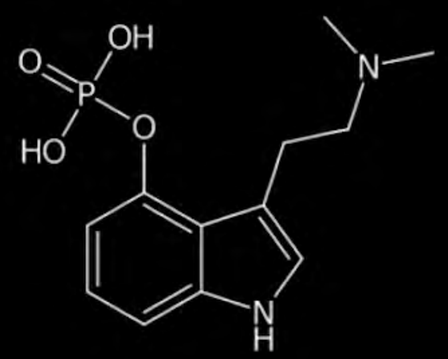
RUNNING... Stop

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

CN(C)CCc1c[nH]c2cccc(OP(=O)(O)O)c12



Thinking...

I have obtained the SMILES representation for Psilocybin. Now, I will find the SMILES for one more psychedelic substance.
Action: Name2SMILES
Action Input: Mescaline

ChemCrow

liacpc11.epfl.ch:8019

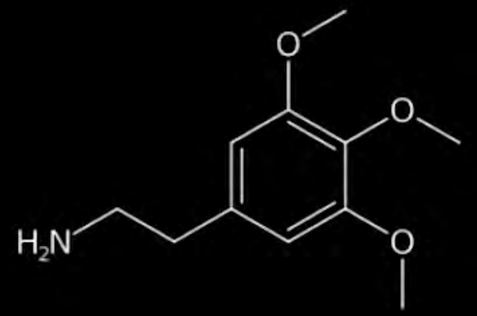
RUNNING... Stop

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

COc1cc(CCN)cc(OC)c1OC



Name2SMILES: Acetic chloride

I have obtained the SMILES representation for Mescaline. Now, I will find the SMILES for acetic chloride to predict the reaction.
Action: Name2SMILES
Action Input: Acetic chloride

ChemCrow

liacpc11.epfl.ch:8019

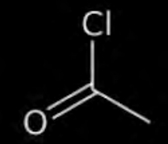
RUNNING... Stop

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

CC(=O)Cl



🤔 **RXNPredict:** CCN(CC)C(=O)[C@@H]1C=C2c3cccc4[nH]cc(c34)C[C@H]2N(C)C1.CC(=O...

Now that I have the SMILES for LSD, Psilocybin, Mescaline, and Acetic chloride, I can predict the reaction products using the RXNPredict tool. I will do this for each psychedelic substance separately.
Action: RXNPredict
Action Input: CCN(CC)C(=O)[C@@H]1C=C2c3cccc4[nH]cc(c34)C[C@H]2N(C)C1.CC(=O)Cl

ChemCrow

liacpc11.epfl.ch:8019

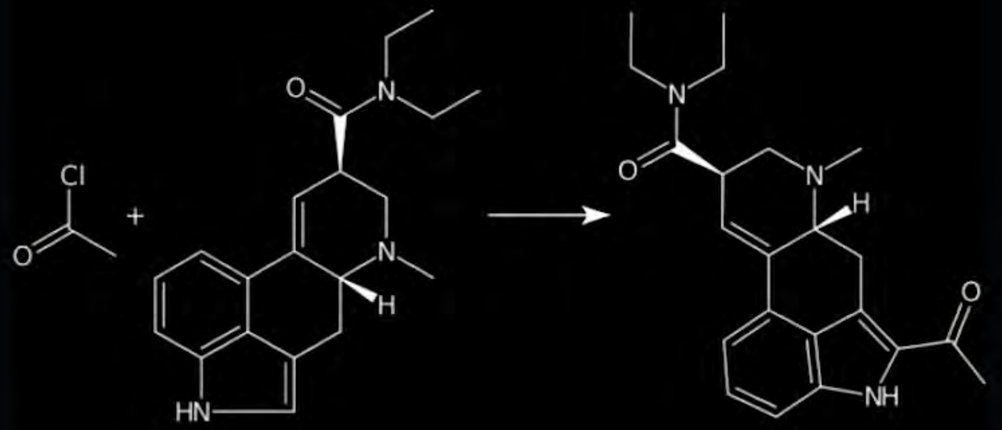
RUNNING... Stop

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

CCN(CC)C(=O)[C@@H]1C=C2c3cccc4[nH]c(C(C)=O)c(c34)C[C@H]2N(C)C1



RXNPredict: CN(C)CCc1c[nH]c2cccc(OP(=O)(O)O)c12.CC(=O)Cl

The reaction of LSD with acetic chloride results in a new compound. Now, I will predict the reaction of Psilocybin with acetic chloride.
Action: RXNPredict
Action Input: CN(C)CCc1c[nH]c2cccc(OP(=O)(O)O)c12.CC(=O)Cl

ChemCrow

liacpc11.epfl.ch:8019

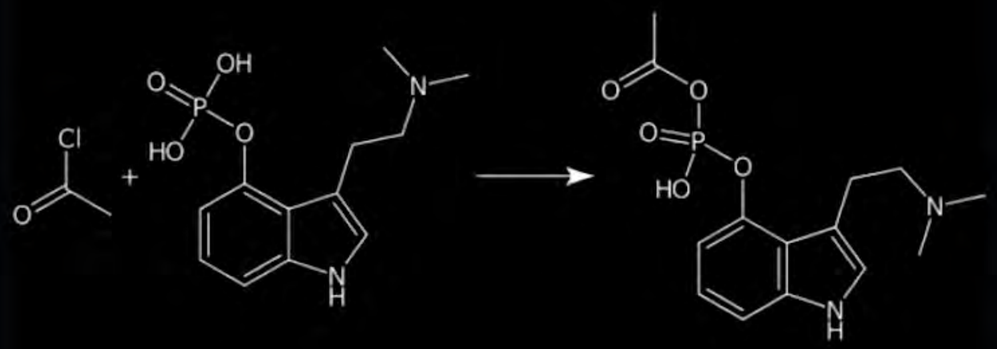
RUNNING... Stop

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

CC(=O)OP(=O)(O)Oc1cccc2[nH]cc(CCN(C)C)c12



Thinking...

The reaction of Psilocybin with acetic chloride results in a new compound. Now, I will predict the reaction of Mescaline with acetic chloride.
Action: RXNPredict
Action Input: COc1cc(CCN)cc(OC)c1OC.CC(=O)Cl

ChemCrow is modular.

Toolset and LLMs can be changed.

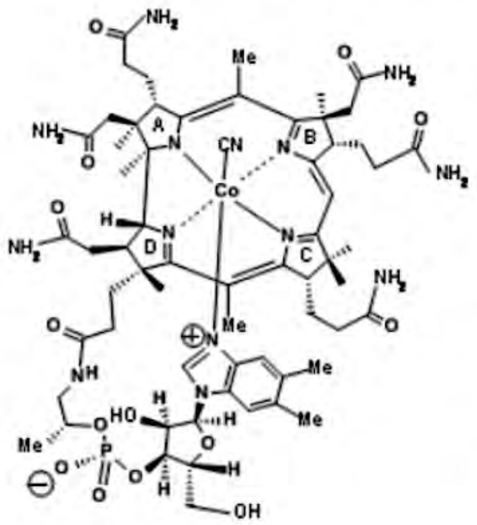
More use cases in the paper.

Code: https://github.com/ur-whitelab/chemcrow-public

Is automated synthesis a solved problem? No...

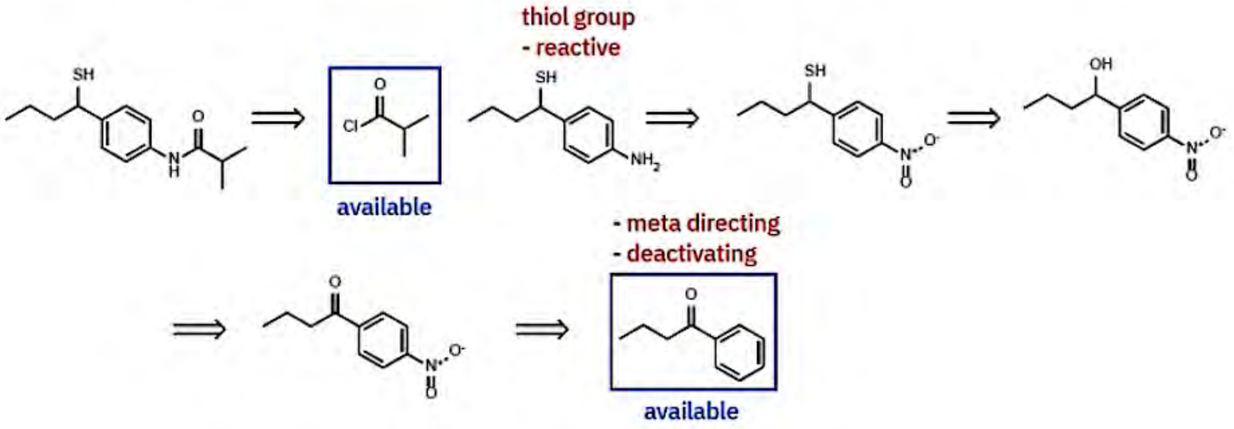
- Supply chain/robotics challenges
- Weak synthesis planning models

Daniel Armstrong & Zlatko Jončev

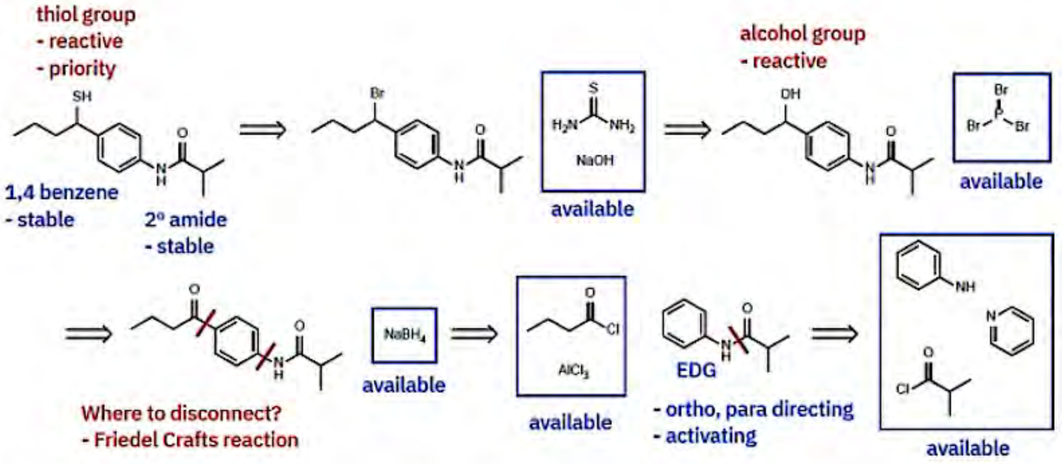


Complexity that an chemist would like to target: Vitamin B12

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



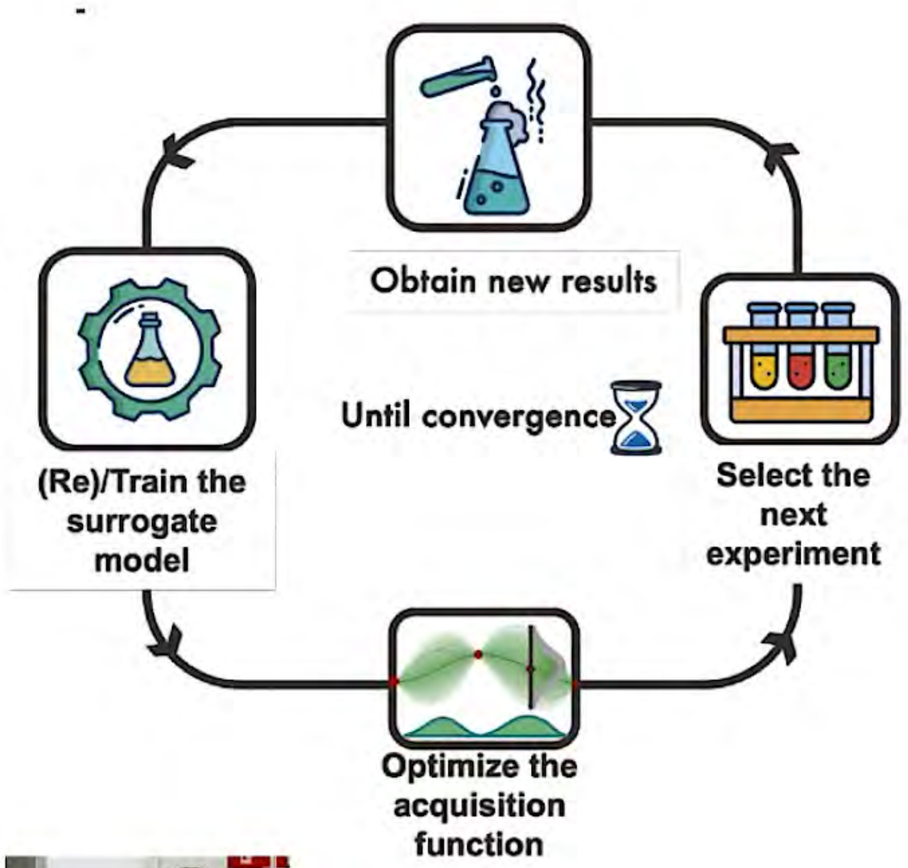
Typical output of unnamed system



More strategic synthesis plan (including reagents)

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

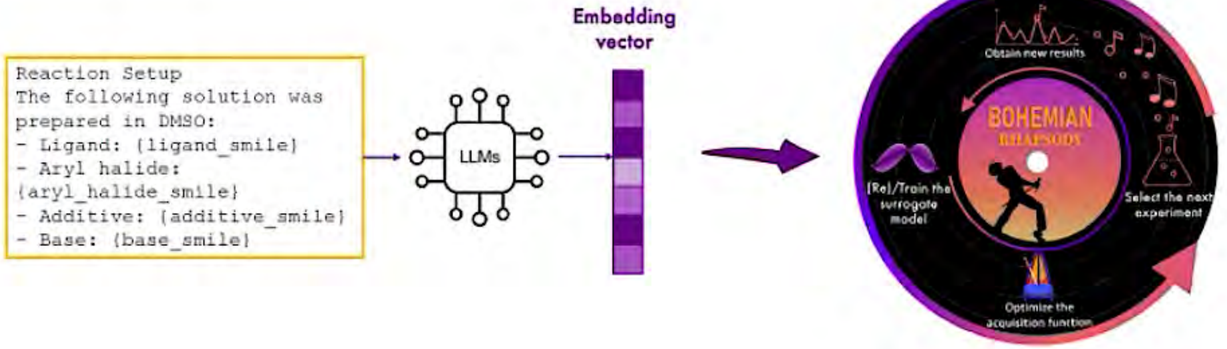
Bayesian Optimization for chemical reactions



Bayesian optimization for additive screening and yield improvements – beyond one-hot encoding

BoChemian: Large Language Model embeddings for Bayesian optimization of chemical reactions

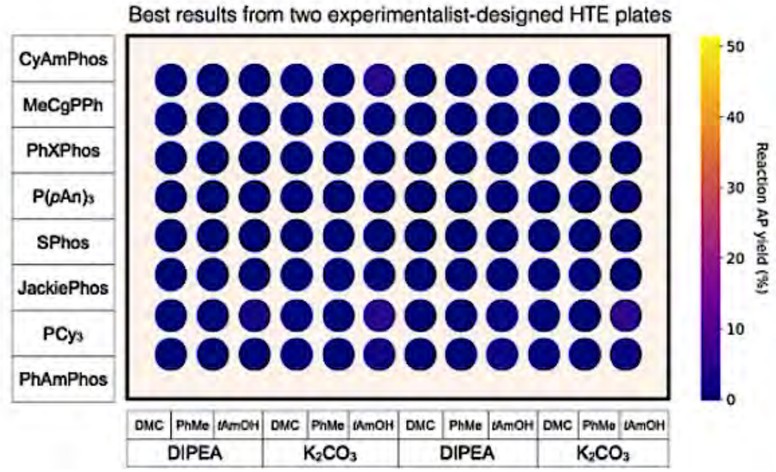
- BOjana Ranković

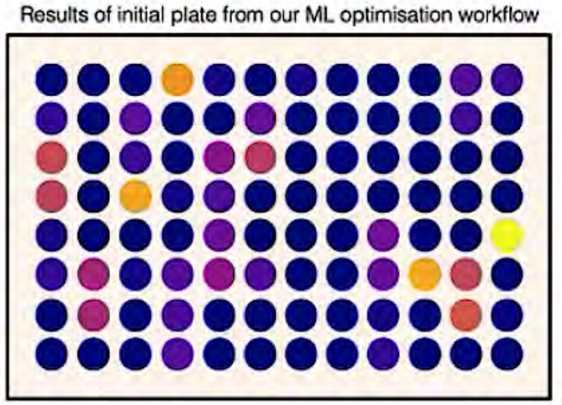


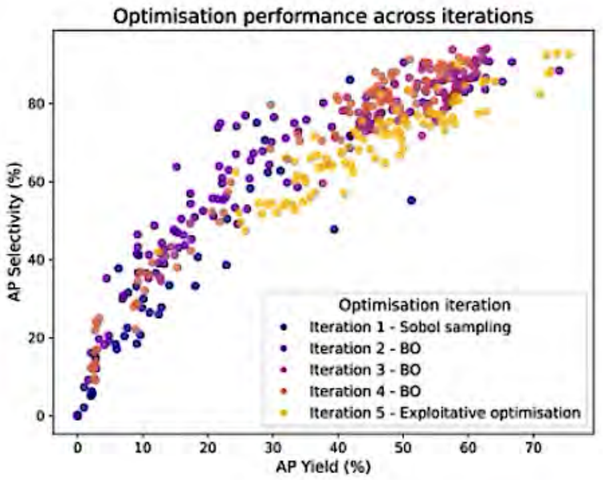
Bayesian Optimization for large batches

Joshua Sin
(with Roche)

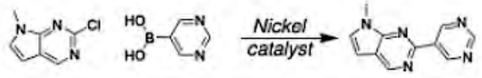
a





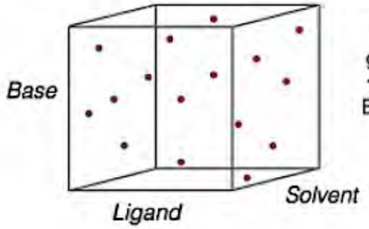


b



Ligand	Base	Solvent	...	Objectives
L1	B1	S1	...	Yield
...	...	...		...

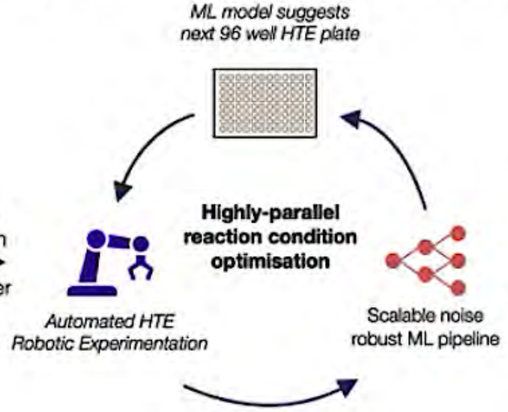
Large, chemist-defined reaction condition space of many possible combinations



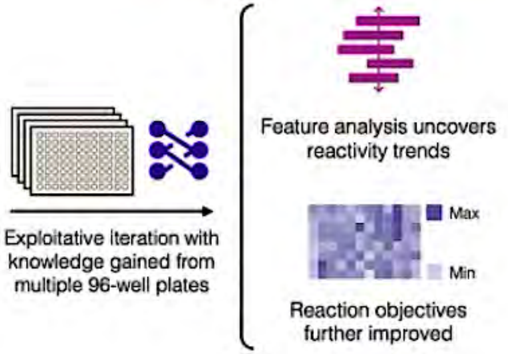
Initial **quasi-random** Sobol guided experiment selection

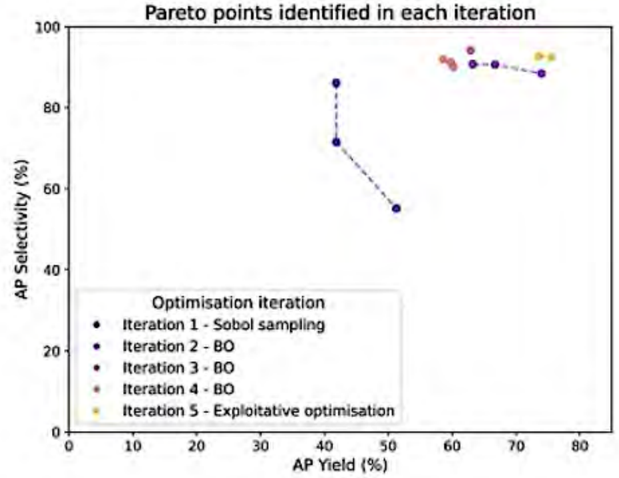
Experiments • diversely cover reaction condition space

Rapid suggestion and execution of batches of 96 parallel experiments at multiple temperatures



- **Optimised selectivity and yield**
- **Unexpected reactivity insights**





Highly Parallel Optimisation of Nickel-Catalysed Suzuki Reactions through Automation and Machine Intelligence

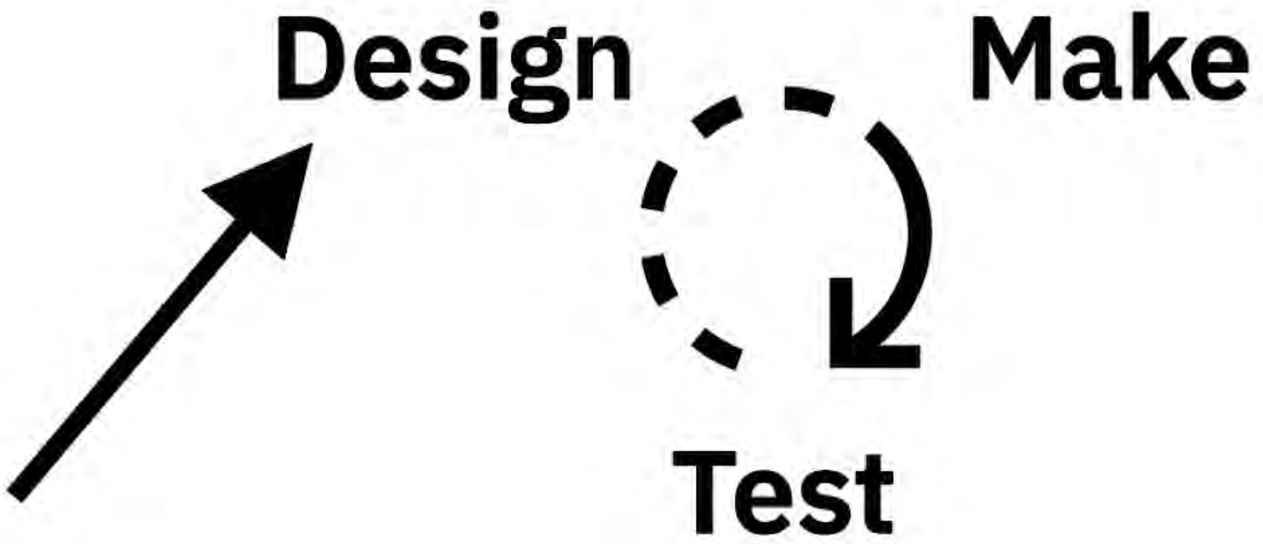
EPFL

What molecule to make?

How to make it?

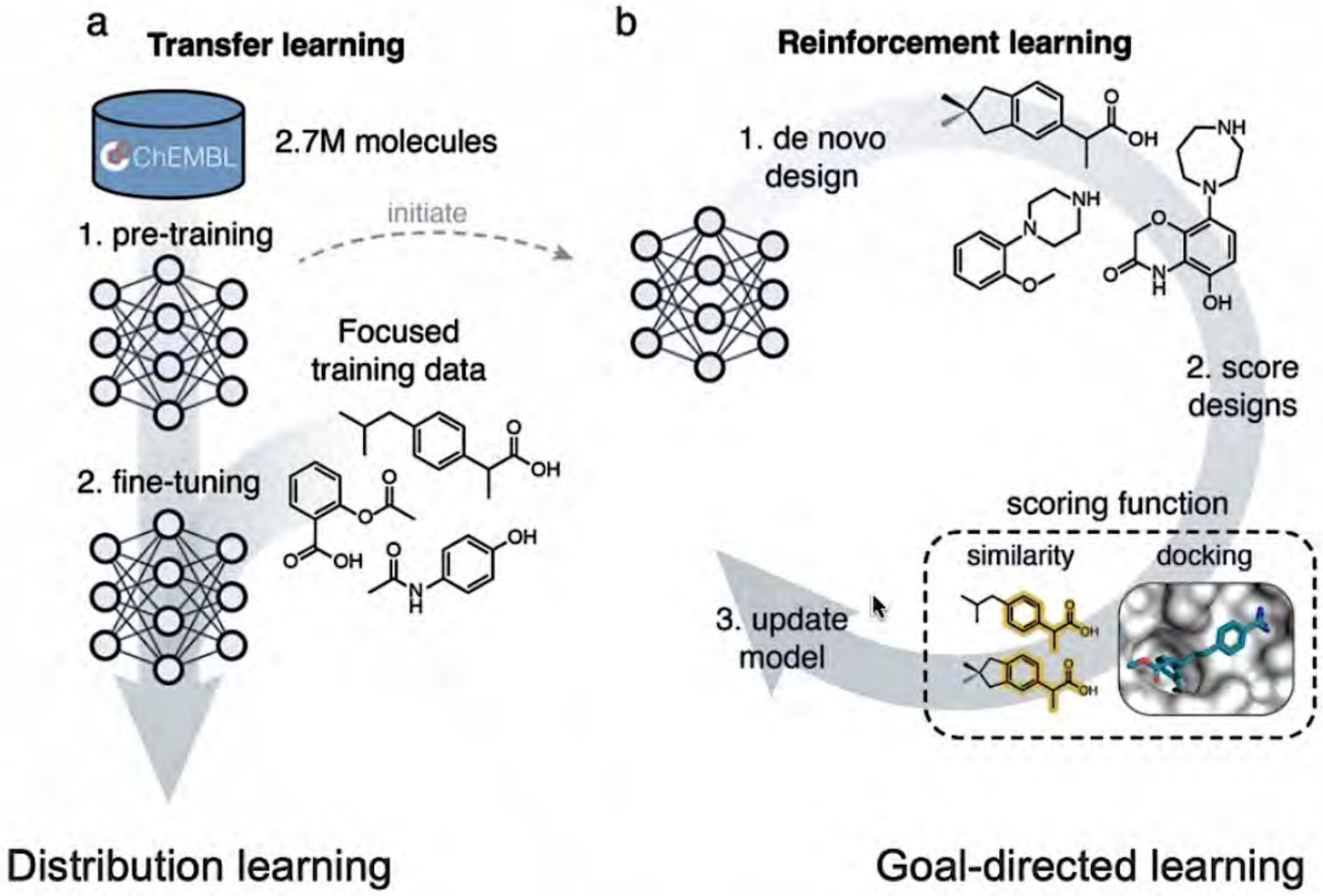
Reaction prediction

Synthesis planning



Experimental validation

Main flavors of Generative De Novo Molecule Design



Van Tilborg,…, Grisoni, https://doi.org/10.26434/chemrxiv-2024-w0wvl

Goal-directed Generative Molecular Design

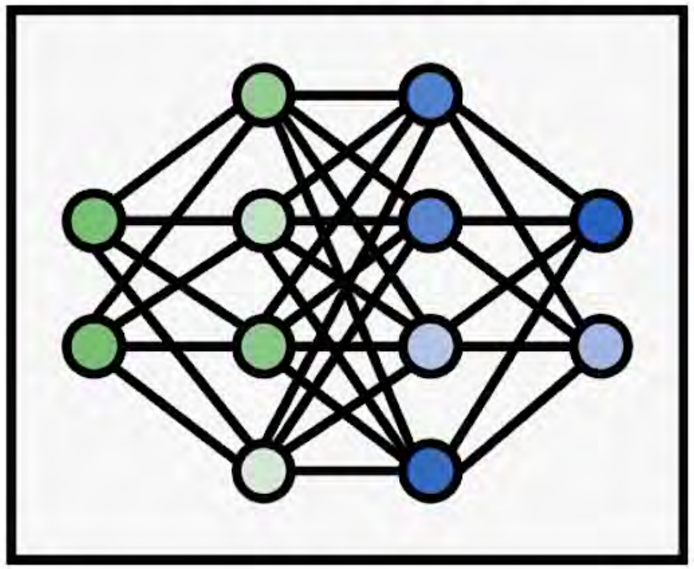
Objective

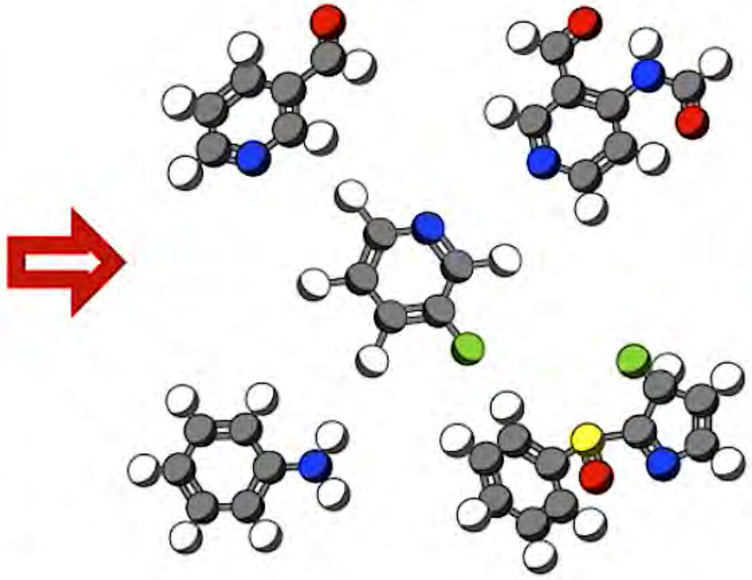
Design molecules satisfying a desired property profile

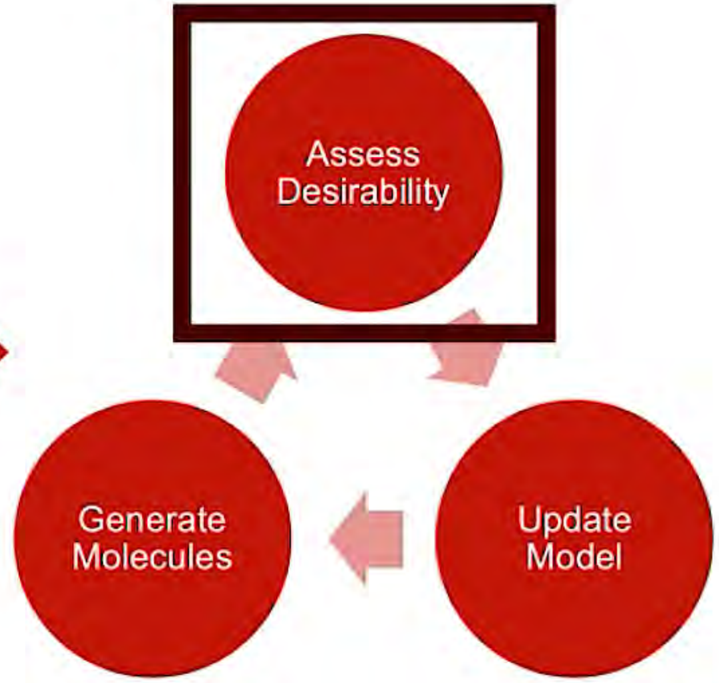
Key Consideration

Is the *in silico* predictor even correlated with the desired end-point?

Generative Model







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

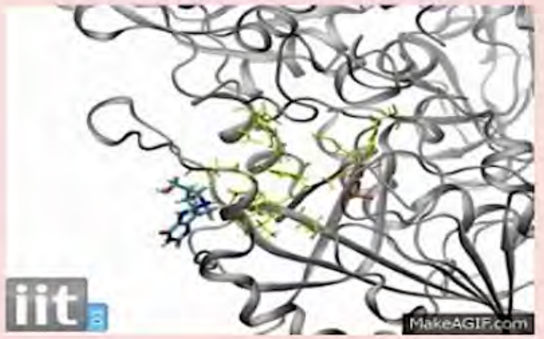
Molecular Docking

OpenEye SCIENTIFIC

Schrödinger

MMPB(GB)SA

Free Energy Calculations

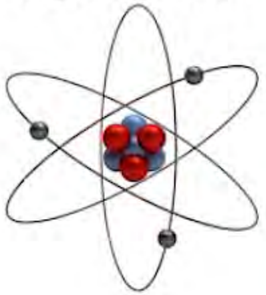


Materials and Catalyst Design

Semi-empirical QM

xTB

Single-point DFT



DFT – Varying Functionals / MD

Our work on improving sample efficiency

Augmented Memory

- Combines data augmentation with experience replay
- New state-of-the-art on the PMO benchmark

Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization

Wenhao Gao[1*], Tianfan Fu[2*], Jimeng Sun[3,4], Connor W. Coley[1,5]
[1]Department of Chemical Engineering, Massachusetts Institute of Technology,
[2]Department of Computational Science and Engineering, Georgia Institute of Technology,
[3] Department of Computer Science, University of Illinois at Urbana-Champaign,
[4] Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign,
[5]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology,
*Equal Contributions

{whgao,ccoley}@mit.edu, tfu42@gatech.edu, jimeng@illinois.edu

- Benchmarks 25 models across 23 optimization tasks (imposing a **fixed** oracle budget **-> 10k**)

Gao, W., Fu, T., et al. NeurIPS 2022 Benchmark and Datasets Track.
Guo, J. & Schwaller, P. *JACS Au*, **2024**.

PMO (Sample efficiency) Benchmark

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

10k oracle calls

AUC Top-10

The benchmark includes all recent de novo design algorithms.

All within 1 point

Big ML conferences models

Recent works benchmarking Augmented Memory

Published at the GEM workshop, ICLR 2024

DyMol: Dynamic Many-Objective Molecular Optimization with Objective Decomposition and Progressive Optimization

Dong-Hee Shin†, Young-Han Son†, Deok-Joong Lee, Ji-Wung Han & Tae-Eui Kam*
Department of Artificial Intelligence, Korea University
{dongheeshin,yhson135,deokjoong,danielhan,kamte}@korea.ac.kr

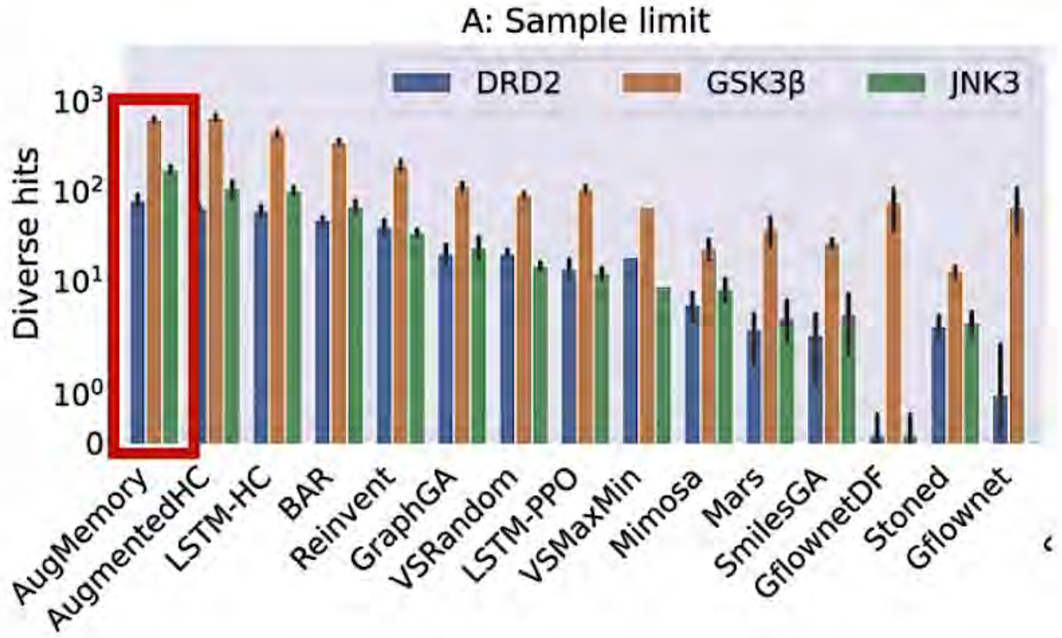
	Four objectives		Five objectives		Six objectives	
Method	HV(↑)	R2(↓)	HV(↑)	R2(↓)	HV(↑)	R2(↓)
Random ZINC	0.065±0.013	4.809±0.243	0.005±0.003	8.867±0.202	0.001±0.000	14.456±0.344
SMILES-VAE	0.073±0.023	4.839±0.435	0.004±0.001	9.226±0.512	0.001±0.000	15.109±0.548
GFlowNet	0.063±0.011	4.952±0.196	0.011±0.004	9.011±0.314	0.004±0.001	14.462±0.422
MIMOSA	0.082±0.028	4.905±0.410	0.016±0.012	9.195±0.761	0.005±0.005	15.454±1.701
LaMBO	0.123±0.006	4.496±0.148	0.009±0.001	9.302±0.159	Out-of-memory	Out-of-memory
HN-GFN	0.120±0.000	4.013±0.060	0.004±9.861	9.861±0.096	0.002±0.000	15.349±0.102
MOEA/D	0.176±0.123	4.615±1.193	0.094±0.052	8.105±1.561	0.025±0.018	15.054±2.028
NSGA-III	0.234±0.107	3.477±0.837	0.071±0.047	7.130±1.349	0.016±0.020	14.366±2.393
GraphGA	0.254±0.069	3.379±0.666	0.100±0.057	7.676±1.312	0.051±0.030	11.814±1.787
GPBO	0.275±0.091	3.311±0.757	0.091±0.031	7.670±0.761	0.026±0.024	12.840±1.811
REINVENT BO	0.309±0.021	2.795±0.103	0.071±0.014	7.537±0.620	0.033±0.019	10.929±1.019
REINVENT	0.338±0.030	2.770±0.116	0.099±0.054	7.578±1.187	0.062±0.028	10.032±0.922
AugMem	0.395±0.038	2.496±0.192	0.090±0.043	8.011±0.955	0.071±0.072	12.472±3.758
DyMol (Ours)	**0.422±0.023**	**2.297±0.095**	**0.247±0.087**	**4.943±0.990**	**0.143±0.056**	**8.842±1.632**

- DyMol = REINVENT + “progressive optimization”
- Augmented Memory is 2nd most sample-efficient

Published at the GEM workshop, ICLR 2024

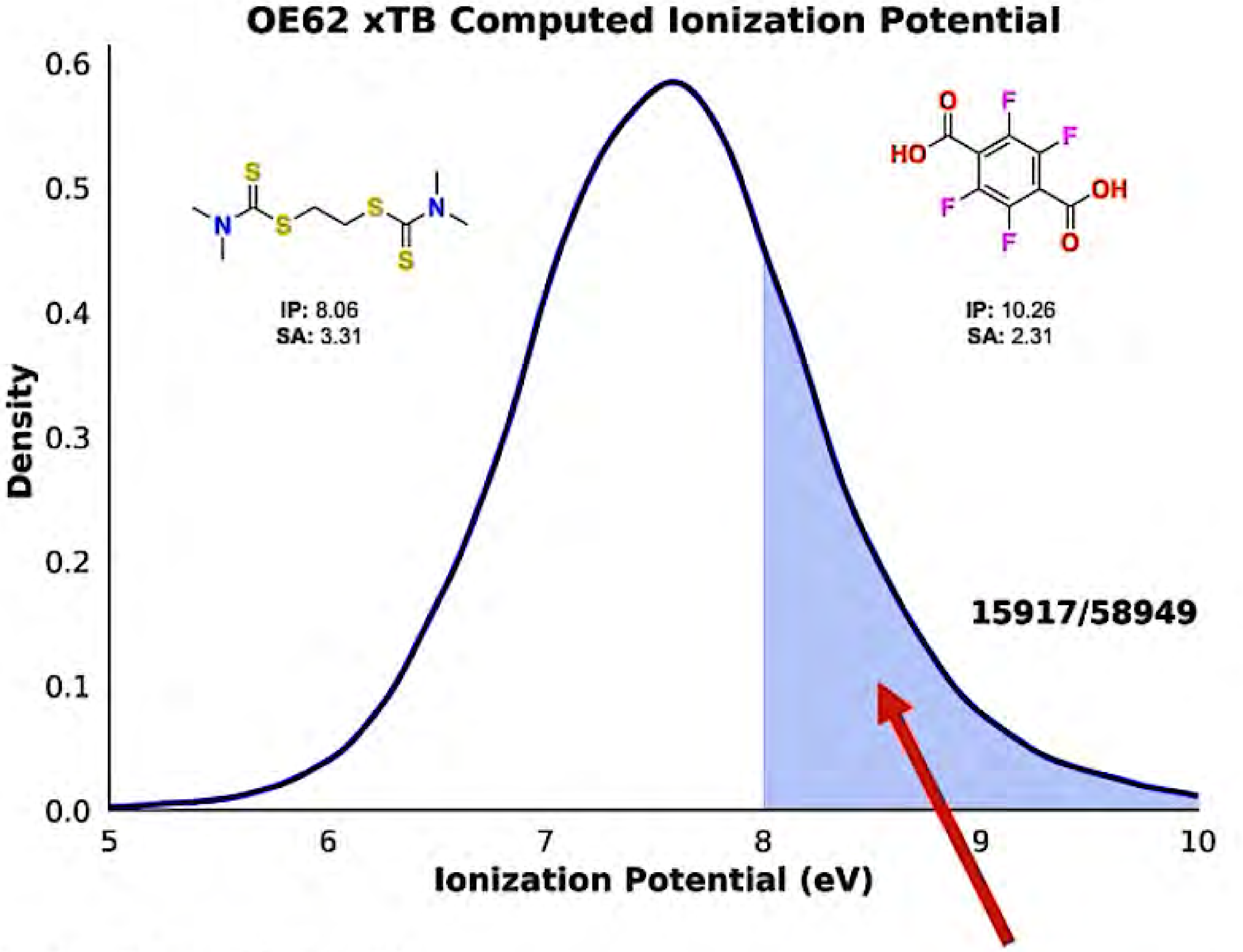
Diverse Hits in de Novo Molecule Design: A Diversity-based Comparison of Goal-directed Generators

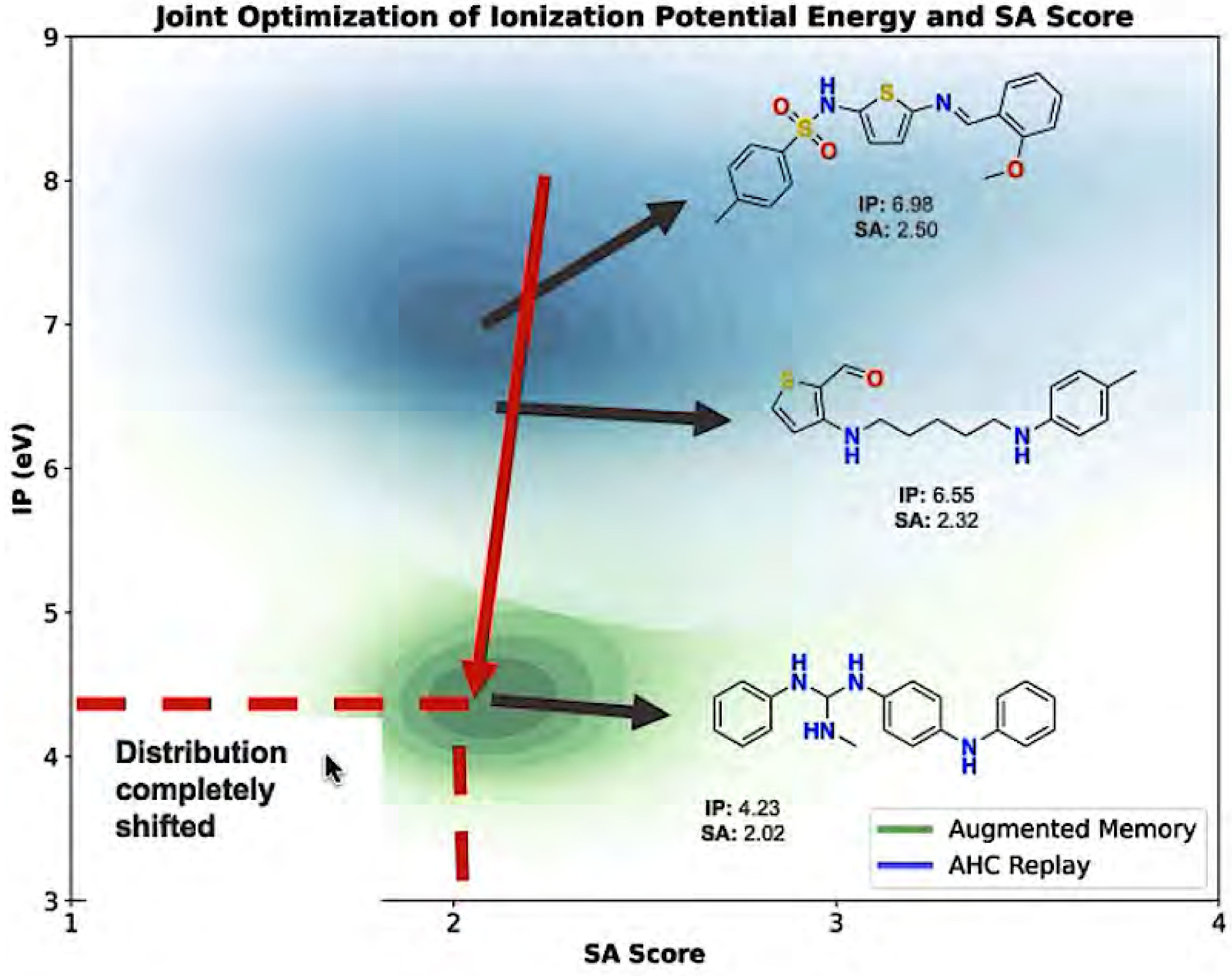
Philipp Renz **Sohvi Luukkonen** **Günter Klambauer**



Augmented Memory is most sample-efficient

Semiconductor design – shifting the distribution away from the training set



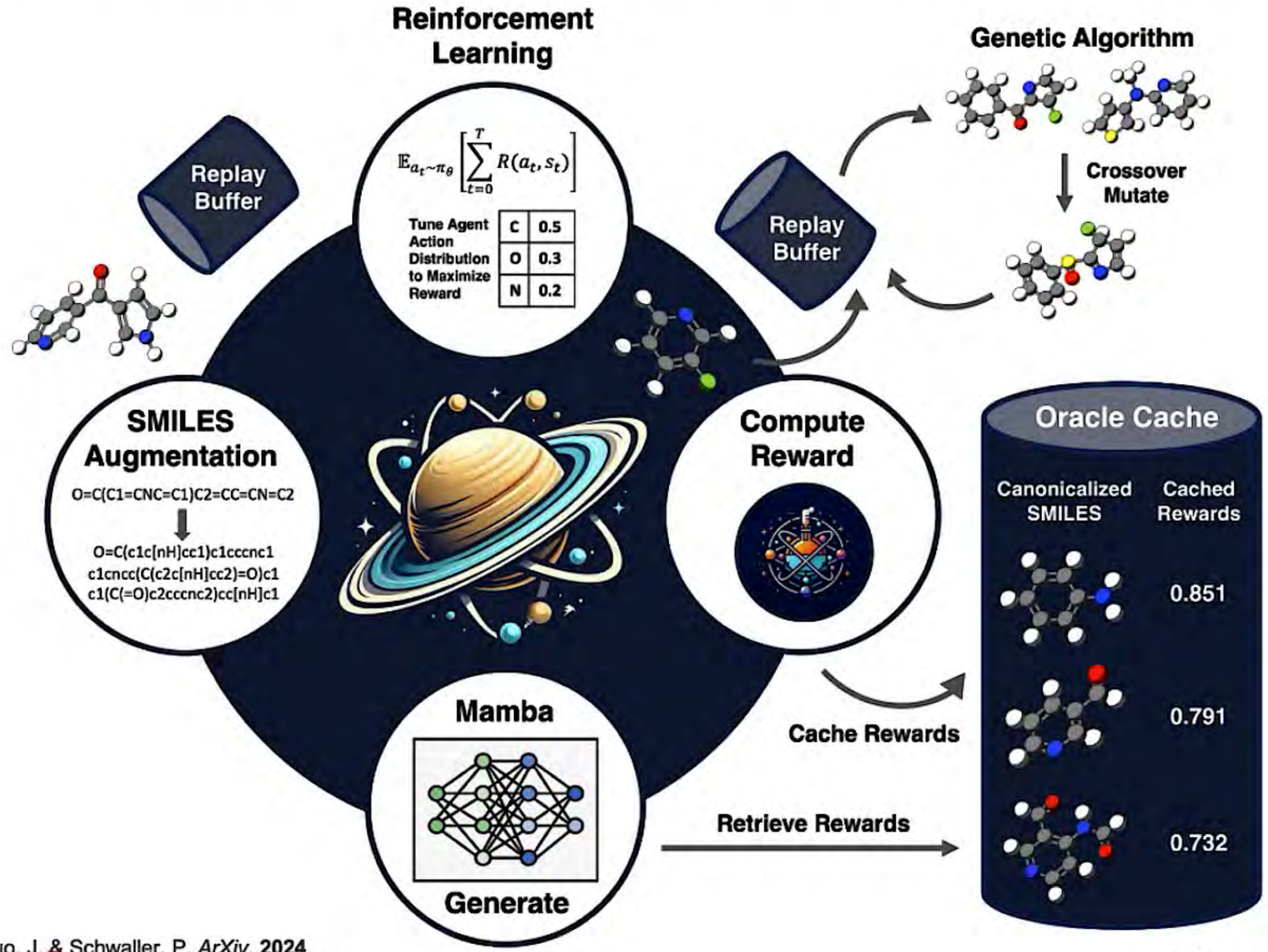


Optimization Objective
1. Minimize IP
2. Minimize SA

Training molecules

Guo, J.; Schwaller, P. *JACS Au*, **2024**.

EPFL Saturn – Sample-efficient generative design framework



- Sample efficiency ***much*** better than *baseline* **Augmented Memory**
- State-of-the-art on drug discovery optimization tasks

Jeff Guo

Guo, J. & Schwaller, P. *ArXiv*, **2024**.
Gu A. & Dao. T. arXiv:2312.00752, ***2023***

https://github.com/schwallergroup/saturn

Saturn – Sample-efficient generative design framework

Target	Model	Yield (↑)	IntDiv1 (↑)	Scaffolds (↑)	OB 1 (↓)	OB 10 (↓)	OB 100 (↓)
DRD2	Augmented Memory	22 ± 7	0.774 ± 0.019	22 ± 7	$143 \pm 75(10)$	$733 \pm 120(10)$	Failed
	Saturn	369 ± 62	0.671 ± 0.050	310 ± 70	$93 \pm 53(10)$	$391 \pm 56(10)$	$663 \pm 55(10)$
AChE	Augmented Memory	173 ± 19	0.843 ± 0.009	170 ± 18	$57 \pm 2(10)$	$189 \pm 52(10)$	$776 \pm 58(10)$
	Saturn	480 ± 79	0.757 ± 0.020	400 ± 96	$32 \pm 24(10)$	$185 \pm 82(10)$	$508 \pm 80(10)$
MK2	Augmented Memory	0.2 ± 0.4	—	0.2 ± 0.4	$836 \pm 186(2)$	Failed	Failed
	Saturn	14.9 ± 14.1	0.454 ± 0.212	14.1 ± 13.2	$677 \pm 186(9)$	$861 \pm 108(6)$	Failed

Multi-objective docking optimization (docking score, MW < 500 Da, maximize QED), reward threshold 0.8, **1000 oracle calls**.

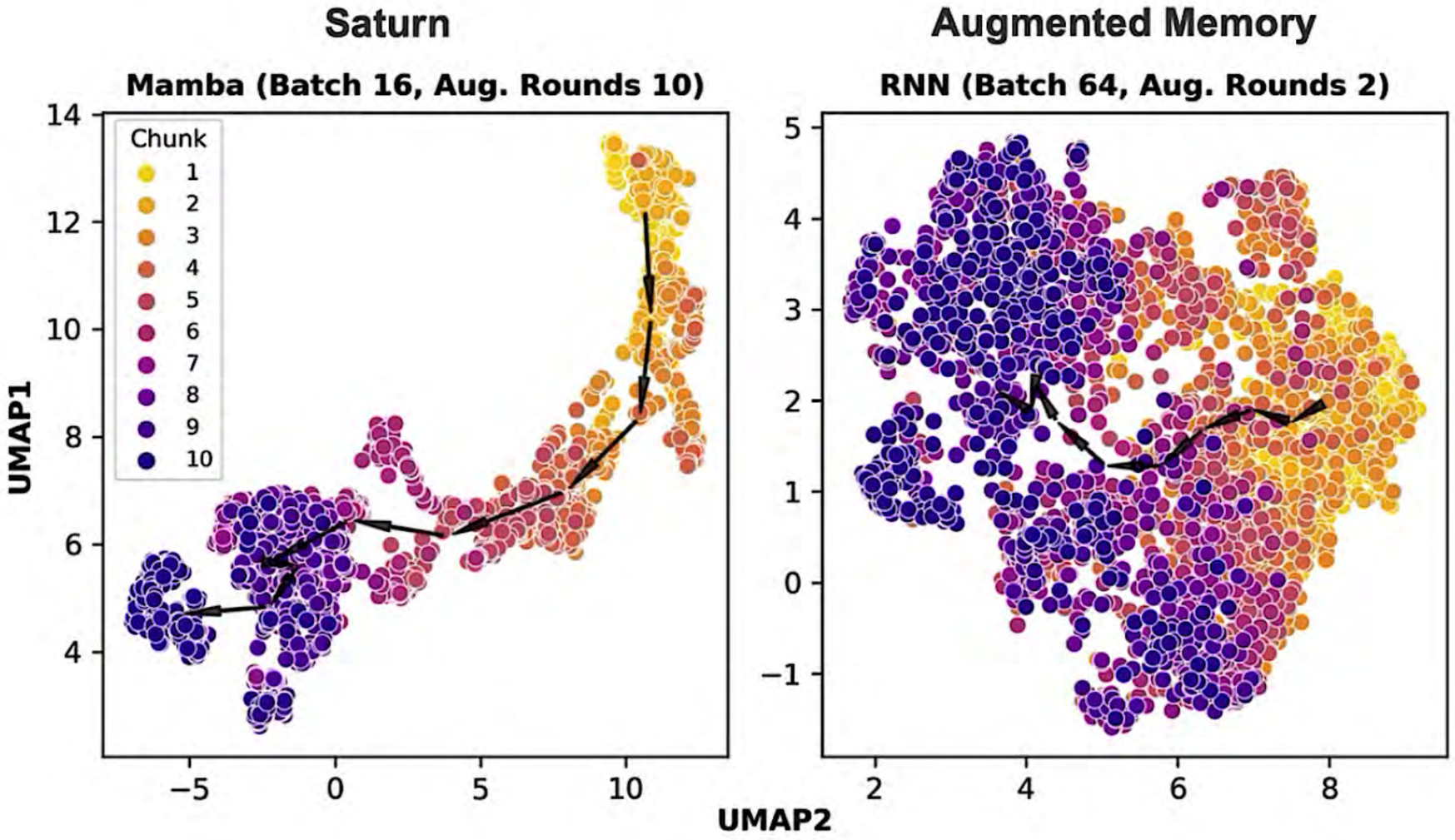
With Saturn (compared to Augmented Memory)

- Diversity decreases slightly
- **Yield**, number of molecules above reward threshold, **increases**
- **Oracle Burden (OB)**, number to get to X molecules above threshold, **decreases**

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

https://github.com/schwallergroup/saturn

EPFL Saturn – Hop and explore behavior



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

https://github.com/schwallergroup/saturn

Synthesizability constrained generation

Problem Statement

Generated molecules need to be synthesizable for ***tangible real-world impact***

Synthesizability-constrained Generation

Generate molecules starting from available building blocks stitched together by ***known***, feasible chemical transformations (in 2024, Swanson et al., Nat. Mach. Intell. / Cretu et al., *arXiv* / Koziarski, et al., *arXiv)*

Why not directly optimize for synthesizability?

Inference time can make this ***impractical*** due to generative models' ***insufficient*** optimization efficiency

→ ***Sample efficiency problem***

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

Synthesizability constrained generation

Problem Statement

Generated m
tangible re

Synthesiza

Generate m
blocks stitch
transformati
Intell. / Cre

Why not directly optimize for synthesizability?

actical due to
timization

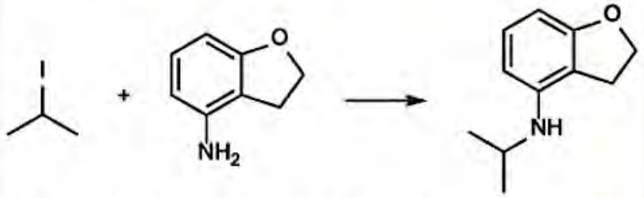
Method	Modes (Yield)	Mol. weight (↓)	QED (↑)	SA score (↓)	AiZynth (↑)	Oracle calls (Wall time)
Previous work		top-500	top-500	top-500	top-100	
GraphGA [68]	NR	521.0 ± 31.8	0.32 ± 0.07	4.14 ± 0.51	0.00	400,000 (NR)
SyntheMol [27]	NR	458.2 ± 60.7	0.45 ± 0.16	2.86 ± 0.56	0.56	100,000[a] (72h)
FGFN [69]	NR	548.6 ± 42.9	0.22 ± 0.03	2.94 ± 0.54	0.25	400,000 (NR)
RGFN [29]	NR	526.2 ± 37.6	0.23 ± 0.04	2.83 ± 0.22	0.65	400,000 (72h)
$R_{All\ MPO}$ **(ours)**	4 objectives (Docking, QED, SA, AiZynth)					
Saturn-ChEMBL (10)	4 ± 1 (5 ± 3)	367.7 ± 15.7	0.70 ± 0.13	2.11 ± 0.19	0.91 ± 0.11	1,000 (2.9h ± 34m)
Saturn-GA-ChEMBL (10)	7 ± 6 (10 ± 9)	373.3 ± 20.9	0.67 ± 0.09	2.08 ± 0.23	0.82 ± 0.17	1,000 (2.1h ± 24m)
Saturn-ZINC (9)	6 ± 3 (8 ± 10)	368.7 ± 27.6	0.79 ± 0.08	2.15 ± 0.22	0.87 ± 0.19	1,000 (2.1h ± 29m)
Saturn-GA-ZINC (10)	7 ± 4 (10 ± 7)	382.8 ± 27.9	0.71 ± 0.08	2.10 ± 0.15	0.85 ± 0.17	1,000 (2.0h ± 26m)

Saturn: 1/400th the oracle budget and within 2 hours is sufficient to find at least ***some*** desirable molecules (*across 10 replicates*)

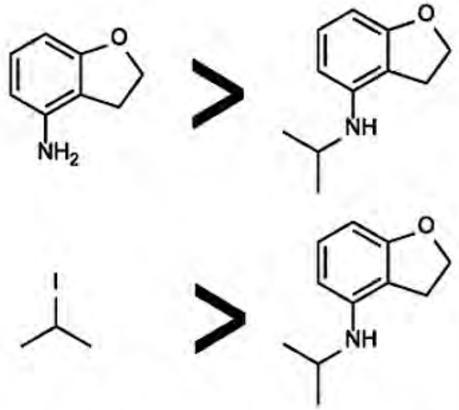
Guo, J. & Schwaller, P. *ArXiv*, **2024**.

FSscore: A Chemist's personalized feasibility score

1. Pretraining

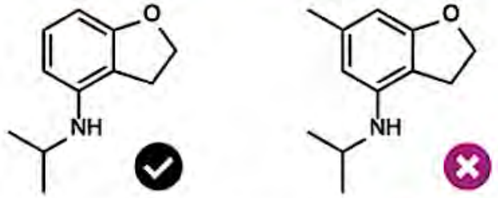


RXNs from USPTO/CHORISO

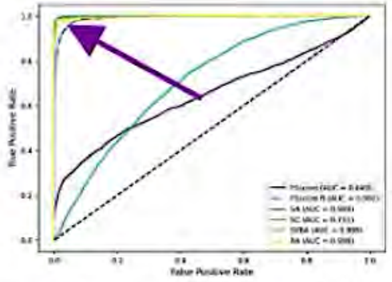


GNN for pair-wise ranking
Inspired by (SCScore, Coley)

2. Fine-tuning



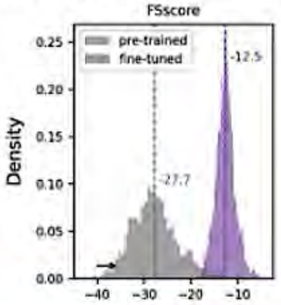
Human preference labelling
- 50 pairs (AUC 0.65 to 0.99)
- 20 pairs (AUC 0.49 to 0.67)



Inspired by MolSkill (Microsoft/ Novartis)

Case studies

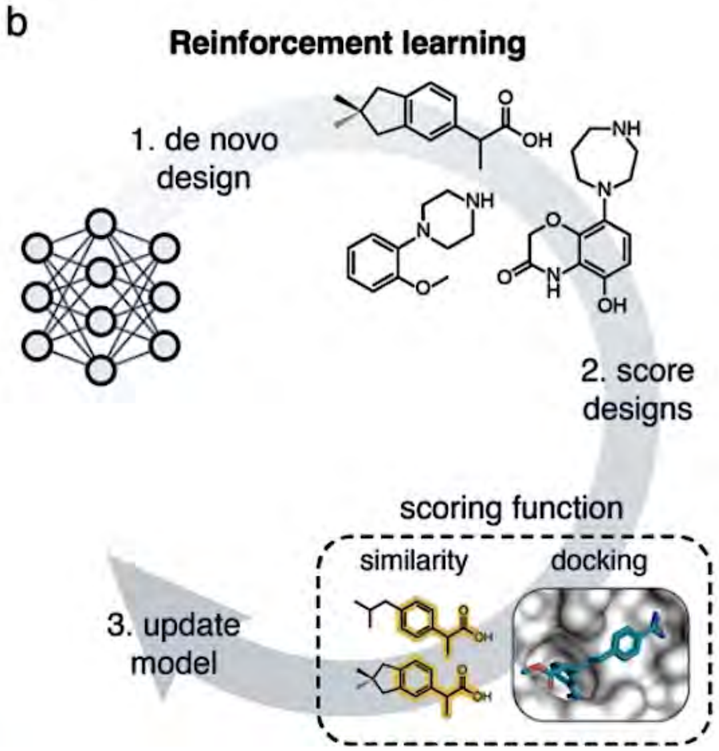
Diverse applications
- Chiral molecules
- Two "hard to synthesize" vs "easy to synthesize" test
- PROTAC degrader molecules (labelled by human expert)
- Natural products (COCONUT) vs known drugs
- DRD2 docking molecules from de novo generation agent



Rebecca Neeser

Neeser, Correia, Schwaller, NeurIPS AI4Science workshop, 2024

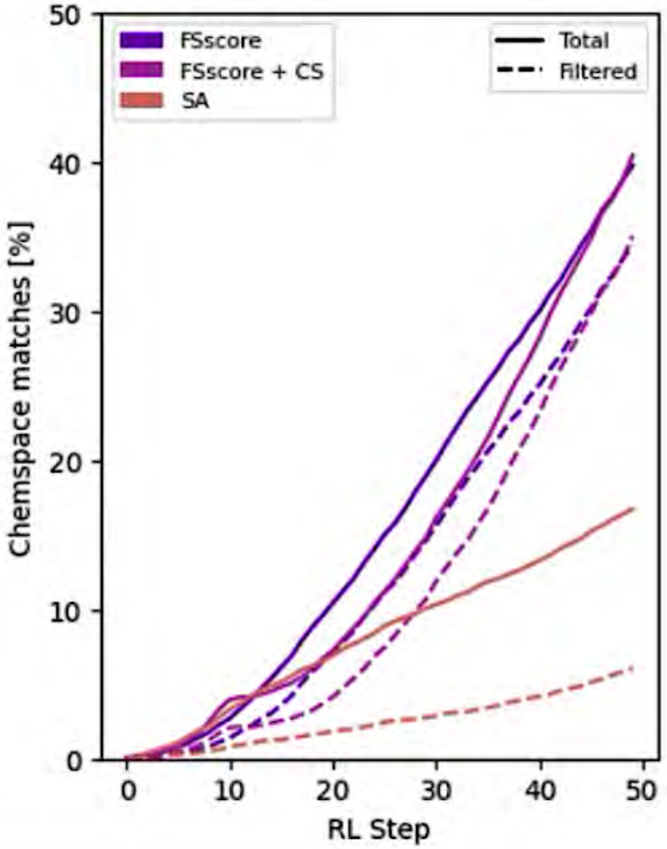
Removing the human out of the loop



Support Ukraine CHEM SPACE

Delivering Discovery Solutions®

Take closest make-on-demand molecule for pair to fine-tune scorer.



40% commercially available molecules (ChemSpace) without human labelling.

Recent review on topic

nature machine intelligence

Review article

Machine learning-ai molecular design

Received: 19 July 2023

Accepted: 24 April 2024

Published online: 18 June 2024

Check for updates

Yuanqi Du[1,10]
Charles Har
Philippe Sch

Machine lea
discovery b
single archi
of medicina
that can ach
remains a cl
have addres
into a series
comprehen
using mach
such as the
and optimiz
practical ap
experiment
efforts. Fina
and empiric
highlight fu
realistic dru

Alex Zhavoronkov · 1st
Founder and CEO at Insilico Medicine
1w ·

Anyone interested in starting or ramping up in generative chemistry should read this paper and share it with their friends, students, and unrelated medicinal and computational chemists to help them get into the field.

This is probably the most concise and useful review of generative AI in small molecule chemistry, which takes 20-30 min for experts to read but provides very deep insights (11 pages, mostly figures). If you want to see what published generative models and platforms worked experimentally for which targets and with what hit rates, you can print it out and get a cheat sheet. Of course, the most commercially tractable examples are rarely published but from what I can see - pretty much every published example was captured and scrutinized. This group reconstructed the timeline and did a gargantuan amount of work that an army of modern LLM agents would not be able to complete with a high degree of accuracy since many of the numbers are in image format and require additional processing.

I was planning to write a review like this one day, but this group beat me to this and did a way better job as I saw some of the models I was not aware of. Yuanqi Du, Arian Jamasb Jeff GuoTianfan Fu Charlie HarrisYingheng W. Chenru DuanPietro Lio'Philippe Schwaller Tom Blundell

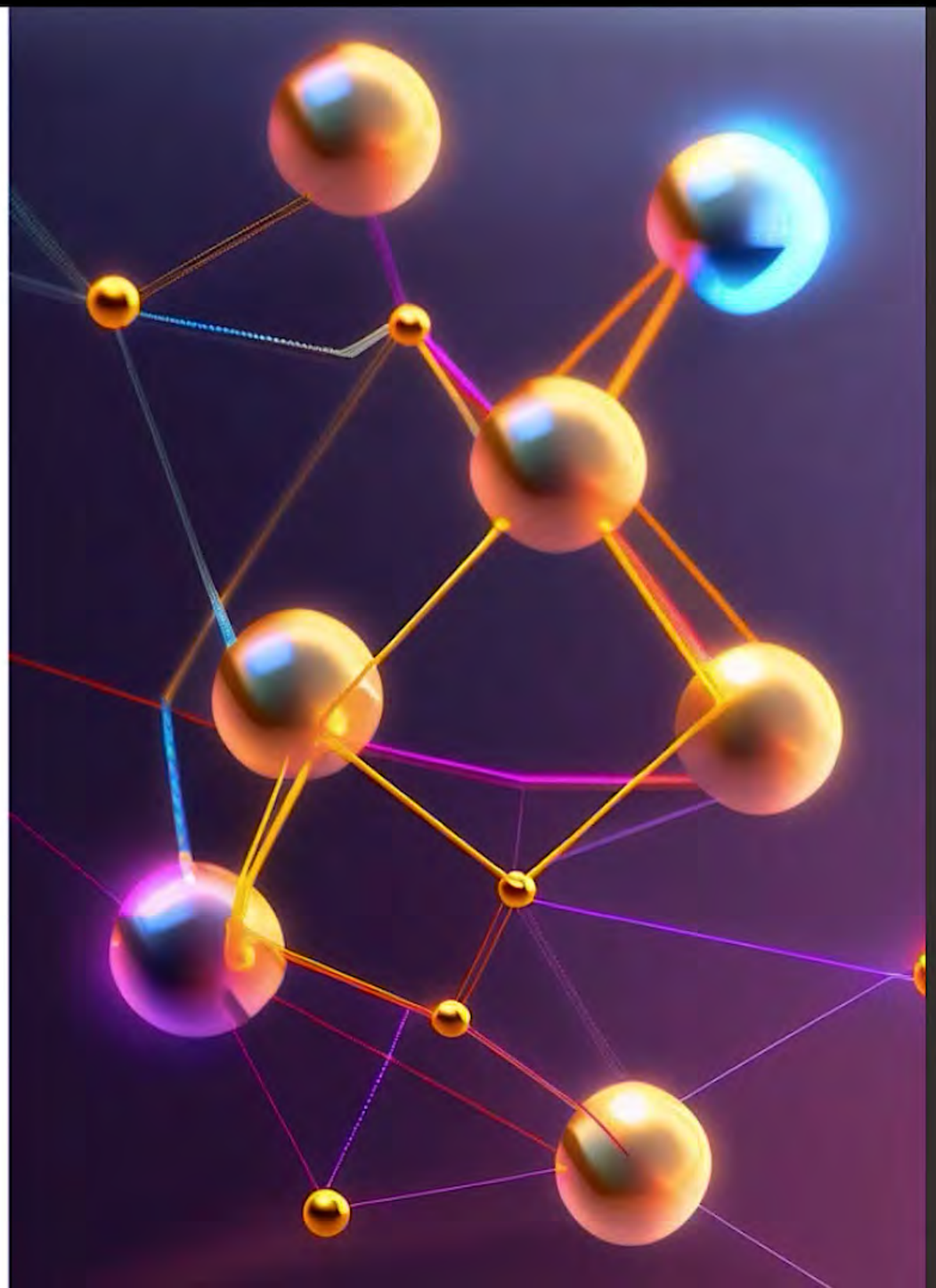
Table 4 | Experimentally validated small-molecule generative design case studies

Model	Input	Output	Design task	Target	Hit rate	Outcome	Publication year
Distribution learning							
LSTM-RNN[100]	SMILES	SMILES	De novo	RXR	4/5 (80%)	nM agonist	2018
LSTM-RNN[100]	SMILES	SMILES	De novo	RXR	2/4 (50%)	μM agonist	2018
				JAK1	7/7 (100%)	nM inhibitor	2021
				LXR	17/25 (68%)	μM agonist	2021
				RORγ	3/3 (100%)	μM agonist	2021
				FLT-3	1/1 (100%)	μM inhibitor	2022
				CDK8	9/43 (21%)	nM inhibitor	2022
				Bacteria	0/1 (0%)[a]	μM inhibitor	2022
				DDR1	2/2 (100%)	nM inhibitor	2021
				MERTK	15/17 (88%)	μM inhibitor	2022
				PI3Kγ	3/18 (17%)	nM inhibitor	2023
				TBK1	1/1 (100%)	nM inhibitor	2023
				CDK2	17/23 (74%)[c]	nM inhibitor (MC)[d]	2023
				Nurr1γ	2/6 (33%)	nM inhibitor	2023
				PPARγ	2/2 (100%)	μM agonist	2023
				Kinases	0[e]	μM inhibitor	2018
				JAK3	1/1 (100%)	μM inhibitor	2018
				DDR1	4/6 (67%)	nM inhibitor[b]	2019
				DDR1	4/6 (67%)	nM inhibitor	2021
				EGFR	4/15 (27%)	nM inhibitor	2022
				RIPK1	4/8 (50%)	nM inhibitor[b]	2022
				CDK20	6/13 (46%)[a]	nM inhibitor	2023
				CDK8	1/1 (100%)	nM inhibitor[d]	2023
				SIK2	6/6 (100%)	nM inhibitor	2023
				KOR	2/5 (40%)	μM antagonist	2023
				PHD enzymes	1/1 (100%)	nM inhibitor[b]	2024
				NLRP3	0[e]	nM inhibitor[d]	2024
				Tuberculosis ClpP	1/6 (17%)[c]	μM inhibitor	2024
				KRAS	1/12 (8%)[c]	μM inhibitor	2024
				MGLL	1/3 (33%)[c]	μM inhibitor	2024
				Pol θ	4/6 (67%)	μM inhibitor[d]	2024
				TNIK	Unknown[f]	nM inhibitor[d]	2024
				Factor Xa	Unknown[g]	μM inhibitor	2024
				HAT1 and YTHDC1	0/2 and 0/3 (0%)[a]	Both μM inhibitor[d]	2024

https://github.com/GuoJeff/generative-drug-design-with-experimental-validation

45 studies with experimentally validated drug candidates

EPFL



AdsMT: Multi-modal Transformer for Predicting Global Minimum Adsorption Energy

AdsMT: A multi-modal transformer for predicting global minimum adsorption energy

07 August 2024, Version 2

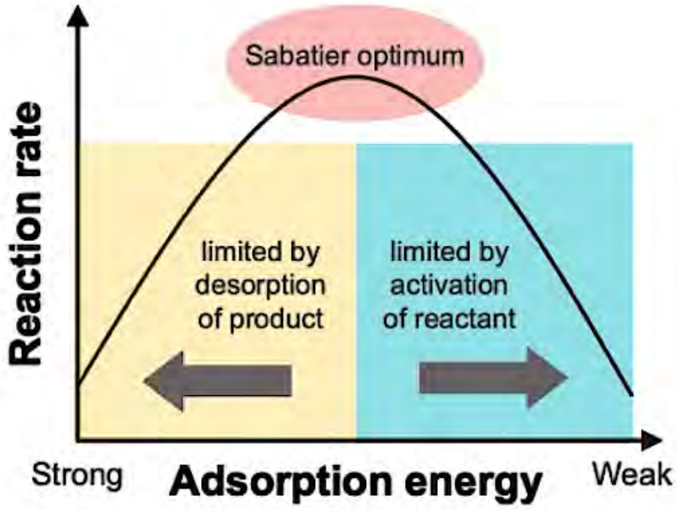
Working Paper

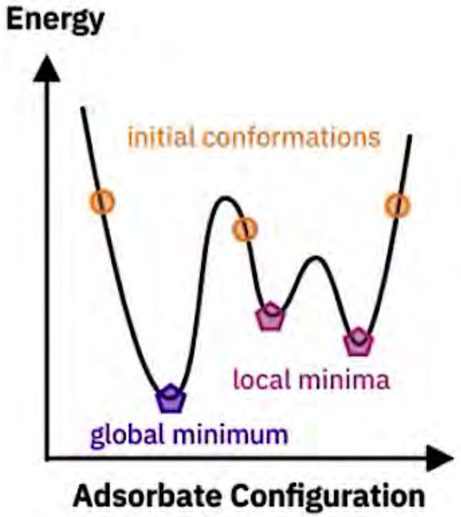
Junwu Chen, Xu Huang, Cheng Hua, Yulian He, Philippe Schwaller

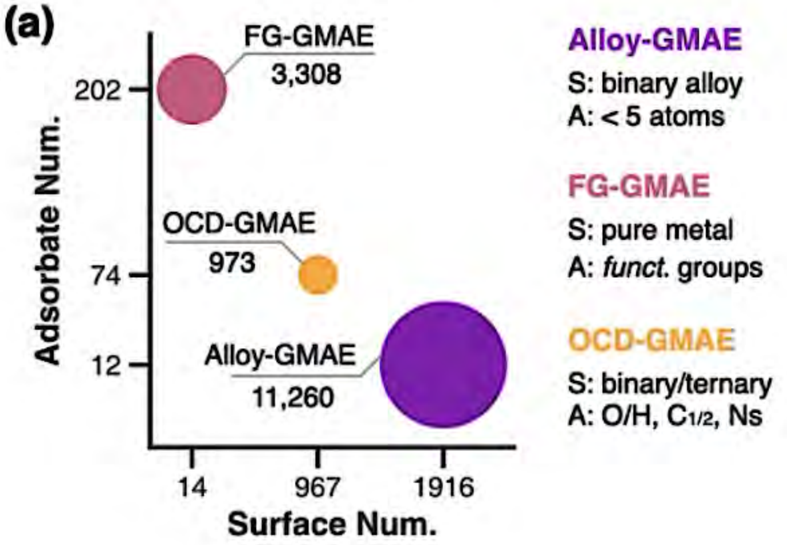
Show author details

I ♥ LIAC

Global Minimum Adsorption Energy Prediction



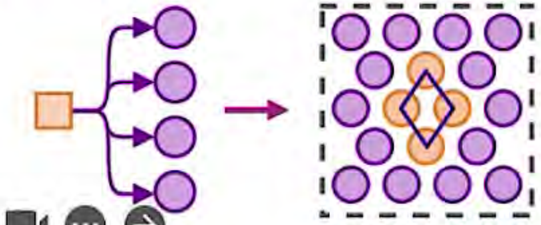


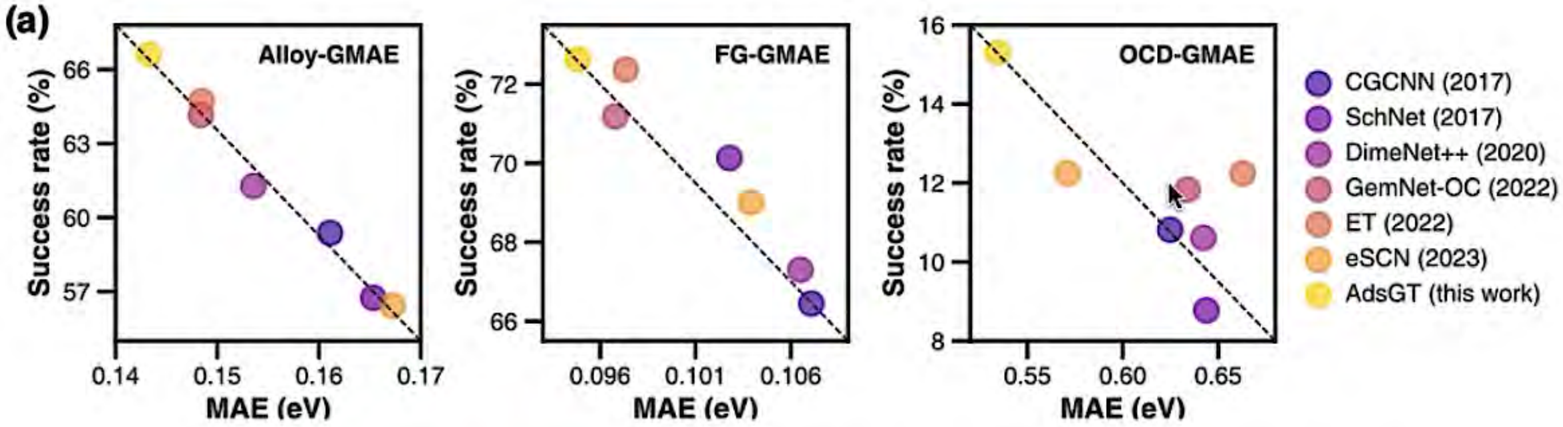


Multi-modal model for GMAE

Attention for adsorption sites

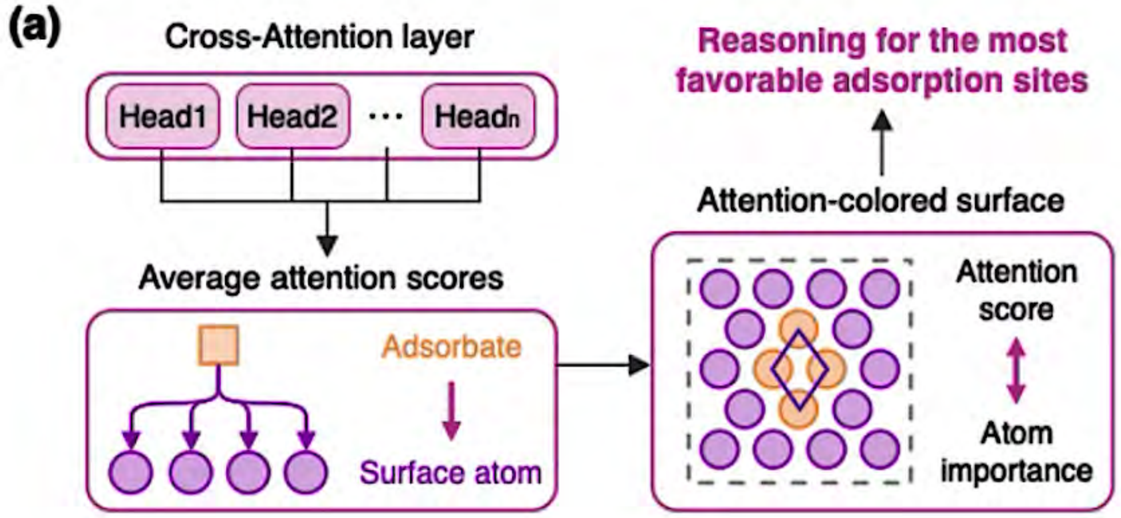
Cross attention Atom importance

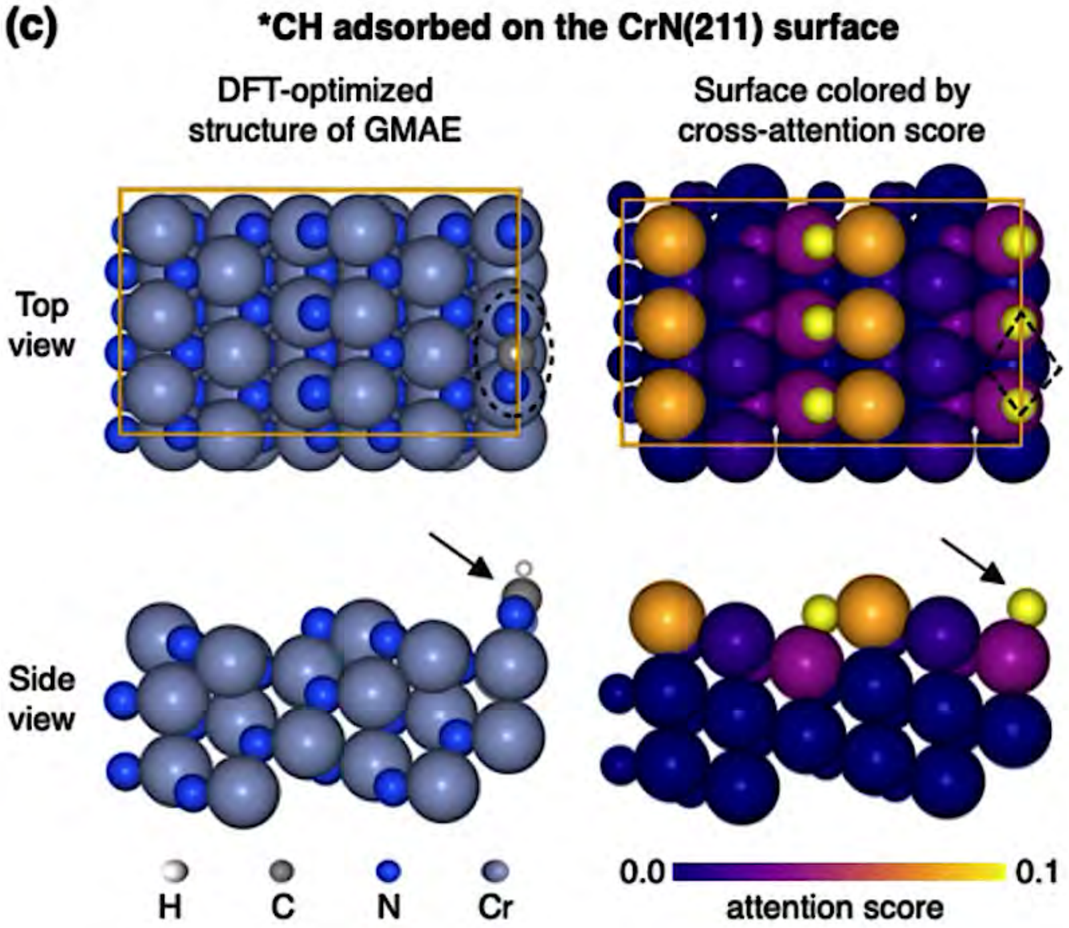




The **success rate** is the percentage of predicted GMAEs within **0.1 eV** of the DFT-computed ground truth GMAEs.

Interpretability: adsorption sites





Conclusions

- **Transformers** capture the language of chemical reactions and more...
- LLM-Powered agents have the potential to **lower access barrier** to ML/computational tools.
- ChemCrow autonomously **solves generic tasks** in chemistry. Evaluation challenging.
- **Augmented Memory/Saturn** is a step towards sample efficient de novo design.
- **FSScore** for tunable and differentiable synthetic feasibility.

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)

>10 nationalities — one team!
https://schwallergroup.github.io

Admin
Annick Delmonaco

Postdocs
Zlatko Jončev
Marta Brucka

Project students
Jenestin Anthonipillai
Remi Schlama
Anna Borisova
Magdalena Lederbauer
Tugdual Kerjan

Funding:

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

Many thanks for your attention!

@pschwllr @SchwallerGroup

https://github.com/pschwllr

https://github.com/schwallergroup

philippe.schwaller@epfl.ch

https://schwallergroup.github.io

Many thanks for your attention!

@pschwllr @SchwallerGroup

https://github.com/pschwllr

https://github.com/schwallergroup

philippe.schwaller@epfl.ch

https://schwallergroup.github.io