Automated and data-driven chemistry
Lecture 12

Closing the loop

7 December 2023

Dr. Patrick Ruch
✉ ruc@zurich.ibm.com

Senior Research Scientist
AI for Scientific Discovery
IBM **Research**

Agenda

1. Scientific discovery as a loop
2. Molecular generation
3. Optimization strategies
4. Closing the loop: case studies
5. Hands-on lab

Scientific discovery as a loop

- Evolution of the scientific method
- AI for science
- State of the art

A new era of accelerated discovery

1st Paradigm	2nd Paradigm	3rd Paradigm	4th Paradigm	5th Paradigm
Empirical Science	**Theoretical Science**	**Computational Science**	**Big data-driven Science**	**Accelerated discovery**
Observations Experimentation	Scientific laws Physics Biology Chemistry	Simulations Molecular dynamics Mechanistic models	Big data Machine learning Patterns Anomalies Visualization	Scientific knowledge at scale AI generated hypotheses Autonomous testing
Pre-Renaissance	~1600s	~1950	~2000	~2020

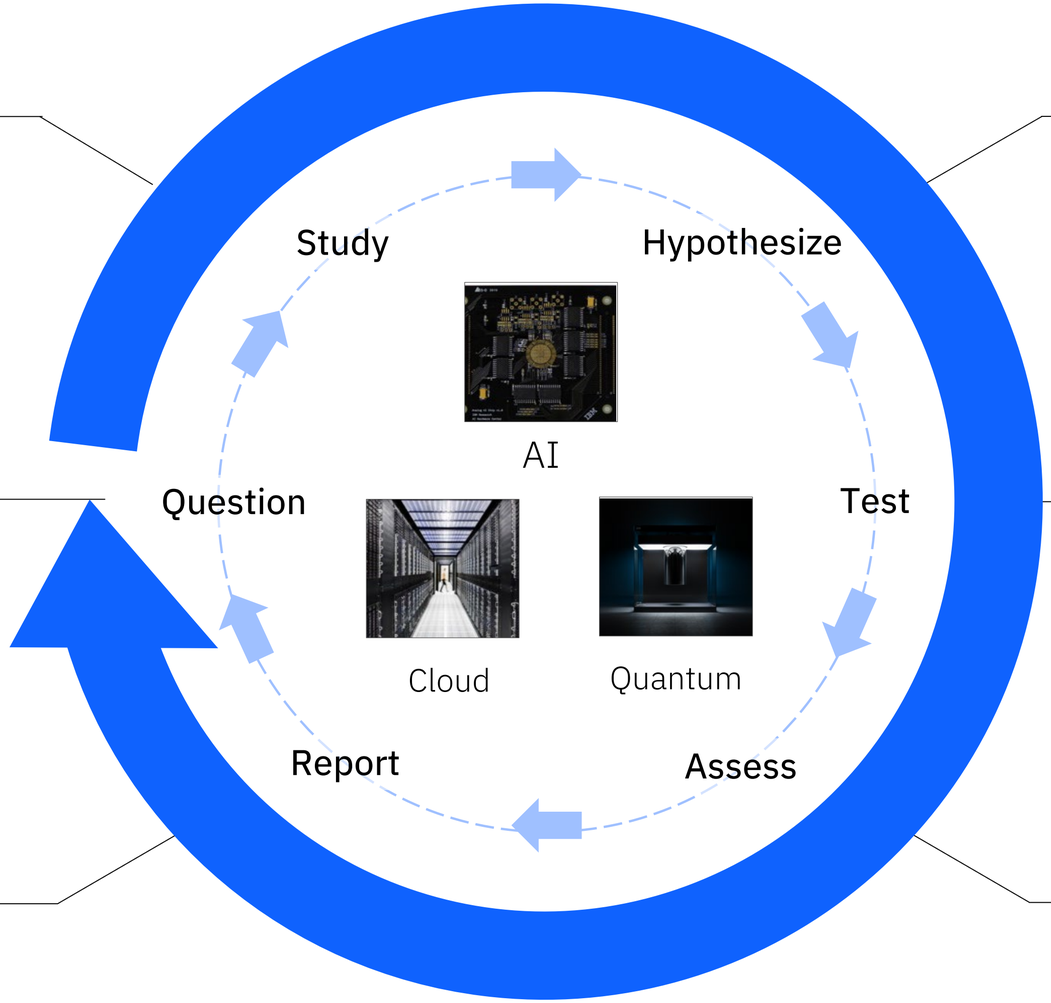
📖 Pyzer-Knapp, E. O. *et al.* Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater* **8**, 84 (2022)

The scientific method is humanity's best model for discovery

Extraction, integration and reasoning with knowledge at scale

Tools help identify new questions based on needs and gaps in knowledge

Machine representation of knowledge leads to new questions



Generative models automatically propose new hypotheses and explore a wider discovery space

Robotic labs automate experimentation and bridge digital models and physical testing

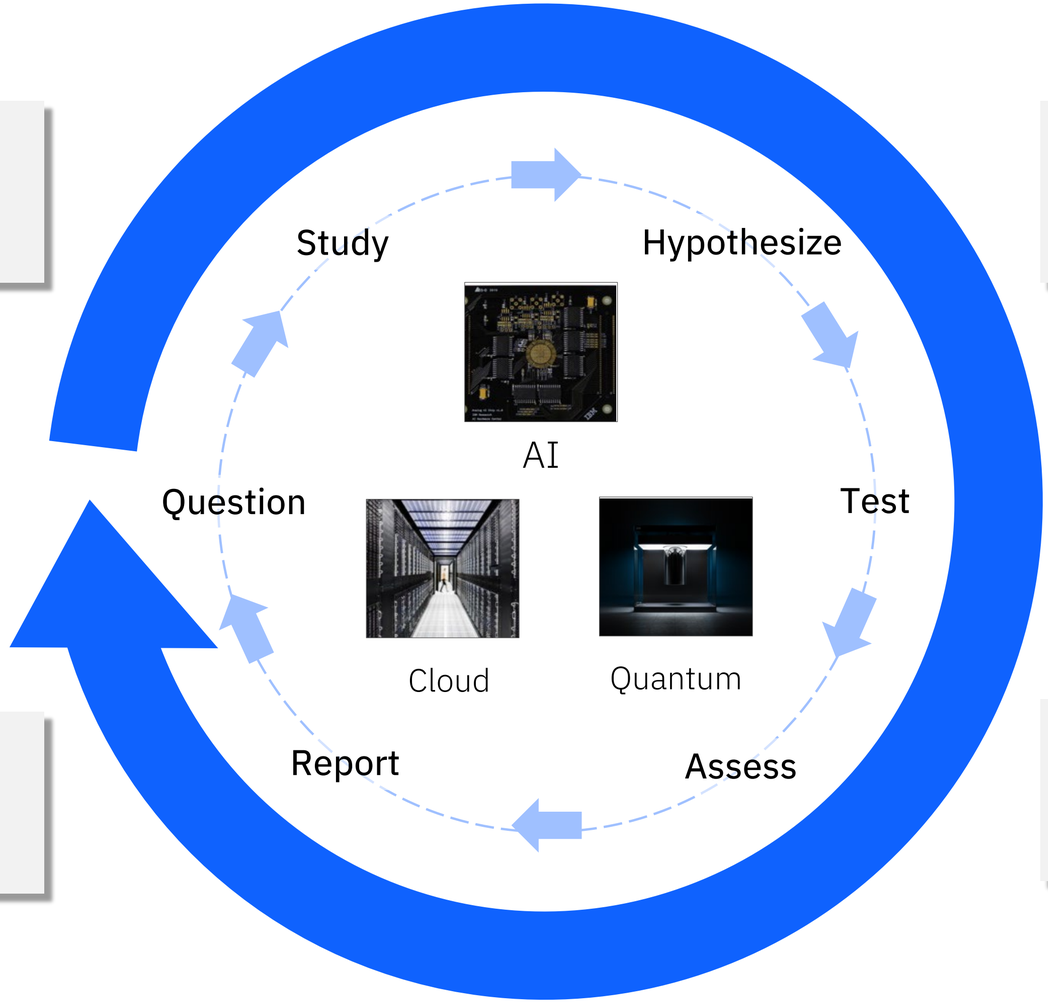
AI surrogate models are integrated with simulation and experimentation to speed up complex workloads

📖 Pyzer-Knapp, E. O. *et al.* Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater* **8**, 84 (2022)

Scientific discovery is accelerated by AI tools

Simulation Toolkit for Scientific Discovery
https://st4sd.github.io/overview/

Generative Toolkit for Scientific Discovery
https://github.com/GT4SD/gt4sd-core

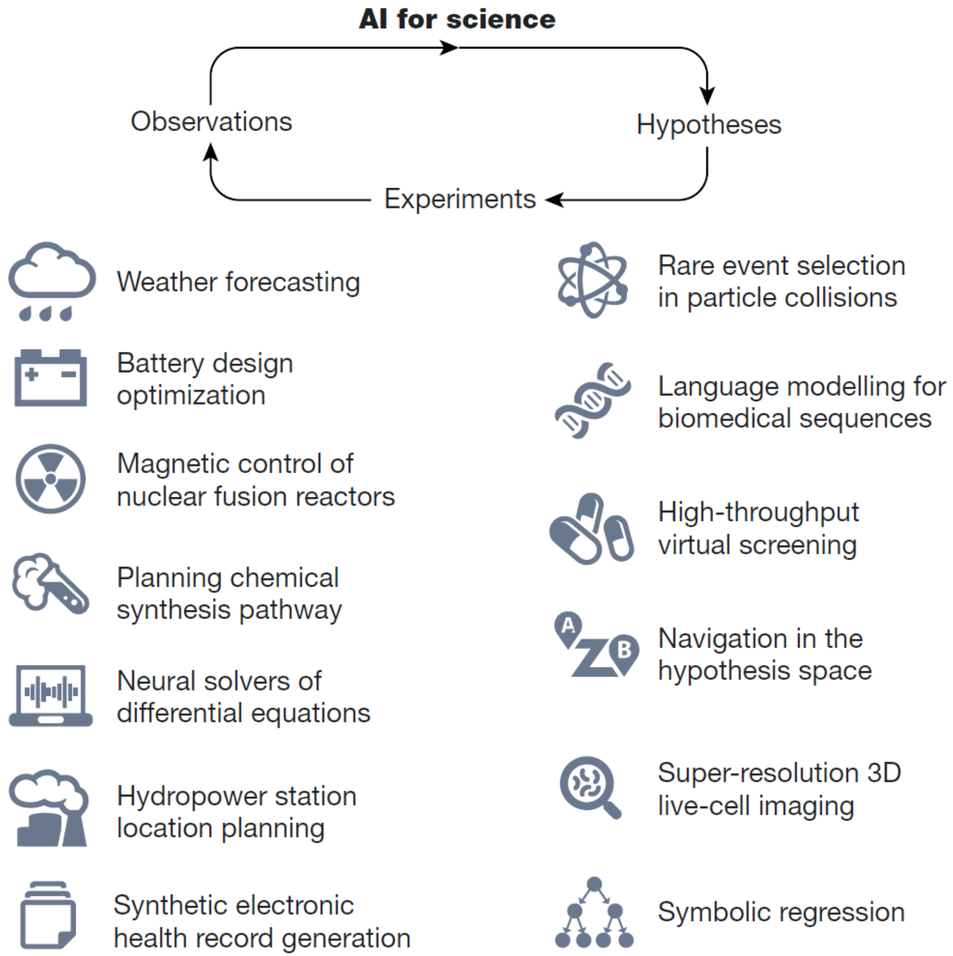


Knowledge Integration
https://ds4sd.github.io

IBM RXN for Chemistry
https://rxn.res.ibm.com

📖 Pyzer-Knapp, E. O. *et al.* Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater* **8**, 84 (2022)

The field of AI for scientific discovery



- AI-aided data collection and curation for scientific research
- Learning meaningful representations of scientific data
- AI-based generation of scientific hypotheses
- AI-driven experimentation and simulation

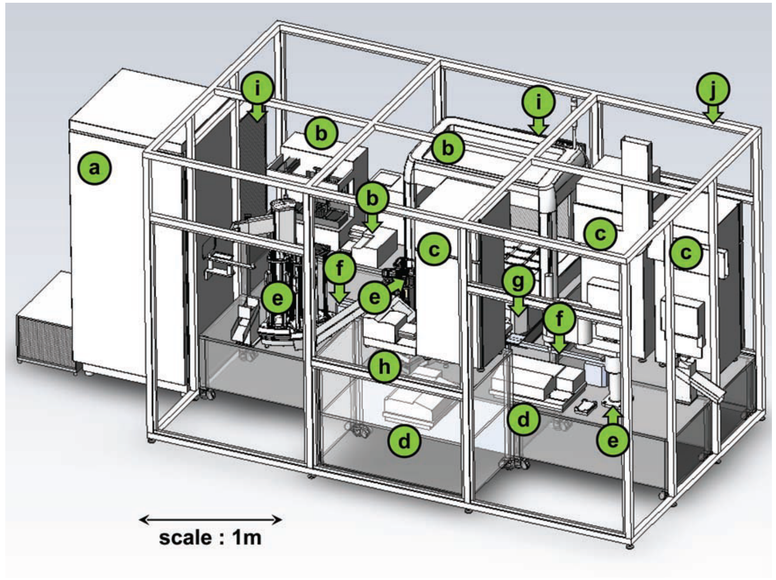
📖 Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023)

SCIENCE VOL 324 3 APRIL 2009

The Automation of Science

Ross D. King,[1]* Jem Rowland,[1] Stephen G. Oliver,[2] Michael Young,[3] Wayne Aubrey,[1] Emma Byrne,[1] Maria Liakata,[1] Magdalena Markham,[1] Pınar Pir,[2] Larisa N. Soldatova,[1] Andrew Sparkes,[1] Kenneth E. Whelan,[1] Amanda Clare[1]

The basis of science is the hypothetico-deductive method and the recording of experiments in sufficient detail to enable reproducibility. We report the development of Robot Scientist "Adam," which advances the automation of both. Adam has autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae* and experimentally tested these hypotheses



📖 King, R. D. *et al.* The Automation of Science. *Science* **324**, 85–89 (2009)

Nature | Vol 583 | 9 July 2020 | 237

Article

A mobile robotic chemist

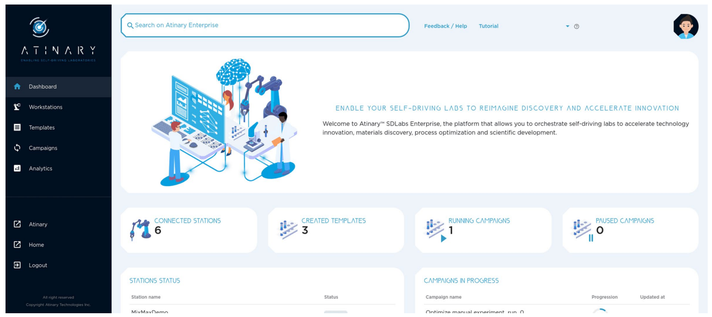
https://doi.org/10.1038/s41586-020-2442-2
Received: 1 November 2019
Accepted: 25 March 2020
Published online: 8 July 2020
Check for updates

Benjamin Burger[1], Phillip M. Maffettone[1], Vladimir V. Gusev[1], Catherine M. Aitchison[1], Yang Bai[1], Xiaoyan Wang[1], Xiaobo Li[1], Ben M. Alston[1], Buyi Li[1], Rob Clowes[1], Nicola Rankin[1], Brandon Harris[1], Reiner Sebastian Sprick[1] & Andrew I. Cooper[1]✉

Technologies such as batteries, biomaterials and heterogeneous catalysts have functions that are defined by mixtures of molecular and mesoscale components. As yet, this multi-length-scale complexity cannot be fully captured by atomistic simulations, and the design of such materials from first principles is still rare[1–5]. Likewise, experimental complexity scales exponentially with the number of variables, restricting most searches to narrow areas of materials space. Robots can assist in experimental searches[6–14] but their widespread adoption in materials research is challenging because of the diversity of sample types, operations, instruments and measurements required. Here we use a mobile robot to search for improved

📖 Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237–241 (2020)

ATINARY
ENABLING SELF-DRIVING LABORATORIES

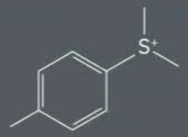


https://youtu.be/SX26XRFx0U0

Making AI for chemistry available to everyone

https://rxn.res.ibm.com/

Synthesizing new molecule



Started: Nov 30 2020, 6:49am PT

Live from IBM RoboRXN

Action 2

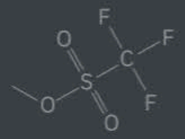
Overview

Adding $C_2H_3F_3O_3S$

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

Methyl trifluoromethane
$C_2H_3F_3O_3S$

2D 3D



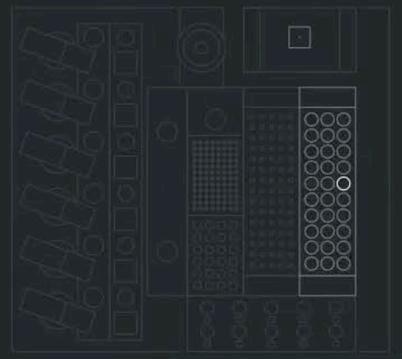
Methyl trifluoromethane sulfonate is a brown liquid. Insoluble in water. This material is a very reactive methylating agent, also known as methyl triflate.

NOW

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

Position of the robot arm

Moving to Vial 61



Live view module

2 Adding $C_2H_3F_3O_3S$

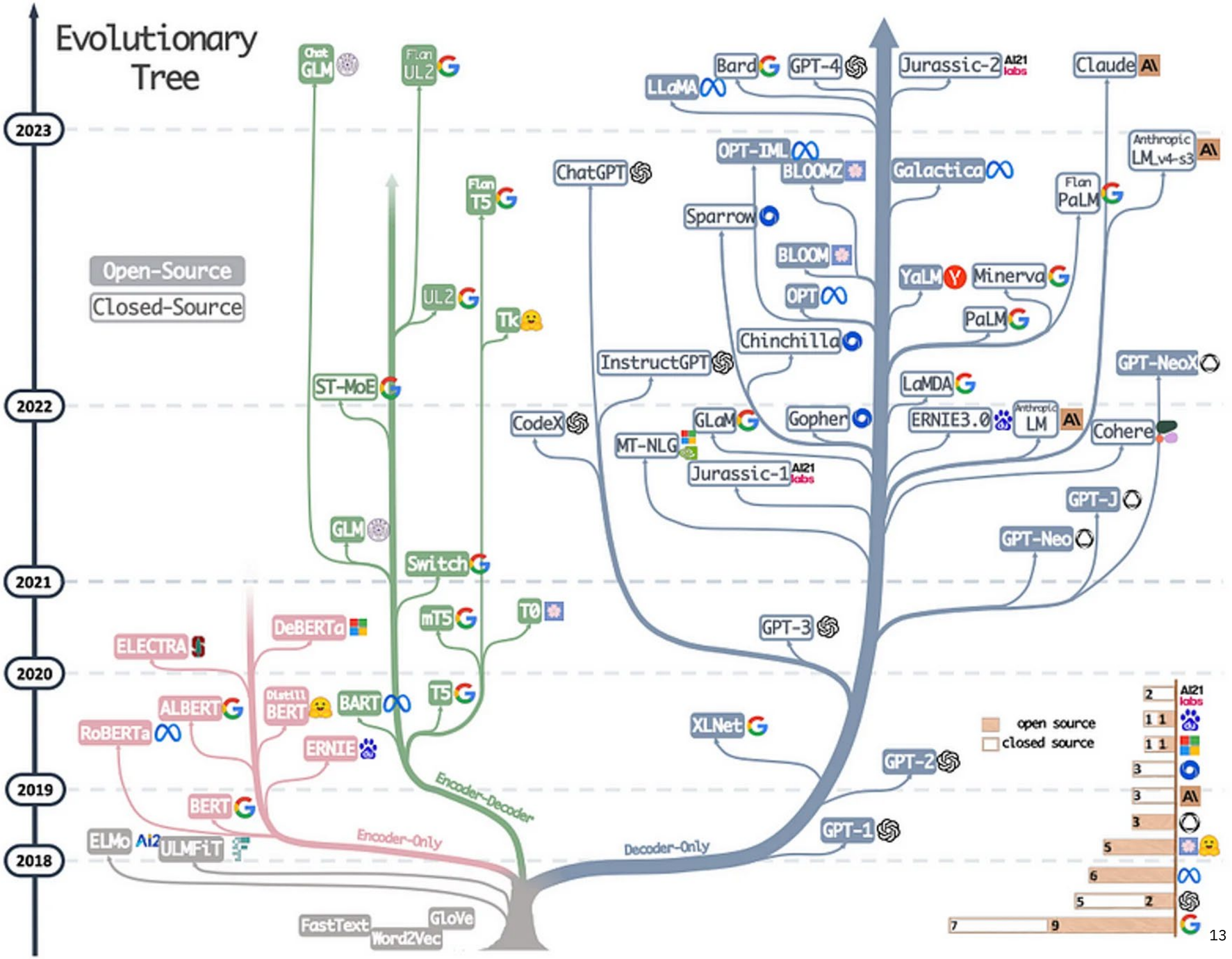
00:06:00

LIVE

Molecular generation

- Language models for molecular discovery
- Generative AI
- Graph neural networks

The world of large language models (LLMs)

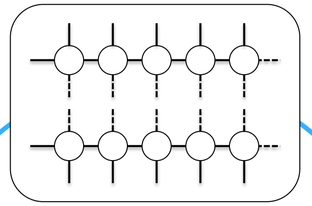


Credit: https://blog.sylphai.com/introduction-to-large-language-models

The same AI breakthroughs for language are changing scientific discovery

Generative modeling and **transformers** are achieving new breakthroughs in scientific disciplines

Natural Language Processing (NLP)



Transformers

AI

Language

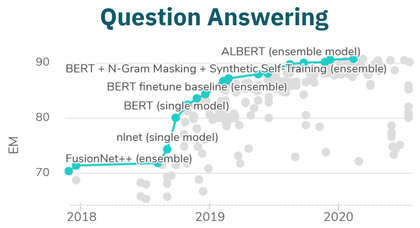
Sentences

Words

Letters

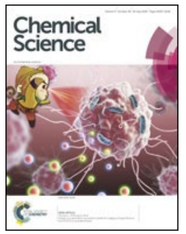
Data

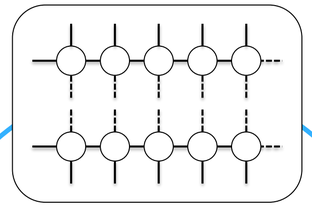
AI Models



Improved Language Tasks

Accelerated Chemistry





Transformers

AI

Chemistry

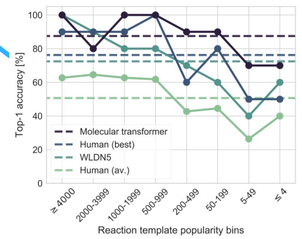
Properties/Reactions

Molecules

Atoms

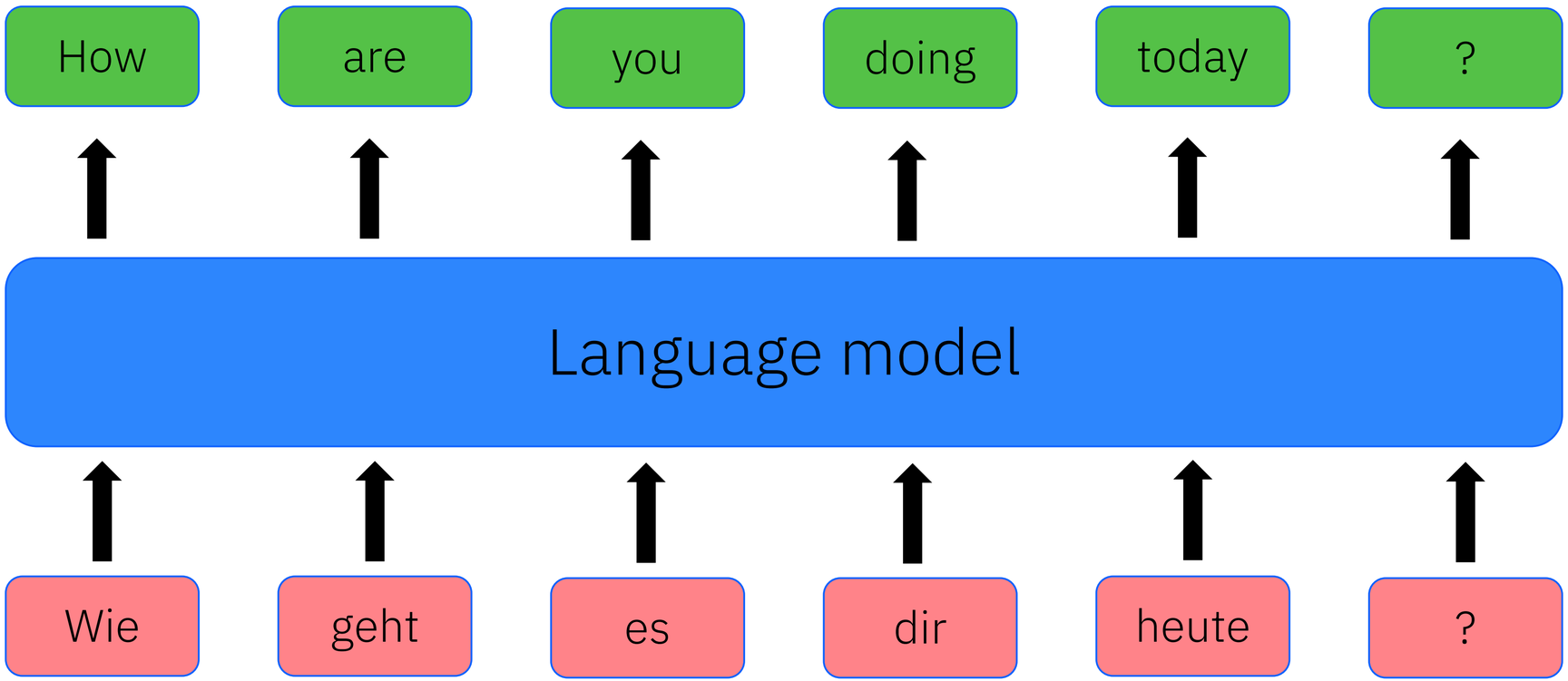
Data

AI Models



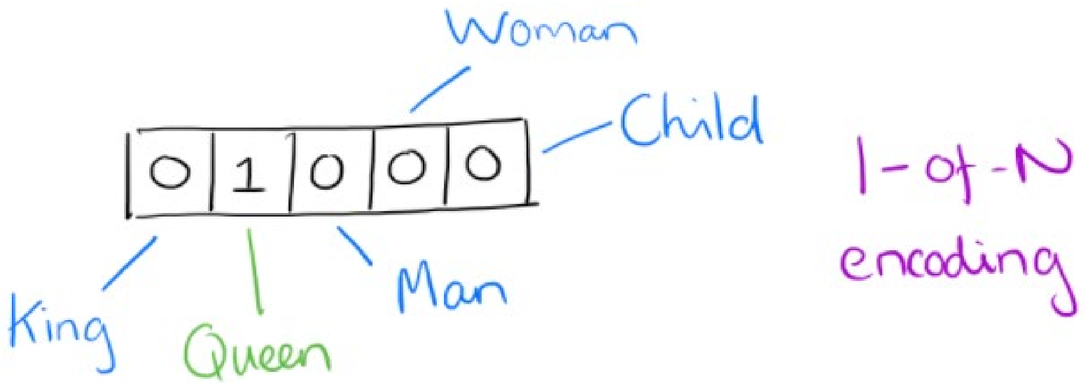
Improved Chemistry Tasks

Translate from German to English

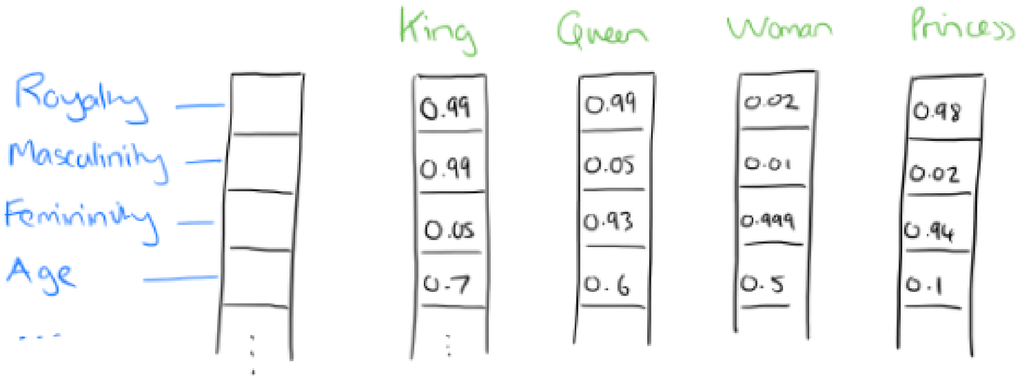


Representing words as input to a neural network

One-hot encoding?



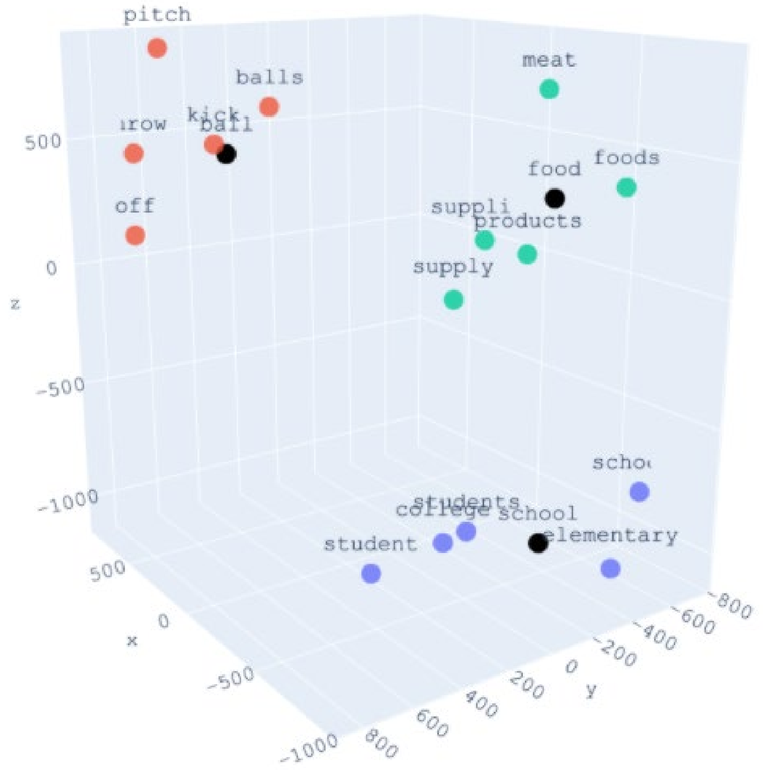
Learned encodings!



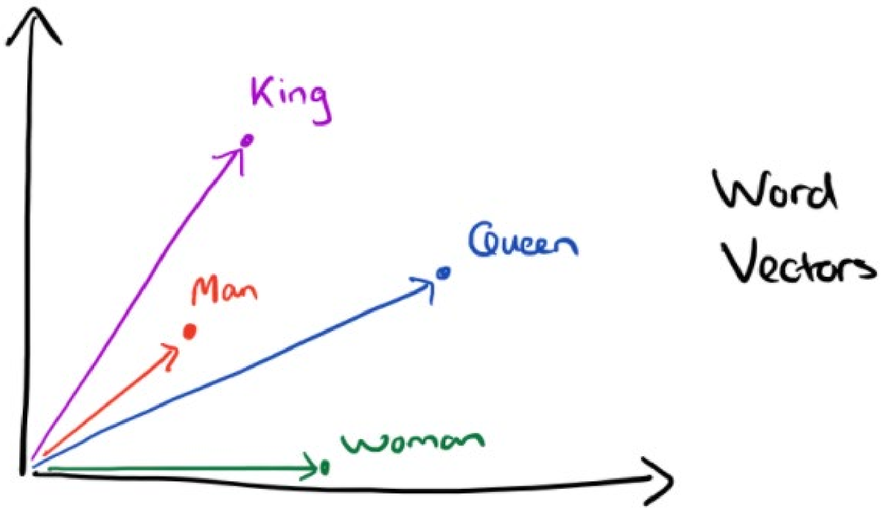
Word2Vec: 🕮 Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546* (2013)

Words as vectors

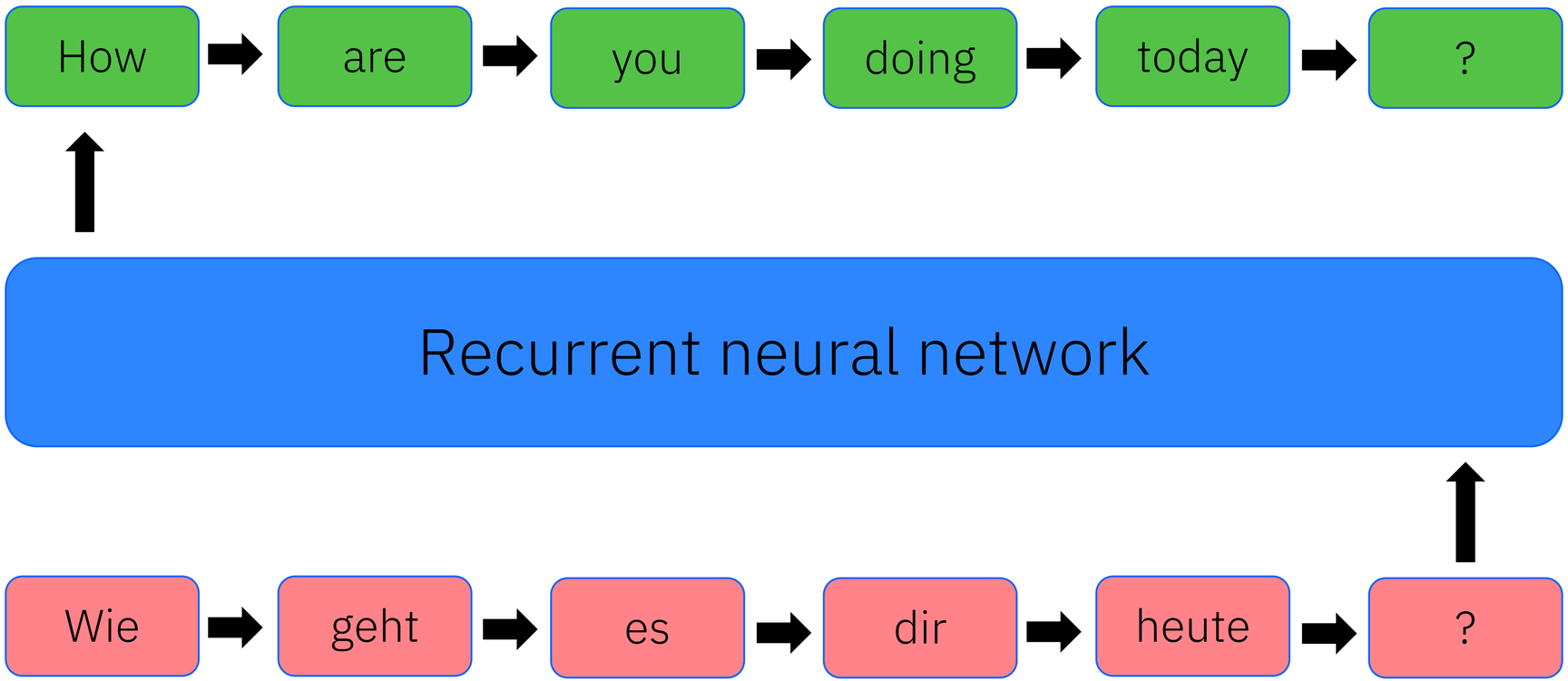
Similar words cluster



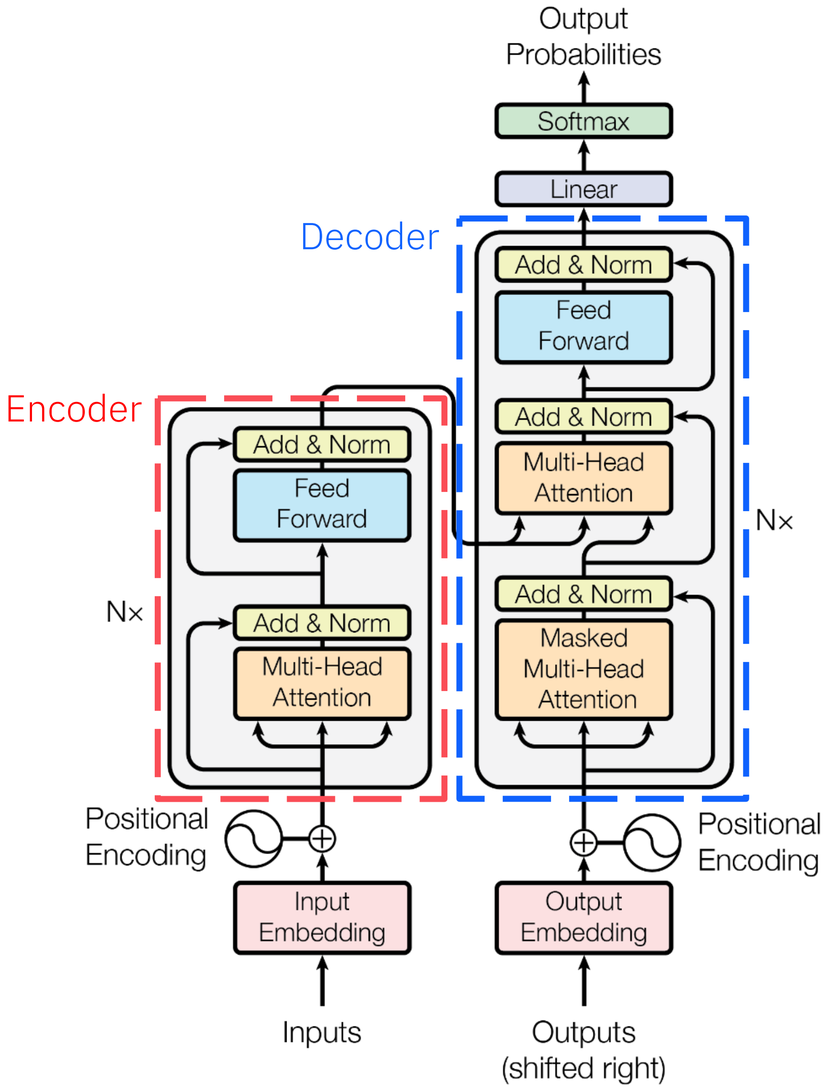
Word arithmetics



Until 2017, language models were based on recurrent neural networks (RNNs)



The transformer model



📖 Vaswani, A. *et al.* Attention is All you Need. *NeurIPS* (2017)

Key concept: **parallel multi-head attention mechanism**

- Tokenizers convert input (such as text) to tokens
- Embedding layer converts tokens and their positions to vector representations
- Transformer layers consisting of alternating attention and feedforward layers extract linguistic information

Superior translation quality and less training effort compared to previous state-of-the-art with recurrent neural networks (RNNs).

Encoder-only models:

Best suited for classification tasks
(e.g. sentence and word classification, entity recognition, extractive Q&A)

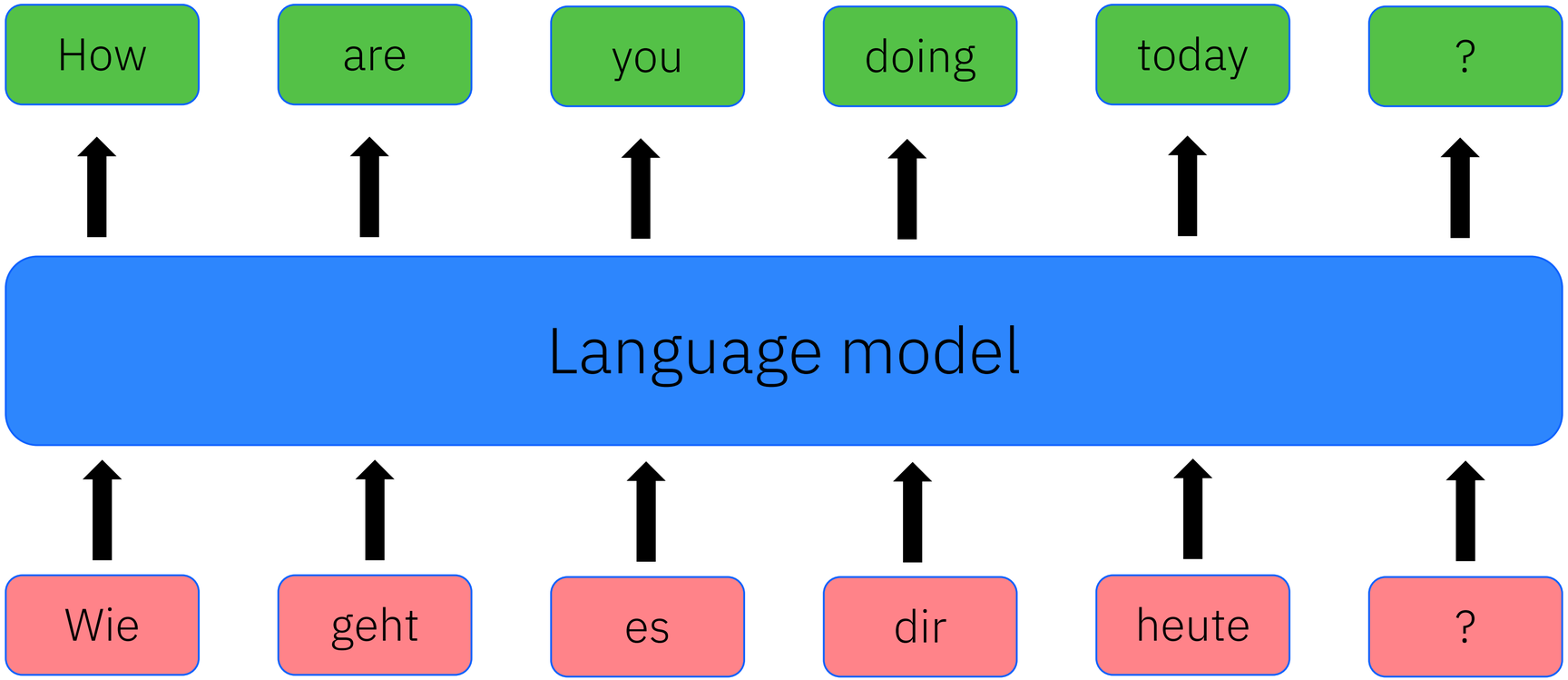
Decoder-only models (also called *auto-regressive models*)

Best suited for generative tasks
(e.g. text generation)

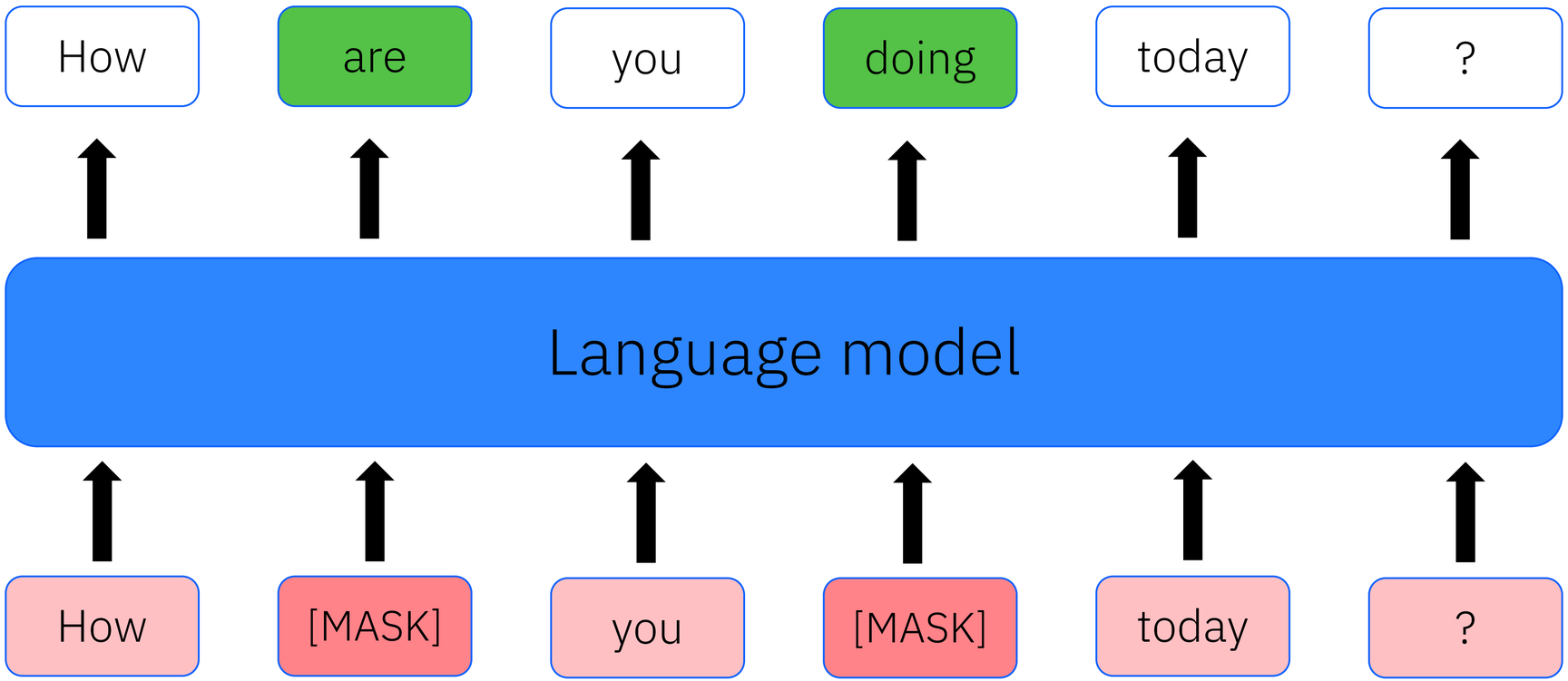
Encoder-decoder models (also called *sequence-to-sequence models*)

Best suited for generative tasks depending on a given input
(e.g. summarization, translation, generative Q&A)

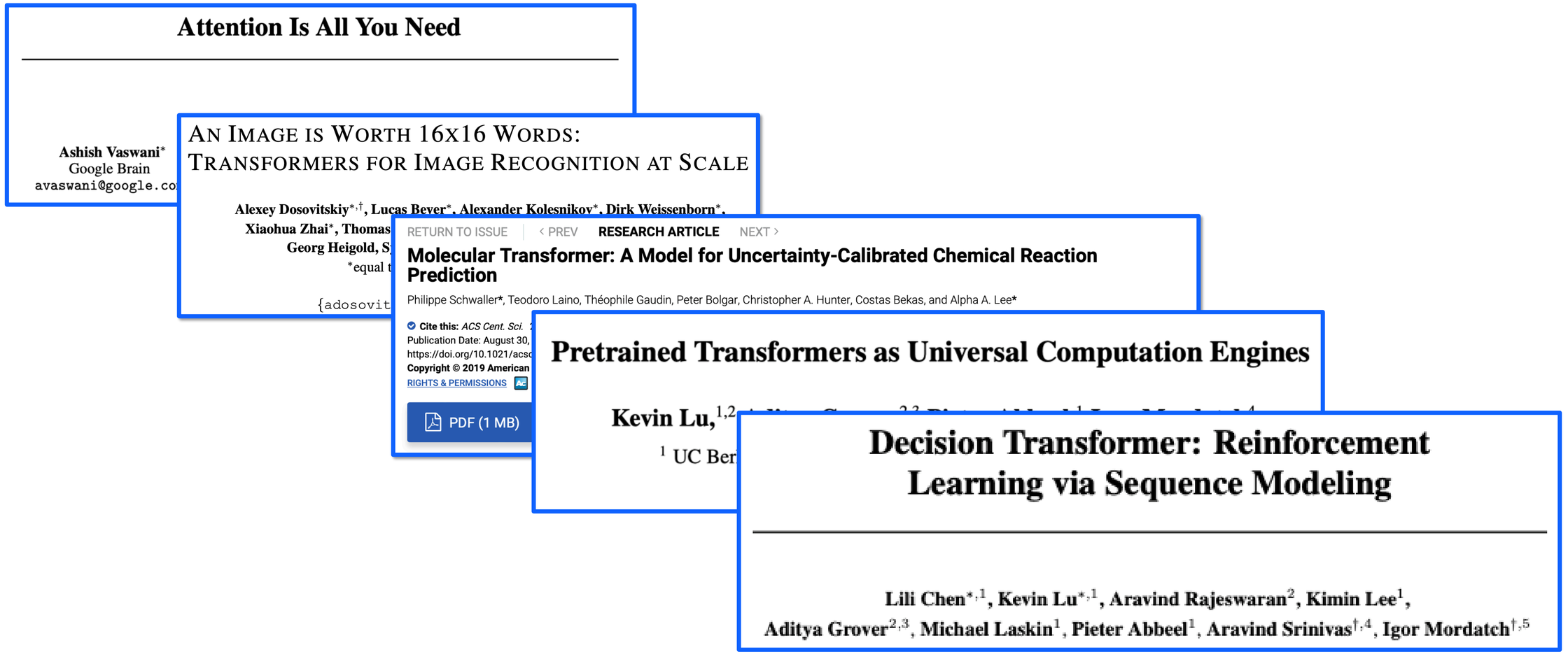
Since 2017, transformers networks are the state-of-the-art for language models



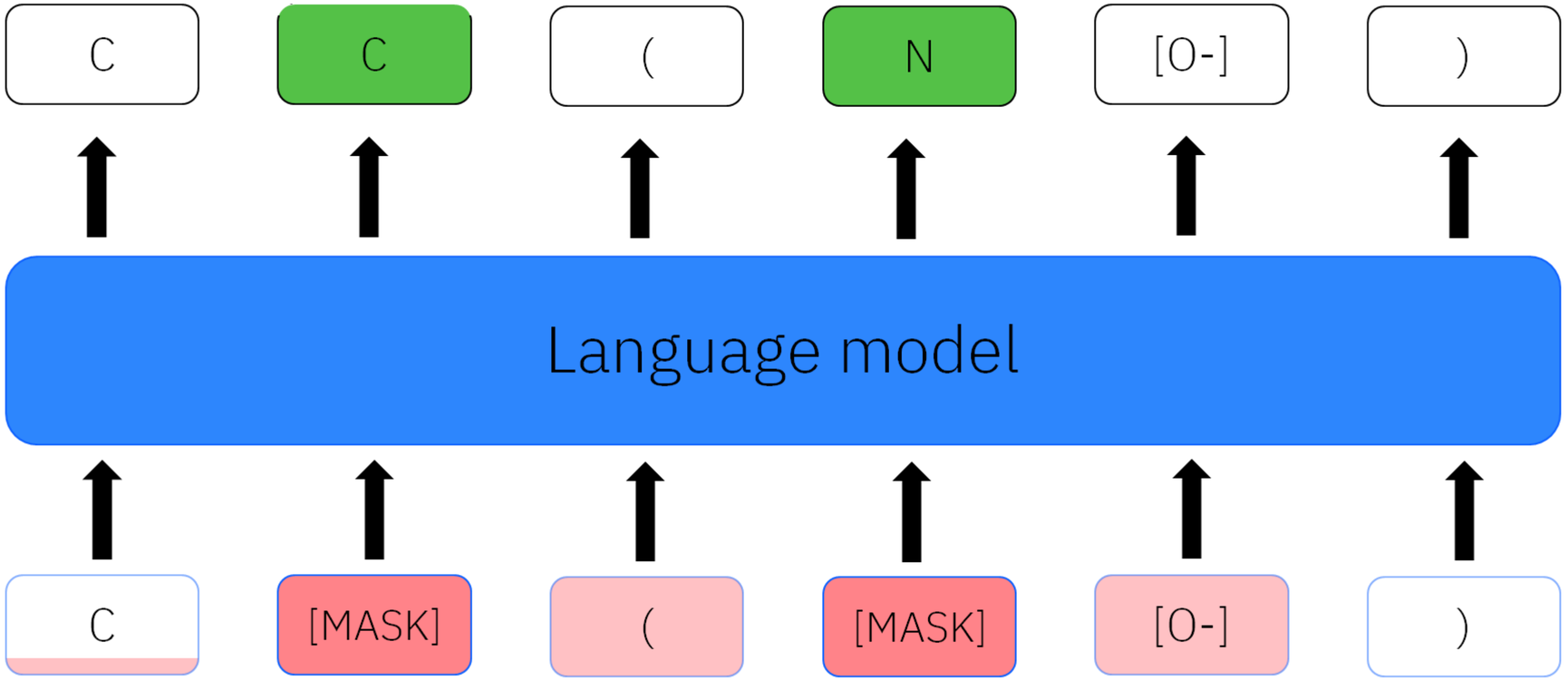
Masked language modelling: training transformers with self-supervision



Transformers impacted multiple application domains

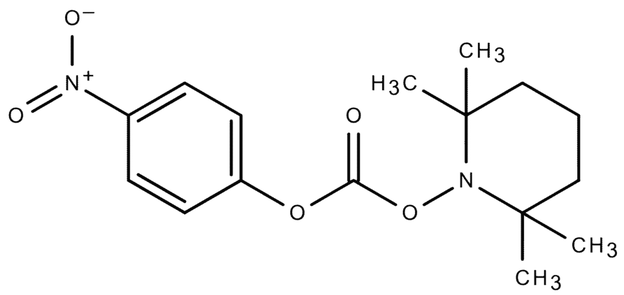


Chemical language models



Learning the language of chemistry

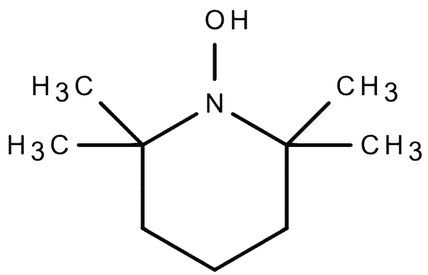
Products

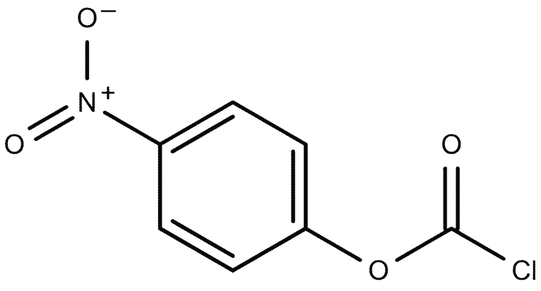


retrosynthesis

synthesis

Reactants + Reagents





Textual representation (SMILES)

CC1(CCCC(N1OC(OC1=CC=C([N+](=O)[O-])C=C1)=O)(C)C)C

CC1(CCCC(N1O)(C)C)C

O(C(=O)Cl)C1=CC=C([N+](=O)[O-])C=C1

“Sentence of atoms”

```
C C 1 ( C C C C ( N 1 O C ( O C 1
= C C = C ( [N+] ( = O ) [O-] ) C
    = C 1 ) = O ) ( C ) C ) C
```

⟶⟵

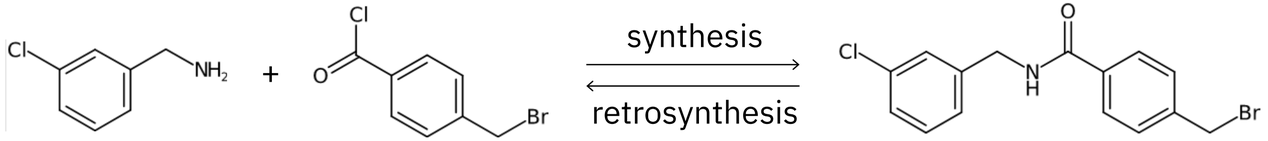
```
C C 1 ( C C C C ( N 1 O ) ( C ) C ) C . O ( C ( = O ) Cl )
      C 1 = C C = C ( [N+] ( = O ) [O-] ) C = C 1
```

Translation

English ⟶⟵ Spanish

Automating lab synthesis and experimentation
with help of language models

1. AI-based chemical reaction prediction



N C c 1 c c c c (Cl) c 1 . O = C (Cl) c 1 c c c (C Br) c c 1 ⟷ O = C (N C c 1 c c c c (Cl) c 1) c 1 c c c (C Br) c c 1

"translation"
transformer

📖 Schwaller et al., *Chem. Sci.* **9** (2018) 6091
📖 Schwaller et al., *ACS Cent. Sci.* **5** (2019) 1572
📖 Schwaller et al., *Chem. Sci.* **11** (2020) 3316

2. Chemical procedures from text (Paragraph2Actions)

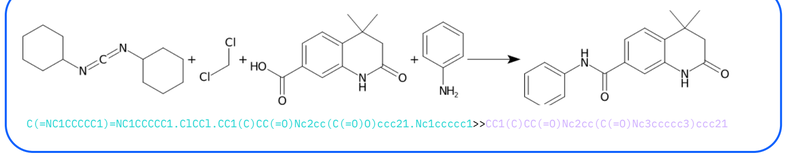
```
The TFA was removed in vacuo and a
saturated solution of NaHCO3 was added.
```

translation →

```
Concentrate(),
Add(name='saturated solution of NaHCO3')
```

📖 Vaucher et al., *Nat. Comm.* **11** (2020) 3601

3. Chemical procedures from reactions (Smiles2Actions)



1. ADD 1
2. ADD 4
3. ADD 2
4. ADD 3
5. STIR for @3@ at #4#
6. FILTER keep precipitate
7. RECRYSTALLIZE from ethanol
8. YIELD -1

📖 Vaucher et al., *Nat. Comm.* **12** (2022) 2573

Automated synthesis via cloud + robotic lab

https://rxn.res.ibm.com/rxn/robo-rxn

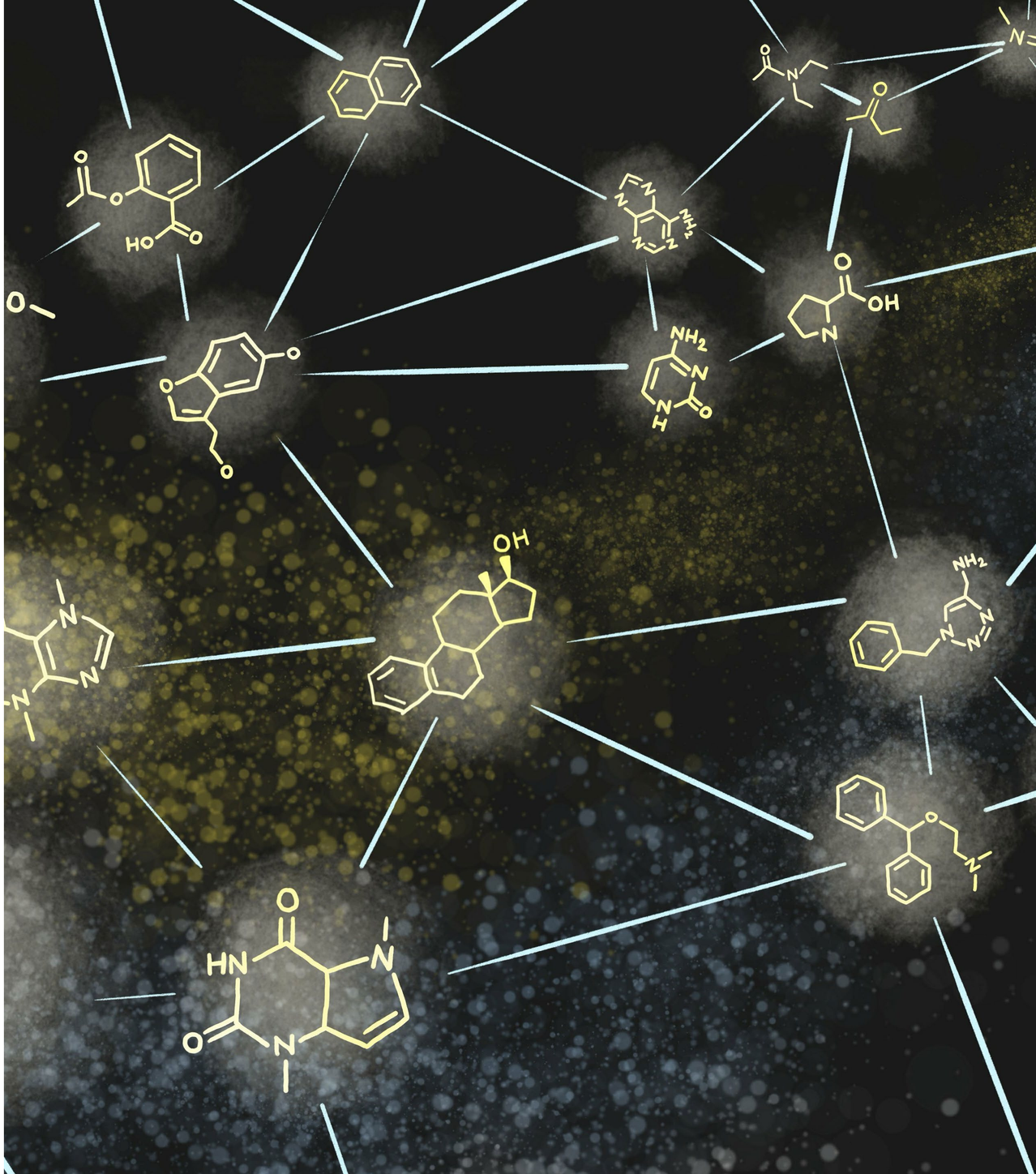
30k+
global users
via cloud

10+ million
reaction
predictions

The size of chemical space

Number of compounds	Reference
$6{,}2 \times 10^{13}$	Henze and Blair [4]
$1{,}3 \times 10^{15}$	Blair and Henze [5]
10^{21}	Weaver and Weaver [8]
10^{23}	Ertl [7]
10^{26}	Ogata et al. [24]
10^{33}	Weininger [23]
10^{33}	This work (see ref. below)
10^{60}	Bohacek et al. [6]
10^{100}	Walters et al. [26]
10^{180}	Weininger [27]

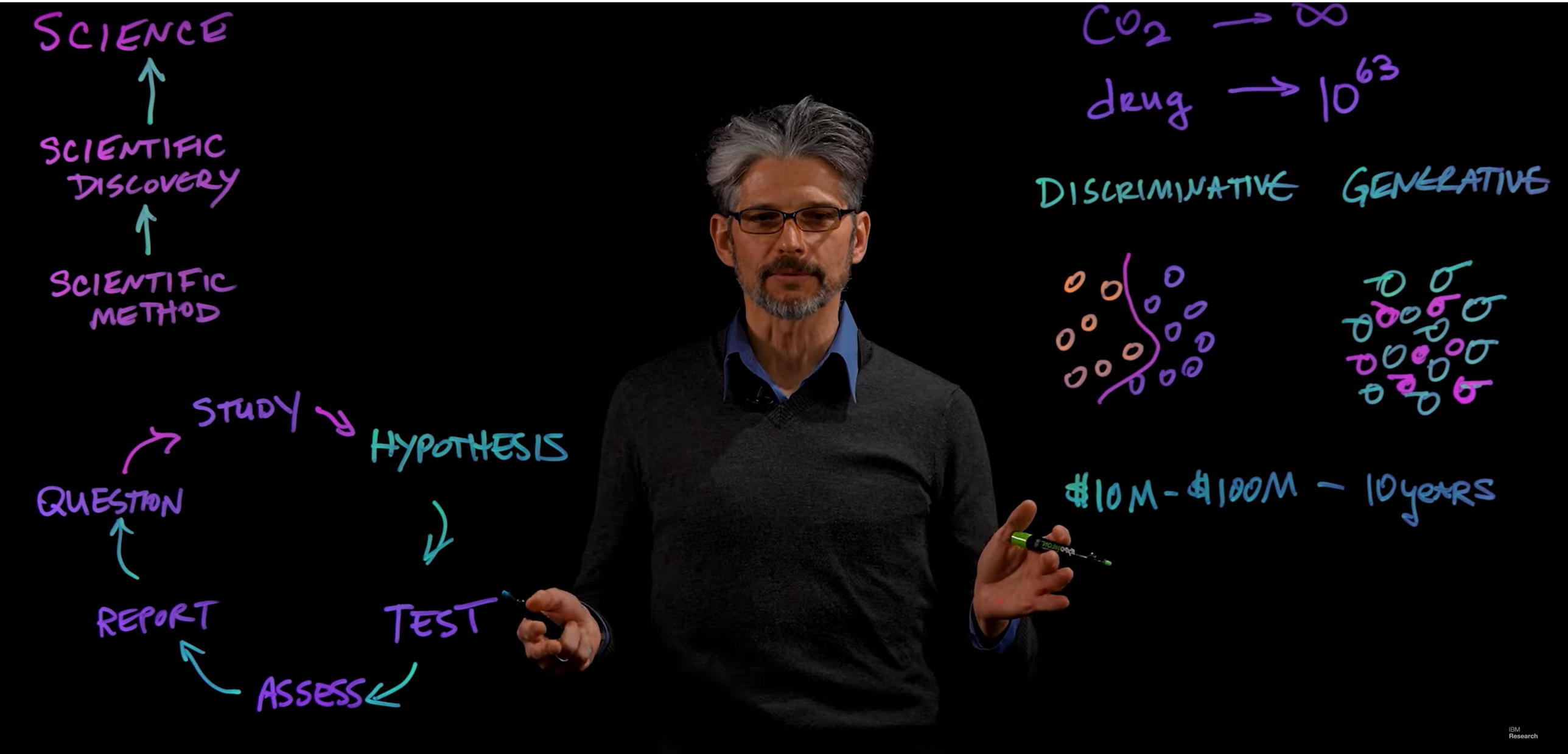
📖 Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aided Mol Des* **27**, 675–679 (2013)



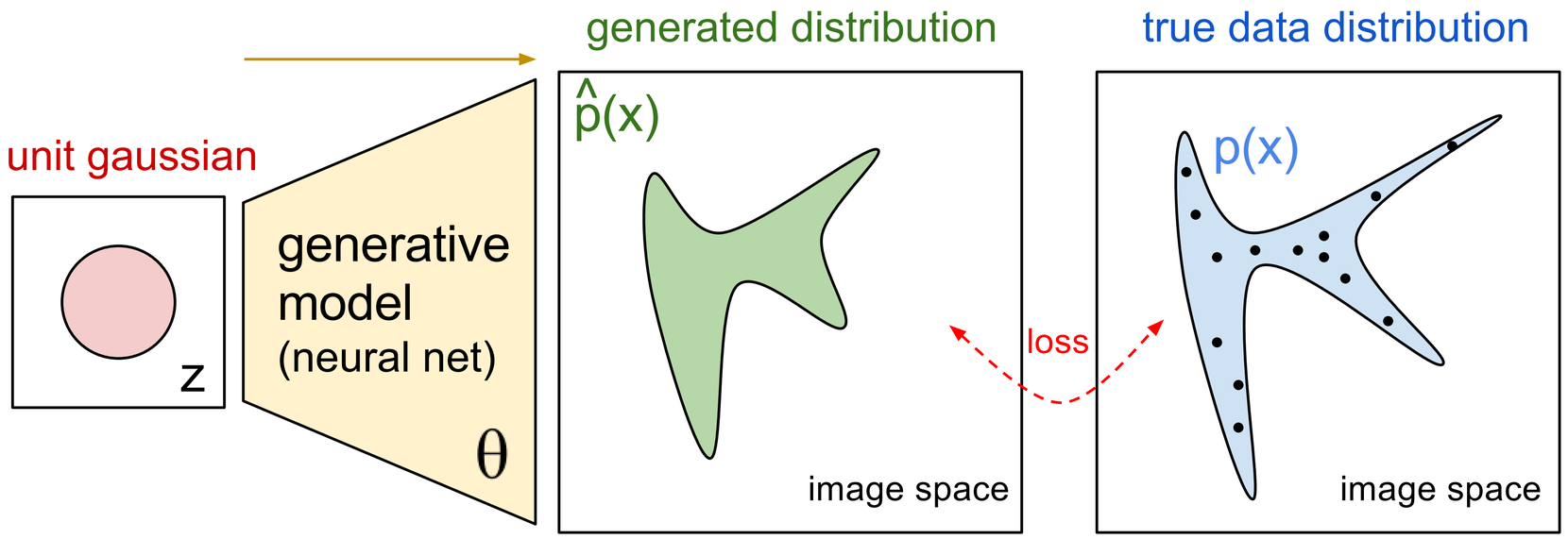
Chemical Space | Naomi Johnson, Lee Cronin

Generative AI for discovery

How can generative models fuel scientific discovery?
https://youtu.be/xsx14v1CRu8?si=jWg2BIu-mADyr3rM
John Smith, IBM Fellow (2022)



Principle of generative models



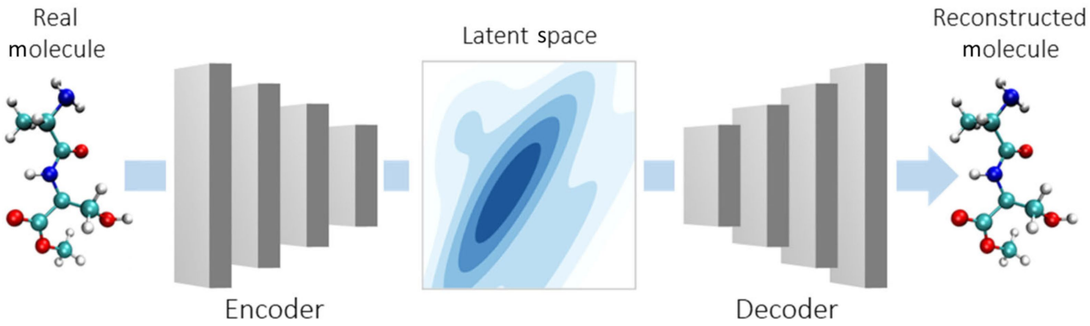
Credit: https://openai.com/research/generative-models

Example approaches to generative models:

- Variational autoencoders (VAEs)
- Generative adversarial networks (GANs)
- Generative flow networks (GFNs)
- Diffusion models (DMs)

Examples of generative models

Variational autoencoders

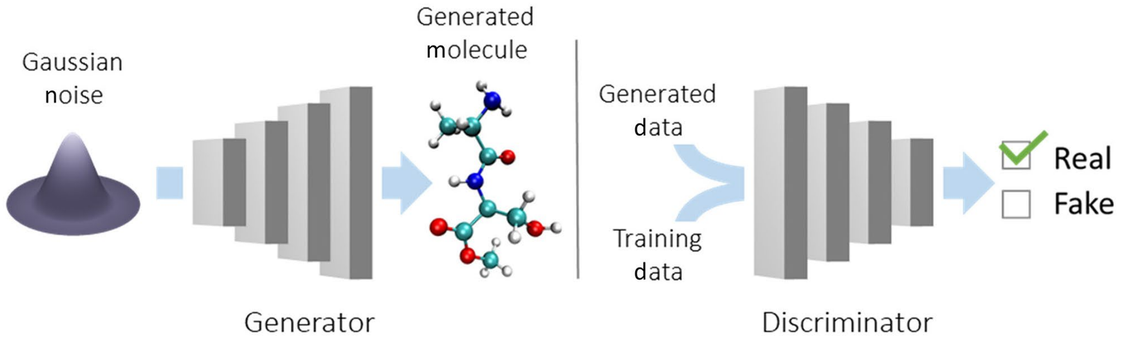


Credit: Bilodeau *et al.* (2022)

- Consist of an encoder and a decoder
- Learn the best encoding-decoding scheme through iterative optimization
- Loss function with two terms: reconstruction loss and regularization term

📖 Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114v11* (2022)

Generative adversarial networks

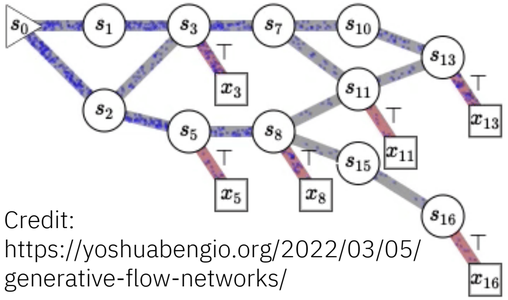


Credit: Bilodeau *et al.* (2022)

- Consist of a generator and a discriminator
- Adversarial training of generator to trick discriminator
- Tends to produce high fidelity, risk of collapsing to low diversity

📖 Goodfellow, I. *et al.* Generative Adversarial Nets. *NeurIPS* (2014)

Generative flow networks

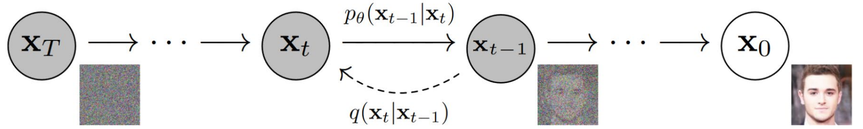


Credit: https://yoshuabengio.org/2022/03/05/generative-flow-networks/

- Allow neural nets to model distributions over data structures like graphs
- Generates a series of actions at a frequency proportional to their reward
- Improves samples diversity and provides non-iterative sampling mechanism

📖 Bengio, E., Jain, M., Korablyov, M., Precup, D. & Bengio, Y. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. *NeurIPS* (2021)

Diffusion models



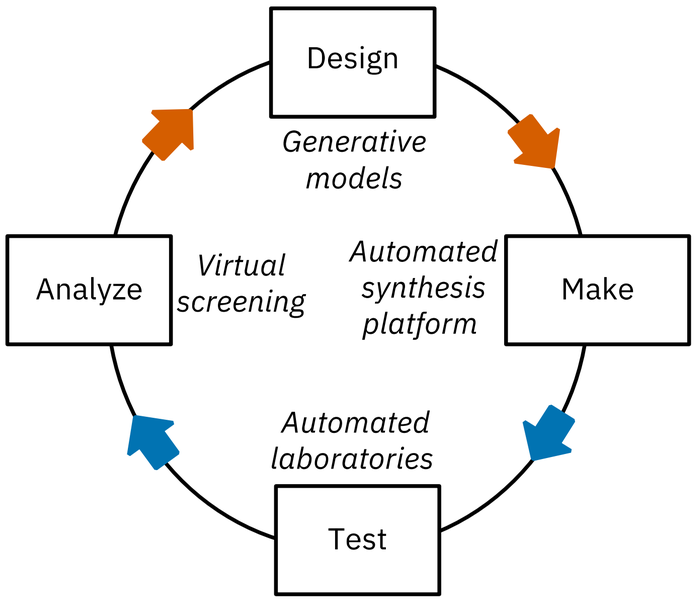
Credit: Ho *et al.* (2022)

- Learn high-dimensional distributions by denoising data at multiple scales
- Fixed forward diffusion process and learnable reverse diffusion process
- High fidelity and diversity samples, but slow sample generation

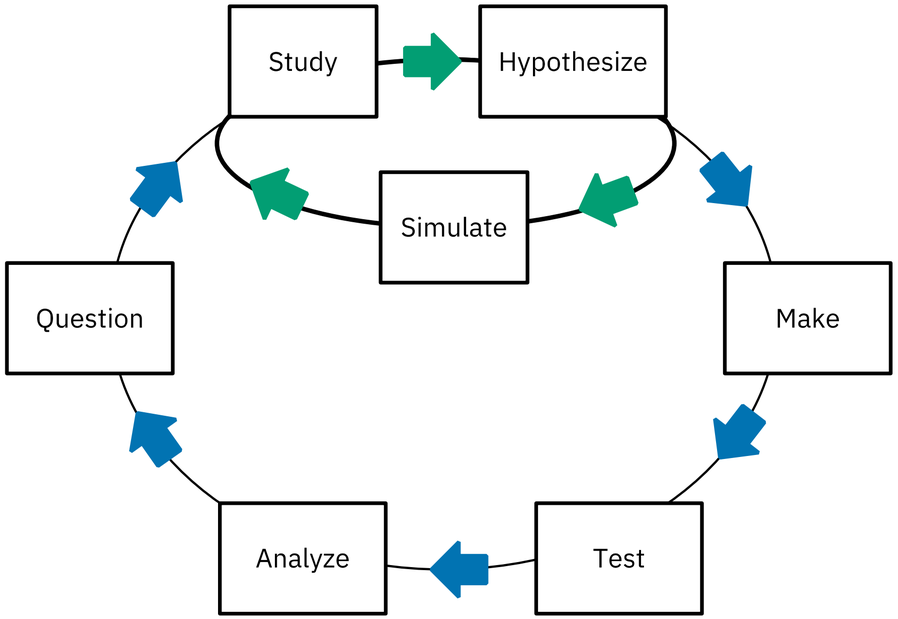
📖 Bengio, E., Jain, M., Korablyov, M., Precup, D. & Bengio, Y. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. *NeurIPS* (2021)

Generative models in molecular discovery

Design-Make-Test-Analyze cycle in chemistry



Accelerated molecular discovery



Generative Toolkit for Scientific Discovery (GT4SD): an open-source library to accelerate hypothesis generation in scientific discovery

GT4SD makes generative AI algorithms and models easier to use in scientific discovery
https://github.com/GT4SD/gt4sd-core

1. Train generative models

```
gt4sd-trainer  --training_pipeline_name paccmann-vae-trainer --epochs 250 t
```

2. Create inference pipelines

```
gt4sd-saving --training_pipeline_name paccmann-vae-trainer --model_path /t t
```

3. Run inference pipelines

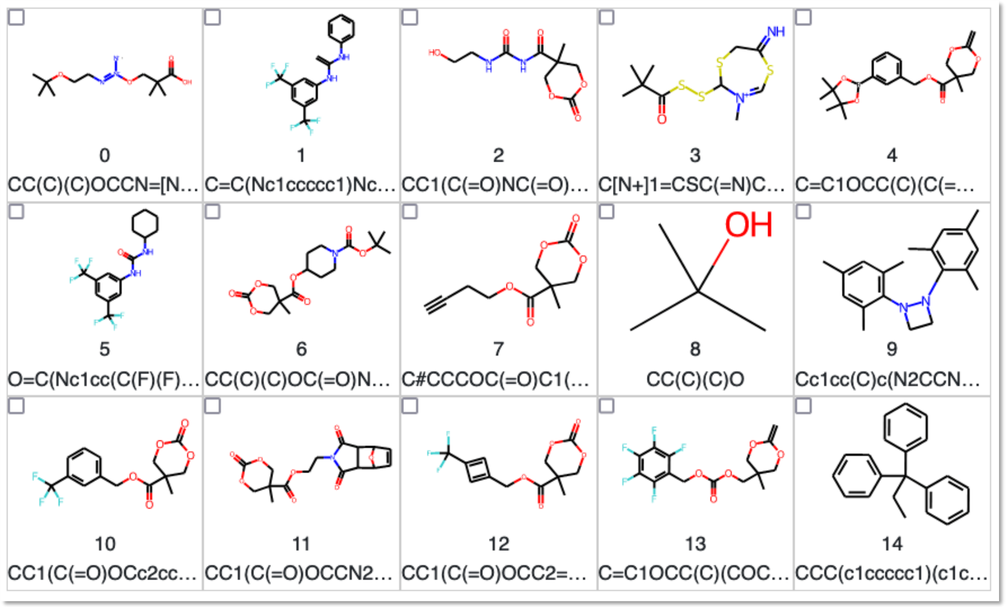
```
gt4sd-inference --algorithm_name PaccMannGP --algorithm_application PaccMa G
```

📖 Manica *et al.*, *npj Comput. Mater.* **9**, 69 (2023)

Study → **Hypothesize** → Test

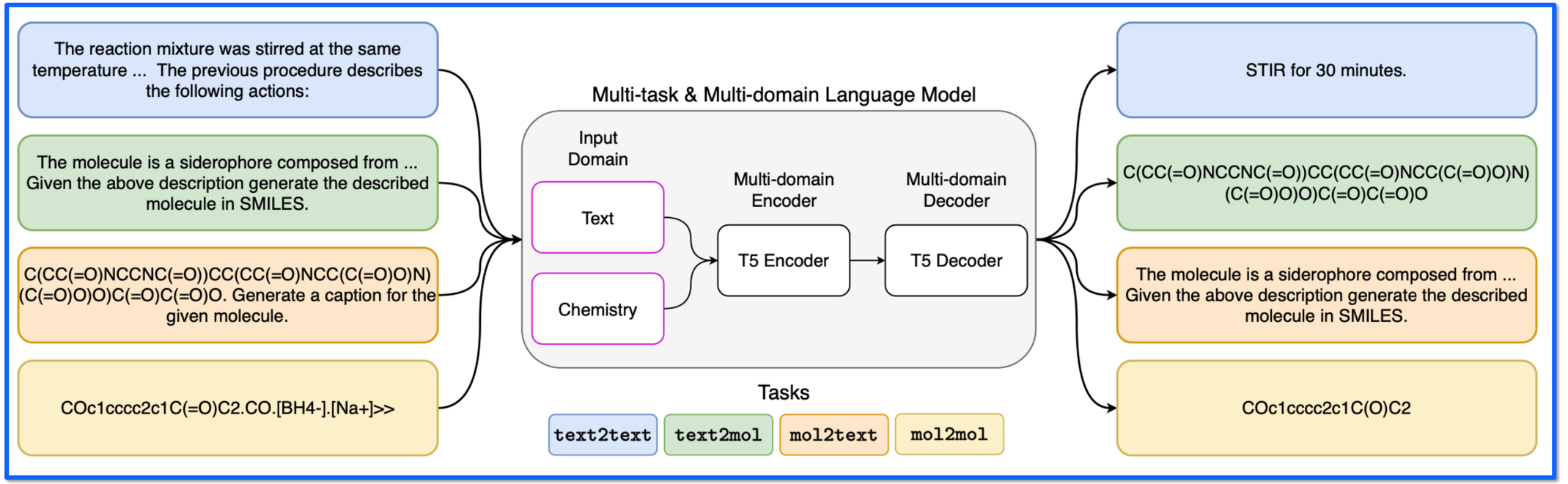
Applications include hypothesis generation for inverse design and discovery of materials and therapeutics like antivirals and antimicrobials

Example molecules generated using GT4SD



Multi-modal foundation models will accelerate fundamental research tasks

An example from chemistry…



📖 Christofidellis *et al.*, ICML (2023)

End-to-end discovery workflow using multi-task text and chemistry T5 model

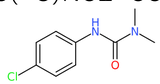
De novo design → Retrosynthesis → Actions

Input

Write in SMILES the described molecule: Give me a member of the class of phenylureas that is urea in which one of the nitrogens is substituted by a p-chlorophenyl group while the other is substituted by two methyl groups. It has a role as a herbicide, a xenobiotic and an environmental contaminant. It should be a member of monochlorobenzenes and a member of phenylureas.

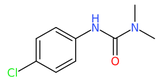
Target

CN(C)C(=O)NC1=CC=C(C=C1)Cl



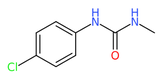
Text + chemistry T5

CN(C)C(=O)NC1=CC=C(C=C1)Cl



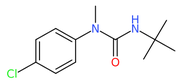
Galactica

1-(4-Chlorophenyl)-3-methylurea



ChatGPT

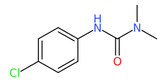
CC(NC(=O)N(C)C1=CC=C(Cl)C=C1)(C)C



Input

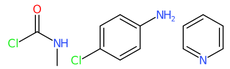
Predict the reaction that produces the following product:

CN(C)C(=O)NC1=CC=C(C=C1)Cl



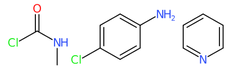
Target

CN(C)C(=O)Cl.Nc1ccc(Cl)cc1.c1ccncc1



Text + chemistry T5

CN(C)C(=O)Cl.NC1=CC=C(C=C1)Cl.c1ccncc1



RXN confidence = 1.0

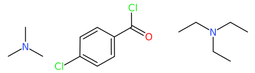
Galactica

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (o) (p) (q) (r) (s) (t) (u) (v) (w)

Invalid reaction

ChatGPT

CN(C)C + 4-chlorobenzoyl chloride + base



RXN confidence = 0.6

Input

Which actions are described in the following paragraph:

4-Chloro-aniline (1.0 g) was dissolved in pyridine and cooled to 0°C before adding dimethyl carbamyl chloride (0.683 g, 6.35 mmol). The reaction mixture was stirred at room temperature for 2 h and then heated overnight at 60°C under argon. The solution was cooled to room temperature, poured into ice water and extracted with EtOAc. The organic extract was dried over MgSO4, filtered and concentrated to a residue to afford 3-(4-Chlorophenyl)-1,1-dimethylurea.

Target

1. ADD 4-Chloroaniline (1.0 g)
2. ADD Pyridine
3. SET TEMPERATURE 0°C
4. ADD dimethyl carbamyl chloride (0.683 g, 6.35 mmol)
5. STIR for 2 h at room temperature
6. STIR for overnight at 60°C under argon
7. SETTEMPERATURE room temperature
8. ADD ice water
9. EXTRACT with EtOAc
10. COLLECTLAYER organic
11. DRYSOLUTION over MgSO4
12. FILTER keep filtrate
13. CONCENTRATE
14. YIELD 3-(4-Chlorophenyl)-1,1-dimethylurea

Text + chemistry T5

1. ADD 4-Chloroaniline (1.0 g);
2. ADD pyridine;
3. SETTEMPERATURE 0°C;
4. ADD dimethyl carbamyl chloride;
5. STIR for overnight at 60°C;
6. SETTEMPERATURE 60°C;
7. ADD water;
8. EXTRACT with EtOAc;
9. DRYSOLUTION over MGSO4
10. FILTER keep filtrate;
11. CONCENTRATE;
12. YIELD product

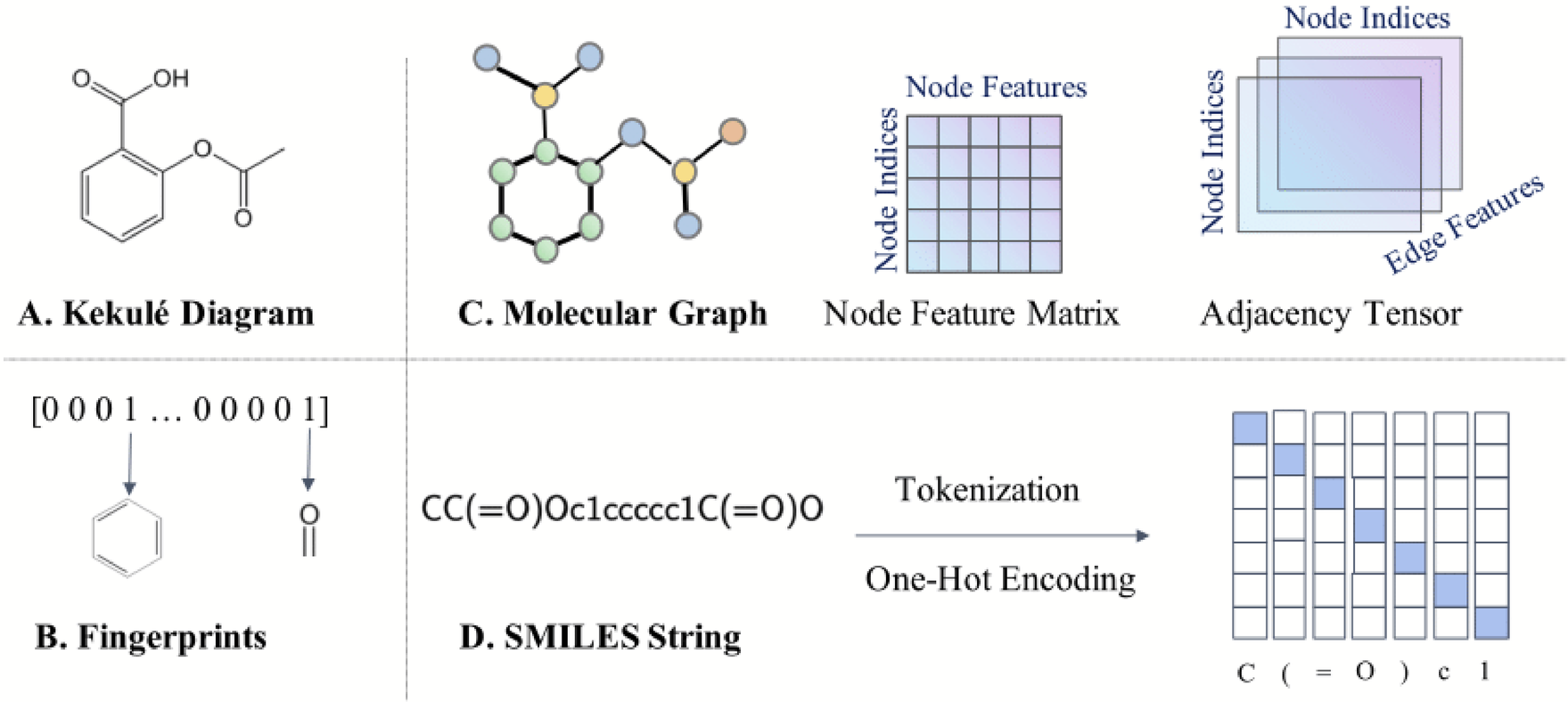
Galactica

1. 3-(4-Chlorophenyl)-1,1-dimethylurea (1.0 g, 4.5 mmol) was dissolved in dry THF (10 mL) and cooled to 0° C. before adding n-BuLi (1.6 M in hexane, 1.6 mL, 2.5 mmol).
2. The reaction mixture was stirred at 0° C. for 10 min and then at room temperature for 1 h.
3. The reaction mixture was quenched with water and extracted with EtOAc.
4. The organic extract was dried over MgSO4, filtered and concentrated to a residue.
5. The residue was purified by flash chromatography (silica gel, 10% EtOAc in hexane) to afford 3-(4-Chlorophenyl)-1,1-dimethylurea (1.0 g, 4.5 mmol) as a white solid (90%)

ChatGPT

1. Dissolving 4-chloro-aniline in pyridine and cooling the mixture to 0°C.
2. Adding dimethyl carbamyl chloride to the mixture.
3. Stirring the reaction mixture at room temperature for 2 hours.
4. Heating the reaction mixture overnight at 60°C under argon.
5. Cooling the solution to room temperature and pouring it into ice water.
6. Extracting the mixture with ethyl acetate.
7. Drying the organic extract over magnesium sulfate.
8. Filtering the extract and concentrating it to a residue.

Different molecular representations for machine learning



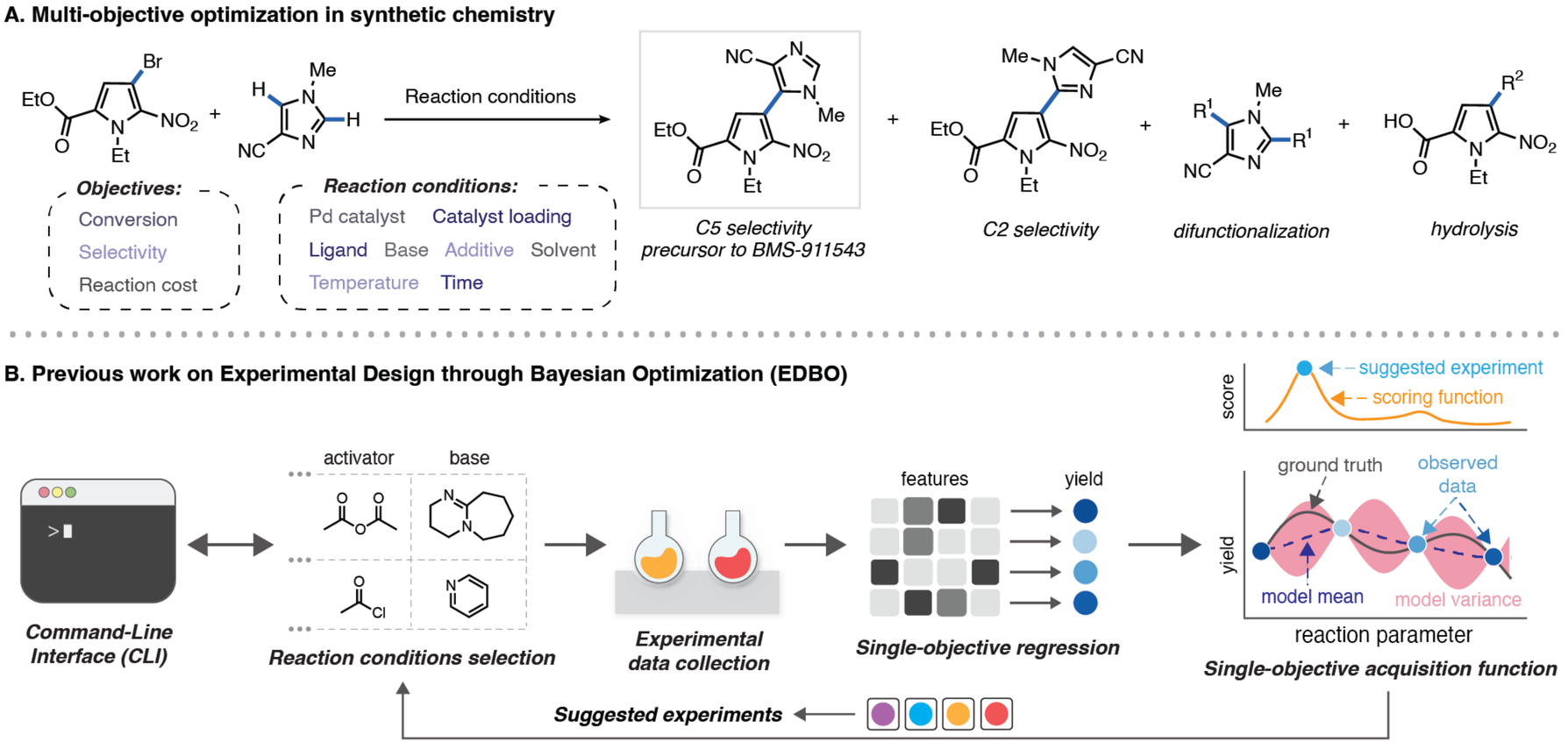
📖 Deng, J., Yang, Z., Ojima, I., Samaras, D. & Wang, F. Artificial intelligence in drug discovery: applications and techniques. *Briefings in Bioinformatics* **23**, bbab430 (2022)

📖 Wigh, D. S., Goodman, J. M. & Lapkin, A. A. A review of molecular representation in the age of machine learning. *WIREs Computational Molecular Science* **12**, e1603 (2022)

Optimization strategies

- Design of Experiment
- Bayesian Optimization
- Reinforcement Learning

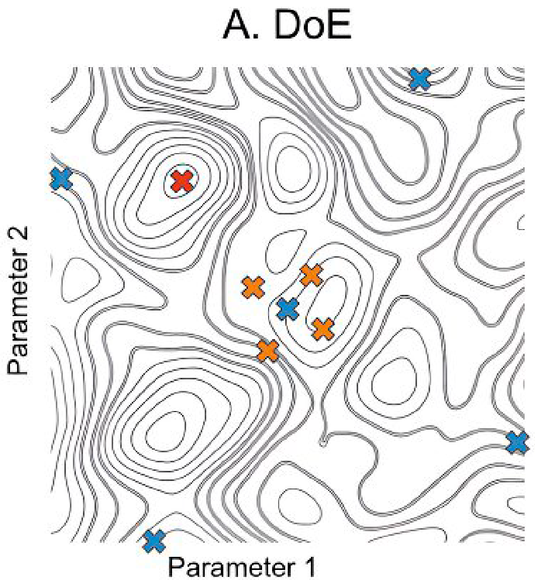
Optimization in chemical reaction planning



📖 Torres, J. A. G. *et al.* A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **144**, 19999–20007 (2022)

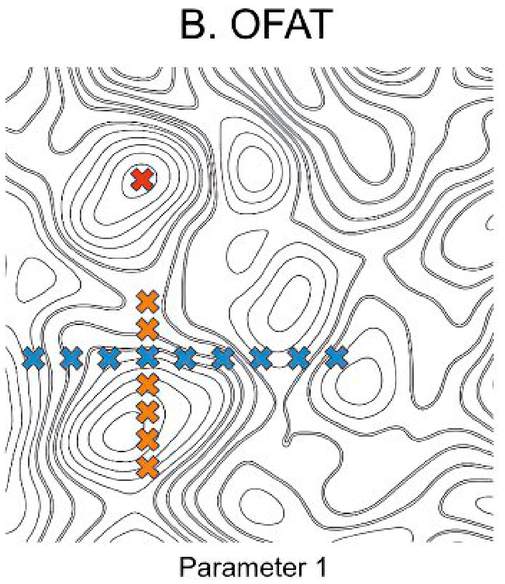
Comparison of approaches to experiment planning

Design of experiment



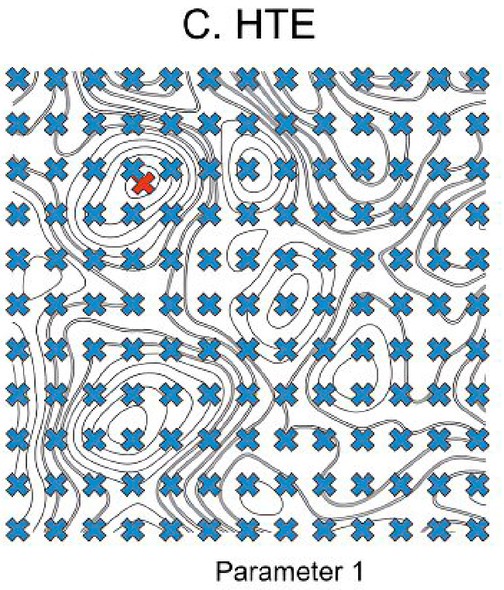
- Identify factors affecting response
- Relies heavily on initial design
- Number of experiments grows exponentially with number of parameters

One-factor-at-a-time



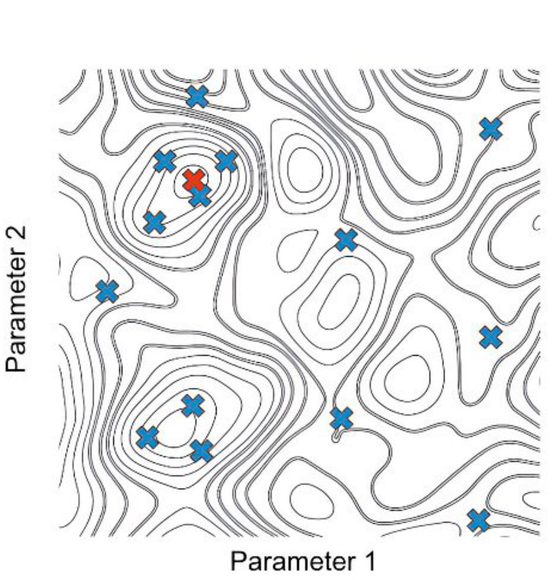
- Identify effect of single factor
- Time consuming for many factors
- Difficult to isolate confounding factors and potential interactions between factors

High-throughput experimentation



- Grid of possible values
- Exhaustive exploration
- Experimentally expensive
- Sensitive to choice of parameters

Machine learning driven



- Adaptive methods that learn from experiments in real time
- Recommend experiments where there is little information (**exploration**) or where better results are likely to occur (**exploitation**)

📖 Gutierrez, D. P., Folkmann, L. M., Tribukait, H. & Roch, L. M. How to Accelerate R&D and Optimize Experiment Planning with Machine Learning and Data Science. *Chimie* **77**, 7–16 (2023)

Key concepts in Bayesian optimization

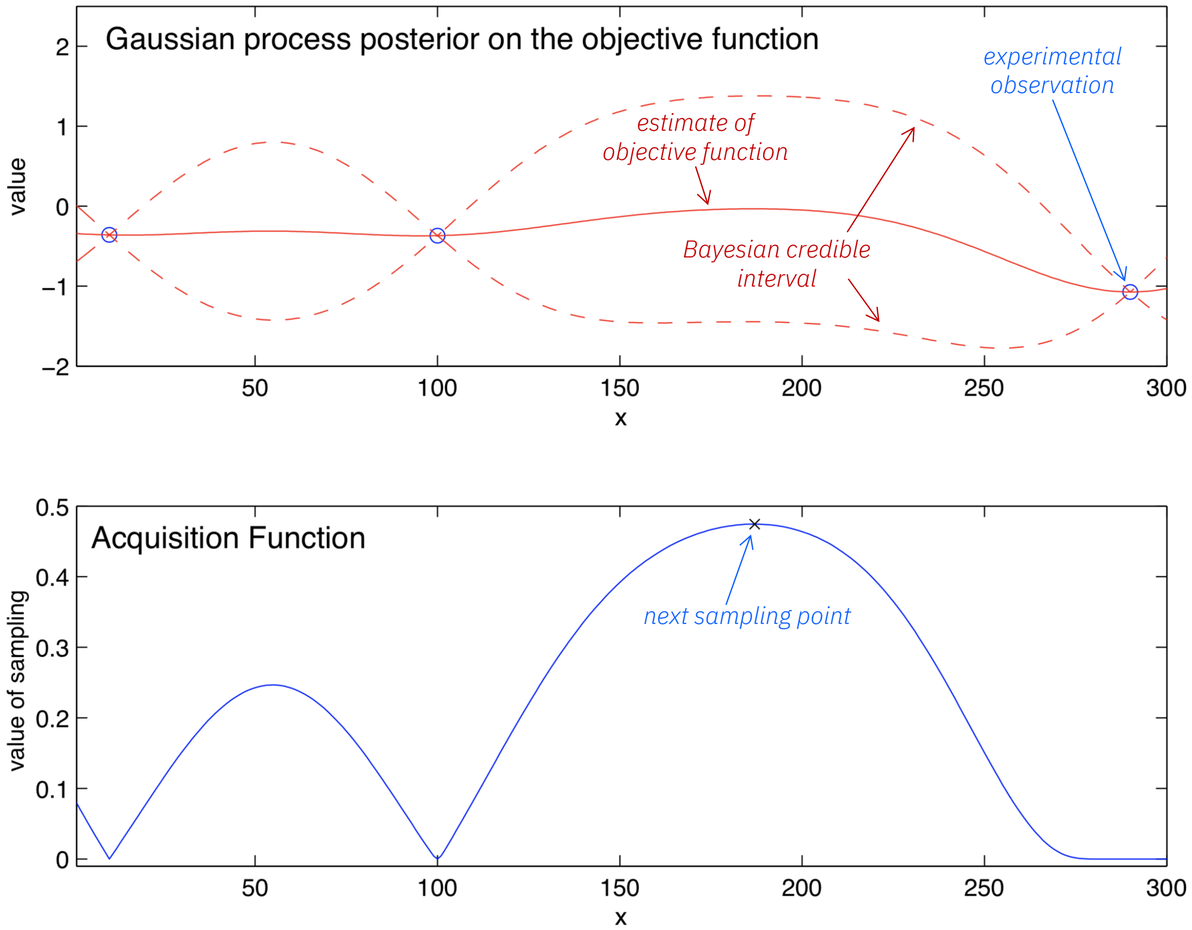
Bayesian optimization is a global optimization algorithm that reduces the need for many experiments.

Surrogate model:

- Probabilistic model to past observations
- Example function: Gaussian process

Acquisition model:

- Selection policy to decide which point to evaluate next
- Determines tradeoff between exploration and exploitation
- Example function: Expected improvement



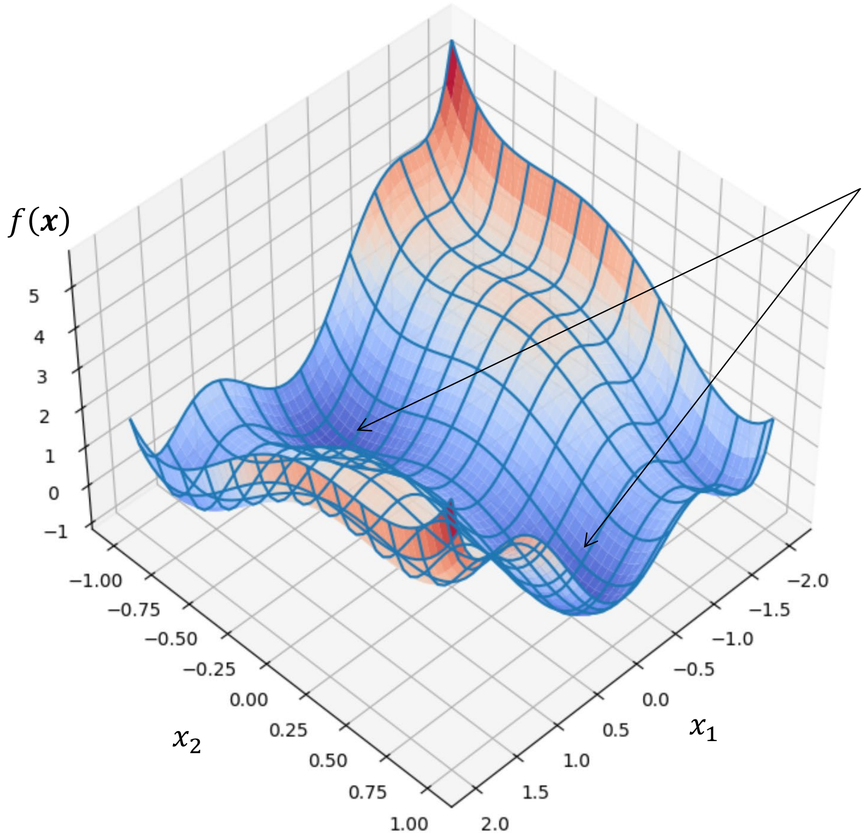
📖 Frazier, P. I. A Tutorial on Bayesian Optimization. arXiv:1807.02811 (2018)

Example using IBM Bayesian Optimization Accelerator

https://www.ibm.com/docs/en/boa/1.1.1?topic=boa-sdk-guide

Six-hump camelback function

$$f(\mathbf{x}) = \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2$$



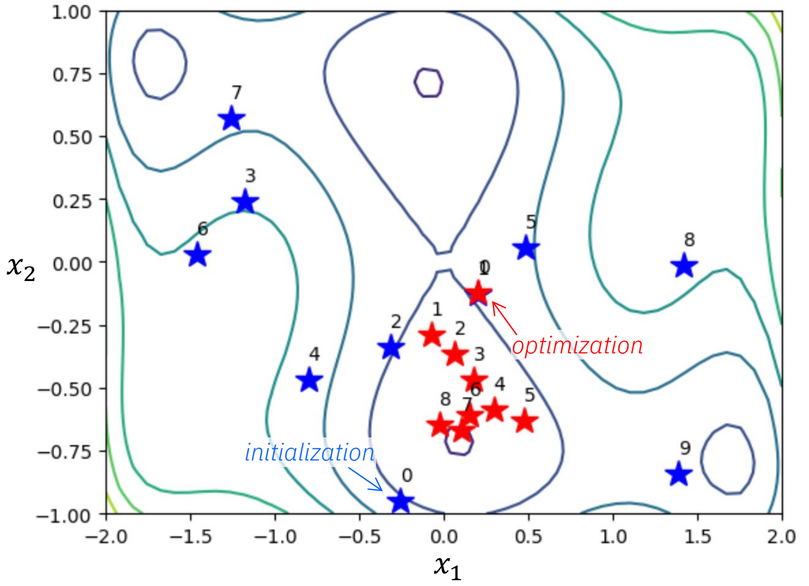
Global minima:

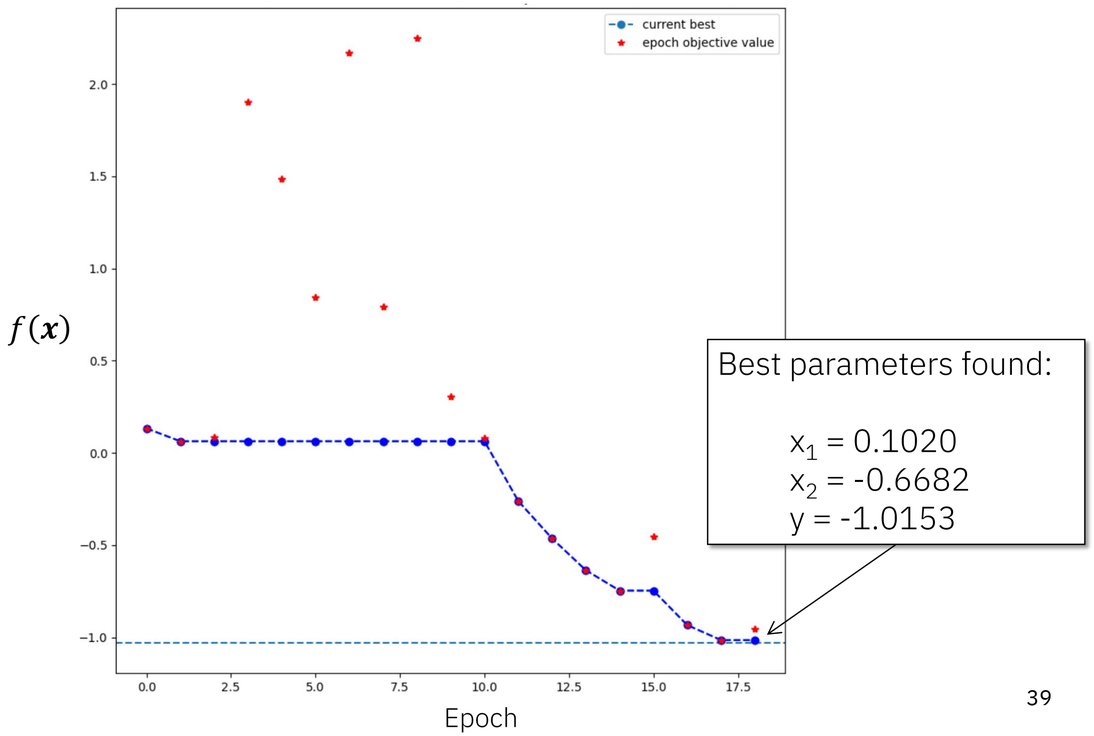
x_1 = 0.0898
x_2 = -0.7126
y = -1.0316

and

x_1 = -0.0898
x_2 = 0.7126
y = -1.0316

Exploration and exploitation in 10 epochs



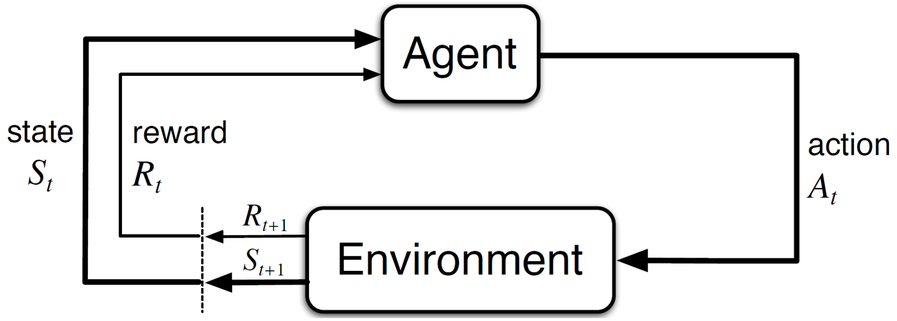


Best parameters found:

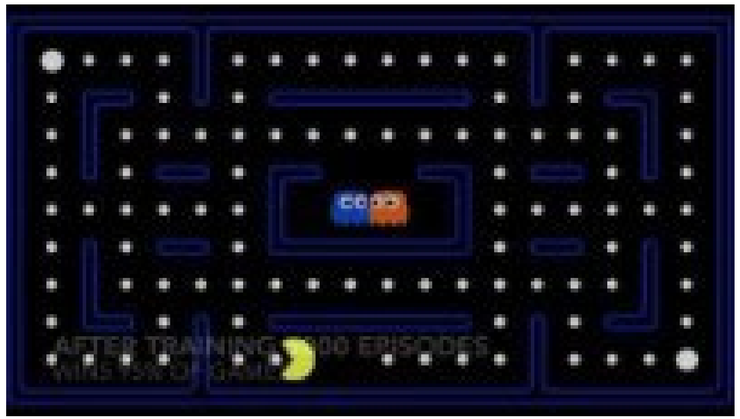
x_1 = 0.1020
x_2 = -0.6682
y = -1.0153

Other interactive examples online:
https://distill.pub/2020/bayesian-optimization/

Reinforcement learning (RL) to learn complex behaviors through trial-and-error interactions



📖 Sutton & Barton, Reinforcement Learning: An Introduction (2018)



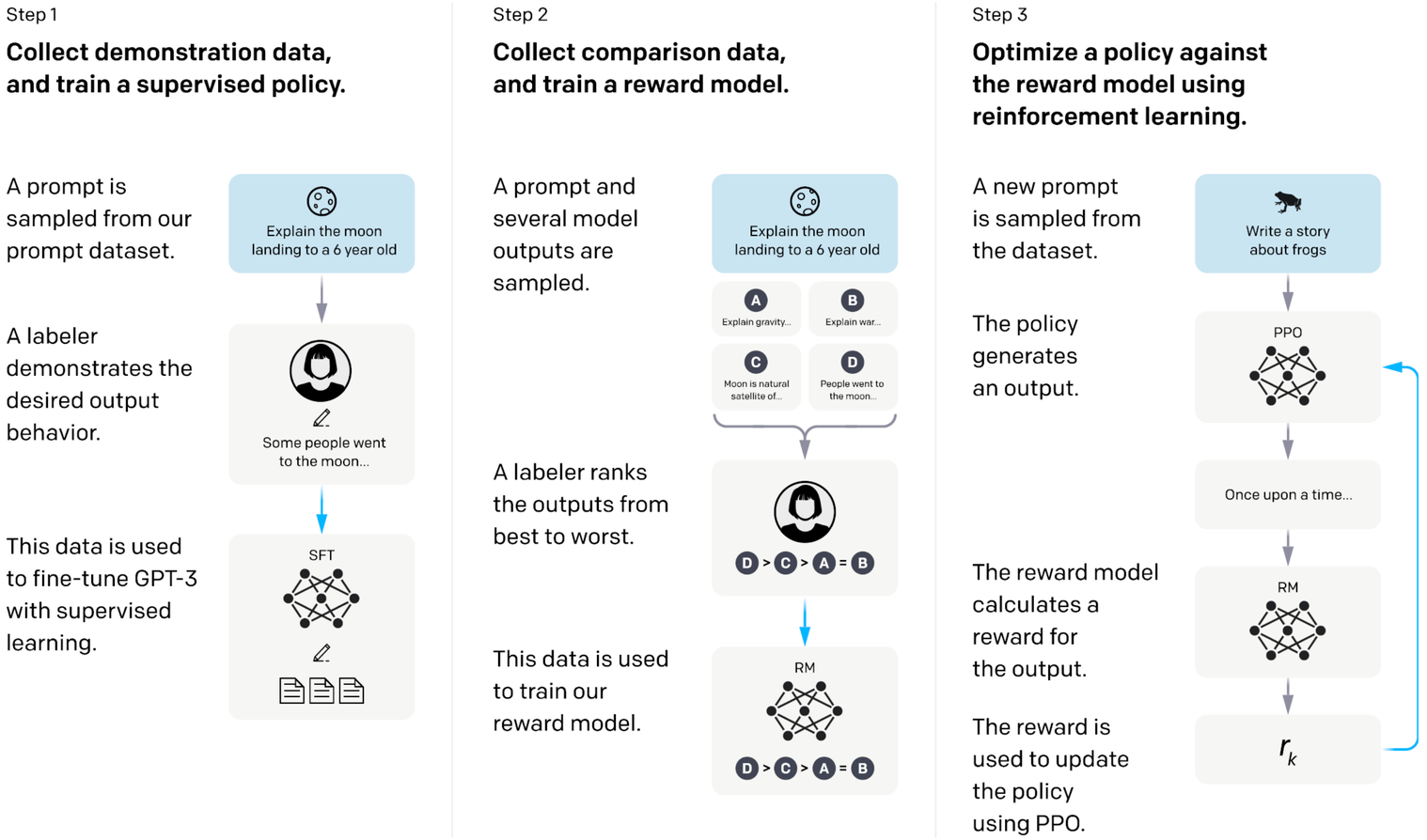
PacMan agent using deep reinforcement learning
https://youtu.be/QilHGSYbjDQ?si=wyP1NDdrymngUb7j

- Reinforcement learning uses punishments and rewards as signals for positive and negative behavior
- Find suitable action model to maximize the total cumulative reward
- Exploration vs. exploitation trade-off
- Two main approaches in deep reinforcement learning:
  - Value functions (estimate expected return, e.g. through neural networks)
  - Policy search (directly find policies, e.g. through gradient-free or gradient-based methods)

📖 Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine* **34**, 26–38 (2017)

📖 Azar, M. G. *et al.* A General Theoretical Paradigm to Understand Learning from Human Preferences. arXiv:2310.12036 (2023)

Reinforcement learning with human feedback (RLHF) to align model with user intent



📖 Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *arXiv:2203.02155* (2022)

- Common approach to RLHF is based on training a reward model (classifier) and then fine-tuning a policy to maximize the reward
- Recent approaches bypass reward models to directly train policies through Direct Preference Optimization (DPO) or pairwise preferences

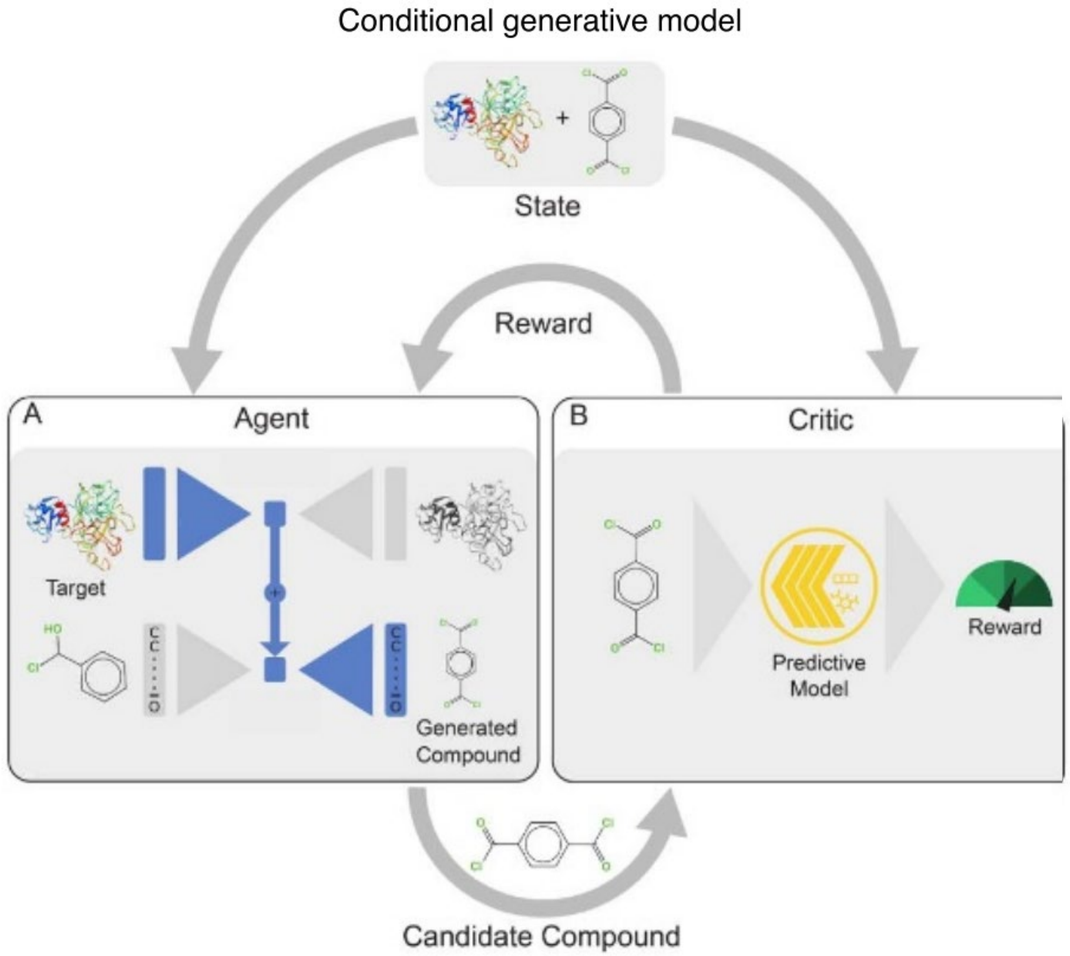
📖 Rafailov, R. *et al.* Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 (2023)

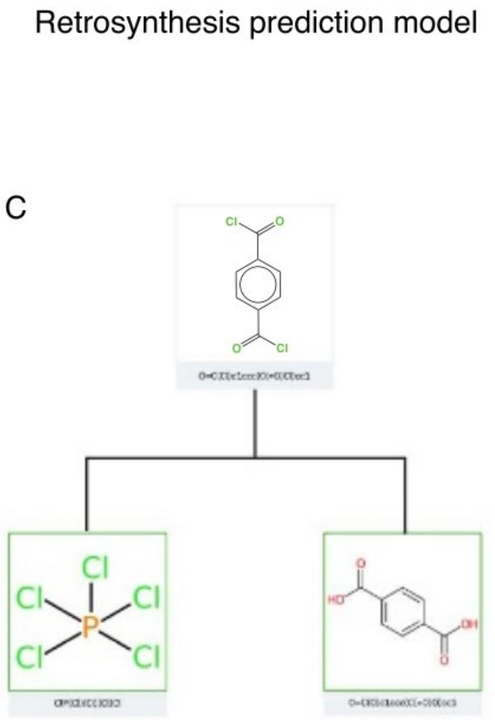
📖 Azar, M. G. *et al.* A General Theoretical Paradigm to Understand Learning from Human Preferences. arXiv:2310.12036 (2023)

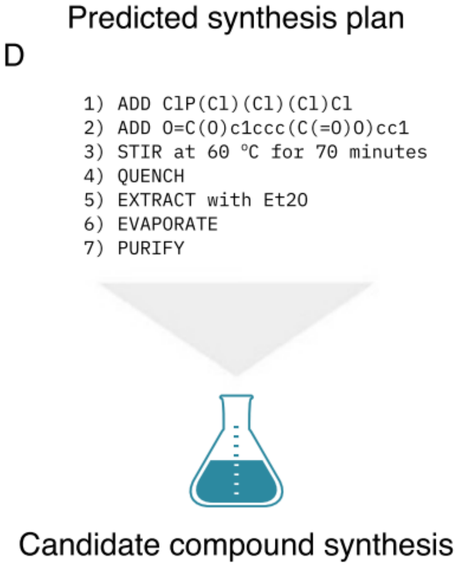
Closing the loop

- **Case studies**

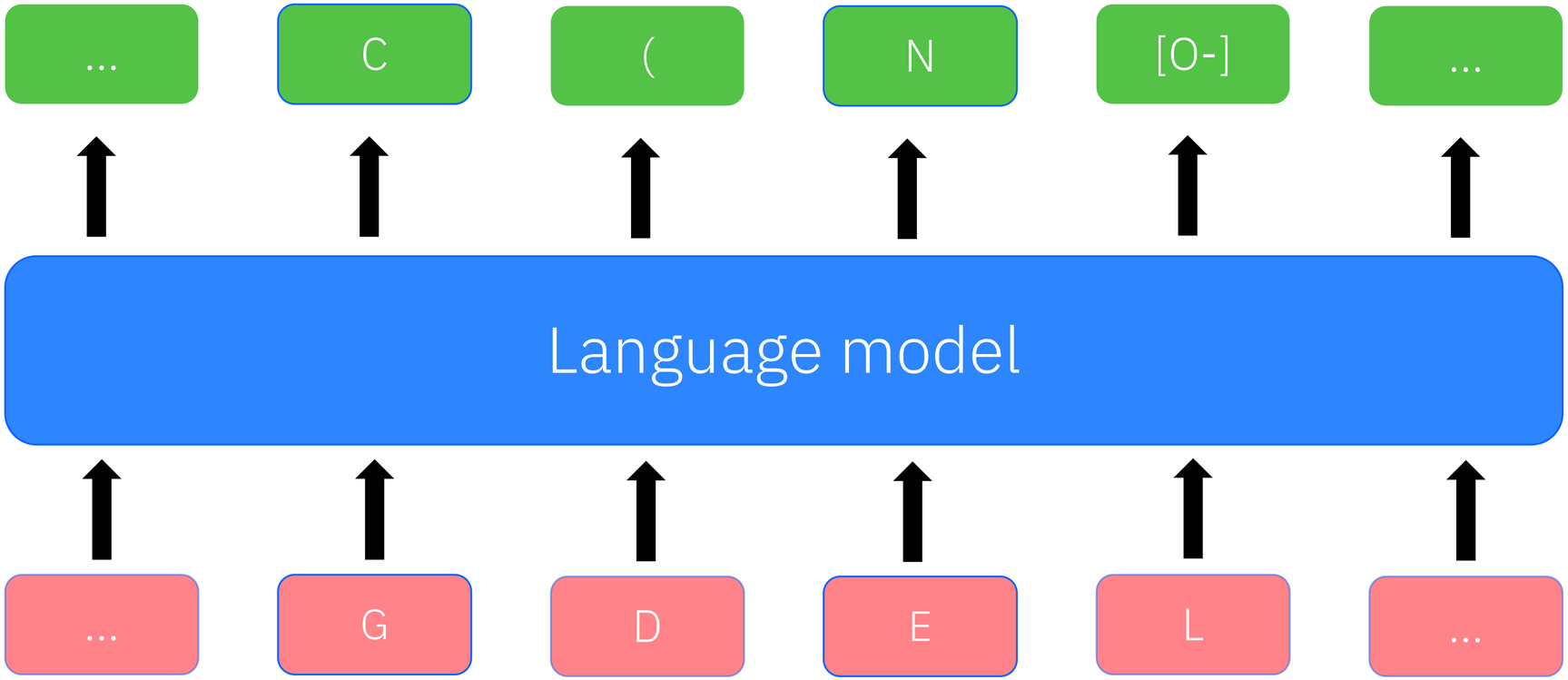
Toward closed-loop autonomous molecular discovery: designing novel antiviral candidates against SARS-CoV-2

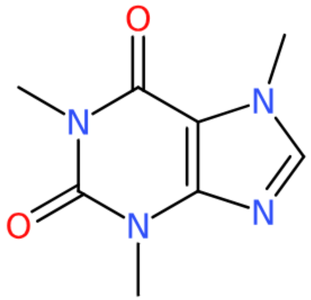


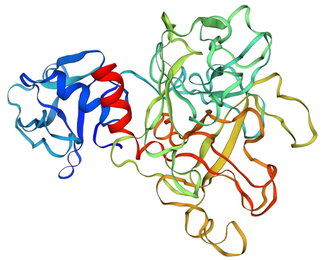




Translate from amino acid sequence to potentially binding ligands

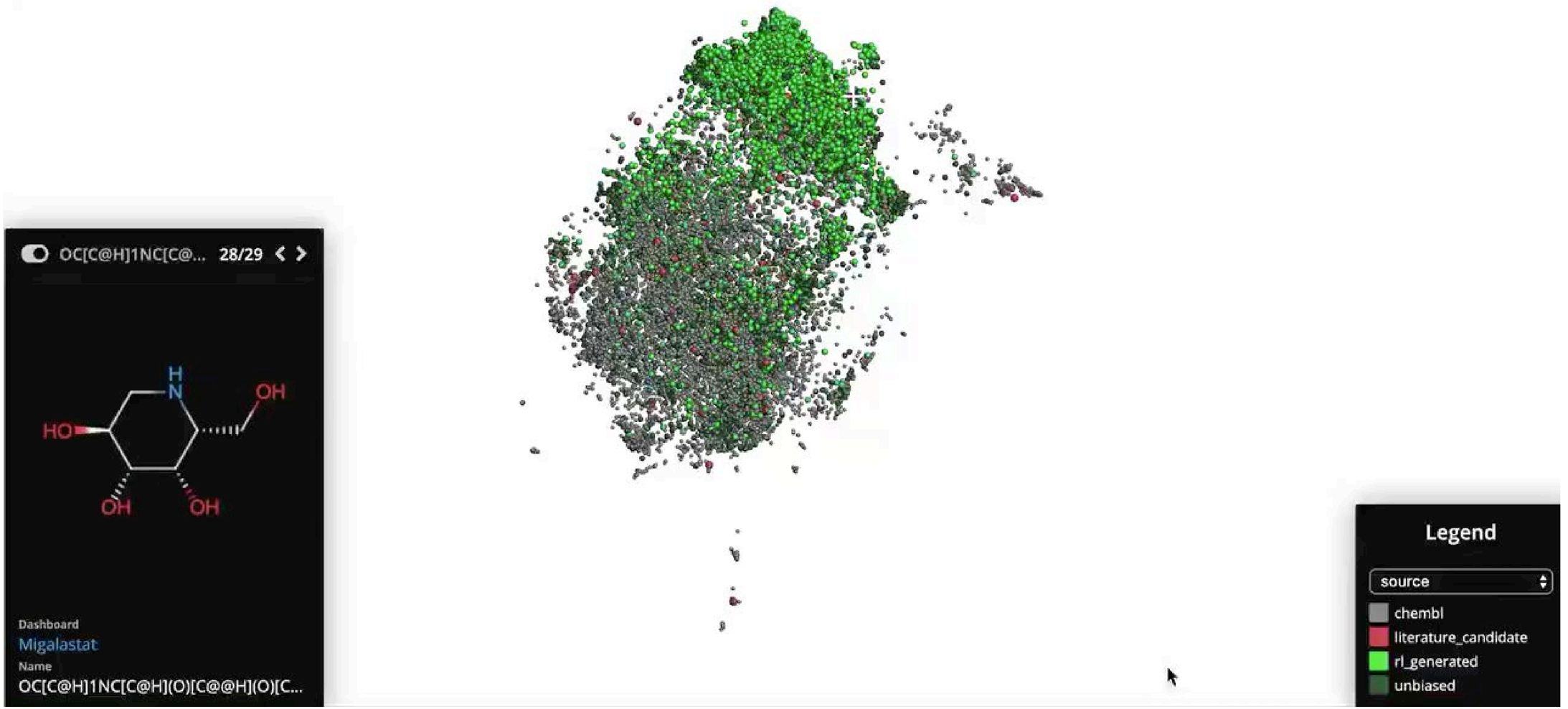






📖 Born, J. *et al.* Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Mach. Learn.: Sci. Technol.* **2**, 025024 (2021)

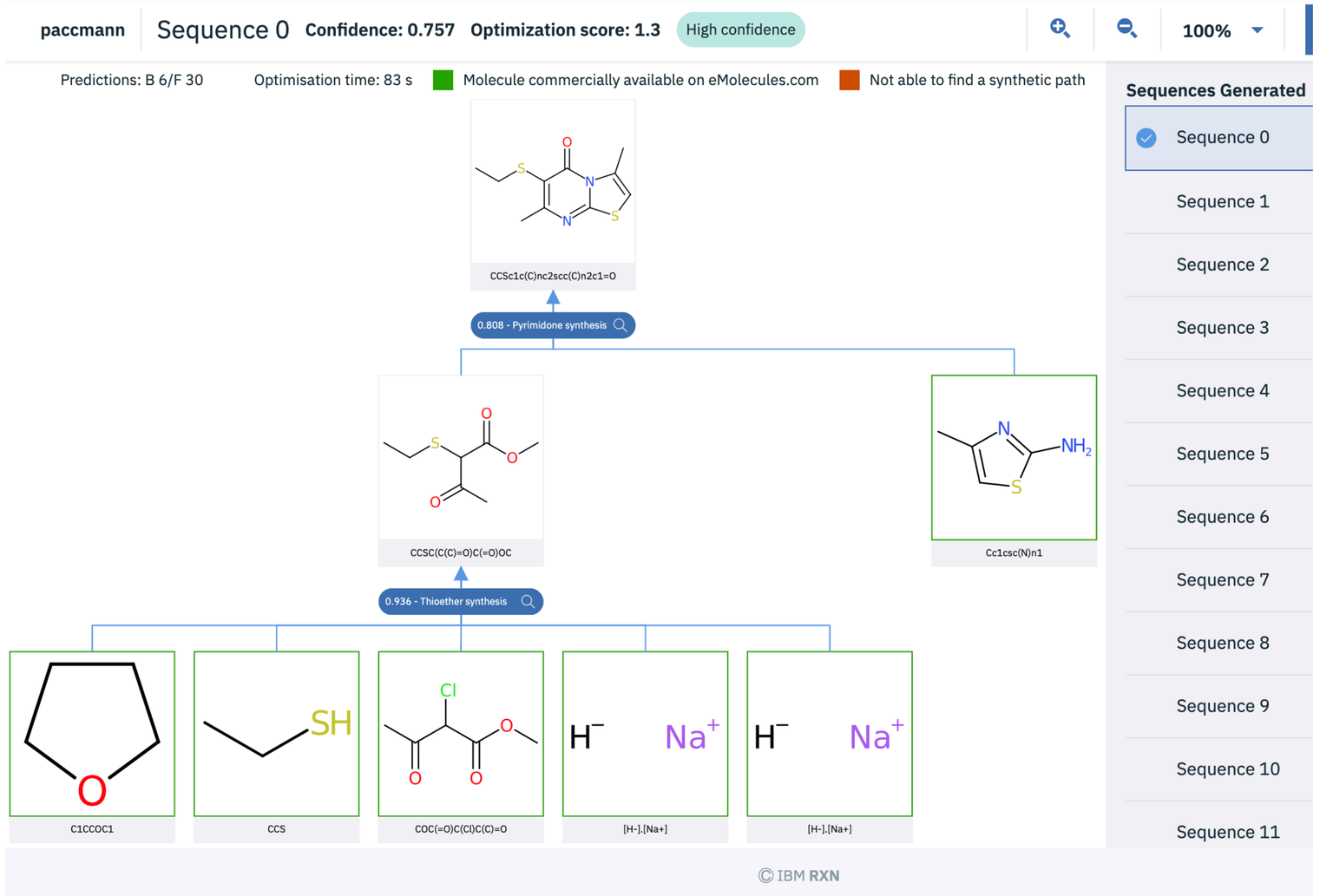
Visualizing the learned chemical space



📖 Born, J. *et al.* Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Mach. Learn.: Sci. Technol.* **2**, 025024 (2021)

Find a viable synthetic route

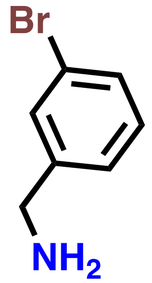
- Web service: rxn.res.ibm.com
- Retrosynthesis route for most promising compounds
- Reaction prediction and retro-synthesis models (Transformer)
- Synthesis routes found for ca. 30% of the best candidates



📖 Born, J. *et al.* Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Mach. Learn.: Sci. Technol.* **2**, 025024 (2021)

Selection of synthesis candidate

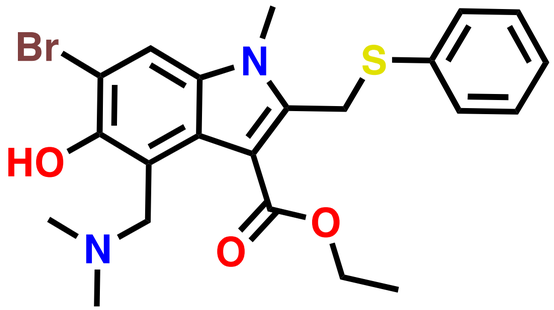
3-Bromobenzylamine



3-Bromobenzylamine:

- Full substructure of Arbidol
- Generated by our model to target ACE2 receptor
- Presence of bromine is key for efficacy of Arbidol

Arbidol



Arbidol:

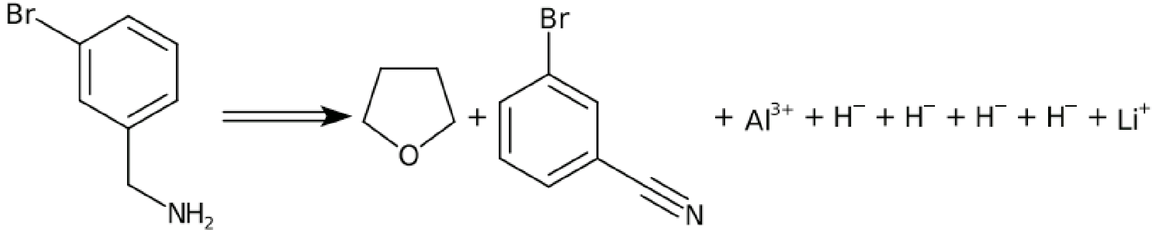
- Approved, broad-spectrum antiviral drug
- Positive evidence for COVID-19 treatment
- Interacts with the ACE2 receptor

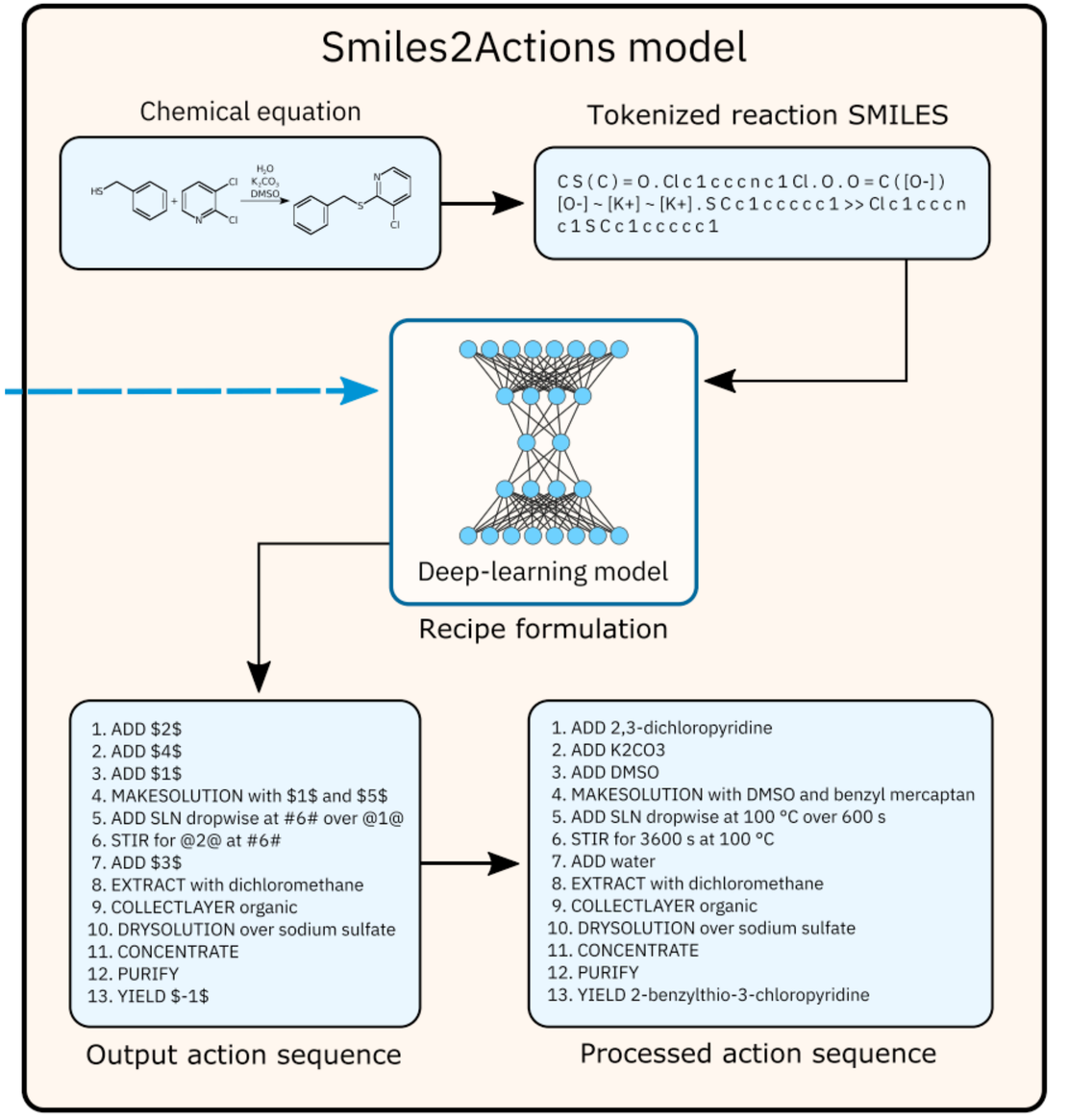
📖 Born, J. *et al.* Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Mach. Learn.: Sci. Technol.* **2**, 025024 (2021)

Stepwise synthesis execution plan

Type: Nitrile reduction, Confidence: 0.985

C1CCOC1.N#Cc1cccc(Br)c1.[Al+3].[H-].[H-].[H-].[H-].[Li+]>>BrC1C=CC=C(CN)C=1



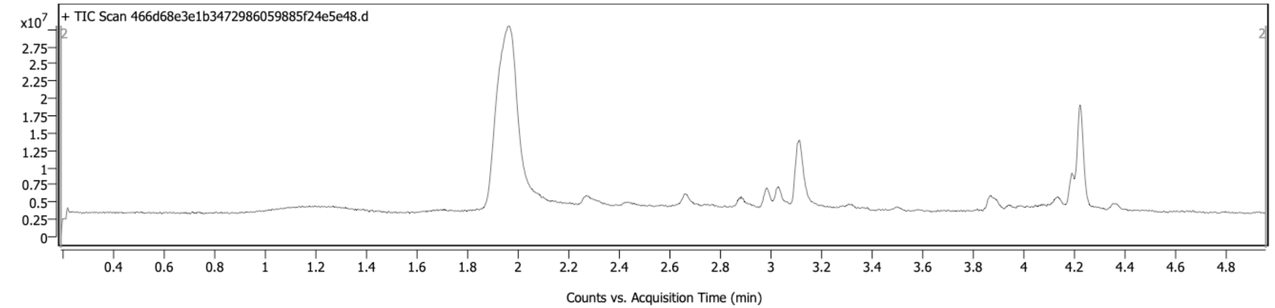


📖 Born, J. *et al.* Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Mach. Learn.: Sci. Technol.* **2**, 025024 (2021)

Molecular synthesis using IBM RoboRXN platform

Automated synthesis

LC/MS chromatogram



Compound Summary

Cpd	Name	Formula	CAS	RT	Mass	Mass (Tgt)	Diff (Tgt, ppm)	Score	Algorithm
1		C7 H8 Br N		2.882	184.9858	184.9840	9.55	90.36	FBF
2		C7 H8 Br N		1.970	184.9837	184.9840	-1.64	99.44	FBF
3		C4 H8 O		3.515	72.0580	72.0575	6.74	83.13	FBF
4		C4 H8 O		3.439	72.0571	72.0575	-6.36	80.60	FBF

Compound Details

Cpd. 1: C7 H8 Br N

Name	Formula	RT	RI	Mass	Diff (Tgt, ppm)	CAS	ID Source	Score	Algorithm
	C7 H8 Br N	2.882		184.9858	9.55		FBF	90.36	FBF

Species	m/z	Score (Tgt)	Score (Lib)	Score (DB)	Score (MFG)	Score (RT)
(M+H)+ (2M+Na)+	185.9902 392.9582	90.36				

MassHunter Qualitative Analysis Page 1 of 4 Generated at 23:01 on 25/08/2020

📖 Born, J. *et al.* Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Mach. Learn.: Sci. Technol.* **2**, 025024 (2021)

In the news last week...

nature

Explore content | About the journal | Publish with us | Subscribe

nature > news > article

NEWS | 29 November 2023

Google AI and robots join forces to build new materials

Tool from Google DeepMind predicts nearly 400,000 stable substances, and an autonomous system learns to make them in the lab.

Article

Scaling deep learning for materials discovery

https://doi.org/10.1038/s41586-023-06735-9
Received: 8 May 2023
Accepted: 10 October 2023
Published online: 29 November 2023
Open access
Check for updates

Amil Merchant[1,3]✉, Simon Batzner[1,3], Samuel S. Schoenholz[1,3], Muratahan Aykol[1], Gowoon Cheon[2] & Ekin Dogus Cubuk[1,3]✉

Novel functional materials enable fundamental breakthroughs across technological applications from clean energy to information processing[1–11]. From microchips to batteries and photovoltaics, discovery of inorganic crystals has been bottlenecked by expensive trial-and-error approaches. Concurrently, deep-learning models for

Article

An autonomous laboratory for the accelerated synthesis of novel materials

https://doi.org/10.1038/s41586-023-06734-w
Received: 16 May 2023
Accepted: 10 October 2023
Published online: 29 November 2023
Open access
Check for updates

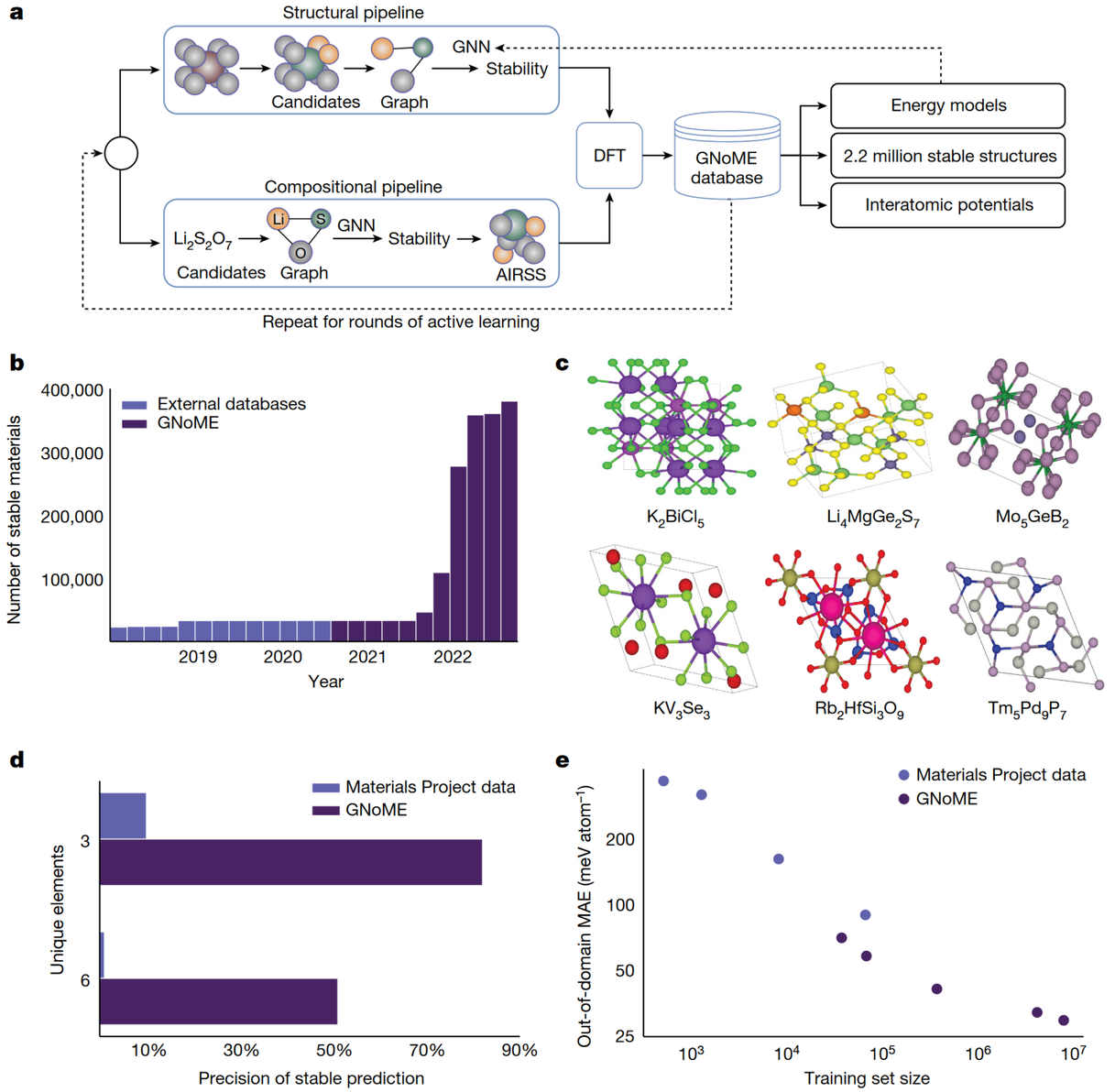
Nathan J. Szymanski[1,2,5], Bernardus Rendy[1,2,5], Yuxing Fei[1,2,5], Rishi E. Kumar[3,5], Tanjin He[1,2], David Milsted[2], Matthew J. McDermott[1,2], Max Gallant[1,2], Ekin Dogus Cubuk[4], Amil Merchant[4], Haegyeom Kim[2], Anubhav Jain[3], Christopher J. Bartel[2], Kristin Persson[1,2], Yan Zeng[2]✉ & Gerbrand Ceder[1,2]✉

To close the gap between the rates of computational screening and experimental realization of novel materials[1,2], we introduce the A-Lab, an autonomous laboratory for the solid-state synthesis of inorganic powders. This platform uses computations,

📖 Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023)
📖 Szymanski, N. J. *et al.* An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023)

Generating a large of database stable crystal structures by deep learning

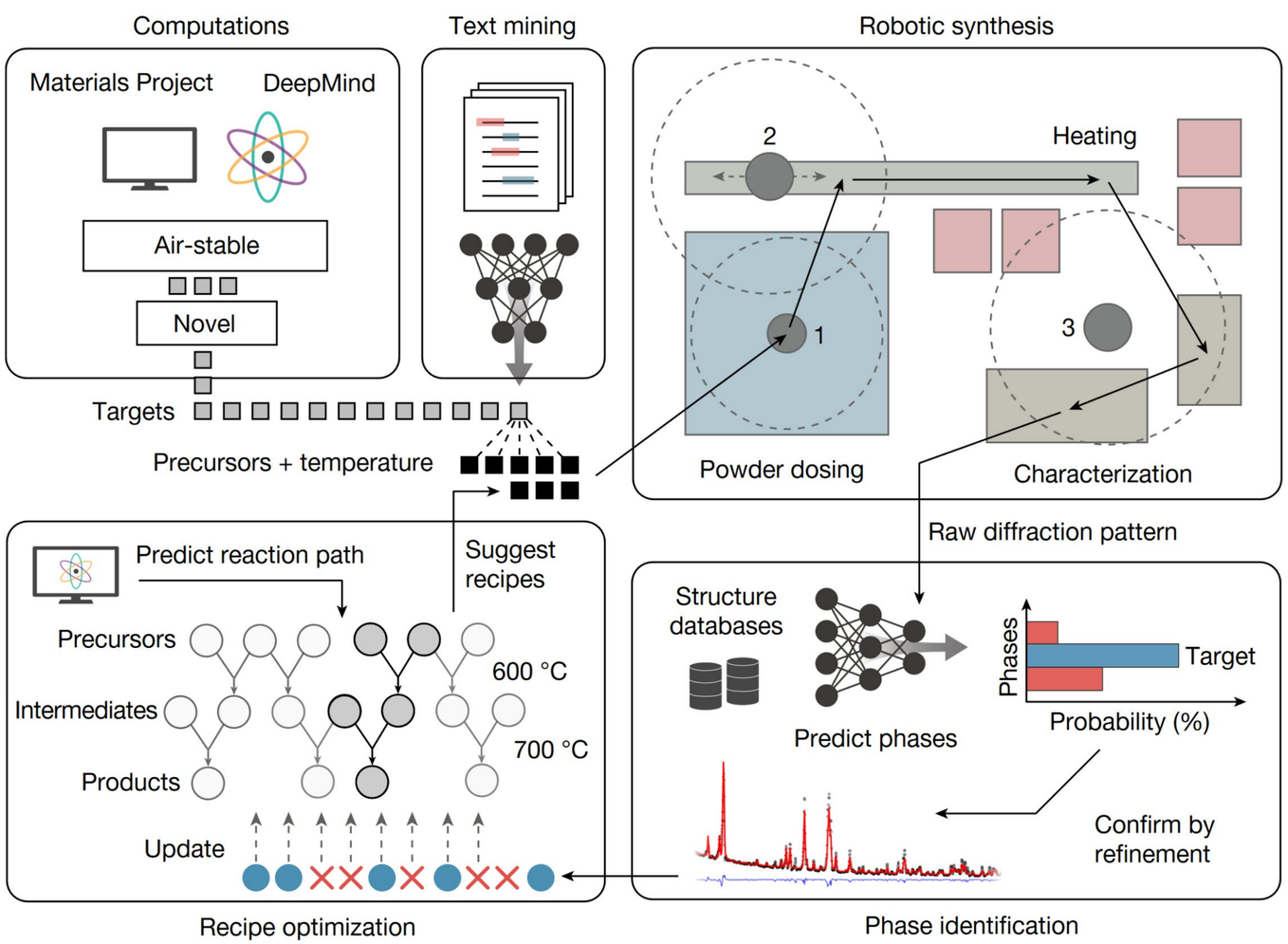
- Generate diverse candidate structures leveraging symmetry-aware partial substitutions and random structure search
- Graph neural networks (GNNs) used to model materials properties and filter candidate materials → graph networks for materials exploration (GNoME)
- Energy of filtered candidates is computed using density functional theory (DFT) and used to improve GNoME models through active learning



📖 Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023)
📖 Szymanski, N. J. *et al.* An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023)

Performing automated experimental validation of new crystal structures

- Synthesis recipes based on similar materials are generated by a language model trained on solid-state synthesis conditions from journal data
- Continuous experimentation with active learning to learn reaction pathways based on fixed set of precursors
- Phase and weight fractions extracted from X-ray diffraction (XRD) patterns using a convolutional neural network (CNN), automated Rietveld refinement, and manual analysis

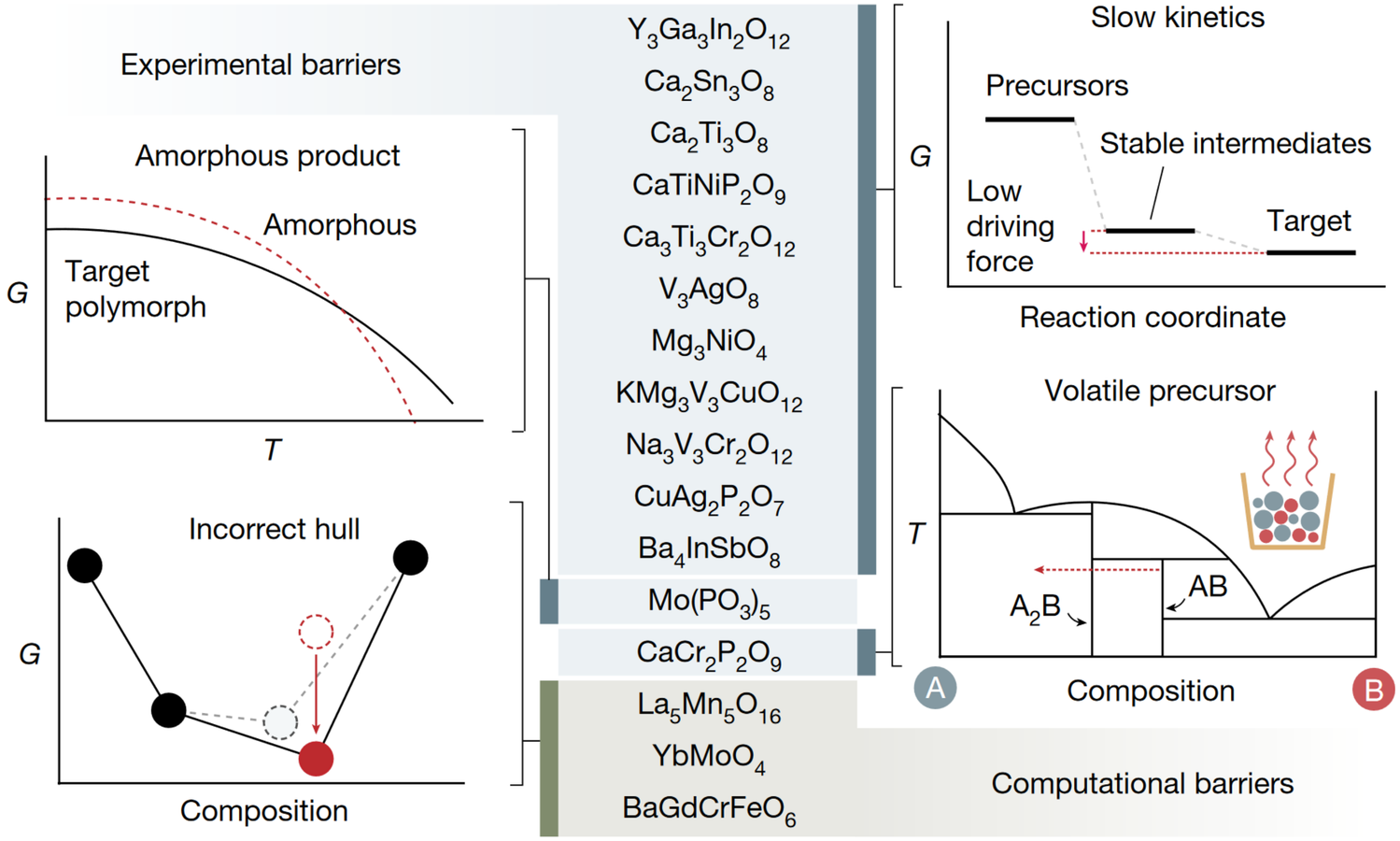


📖 Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023)
📖 Szymanski, N. J. *et al.* An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023)

Lessons learned from failures to synthesize target product

17 out of 58 targets were not obtained at the end of the active learning cycle:

1. 11 failures due to sluggish reaction kinetics
2. Decomposition and evaporation of ammonium phosphate precursors
3. Crystallization may be inhibited upon melting of samples at high temperatures
4. Underestimation of energy by DFT calculation



📖 Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023)
📖 Szymanski, N. J. *et al.* An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023)

Hands-on lab

- Design and synthesize your own molecule

Design and synthesize your own molecule

1. Generate novel molecules and pick your favorite one!
   https://huggingface.co/spaces/GT4SD/regression_transformer

2. Predict its molecular properties:
   https://huggingface.co/spaces/GT4SD/regression_transformer

3. Find synthesis routes with IBM RXN:
   https://rxn.res.ibm.com

4. Press a button to synthesize your own compound ☺

Summary

1. Scientific discovery as a loop
    - Evolution of the scientific method
    - AI for science
    - State of the art

2. Molecular generation
    - Language models for molecular discovery
    - Generative AI
    - Graph neural networks

3. Optimization strategies
    - Design of Experiment
    - Bayesian Optimization
    - Reinforcement Learning

4. Closing the loop: case studies

5. Hands-on lab