

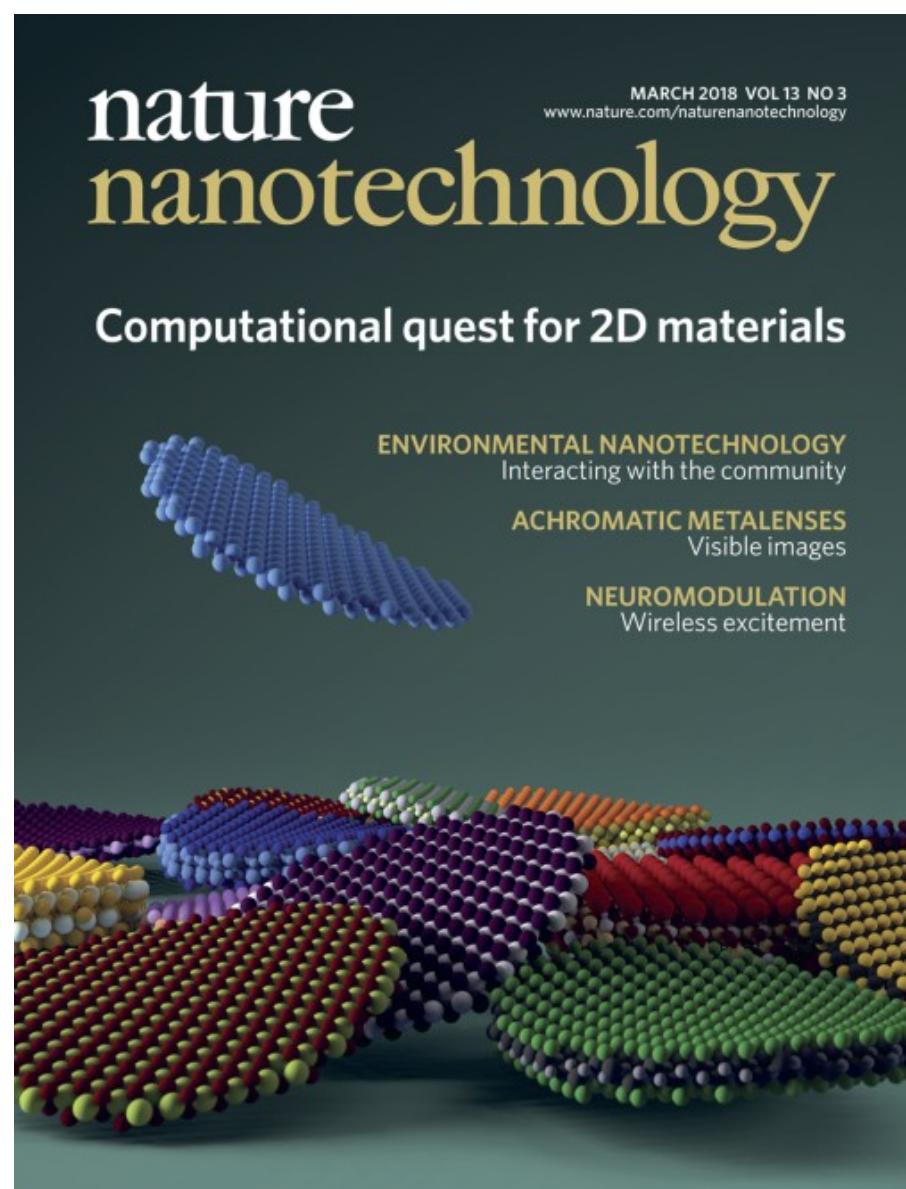
# Introduction

Philippe Schwaller  
Laboratory of Artificial  
Chemical Intelligence  
(LIAC)

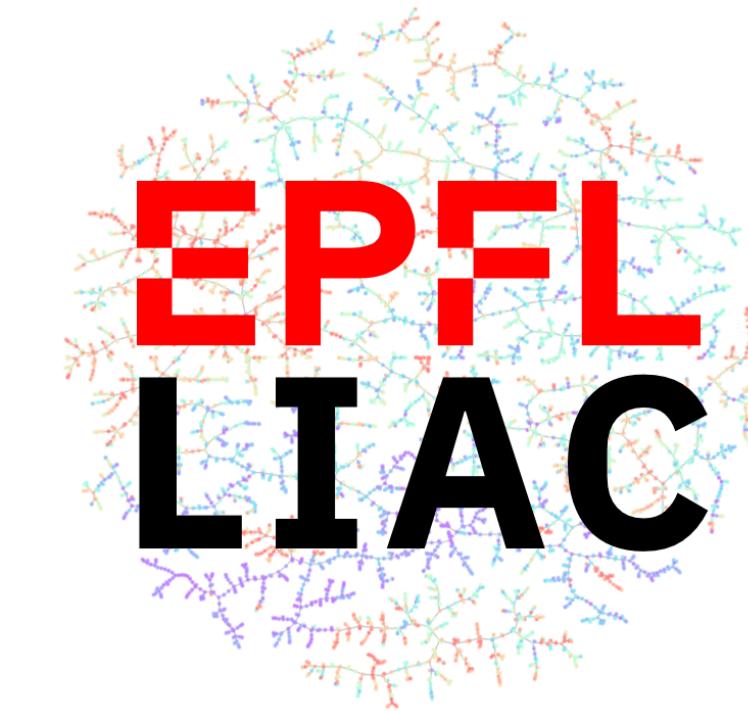


Grew up in Fribourg,  
Switzerland  
- French  
- Swiss German / German

# EPFL



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

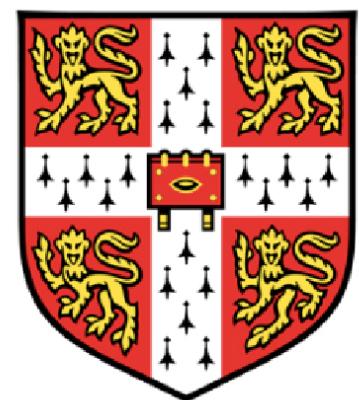


Collaboration with synthetic chemists

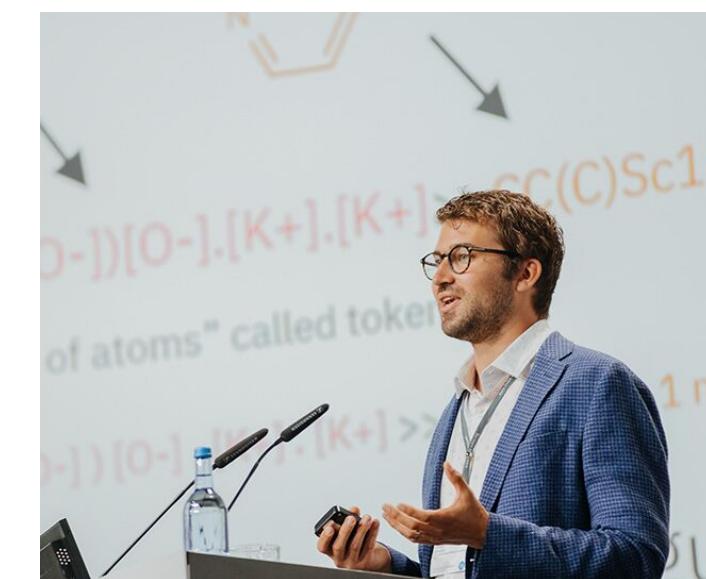
*u*<sup>b</sup>

<sup>b</sup>  
UNIVERSITÄT  
BERN

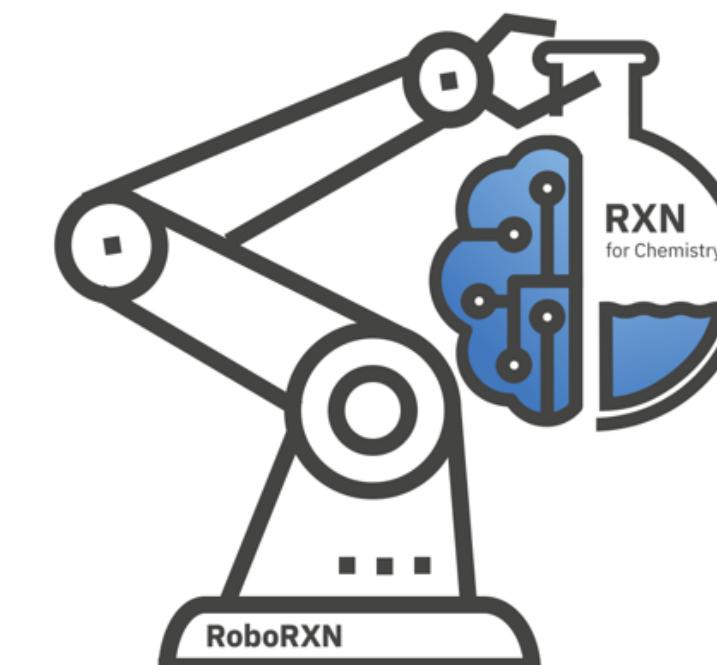
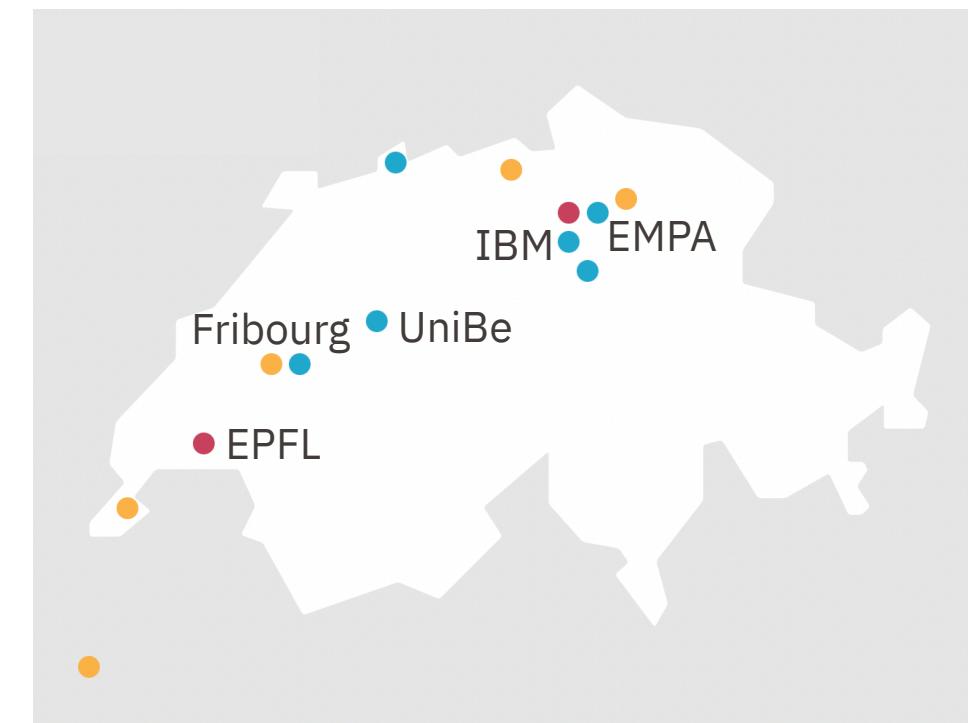
PhD in Chemistry  
and Molecular Sciences ('21)  
Prof Jean-Louis Reymond



UNIVERSITY OF  
CAMBRIDGE



MPhil in Physics ('19)  
Dr Alpha Lee



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

# IBM Research Europe

Materials Science &  
Engineering

Virtual screening &  
simulation workflows  
Prof Nicola Marzari

BSc ('14)/MSc ('16)



Materials Science and Technology

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



Working group:



Goal:

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

Serendipity is our friend.



# This class

- What are your expectations?
- How many of you are computational?
- How many of you are more experimental?
  
- First-time
- Lots of the material does not yet exist
- Crash-course
- Focus on applications, not math behind



# You will learn about

**Martin Vetterli** @MartinVetterli · Dec 5, 2017

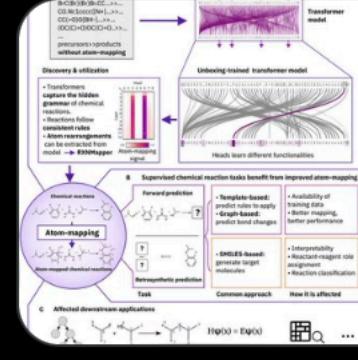
Congrats to @EPFL\_en alumni @phisch124 and @TheophileGaudin for their work at @IBMResearch! Applying state-of-the-art neural machine translation methods to predicting the outcome of organic chemistry reactions [arxiv.org/abs/1711.04810](https://arxiv.org/abs/1711.04810) [technologyreview.com/the-download/6...](https://technologyreview.com/the-download/6...) #NIPS2017



2 15 34

**Marwin Segler** @marwinsegler

Self-supervised learning is a powerful paradigm in chemistry: Training a masked language model on chemical reactions allows to discover which atoms and bonds are changed in the course of reaction (atom-mapping), without using any manual labeling! 1/2



science.org  
Extraction of organic chemistry grammar from unsupervised...  
RXNMapper constructs coherent atom-mapping rules from raw chemical reactions using unsupervised training of ...

2:12 AM · Apr 11, 2021

100 Retweets 5 Quote Tweets 529 Likes

**Marwin Segler** @marwinsegler · Apr 11, 2021

Replying to @marwinsegler

Cases where ML seems to work almost like magic are rare, but this is one of them, a breakthrough in my opinion. And even better: it can be completely rationalized why it works, with 🔥 visualisations! Great work

**Berend Smit** @SmitBerend

The most controversial article I have ever written! @kmjablonka shows that making a #MachineLearning model for different problems in #Chemistry with #GPT3 is as easy as a literature search. The scary part; it outperforms all the conventional ML models [dx.doi.org/10.26434/chemr...](https://dx.doi.org/10.26434/chemr...) 1/4

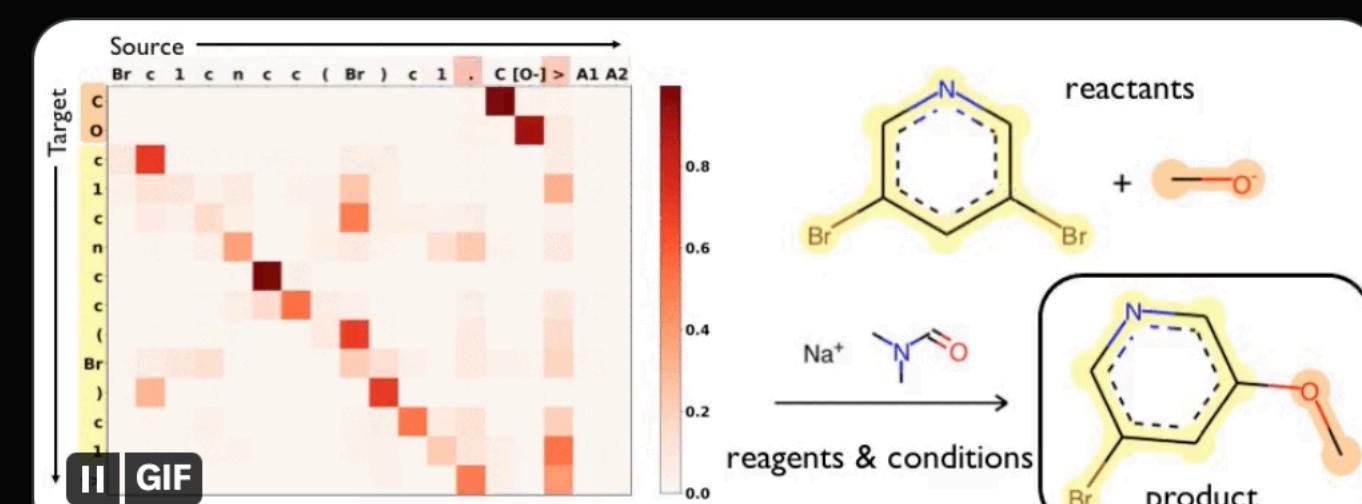


10:24 PM · Feb 20, 2023 · 100.5K Views

114 Retweets 14 Quote Tweets 586 Likes

**IBM Research** @IBMResearch · Aug 19, 2018

Attention chemists and chemistry students: IBM just launched the first, free #AI web service in the @IBMcloud for predicting chemical reactions. Try it now [rxn.res.ibm.com](https://rxn.res.ibm.com) @AmerChemSociety #ACSBoston

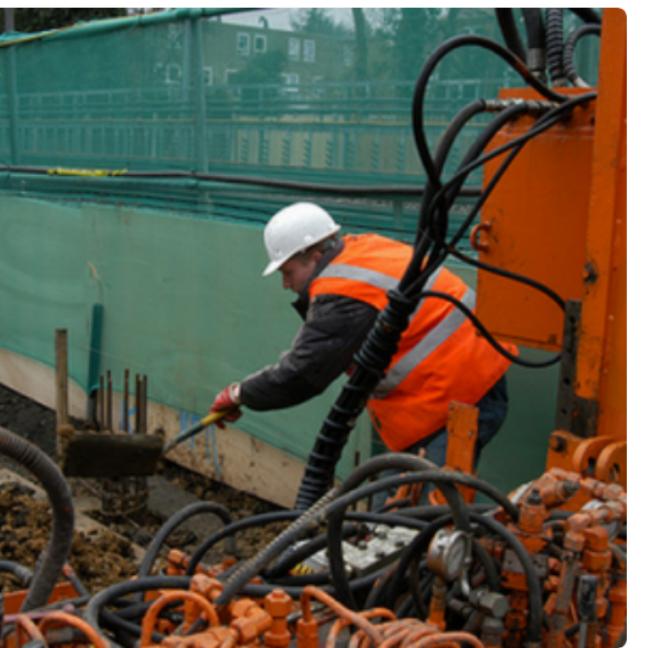


24 827 1,079

## And many more ...



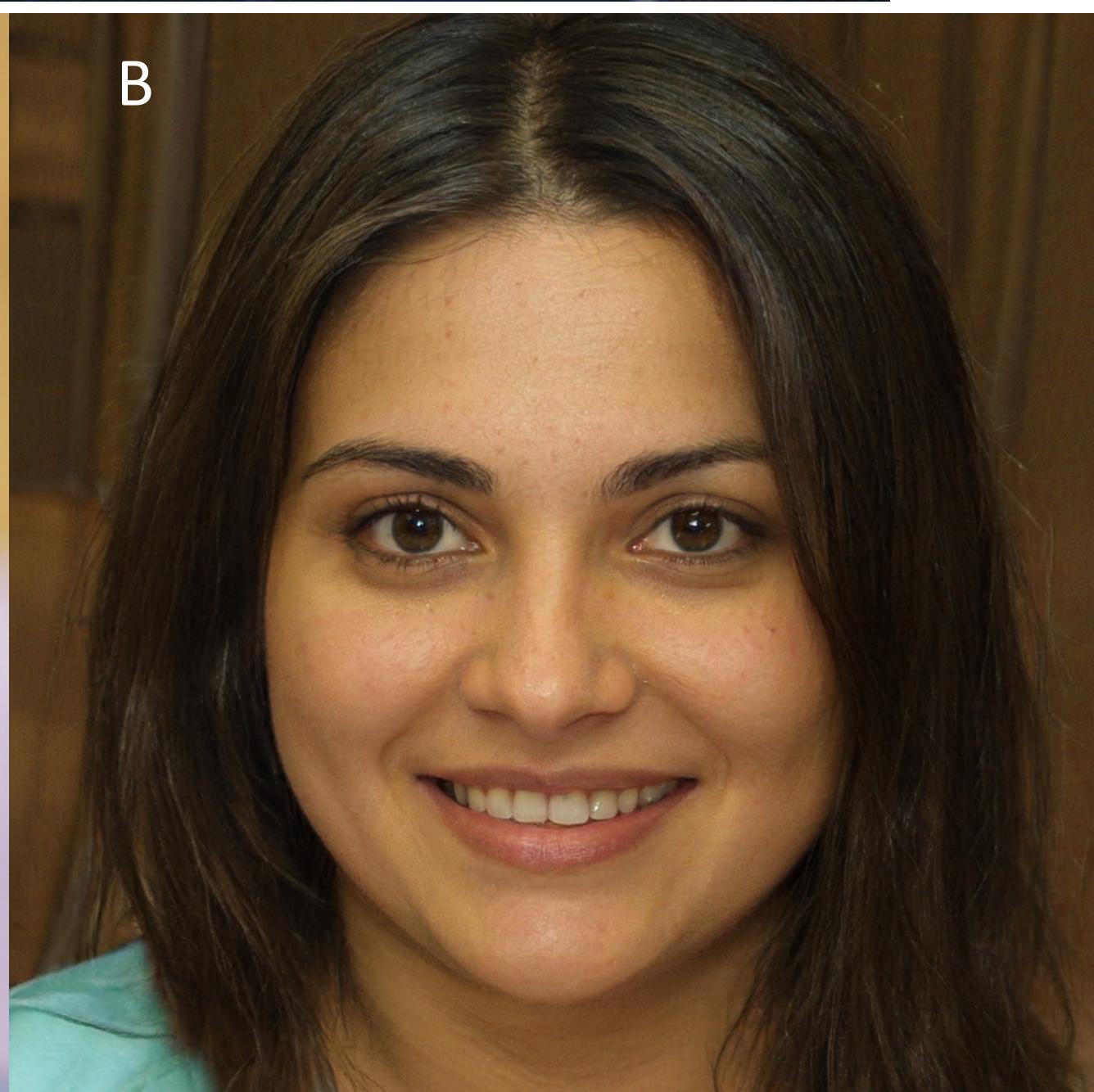
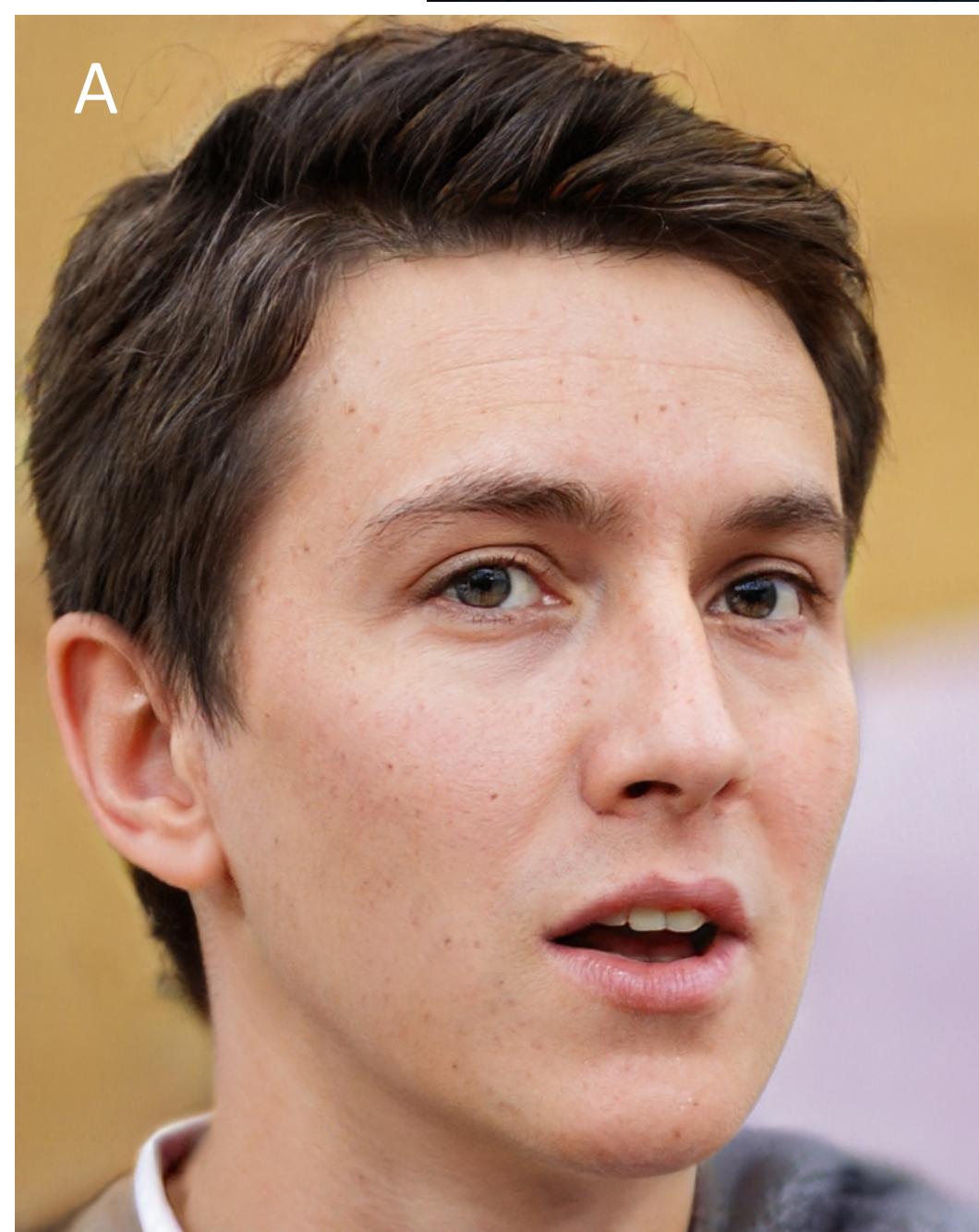
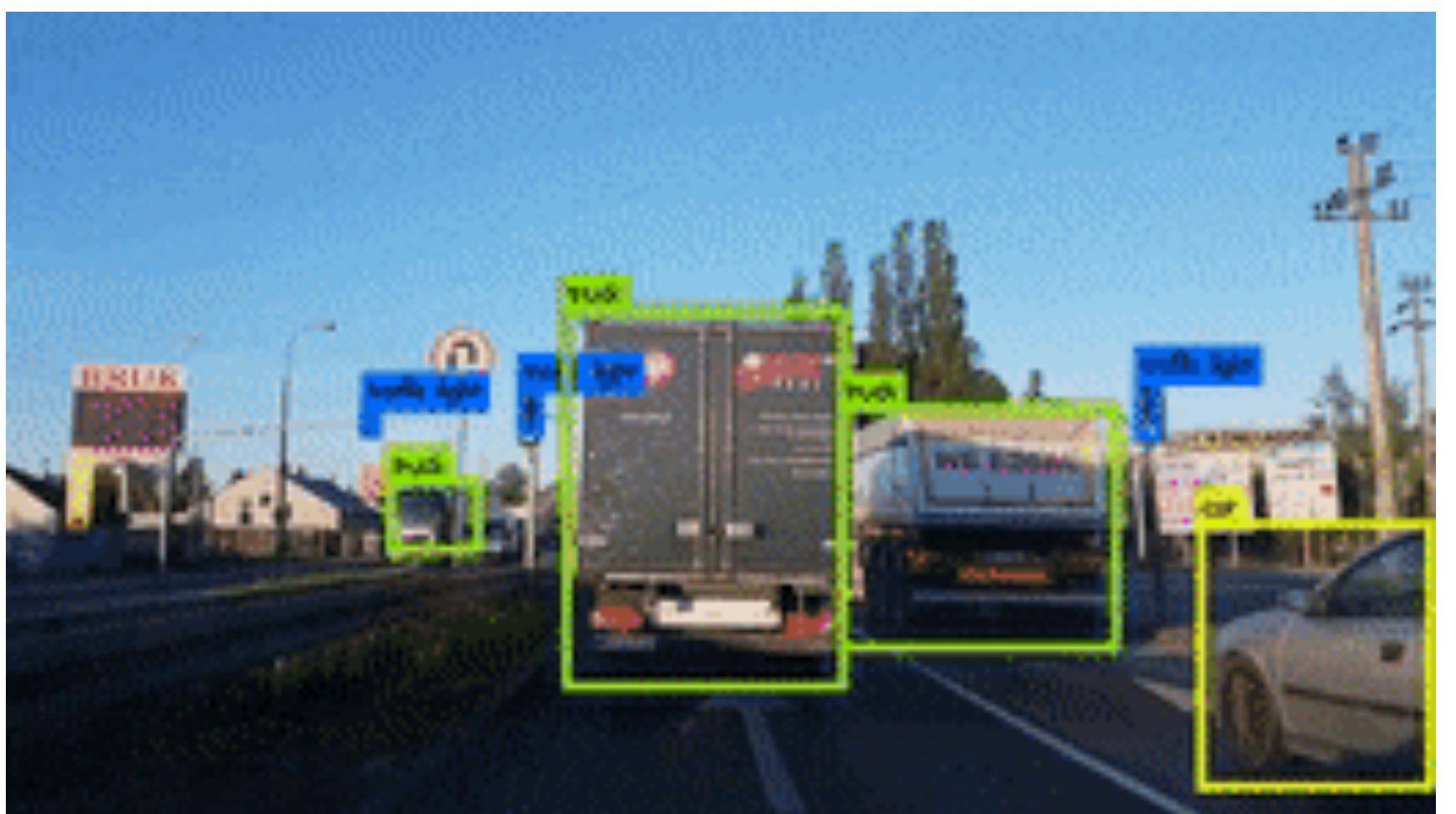
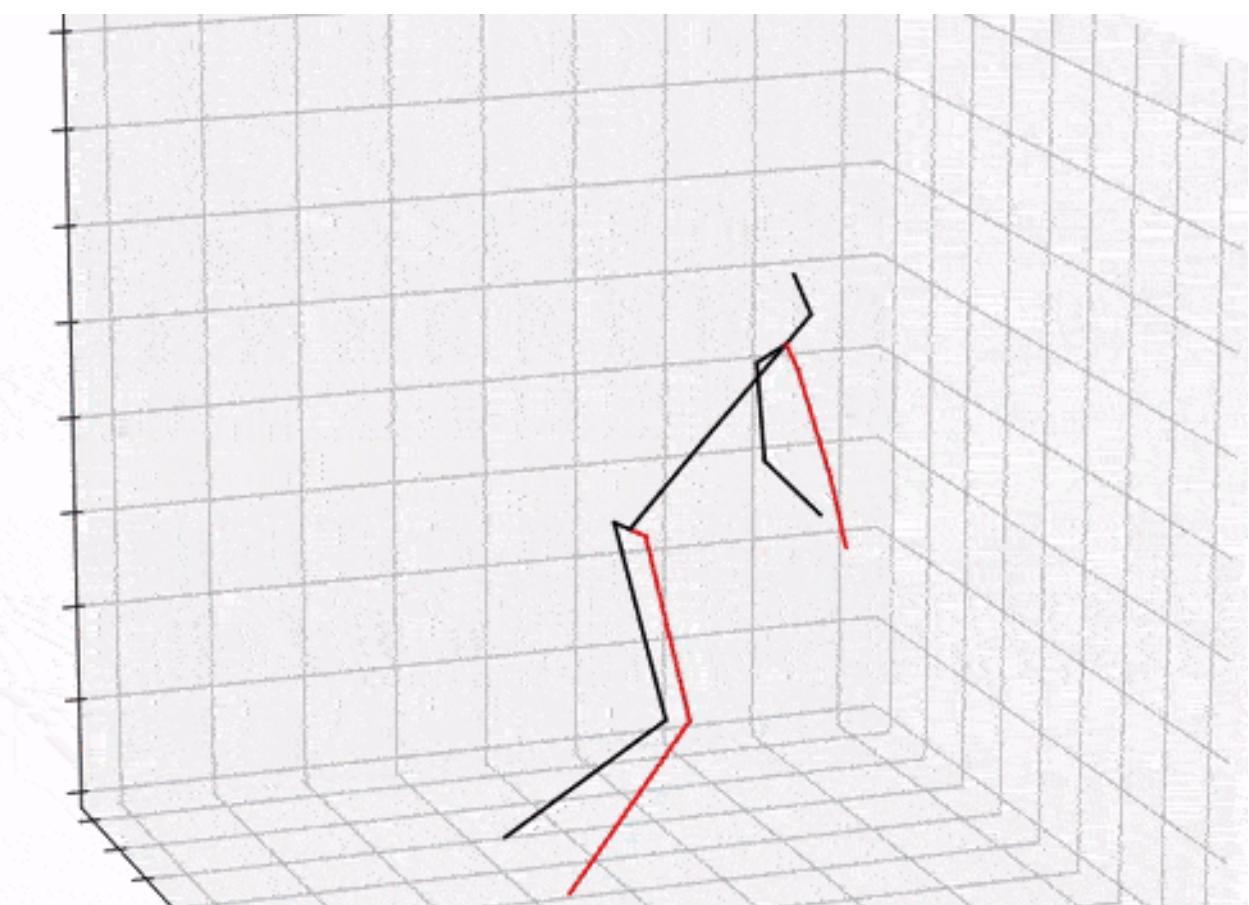
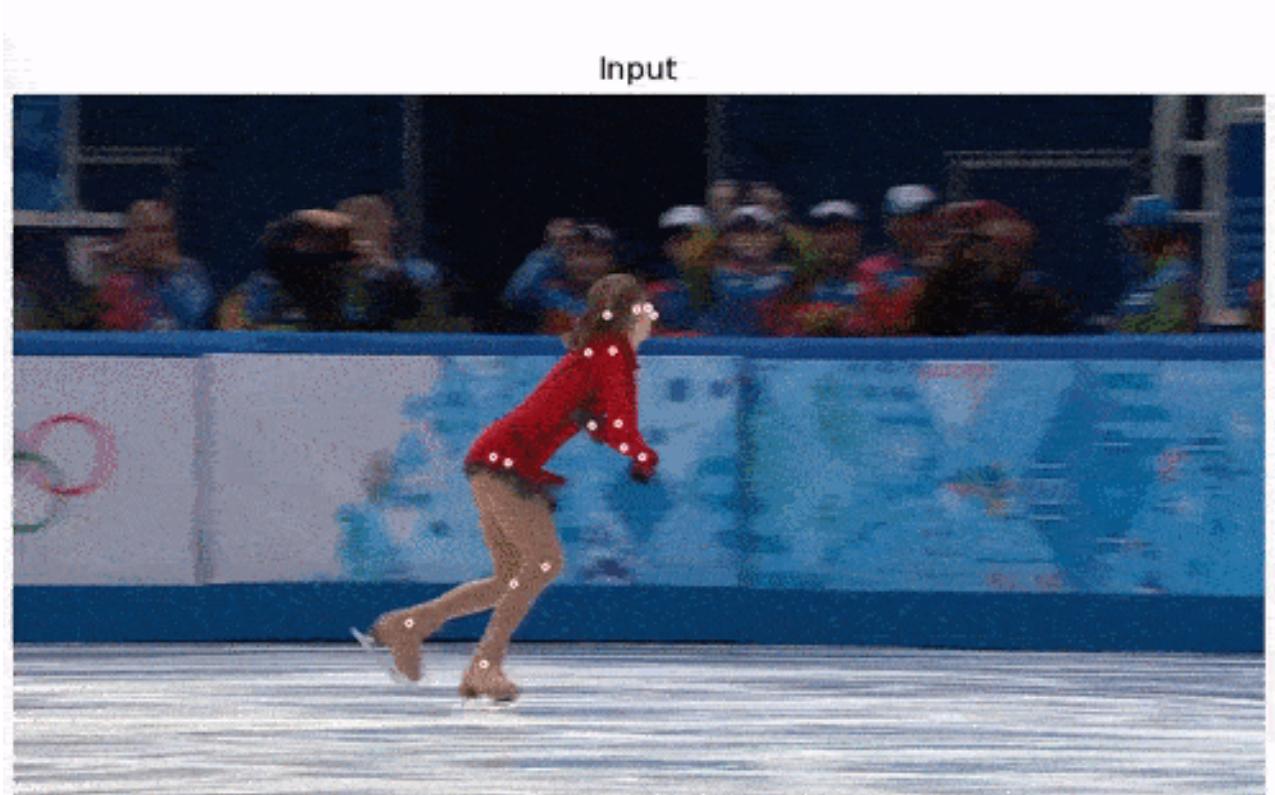
"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



KINYARWANDA - DETECTED	KINYARWANDA	ODIA (ORIYA)	TATAR	ENGLISH	TURKMEN	UYGHUR
Gutuma buri wese, aho ari ho hose amenya isi akanavuga ibyo atekereza mu ndimi zitandukanye.	x	Let everyone, anywhere, know the world and express their thoughts in different languages.	★			



Write With Transformer gpt2 ⓘ

gpt2 

i

 Shuffle initial text



### Shuffle initial text



Trigger autocomplete or tab

Select suggestion   and 

Cancel suggestion **esc**

## Save & Publish



TEXT PROMPT

an illustration of a baby panda in a wizard hat playing a guitar



AI-GENERATED  
IMAGES

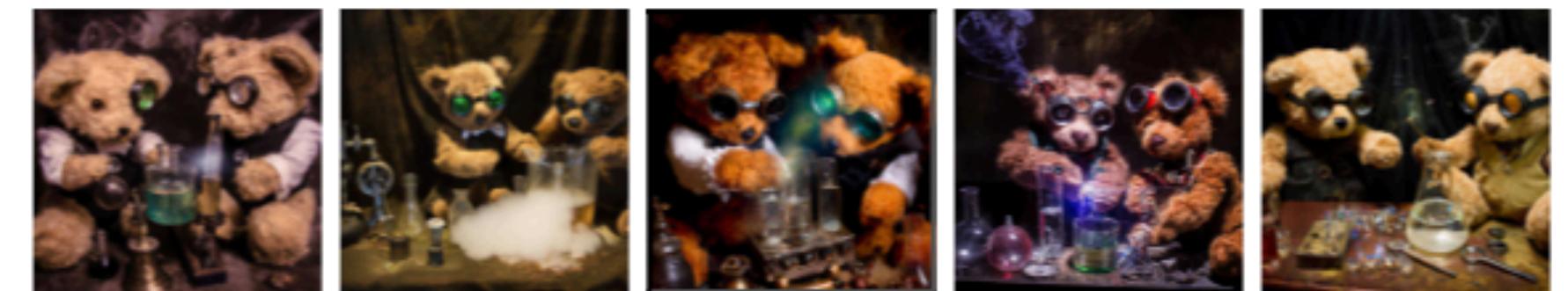


<https://openai.com/blog/dall-e/>, Jan 2021

TEXT DESCRIPTION

An astronaut **Teddy bears** A bowl of  
soup  
  
mixing sparkling chemicals as mad  
scientists shopping for  
groceries working on new AI research  
  
as a 1990s Saturday morning cartoon as  
digital art **in a steampunk style**

DALL-E 2



<https://openai.com/dall-e-2/>, April 2022

EP



P

Act as a professional ML researcher in chemistry, and teach me, a beginner, how to write AI for Chemistry code in Python:

## 1. Basics of machine learning

## 2. Libraries like scikit-learn and PyTorch



As a professional ML researcher in chemistry, I recommend starting by familiarizing yourself with the basics of machine learning, as well as the specific challenges and considerations that come with using ML in the field of chemistry.

In terms of writing AI for chemistry code in Python, there are several popular libraries and frameworks that are widely used in the field, including scikit-learn, TensorFlow, and PyTorch.

To get started, I recommend installing these libraries and experimenting with some simple tutorials and examples to gain a basic understanding of how they work. Additionally, I recommend studying the application of machine learning in chemistry, such as QSAR, QSPR, and molecule generation.

## 4. Datasets for molecular property prediction, e.g. QM9 and ZINC

Once you have a basic understanding of the libraries and frameworks, you can start working on your own projects. A good starting point could be to use pre-existing datasets such as QM9 and ZINC to train a model to predict chemical properties or generate novel molecules.

It is also a good idea to consult with the chemical community to understand the problem and get feedback on your work. Many conferences and journals have dedicated sessions on Machine Learning in Chemistry, and open-source platforms such as GitHub and Kaggle have many chemistry related challenges that one could take part in.

Remember that this is an active research area and new techniques, frameworks, and applications are being developed all the time, so it is important to keep up-to-date with the latest developments.

## 3. Simple tutorials for Quantitative structure-activity relationship (QSAR), Quantitative structure property relationships (QSPR), and molecule ■ generation

## 5. Interact with the #compchem community, get examples from GitHub and Kaggle (or also on Twitter)



- ChatGPT -> Text
- Midjourney -> Image
- [elevenlabs.io](https://elevenlabs.io)  
-> text to speech
- D-ID  
-> image, speech to video

Total time 15 minutes!

# Last year - Sora (text to video, OpenAI)



One year ago

Prompt: Photorealistic closeup video of two pirate ships battling each other as they sail inside a cup of coffee.

# Same prompt - generated by myself with Qwen Chat.



PROMPT: A diver in a diving suit explores the secrets of a sunken ship in the deep sea, holding a searchlight that illuminates the dark underwater world.



8. A worker is striking an iron block with a hammer, and sparks are flying.



<https://x.com/minchoi/status/1885060598188159212>

## The Nobel Prize in Physics 2024

### John Hopfield

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"



John Hopfield. Ill. Niklas Elmehed © Nobel Prize Outreach

### Geoffrey Hinton

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"



Geoffrey Hinton. Ill. Niklas Elmehed © Nobel Prize Outreach

# 2024 - two Nobel Prizes for AI

## The Nobel Prize in Chemistry 2024

### David Baker

"for computational protein design"



David Baker. Ill. Niklas Elmehed © Nobel Prize Outreach

### Demis Hassabis

"for protein structure prediction"



Demis Hassabis. Ill. Niklas Elmehed © Nobel Prize Outreach

### John Jumper

"for protein structure prediction"



John Jumper. Ill. Niklas Elmehed © Nobel Prize Outreach

# What is Artificial Intelligence (AI)?



- AI enables computers to **mimic human behaviour**.
- Perform tasks that usually require human intelligence
  - Visual perception
  - Speech recognition
  - Decision-making
  - Language translation
  - And many tasks in Chemistry!
- Machine Learning (ML) and expert systems (rule-based) are subfields of AI

# What is Machine Learning?



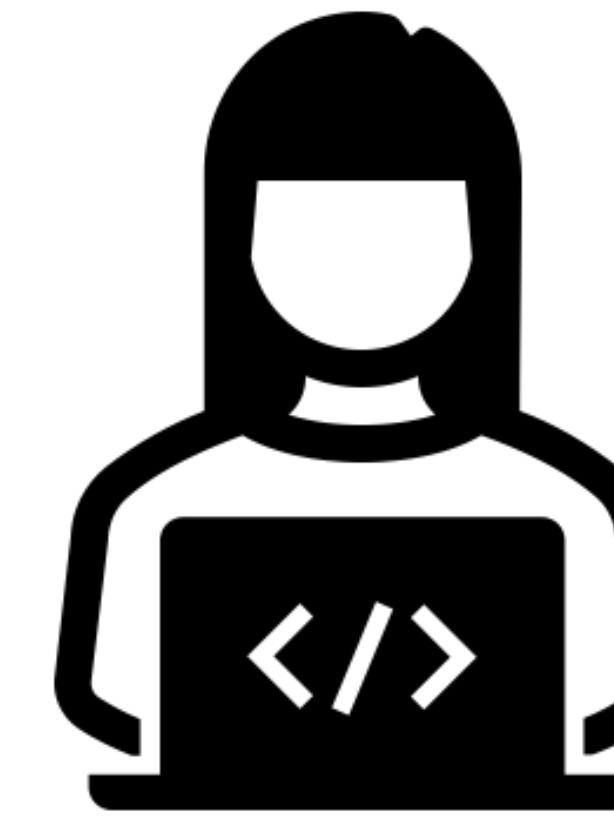
- ML is a subset of AI that **enables machines to learn** and improve from experience, **without being explicitly programmed**.
- Learn from data, identify patterns and make decisions
- 3 main types:
  - **Supervised learning** (labeled data)
  - **Unsupervised learning** (structure in unlabeled data)
  - **Reinforcement learning** (rewards for actions)

# Traditional programming ("Expert system")

Input x

**5.5**

**3.5**



Output y

**Pass**

**Fail**

If grade 4 or higher ( $x \geq 4$ ),  
student passes the course.

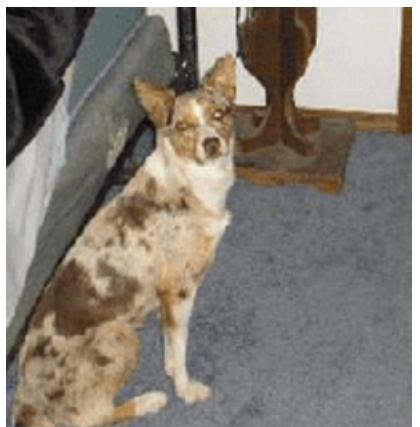
Predefined human-  
written rules  
(knowledge base)

Input  $x$

Data (training + validation)

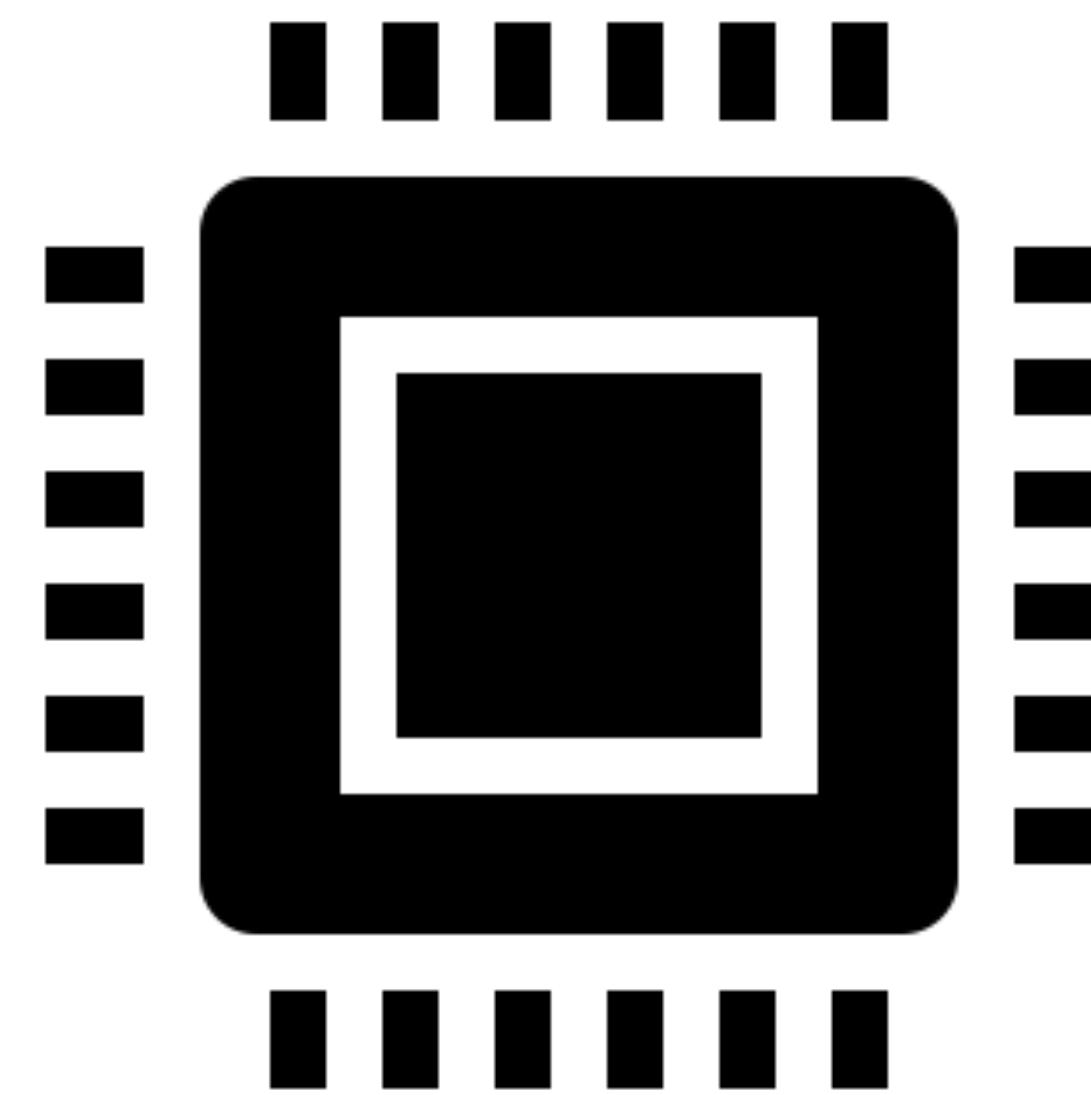


Unseen examples (test)



Model

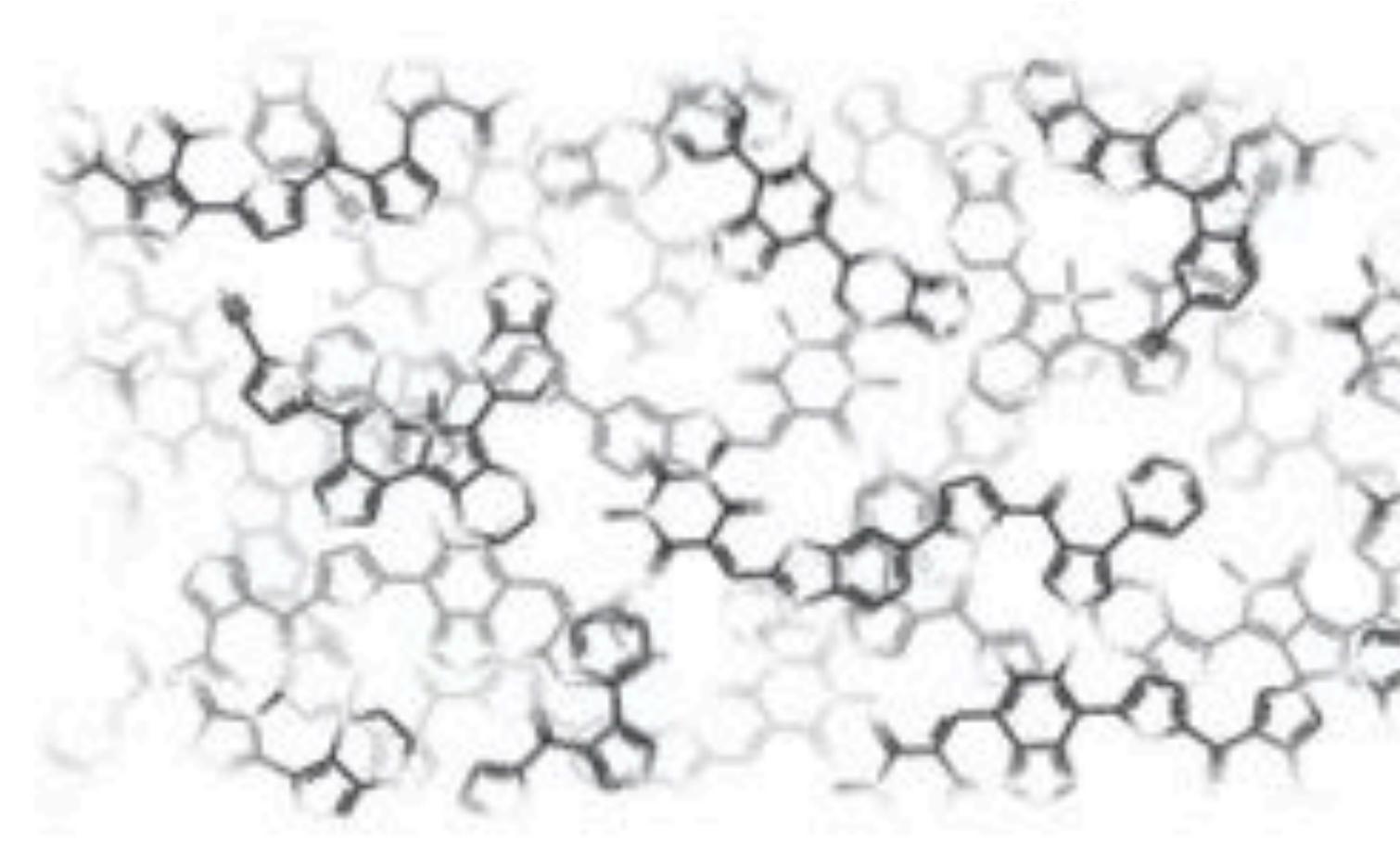
Output  $y$



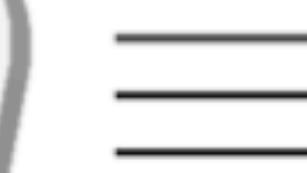
dog  
cat  
cat

dog

# Machine Learning in Chemistry



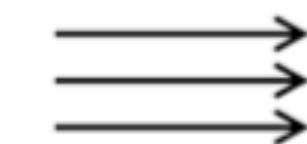
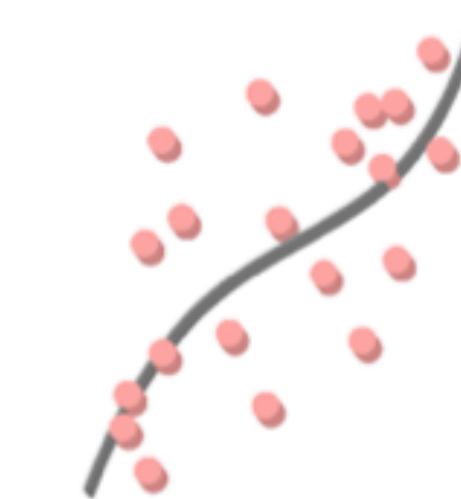
An unknown mechanism



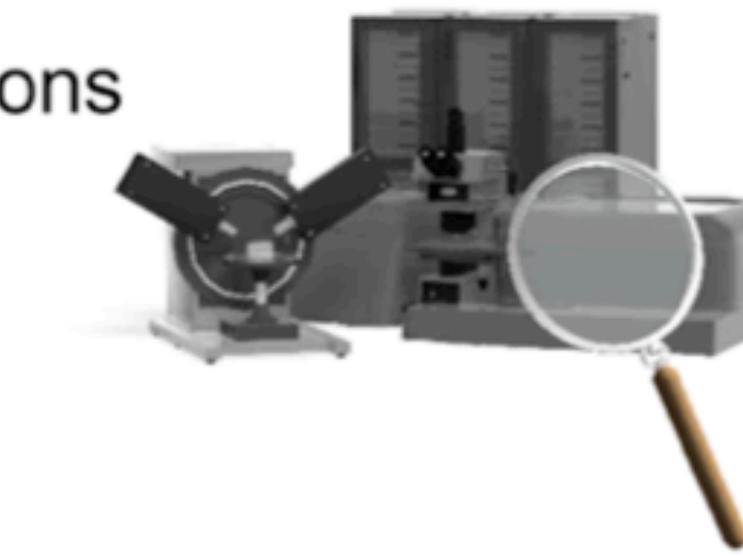
Observations  
(data)



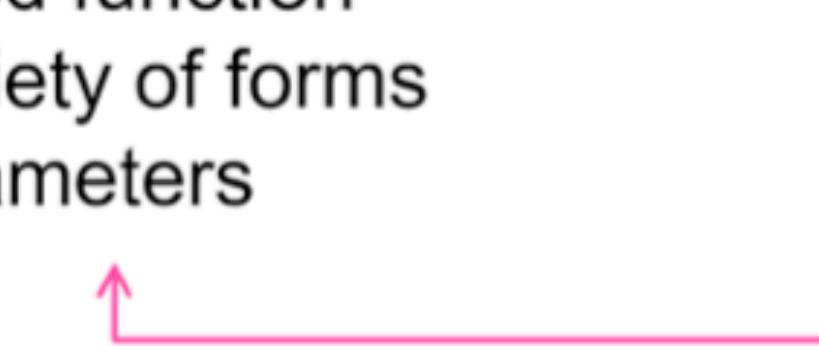
- Huge chemical space
- Challenging to write rules



Observations  
(data)



A versatile unfixed function  
that can take a variety of forms  
by tuning parameters



Function is best fitted to data by tuning parameters

# Key ingredients for Machine Learning

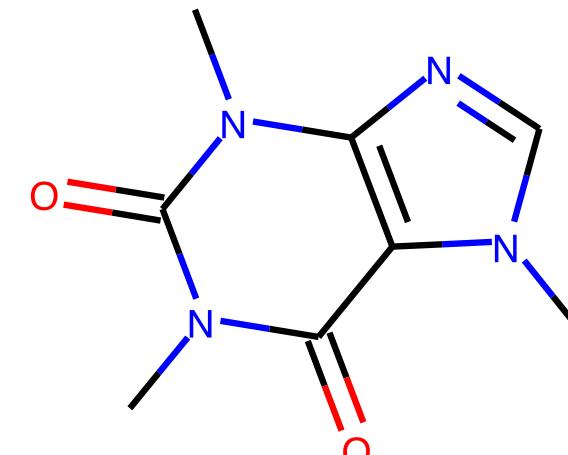
## Molecular fingerprints

000010000....0100

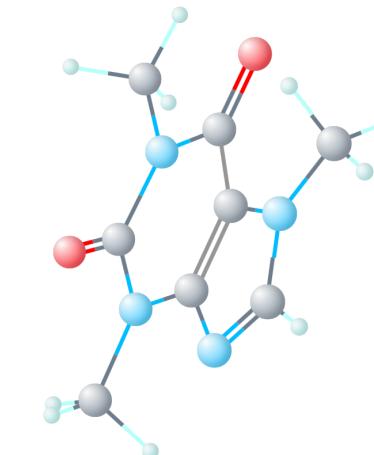
## Text-based representations

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

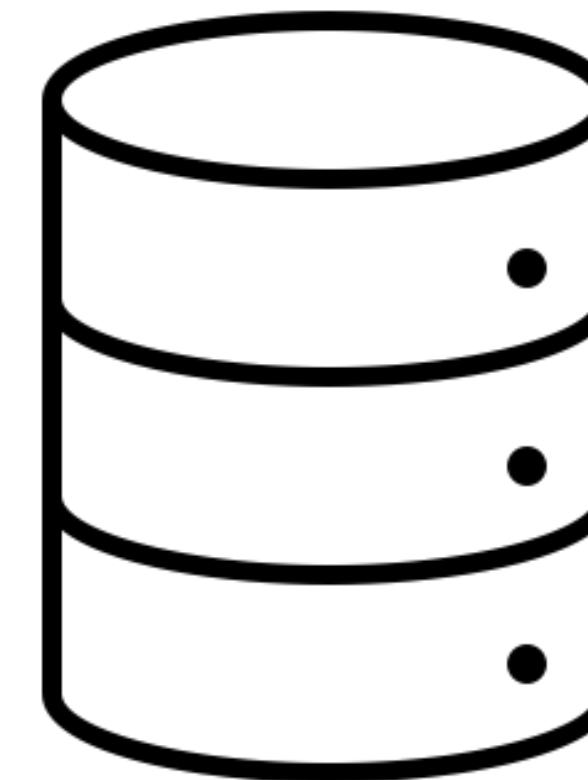
## Graph-based representations



## 3D coordinates & surface



Representations  
(machine-readable)

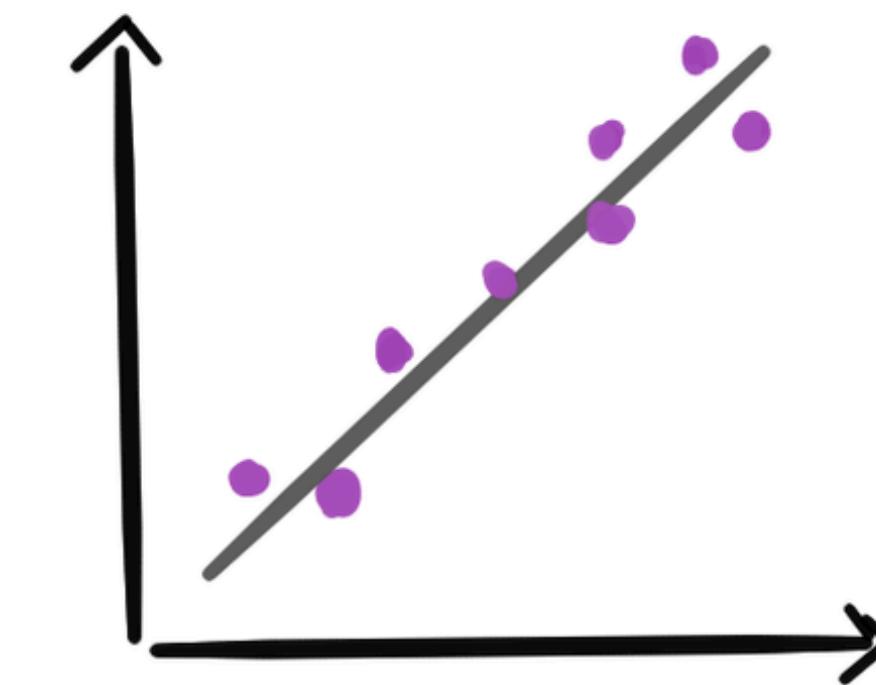


Examples are:

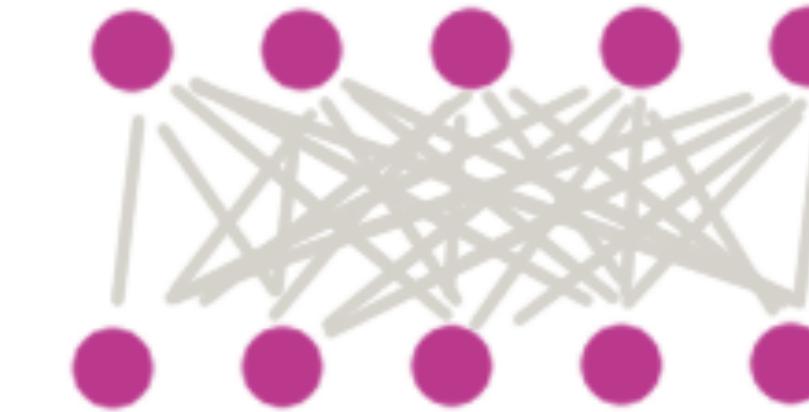
- Molecules & properties
- Chemical reactions
- Synthesis procedures

Data  
(garbage in = garbage out)

## Linear regression model



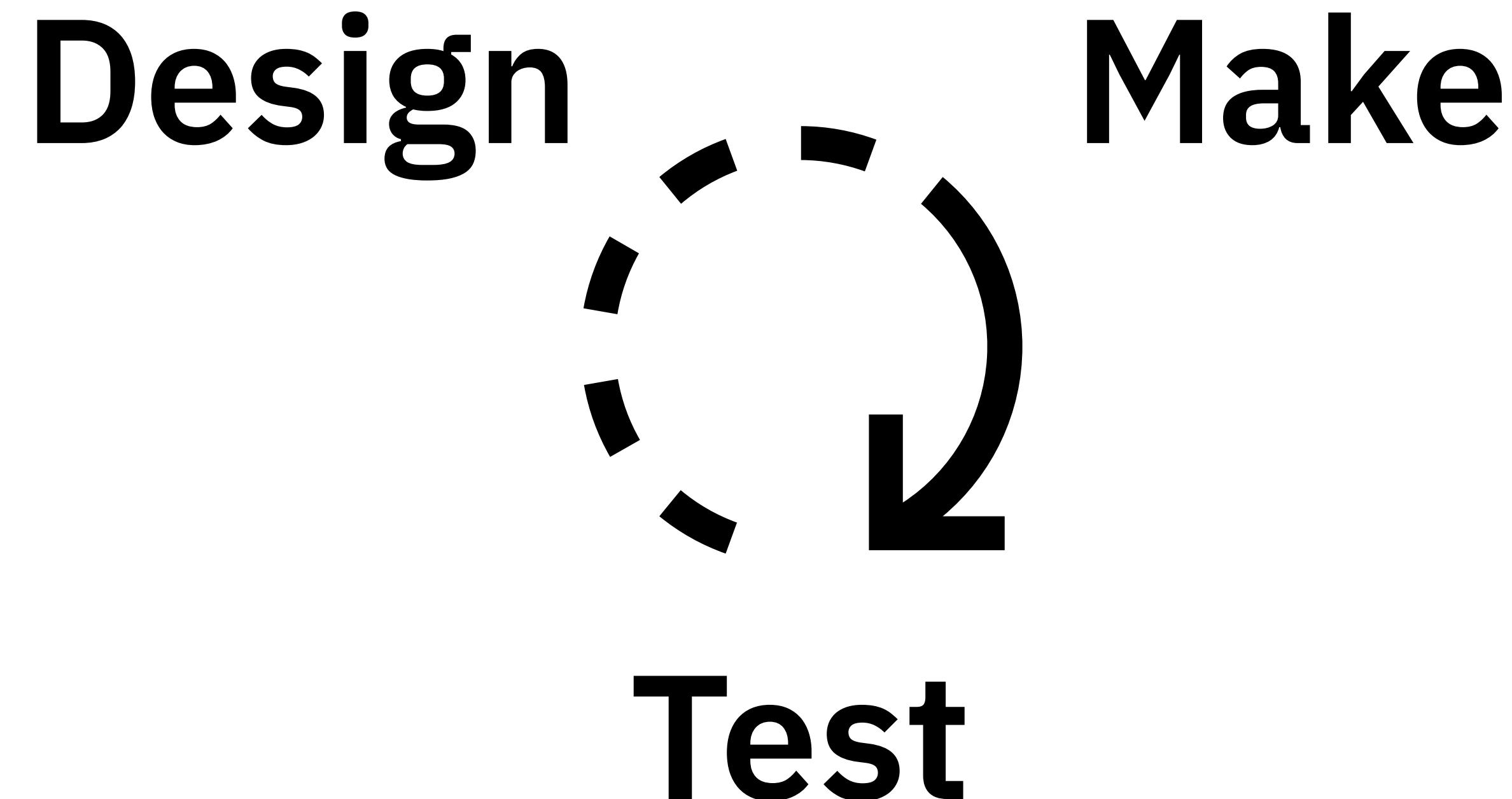
## Neural networks



And many more..

Models/algorithms

1. Chemistry-focused python fundamentals
2. Machine learning basics
3. AI-accelerated molecular discovery and synthesis



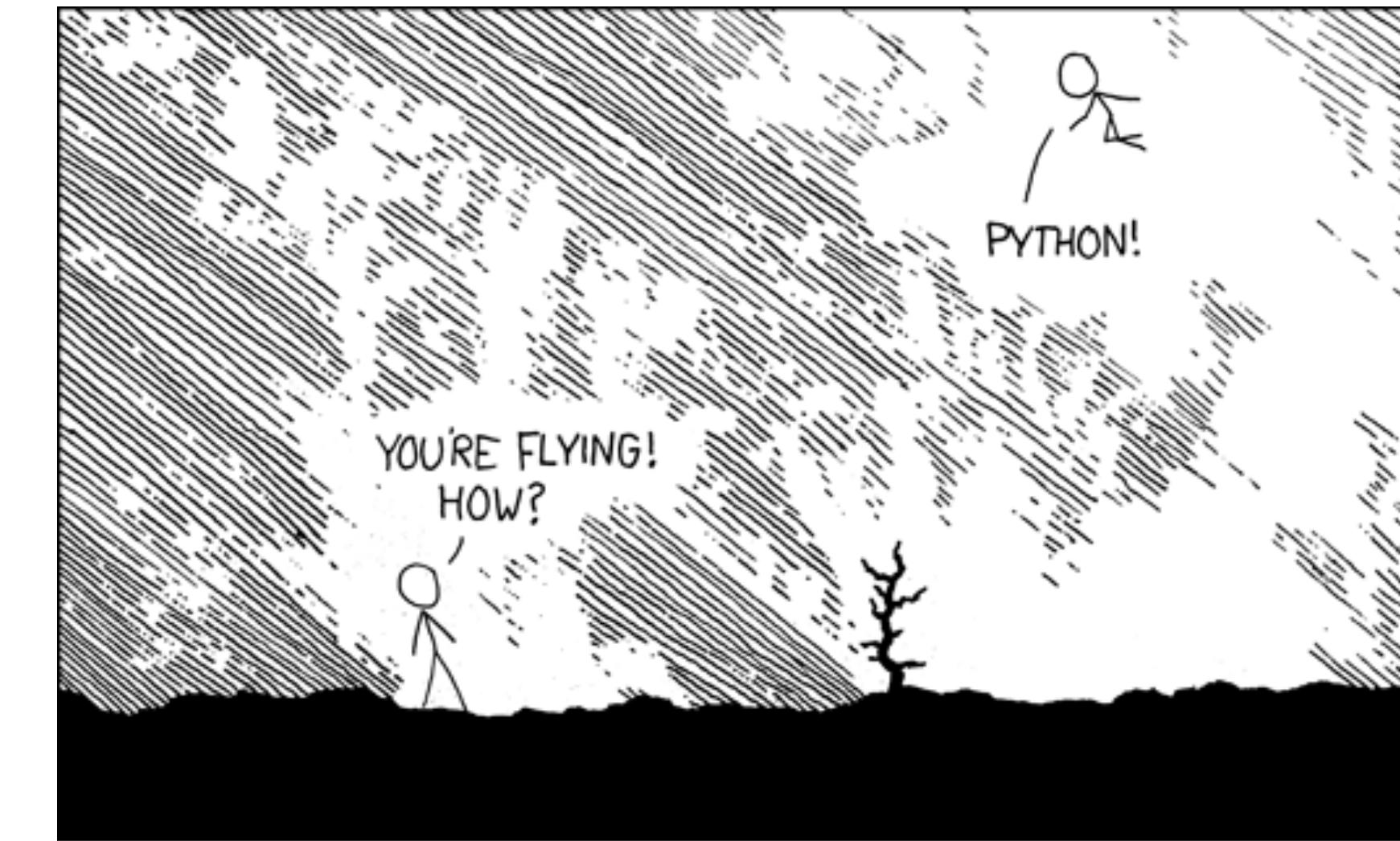
- 1: Intro / molecular representations
- 2: Traditional ML (supervised)
- 3: Deep Learning
- 4: Unsupervised ML
- 5: De novo molecule generation
- 6: De novo molecule generation
- 7: Chemical reactions (prediction)
- 8: Synthesis planning
- 9: Other reaction tasks
- 10: Bayesian Optimisation for Reactions
- 11: Open Data / FAIR / sharing models / code
- 12: LLMs in Chemistry (new lecture)
- 13: Free topic (guest speaker?)
- 14: Project presentations (the week before the end of the semester)

- Learn the importance of data and machine-readable representations
- Learn how to manipulate chemical data on a computer
- Learn machine learning for chemistry fundamentals
- Get an overview of state-of-the-art methods for the molecular design cycle
  - Property prediction (regression/classification)
  - Molecule design
  - Reaction prediction
  - Synthesis planning

- Groups of up to 3 (mix of computationally / experimentally-oriented students)
- ML project in chemistry
  - Collect data from literature / from a project you have worked on
  - Train ML models (GitHub repo, 30%)
  - Write a 4-page paper including intro/task/data/methods/results (40%)
  - Present outcomes in the last course session (time ~135 min/# of groups, 30%)
- Extra points for interpretations/visualisations/web app for model

# Programming language - Python

- Python is beginner-friendly
  - Like English
  - Concise and easy to read
- Versatile
  - Libraries for chemistry, ML, web
- Excellent community
  - Questions -> [https://  
stackoverflow.com/](https://stackoverflow.com/)
- or ChatGPT (as your personal tutor)
- High-in-demand



- <https://github.com/sib-swiss/first-steps-with-python-training>
- <https://realpython.com/>
- <https://schwallergroup.github.io/practical-programming-in-chemistry/> (work in progress)
- Ask ChatGPT any question you have, or if you prefer an open source large language model (<https://labs.perplexity.ai/>, code llama)

✓ sonar-pro  
sonar  
sonar-reasoning-pro  
sonar-reasoning

- Python version 3.11.11 (default on Colab)
- Notebooks either in Google Colab (<https://colab.research.google.com/>, gives you access to GPUs) or locally in a Jupyter lab environment



Interactive programming

- Interactive documents
- Made of cells (for text, equations, code, visualisations)
- You can run cells with “Run” or by pressing **Shift + Enter**
- Output is displayed below the cell

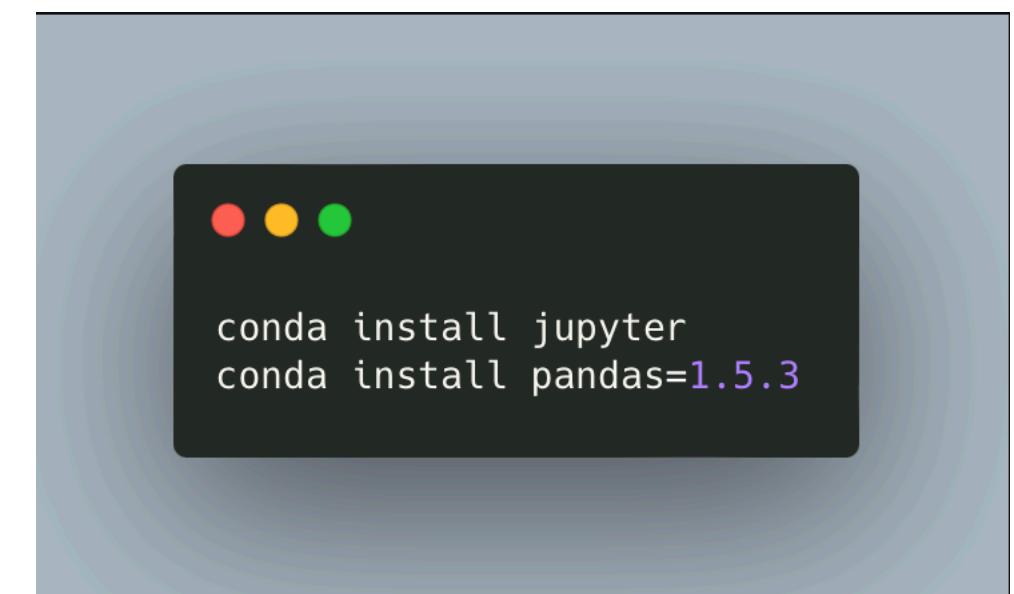
## Exercises

### Weekly schedule

Week	Topic	Link to Colab
1	Python and Jupyter	 Open in Colab
	Pandas	 Open in Colab
	Plotting data	 Open in Colab
	Intro to RDKit	 Open in Colab
2		

[https://schwallergroup.github.io/ai4chem\\_course/](https://schwallergroup.github.io/ai4chem_course/)

- “pip” (what we will mainly use on collar)
  - Default Python package manager
- “conda” / “miniconda” (if you install stuff locally)
  - Package manager
  - Managing environments (isolate dependencies)



# Virtual environments using Conda (for those who work locally)

- Different projects will require
  - Different python versions
  - Different module versions
  - ...
- They might be incompatible.
- Good practice:  
*one environment per project*

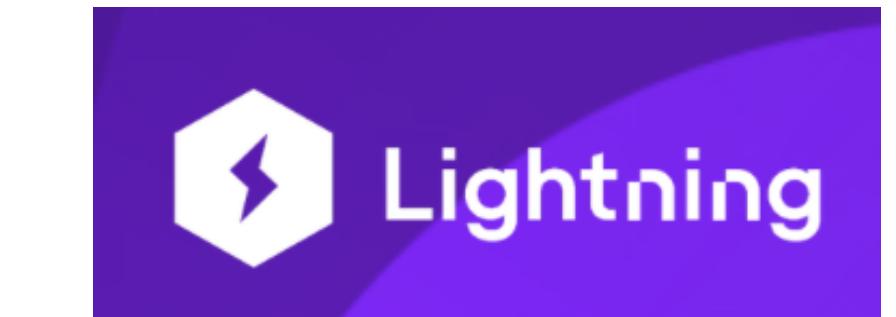


```
# if you are working locally,  
# create the environment using  
conda create -n ai4chem python=3.10  
  
# activate the environment  
conda activate ai4chem  
  
# install more packages  
conda install ...
```

- Important if you work locally. On Google colabs, you will install packages for single notebooks.

# EPFL Python ecosystem

32



- Günter Klambauer
- Greg Landrum (and other RDKit contributors)
- Sebastian Raschka (<https://sebastianraschka.com/blog/2021/ml-course.html>)
- Rocío Mercado