Machine learning for chemistry and materials science
AI in Astrochemistry – COST Nanospace Training School

Miguel A. Caro
Assistant Professor
Department of Chemistry and Materials Science, Aalto University, Finland

Aalto University School of Chemical Engineering
5 August 2025

turbogap.fi **miguelcaro.org**

Aalto University
School of Chemical
Engineering

SUOMEN AKATEMIA
FINLANDS AKADEMI
ACADEMY OF FINLAND

CSC

TurboGAP
Data-driven atomistic simulation

Today's lecture: (incomplete) overview of ML for chemistry and materials science

Today's lecture: (incomplete) overview of ML for chemistry and materials science

- Disclaimers:

Today's lecture: (incomplete) overview of ML for chemistry and materials science

- Disclaimers:
  - I'm not an ML expert! Background is physics/atomistic modeling

Today's lecture: (incomplete) overview of ML for chemistry and materials science

- Disclaimers:
  - I'm not an ML expert! Background is physics/atomistic modeling
  - My perpective is skewed towards my own "journey" in applied ML (ML interatomic potentials)

Today's lecture: (incomplete) overview of ML for chemistry and materials science

- Disclaimers:
  - I'm not an ML expert! Background is physics/atomistic modeling
  - My perpective is skewed towards my own "journey" in applied ML (ML interatomic potentials)
  - Growing more ignorant* every day!

Today's lecture: (incomplete) overview of ML for chemistry and materials science

- Disclaimers:
  - I'm not an ML expert! Background is physics/atomistic modeling
  - My perpective is skewed towards my own "journey" in applied ML (ML interatomic potentials)
  - Growing more ignorant* every day!
- Overview of ML within the chemistry/materials science domain

Today's lecture: (incomplete) overview of ML for chemistry and materials science

- Disclaimers:
  - I'm not an ML expert! Background is physics/atomistic modeling
  - My perpective is skewed towards my own "journey" in applied ML (ML interatomic potentials)
  - Growing more ignorant* every day!
- Overview of ML within the chemistry/materials science domain
- Show some examples of what people have done with ML

Today's lecture: (incomplete) overview of ML for chemistry and materials science

- Disclaimers:
  - I'm not an ML expert! Background is physics/atomistic modeling
  - My perpective is skewed towards my own "journey" in applied ML (ML interatomic potentials)
  - Growing more ignorant* every day!
- Overview of ML within the chemistry/materials science domain
- Show some examples of what people have done with ML
- Shameless self promotion of our own group's work!

Today's lecture: (incomplete) overview of ML for chemistry and materials science

- Disclaimers:
  - I'm not an ML expert! Background is physics/atomistic modeling
  - My perpective is skewed towards my own "journey" in applied ML (ML interatomic potentials)
  - Growing more ignorant* every day!
- Overview of ML within the chemistry/materials science domain
- Show some examples of what people have done with ML
- Shameless self promotion of our own group's work!
- My main goal: to get you **excited** and convince you of the benefits of adding ML to your research workflows

Today's lecture: (incomplete) overview of ML for chemistry and materials science

- Disclaimers:
  - I'm not an ML expert! Background is physics/atomistic modeling
  - My perpective is skewed towards my own "journey" in applied ML (ML interatomic potentials)
  - Growing more ignorant* every day!
- Overview of ML within the chemistry/materials science domain
- Show some examples of what people have done with ML
- Shameless self promotion of our own group's work!
- My main goal: to get you **excited** and convince you of the benefits of adding ML to your research workflows
- Some of these slides are based on a new course at Aalto-CHEM: "AI for Chemistry and Materials Science"

About Aalto University

About Aalto University

- Top technical university in Finland



About Aalto University

- Top technical university in Finland
- Named after famous Finnish architect Alvar Aalto



About Aalto University

- Top technical university in Finland
- Named after famous Finnish architect Alvar Aalto
- Great place to do computational science



Computational Chemistry at Aalto

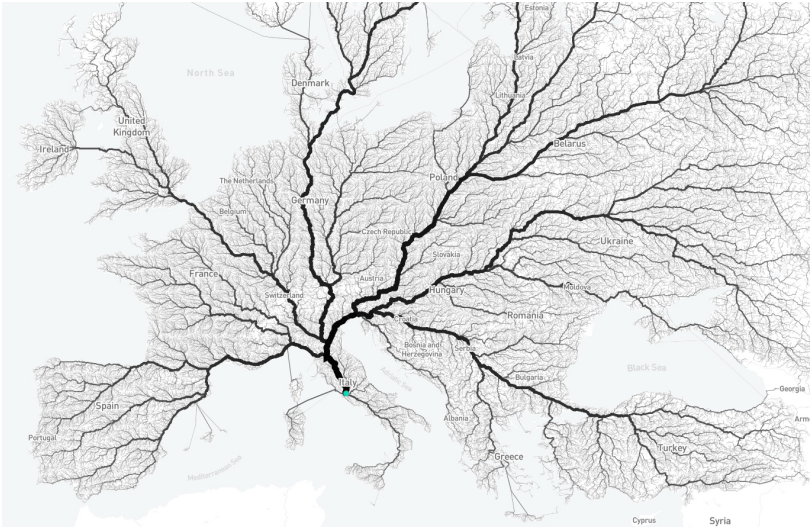
Aalto University (former Helsinki U. Tech.)

- School of Science
- School of Engineering
- School of Arts, Design & Architecture
- School of Business
- School of Electrical Engineering
- **School of Chemical Engineering**
  - Dept. of Bioproducts & Biosystems
  - Dept. of Chemical & Metallurgical Eng.
  - **Dept. of Chemistry & Materials Science**
    - Computational Chemistry (Kari Laasonen)
    - Soft Materials Modelling (Maria Sammalkorpi)
    - Inorganic Materials Modelling (Antti Karttunen)
    - **Data-driven Atomistic Simulation (Miguel Caro)**

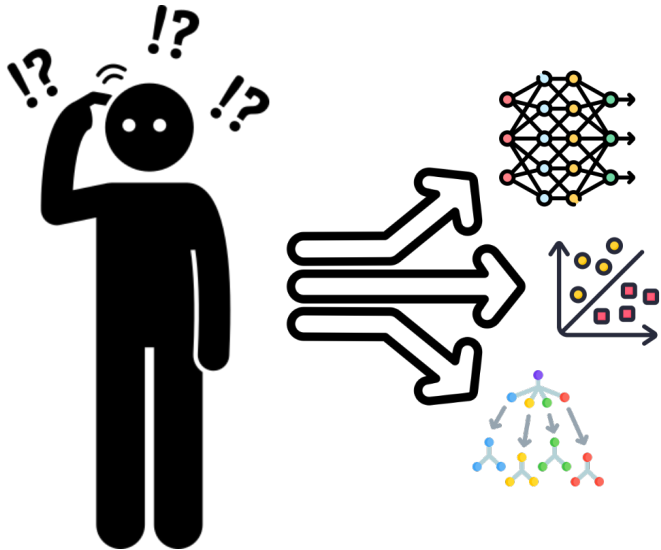


Otaniemi, Espoo

Since antiquity: "All paths lead to Rome"



In 2025: do all (scientific) paths "lead to ~~AI~~ machine learning"?



My path to ML:
Trying to understand the structure of amorphous carbon materials

What do complex/disordered materials look like?

What do complex/disordered materials look like?



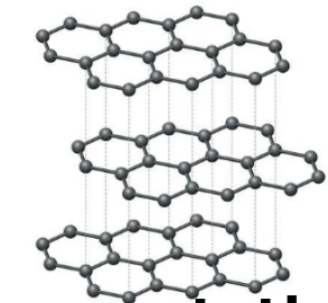
expectation

VS



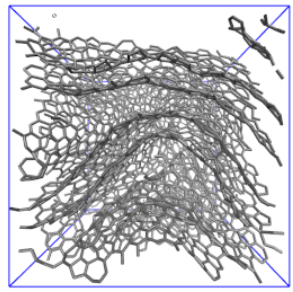
reality

What do complex/disordered materials look like?



expectation

vs



reality

What do complex/disordered materials look like?

Goal:
To infer the **atomic-scale structure** of (disordered) materials.

What do complex/disordered materials look like?

Complication:
Experimental techniques for materials characterization (microscopy, spectroscopy) do not (usually) offer direct access to the atomic positions.

What do complex/disordered materials look like?

(A) Solution:

Use atomistic modeling tools to 1) **generate the structure** and 2) to derive the **same observables** as accessible experimentally. Then, **compare**.

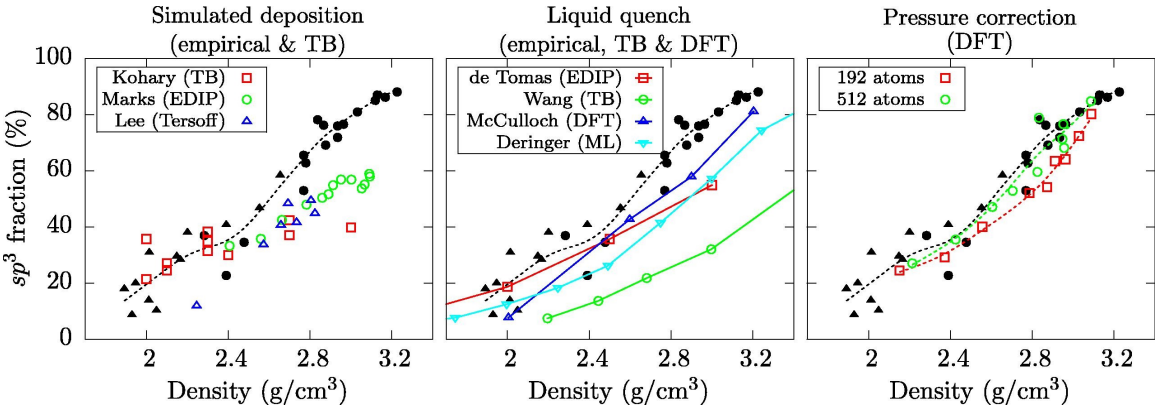
What do complex/disordered materials look like?

Ingredients:

- A good description of the **potential energy surface**
- Efficient/effective sampling/optimization protocols
- A quantitatively accurate method to predict the experimentally accessible observables

1. Let's start "easy":

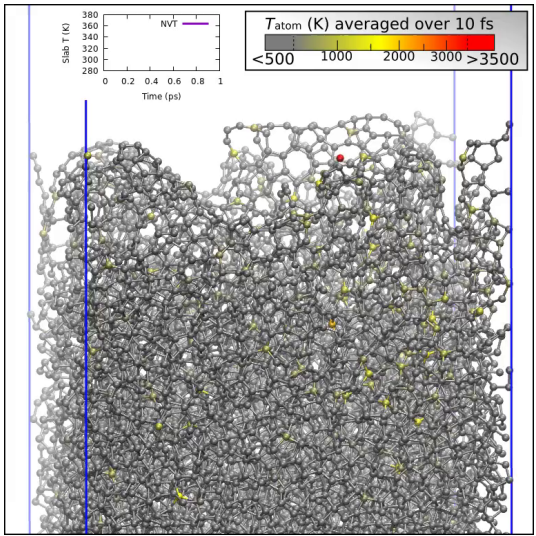
1. Let's start "easy": what does *chemically pure* amorphous carbon look like?



[1] **T. Laurila, S. Sainio, and M. A. Caro, Prog. Mater. Sci. 88, 499 (2017)**

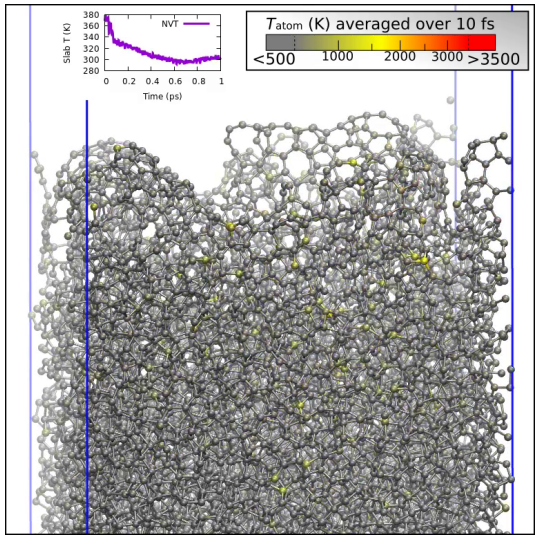
[2] V. L. Deringer and G. Csányi, Phys. Rev. B **95**, 094203 (2017)

Deposition simulations, one atom at a time



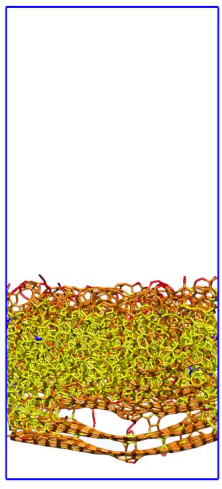
[1] **M. A. Caro, G. Csányi, T. Laurila, and V. L. Deringer, Phys. Rev. B 102, 174201 (2020)**

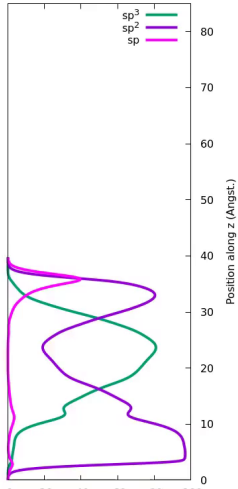
Deposition simulations, one atom at a time



[1] **M. A. Caro, G. Csányi, T. Laurila, and V. L. Deringer, Phys. Rev. B 102, 174201 (2020)**

Amorphous carbon growth from molecular dynamics

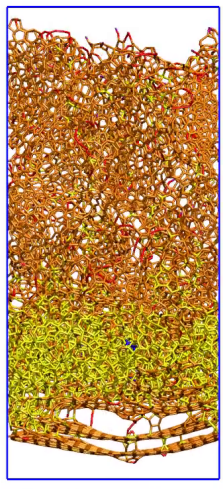


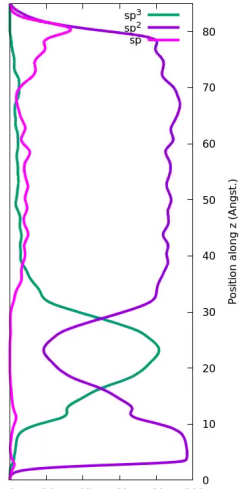


[1] M. A. Caro, V. L. Deringer, J. Koskinen, T. Laurila, and G. Csányi, Phys. Rev. Lett. 120, 166101 (2018)

[2] M. A. Caro, Zenodo DOI:10.5281/zenodo.1133425, (2017)

Amorphous carbon growth from molecular dynamics

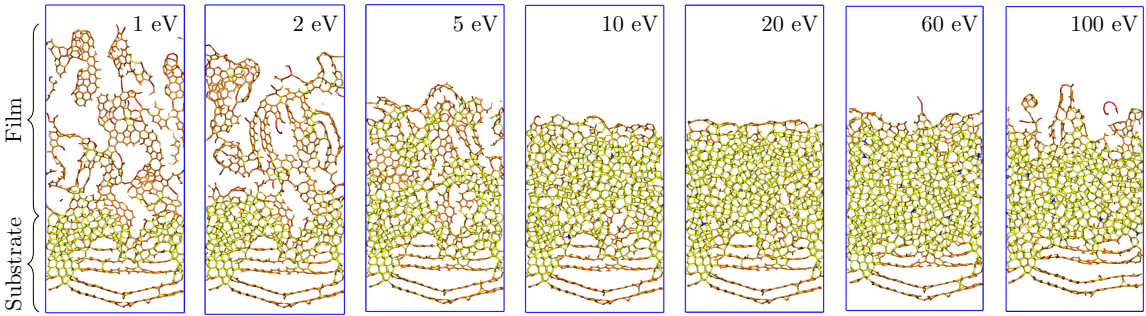




[1] **M. A. Caro, V. L. Deringer, J. Koskinen, T. Laurila, and G. Csányi, Phys. Rev. Lett. 120, 166101 (2018)**

[2] **M. A. Caro, Zenodo DOI:10.5281/zenodo.1133425, (2017)**

Film characteristics



[1] M. A. Caro, V. L. Deringer, J. Koskinen, T. Laurila, and G. Csányi, Phys. Rev. Lett. 120, 166101 (2018)

[2] M. A. Caro, G. Csányi, T. Laurila, and V. L. Deringer, Phys. Rev. B 102, 174201 (2020)

- Bold claim: The structure of chemically pure (elemental) a-C is “solved” with machine learning potentials

- Bold claim: The structure of chemically pure (elemental) a-C is "solved" with machine learning potentials
- What's next: how much other "stuff" (H, O, N) can you put in a-C?

1. ~~Let's start easy: what does *chemically pure* amorphous carbon look like?~~

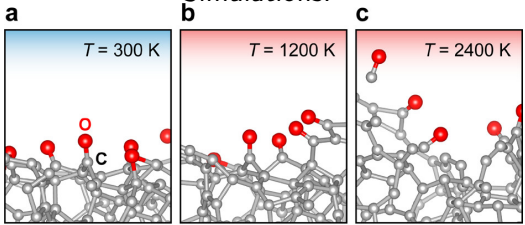
1. ~~Let's start easy: what does *chemically pure* amorphous carbon look like?~~
2. Much harder:

1. ~~Let's start easy: what does *chemically pure* amorphous carbon look like?~~
2. Much harder: how much oxygen can you put in amorphous carbon before it *burns*?

1. ~~Let's start easy: what does *chemically pure* amorphous carbon look like?~~
2. Much harder: how much oxygen can you put in amorphous carbon before it *burns*? and what does it look like?

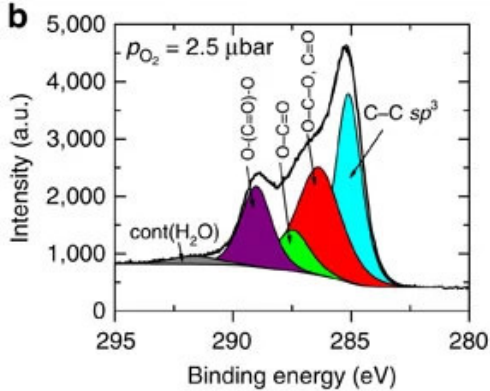
Oxygen-rich amorphous carbon

Simulations:[1]



(follow up simulations gave up to $\sim$ 34% O)

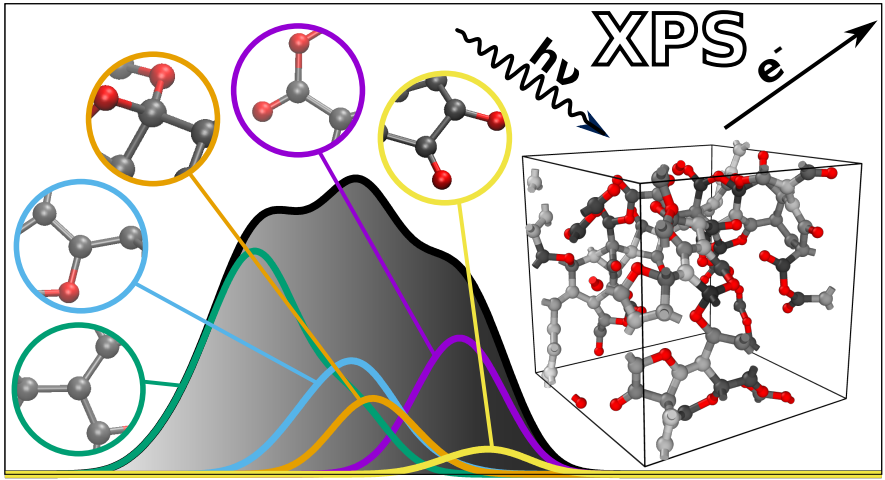
Experimental claim from XPS data on a-C:O of close to 40% O content:[2]



[1] V. L. Deringer, M. A. Caro, R. Jana, et al., Chem. Mater. 30, 7438 (2018)

[2] C. A. Santini, A. Sebastian, C. Marchiori, et al., Nat. Commun. **6**, 1 (2015)

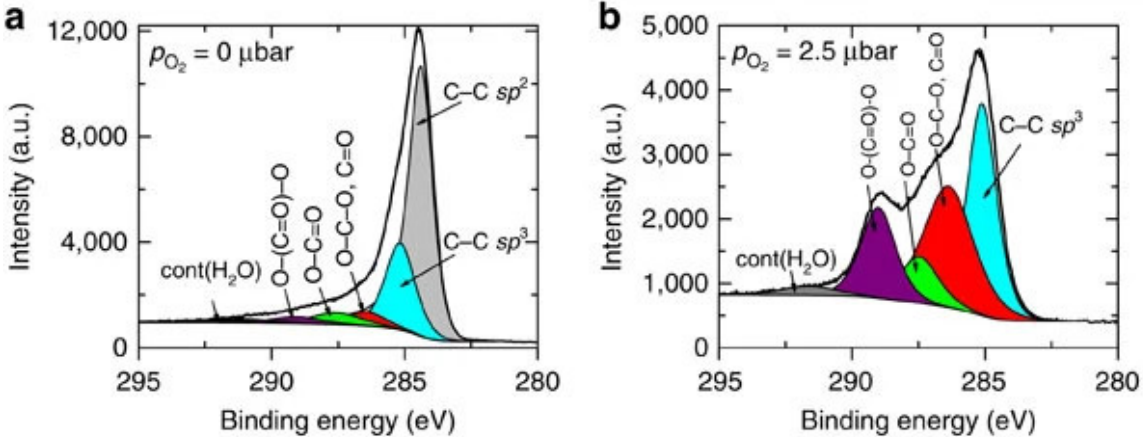
XPS prediction (link between experiment and simulation)



[1] D. Golze, M. Hirvensalo, Hernández-León P., et al., Chem. Mater. 34, 6240 (2022)

How to intepret X-ray spectra (experimentally): peak fitting

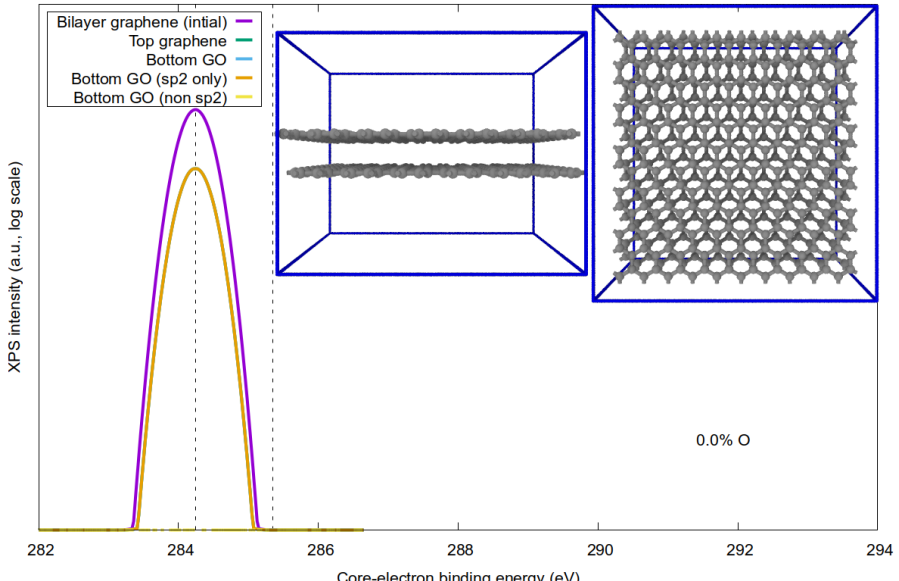
X-ray photoelectron spectroscopy (XPS)



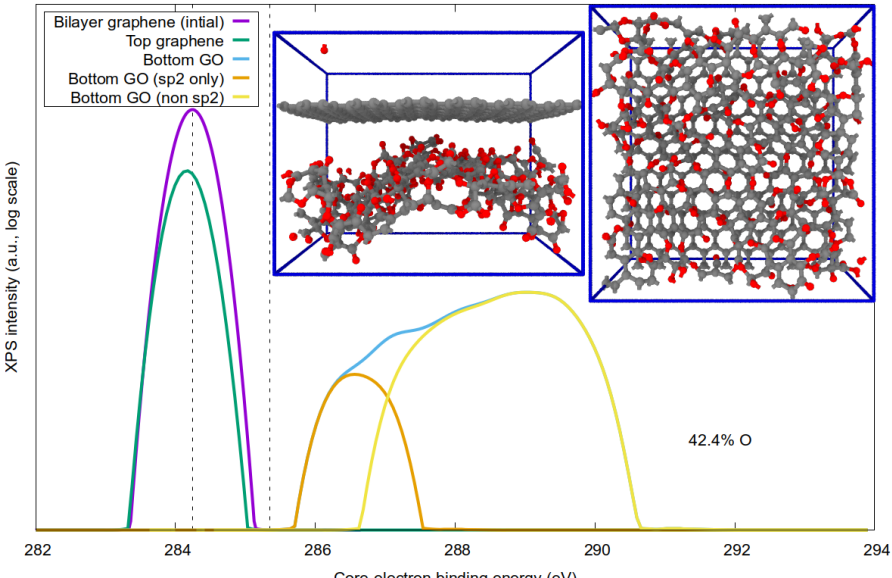
It's very inconclusive!

[1] C. A. Santini, A. Sebastian, C. Marchiori, et al., Nat. Commun. **6**, 1 (2015)

Peaks positions depend on surroundings

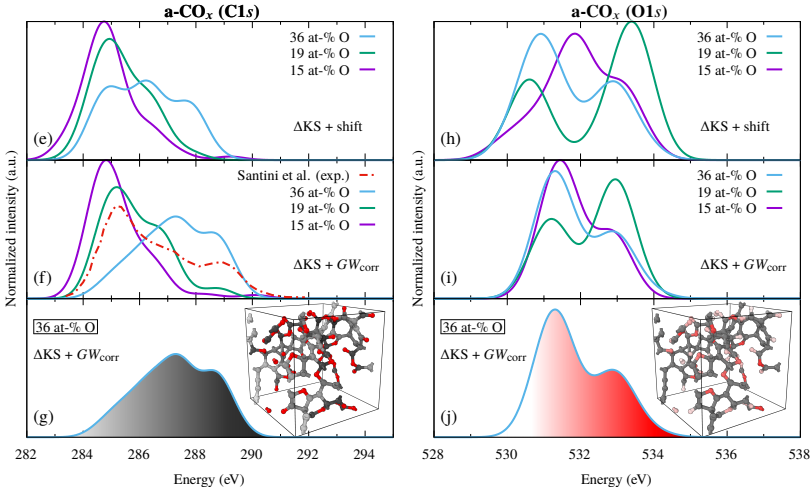


Peaks positions depend on surroundings



XPS prediction

Core-electron binding energy model (to get XPS) with very similar architecture as a GAP MLP



[1] D. Golze, M. Hirvensalo, Hernández-León P., et al., Chem. Mater. 34, 6240 (2022)

How to generate high-O-content samples?

- We have a (cheap) quantitatively accurate method to check whether the a-C:O structure's XPS matches experiment

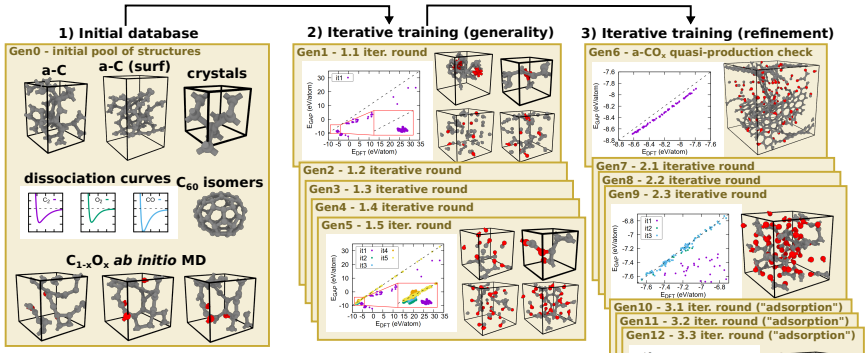
How to generate high-O-content samples?

- We have a (cheap) quantitatively accurate method to check whether the a-C:O structure's XPS matches experiment
- We only have 3 samples painfully generated with DFT: none match experiment

How to generate high-O-content samples?

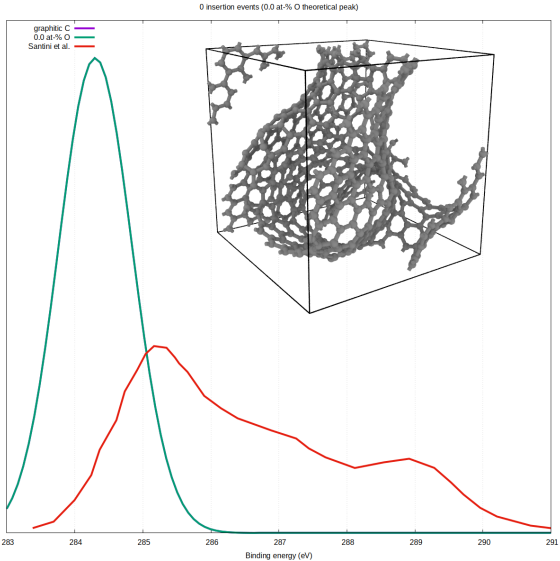
- We have a (cheap) quantitatively accurate method to check whether the a-C:O structure's XPS matches experiment
- We only have 3 samples painfully generated with DFT: none match experiment
- How do we generate more samples? Make a general-purpose **ML potential** for CO!

Making a general-purpose MLP for CO

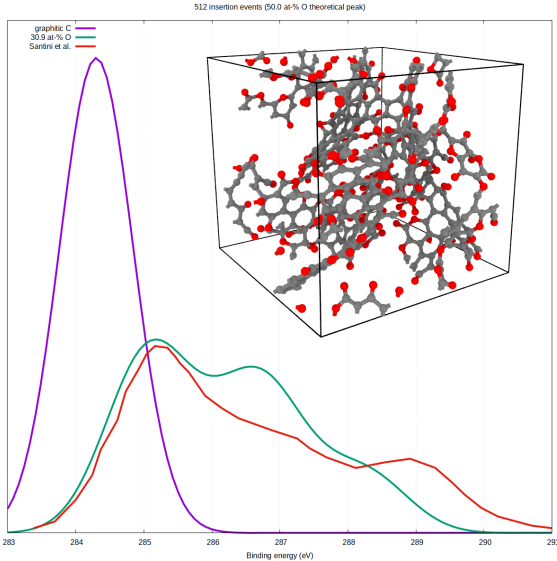


[1] T. Zarrouk, R. Ibragimova, A. P. Bartók, and M. A. Caro, J. Am. Chem. Soc. 146, 14645 (2024)

Cool: using an efficient+accurate ML potential to check the structure



Cool: using an efficient+accurate ML potential to check the structure

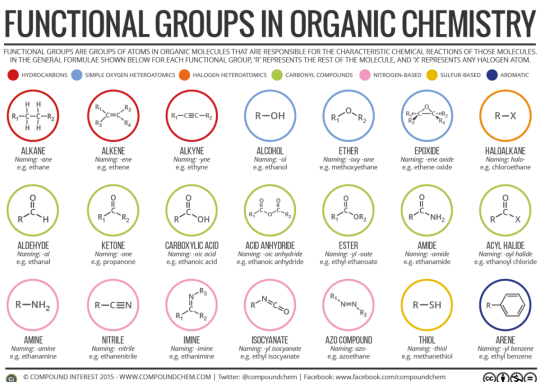


Dang!

- After a lot of simulations the computational and experimental spectra still do not match.

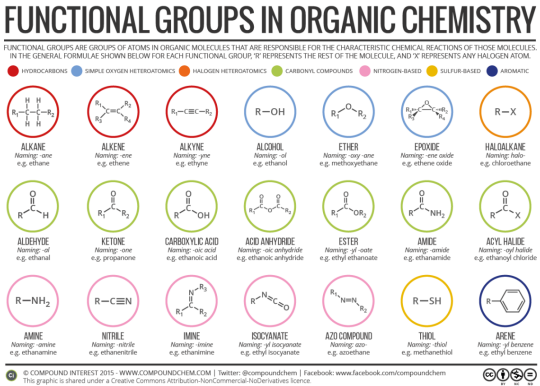
Dang!

- After a lot of simulations the computational and experimental spectra still do not match.
- Missing functional groups at high binding energy are less thermodynamically favorable



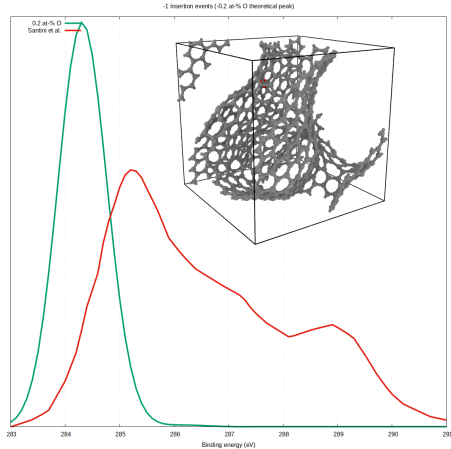
Dang!

- After a lot of simulations the computational and experimental spectra still do not match.
- Missing functional groups at high binding energy are less thermodynamically favorable
- The experimental growth process is non-equilibrium high-energy deposition



Experiment-driven atomistic materials modeling

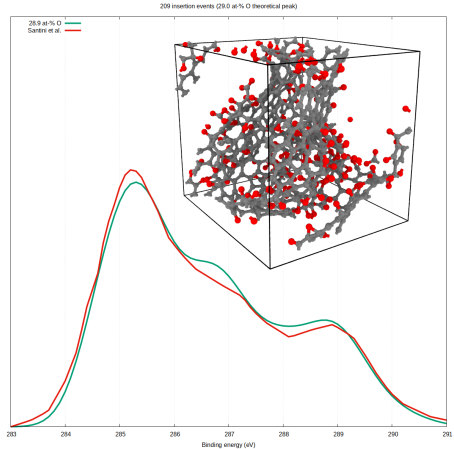
- Physical MLP Hamiltonian is augmented with experimental observables[1]
- Similar to hybrid reverse Monte Carlo but our local-property framework accepts any observable that can be mapped onto an XPS-like problem (from the computational point of view); we've also extended it to variable number of particles
- Other experimental techniques, like XRD, need to be developed and implemented on a case-by-case basis
- Not restricted to one experimental constraint at a time (we can use a linear combination of them!)



[1] T. Zarrouk, R. Ibragimova, A. P. Bartók, and M. A. Caro, J. Am. Chem. Soc. 146, 14645 (2024)

Experiment-driven atomistic materials modeling

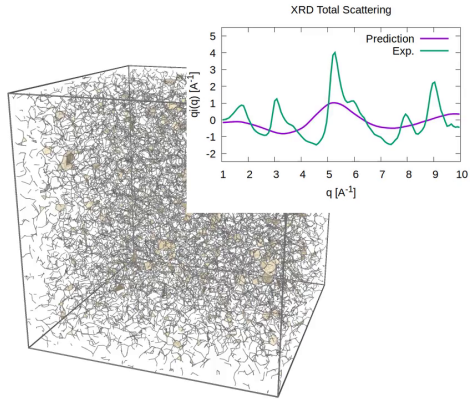
- Physical MLP Hamiltonian is augmented with experimental observables[1]
- Similar to hybrid reverse Monte Carlo but our local-property framework accepts any observable that can be mapped onto an XPS-like problem (from the computational point of view); we've also extended it to variable number of particles
- Other experimental techniques, like XRD, need to be developed and implemented on a case-by-case basis
- Not restricted to one experimental constraint at a time (we can use a linear combination of them!)



[1] T. Zarrouk, R. Ibragimova, A. P. Bartók, and M. A. Caro, J. Am. Chem. Soc. 146, 14645 (2024)

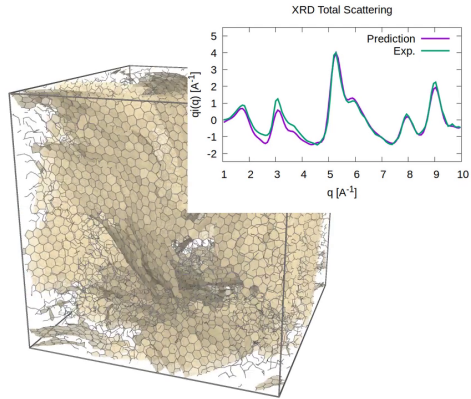
Experiment-driven atomistic materials modeling

- The framework accepts analytical gradients: modified/generalized “forces”
- Molecular Augmented Dynamics (MAD) sampling significantly more efficient than MC, $\mathcal{O}(N)$ vs $\mathcal{O}(N^2)$
- XRD module does not need ML models to work, can be used with any code
- This is a new paradigm in atomistic modeling: atomistic structures that match the experimental observables *by design*

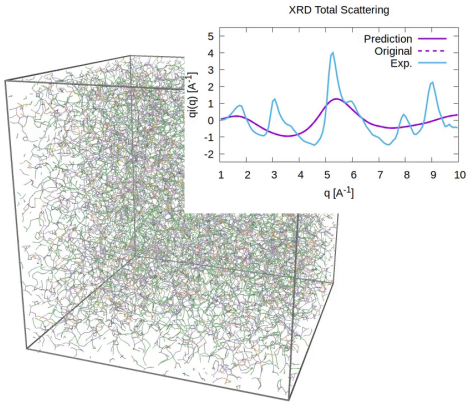


Experiment-driven atomistic materials modeling

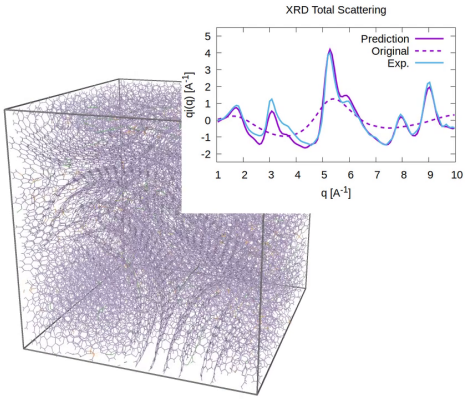
- The framework accepts analytical gradients: modified/generalized “forces”
- Molecular Augmented Dynamics (MAD) sampling significantly more efficient than MC, $\mathcal{O}(N)$ vs $\mathcal{O}(N^2)$
- XRD module does not need ML models to work, can be used with any code
- This is a new paradigm in atomistic modeling: atomistic structures that match the experimental observables *by design*



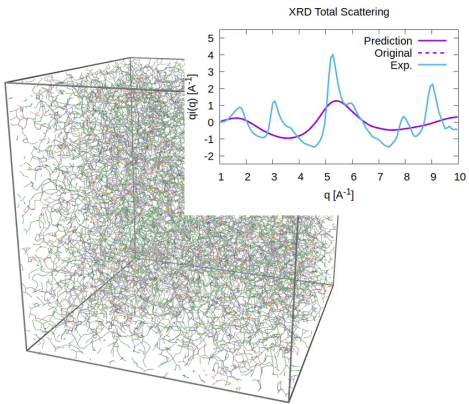
Thermostatting and barostatting!



Thermostatting and barostatting!



Thermostatting and barostatting!





It's MAD!

My path to ML: Trying to understand the structure of amorphous carbon materials

My path to ML: Trying to understand the structure of amorphous carbon materials

- Before 2016, my idea of AI was Skynet from the Terminator movies

My path to ML: Trying to understand the structure of amorphous carbon materials

- Before 2016, my idea of AI was Skynet from the Terminator movies
- I was very skeptical about using ML for my research

My path to ML: Trying to understand the structure of amorphous carbon materials

- Before 2016, my idea of AI was Skynet from the Terminator movies
- I was very skeptical about using ML for my research
- It didn't take a leap of faith: I embraced ML by empirically verifying its advantages

My path to ML: Trying to understand the structure of amorphous carbon materials

- Before 2016, my idea of AI was Skynet from the Terminator movies
- I was very skeptical about using ML for my research
- It didn't take a leap of faith: I embraced ML by empirically verifying its advantages
- More science, better science, bigger and faster

My path to ML: Trying to understand the structure of amorphous carbon materials

- Before 2016, my idea of AI was Skynet from the Terminator movies
- I was very skeptical about using ML for my research
- It didn't take a leap of faith: I embraced ML by empirically verifying its advantages
- More science, better science, bigger and faster
- **Different** science (it's a paradigm shift!)

My path to ML: Trying to understand the structure of amorphous carbon materials

- Before 2016, my idea of AI was Skynet from the Terminator movies
- I was very skeptical about using ML for my research
- It didn't take a leap of faith: I embraced ML by empirically verifying its advantages
- More science, better science, bigger and faster
- **Different** science (it's a paradigm shift!)
- ML enables science that was not *possible* before

What else is happening?

What else is happening?
(a.k.a. “more ignorant every day”)

What is AI?

What is AI?

- AI = “artificial” + “intelligence”

What is AI?

- AI = “artificial” + “intelligence”
- What is “artificial”?

What is AI?

- AI = “artificial” + “intelligence”
- What is “artificial”?
- What is “intelligence”?

What is AI?

- AI = "artificial" + "intelligence"
- What is "artificial"?
- What is "intelligence"?
- A possible definition is "being able to perform tasks without explicit programming"

What is AI?

- AI = "artificial" + "intelligence"
- What is "artificial"?
- What is "intelligence"?
- A possible definition is "being able to perform tasks without explicit programming"
- Example: machine-learning potentials can learn the solution to the Schrödinger equation without solving it

What is AI?

- AI = "artificial" + "intelligence"
- What is "artificial"?
- What is "intelligence"?
- A possible definition is "being able to perform tasks without explicit programming"
- Example: machine-learning potentials can learn the solution to the Schrödinger equation without solving it
- Example 2: football player learns how to score goals without solving the ball's equations of motion

What is AI?



What is AI?

What is AI?

- Humans are extremely efficient at learning from data

What is AI?

- Humans are extremely efficient at learning from data
- ChatGPT trained on tens of trillions of tokens

What is AI?

- Humans are extremely efficient at learning from data
- ChatGPT trained on tens of trillions of tokens
- In a human lifespan, absolute maximum of information we are exposed to: 150 words/minute $\times$ 60 minutes/hour $\times$ 16 wake hours/day $\times$ 365 days/year $\times$ 80 years/lifespan $=\sim$ 4.2 *billion* words

What is AI?

- Humans are extremely efficient at learning from data
- ChatGPT trained on tens of trillions of tokens
- In a human lifespan, absolute maximum of information we are exposed to: 150 words/minute $\times$ 60 minutes/hour $\times$ 16 wake hours/day $\times$ 365 days/year $\times$ 80 years/lifespan $=\sim$ 4.2 *billion* words
- We also need to eat, breathe, avoid danger, etc.!

What is AI?

What is AI?

- AI as a term is imprecise and ambiguous ("intelligence" is imprecise and ambiguous!)

What is AI?

- AI as a term is imprecise and ambiguous ("intelligence" is imprecise and ambiguous!)
- I personally do not like it

What is AI?

- AI as a term is imprecise and ambiguous ("intelligence" is imprecise and ambiguous!)
- I personally do not like it
- It's the generally accepted term and so it's in the name of the School...

What is AI?

- AI as a term is imprecise and ambiguous ("intelligence" is imprecise and ambiguous!)
- I personally do not like it
- It's the generally accepted term and so it's in the name of the School...
- In scientific research "machine learning"

What is AI?

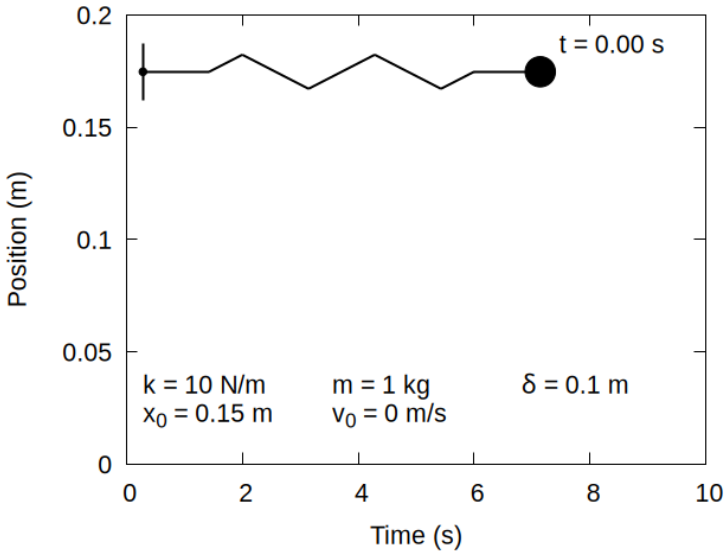
- AI as a term is imprecise and ambiguous ("intelligence" is imprecise and ambiguous!)
- I personally do not like it
- It's the generally accepted term and so it's in the name of the School...
- In scientific research "machine learning"
- ML is a subset of AI where a computer infers relationships between input and output in a training dataset and generalizes outside this dataset

What is AI?

- AI as a term is imprecise and ambiguous ("intelligence" is imprecise and ambiguous!)
- I personally do not like it
- It's the generally accepted term and so it's in the name of the School...
- In scientific research "machine learning"
- ML is a subset of AI where a computer infers relationships between input and output in a training dataset and generalizes outside this dataset
- An ML "**model**" is a set of mathematical objects, algorithms, and data that provides these predictions

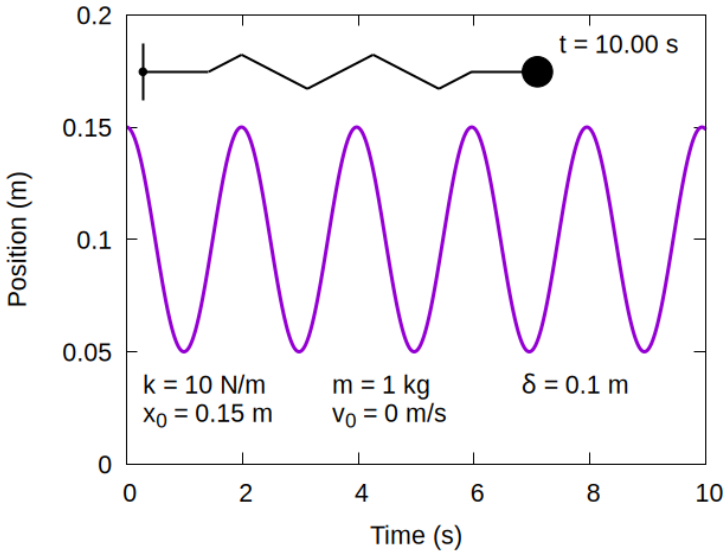
What is a model?

Consider a one-dimensional harmonic oscillator:



What is a model?

Consider a one-dimensional harmonic oscillator:



What is a model?

The one-dimensional harmonic oscillator's motion is governed by Hamilton's equations:

$$H(t) = \frac{1}{2}m\left(v(t)\right)^2 + \frac{1}{2}k\left(x(t) - \delta\right)^2, \tag{1}$$

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial H}{\partial (mv)} = v(t); \qquad \frac{\mathrm{d}(mv)}{\mathrm{d}t} = -\frac{\partial H}{\partial x} = -k\left(x(t) - \delta\right), \tag{2}$$

with analytical solution:

$$x(t) = \delta + (x_0 - \delta)\cos\left(\sqrt{\frac{k}{m}}t + \varepsilon\right), \tag{3}$$

$$\varepsilon = \arctan\left(\frac{v_0}{\delta - x_0}\sqrt{\frac{m}{k}}\right). \tag{4}$$

What is a model?

What is a model?

- The one-dimensional harmonic oscillator is described by a **physical** model

What is a model?

- The one-dimensional harmonic oscillator is described by a **physical** model
- Physical parameters

What is a model?

- The one-dimensional harmonic oscillator is described by a **physical** model
- Physical parameters
- Equations of motion

What is a model?

- The one-dimensional harmonic oscillator is described by a **physical** model
- Physical parameters
- Equations of motion
- With some limitations, we can predict the position x at time t

What is a model?

What is a model?

- How do we describe the one-dimensional harmonic oscillator with an ML model?

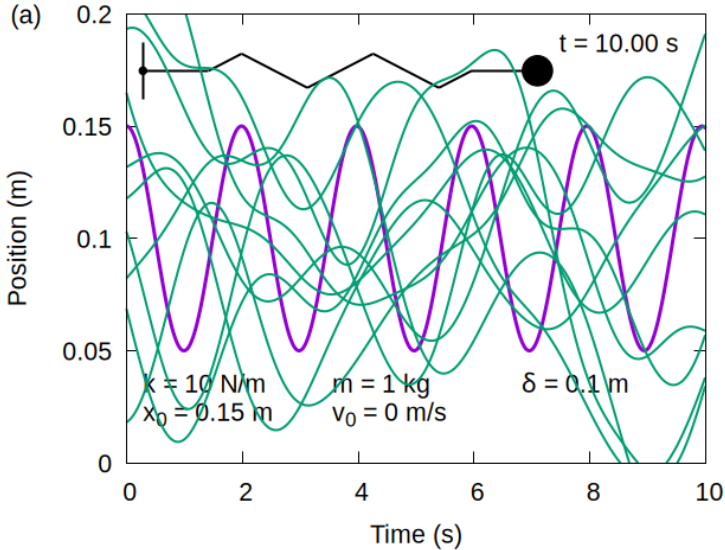
What is a model?

- How do we describe the one-dimensional harmonic oscillator with an ML model?
- Starting point: We don't know physics (or anything about the system)

What is a model?

- How do we describe the one-dimensional harmonic oscillator with an ML model?
- Starting point: We don't know physics (or anything about the system)
- We want our model $x(t)$ to be "smooth"

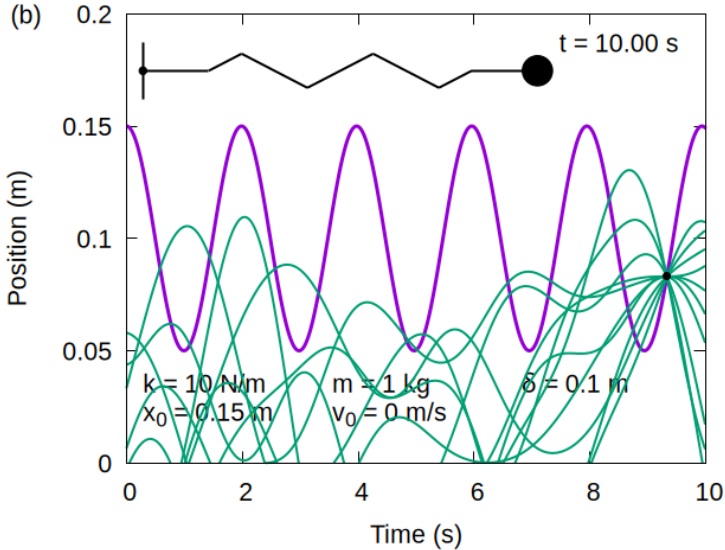
What is a model?



What is a model?

- How do we describe the one-dimensional harmonic oscillator with an ML model?
- Starting point: We don't know physics (or anything about the system)
- We want our model $x(t)$ to be "smooth"
- Instead of physics, let's add **data**

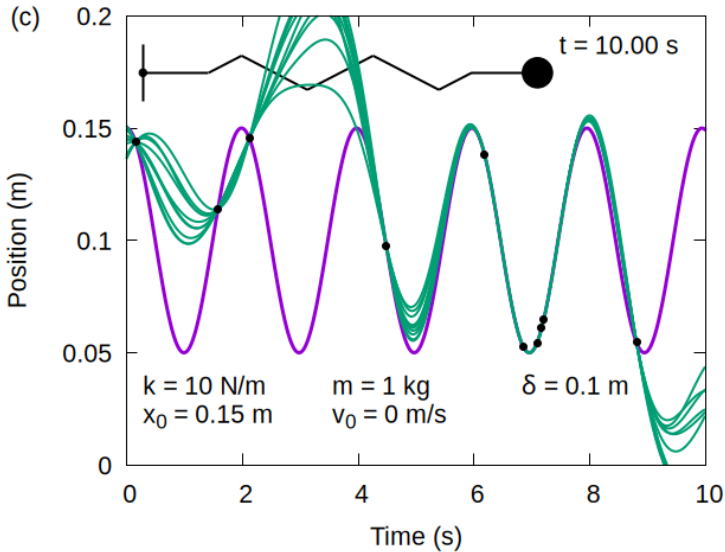
What is a model?



What is a model?

- How do we describe the one-dimensional harmonic oscillator with an ML model?
- Starting point: We don't know physics (or anything about the system)
- We want our model $x(t)$ to be "smooth"
- Instead of physics, let's add **data**
- Let's add **more** data

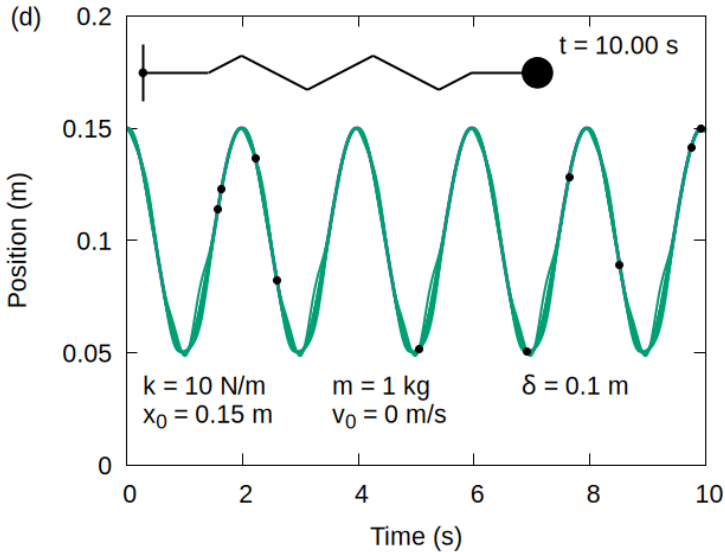
What is a model?



What is a model?

- How do we describe the one-dimensional harmonic oscillator with an ML model?
- Starting point: We don't know physics (or anything about the system)
- We want our model $x(t)$ to be "smooth"
- Instead of physics, let's add **data**
- Let's add **more** data
- Let's add extra information: $T = 2\pi\sqrt{\frac{m}{k}}$

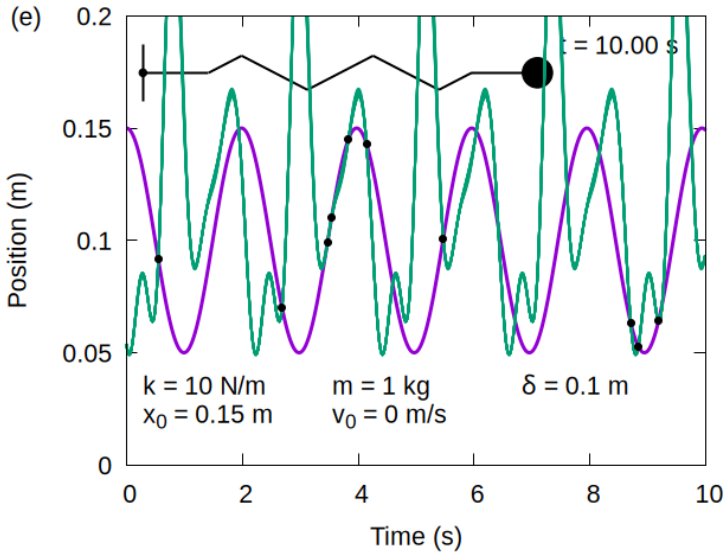
What is a model?



What is a model?

- How do we describe the one-dimensional harmonic oscillator with an ML model?
- Starting point: We don't know physics (or anything about the system)
- We want our model $x(t)$ to be "smooth"
- Instead of physics, let's add **data**
- Let's add **more** data
- Let's add extra information: $T = 2\pi\sqrt{\frac{m}{k}}$
- What if my information is bad? $T = 1.1 \times 2\pi\sqrt{\frac{m}{k}}$

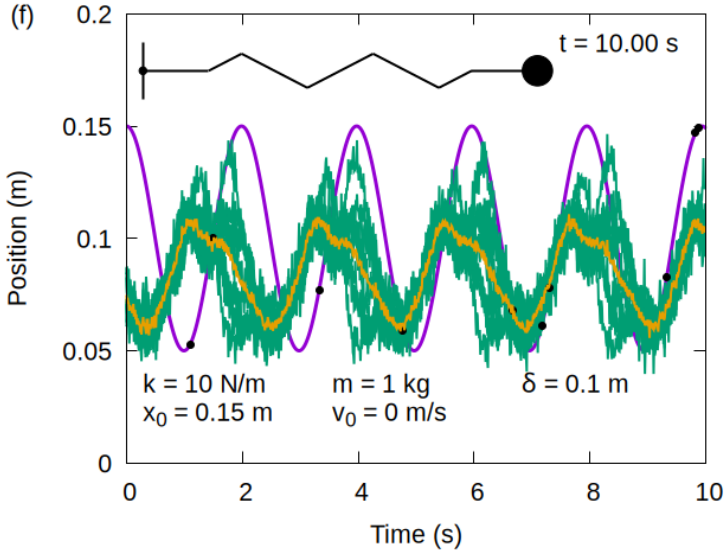
What is a model?



What is a model?

- How do we describe the one-dimensional harmonic oscillator with an ML model?
- Starting point: We don't know physics (or anything about the system)
- We want our model $x(t)$ to be "smooth"
- Instead of physics, let's add **data**
- Let's add **more** data
- Let's add extra information: $T = 2\pi\sqrt{\frac{m}{k}}$
- What if my information is bad? $T = 1.1 \times 2\pi\sqrt{\frac{m}{k}}$
- What if I use some ML tricks (regularization)?

What is a model?



It gets worse before it gets better!

Ingredients to use ML in your research:

Ingredients to use ML in your research:

- Knowing what you want to achieve (seems more obvious than it is!)

Ingredients to use ML in your research:

- Knowing what you want to achieve (seems more obvious than it is!)
- Do you really need ML for this? Can ML really help you with your problem?

Ingredients to use ML in your research:

- Knowing what you want to achieve (seems more obvious than it is!)
- Do you really need ML for this? Can ML really help you with your problem?
- Database

Ingredients to use ML in your research:

- Knowing what you want to achieve (seems more obvious than it is!)
- Do you really need ML for this? Can ML really help you with your problem?
- Database
- An ML algorithm/method ("model architecture")

Ingredients to use ML in your research:

- Knowing what you want to achieve (seems more obvious than it is!)
- Do you really need ML for this? Can ML really help you with your problem?
- Database
- An ML algorithm/method ("model architecture")
- The optimal choice of algorithm depends on what you want to achieve *and* the type/amount of data you have available

Three main types of problems

Three main types of problems

- “Supervised” learning:

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output
  - Example: predicting the temperature in Helsinki (output) as a function of time (input)

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output
  - Example: predicting the temperature in Helsinki (output) as a function of time (input)
- "Unsupervised" learning:

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output
  - Example: predicting the temperature in Helsinki (output) as a function of time (input)
- "Unsupervised" learning:
  - Uses "unlabeled" data: only input is provided

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output
  - Example: predicting the temperature in Helsinki (output) as a function of time (input)
- "Unsupervised" learning:
  - Uses "unlabeled" data: only input is provided
  - Usually understood as a "classification" task: find patterns in the data

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output
  - Example: predicting the temperature in Helsinki (output) as a function of time (input)
- "Unsupervised" learning:
  - Uses "unlabeled" data: only input is provided
  - Usually understood as a "classification" task: find patterns in the data
  - Example: grouping together atomic motifs in a material that are structurally similar to one another

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output
  - Example: predicting the temperature in Helsinki (output) as a function of time (input)
- "Unsupervised" learning:
  - Uses "unlabeled" data: only input is provided
  - Usually understood as a "classification" task: find patterns in the data
  - Example: grouping together atomic motifs in a material that are structurally similar to one another
- "Reinforcement" learning:

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output
  - Example: predicting the temperature in Helsinki (output) as a function of time (input)
- "Unsupervised" learning:
  - Uses "unlabeled" data: only input is provided
  - Usually understood as a "classification" task: find patterns in the data
  - Example: grouping together atomic motifs in a material that are structurally similar to one another
- "Reinforcement" learning:
  - The algorithm performs tasks according to a predefined goal

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output
  - Example: predicting the temperature in Helsinki (output) as a function of time (input)
- "Unsupervised" learning:
  - Uses "unlabeled" data: only input is provided
  - Usually understood as a "classification" task: find patterns in the data
  - Example: grouping together atomic motifs in a material that are structurally similar to one another
- "Reinforcement" learning:
  - The algorithm performs tasks according to a predefined goal
  - Usually understoo as an "optimization" task: find location and value of some maximum/minimum

Three main types of problems

- "Supervised" learning:
  - Uses "labeled" data: both input and output are provided in database
  - Usually understood as a regression task: learn the relationship between input (features/descriptors) and output
  - Example: predicting the temperature in Helsinki (output) as a function of time (input)
- "Unsupervised" learning:
  - Uses "unlabeled" data: only input is provided
  - Usually understood as a "classification" task: find patterns in the data
  - Example: grouping together atomic motifs in a material that are structurally similar to one another
- "Reinforcement" learning:
  - The algorithm performs tasks according to a predefined goal
  - Usually understoo as an "optimization" task: find location and value of some maximum/minimum
  - Example: tuning the composition of a material to optimize a property (e.g., hardness) in a self-driving lab

Three main types of problems?

Three main types of problems?

- Lines separating approaches are blurry

Three main types of problems?

- Lines separating approaches are blurry
- Example: self-supervised learning finds structure in data (unsupervised learning) to generate labels (supervised learning) to train itself; this is used in large language models like ChatGPT

Three main types of problems?

- Lines separating approaches are blurry
- Example: self-supervised learning finds structure in data (unsupervised learning) to generate labels (supervised learning) to train itself; this is used in large language models like ChatGPT
- Example: to optimize a material's composition we can use reinforcement learning to design the data acquisition strategy *and* supervised learning to train an ML surrogate model to make the predictions

Three main types of problems?

- Lines separating approaches are blurry
- Example: self-supervised learning finds structure in data (unsupervised learning) to generate labels (supervised learning) to train itself; this is used in large language models like ChatGPT
- Example: to optimize a material's composition we can use reinforcement learning to design the data acquisition strategy *and* supervised learning to train an ML surrogate model to make the predictions
- On-the-fly data acquisition that optimally improves the model towards achieving a goal is often referred to as "active" learning

Three main types of problems?

- Lines separating approaches are blurry
- Example: self-supervised learning finds structure in data (unsupervised learning) to generate labels (supervised learning) to train itself; this is used in large language models like ChatGPT
- Example: to optimize a material's composition we can use reinforcement learning to design the data acquisition strategy *and* supervised learning to train an ML surrogate model to make the predictions
- On-the-fly data acquisition that optimally improves the model towards achieving a goal is often referred to as "active" learning
- The ML zoo is constantly growing, it's confusing...

Three main types of problems?

- Lines separating approaches are blurry
- Example: self-supervised learning finds structure in data (unsupervised learning) to generate labels (supervised learning) to train itself; this is used in large language models like ChatGPT
- Example: to optimize a material's composition we can use reinforcement learning to design the data acquisition strategy *and* supervised learning to train an ML surrogate model to make the predictions
- On-the-fly data acquisition that optimally improves the model towards achieving a goal is often referred to as "active" learning
- The ML zoo is constantly growing, it's confusing...
- Let's see some examples in chemistry and materials science!

Example 1, databases: The Cambridge Structural Database

Example 1, databases: The Cambridge Structural Database

- Databases are more or less cohesive sets of data

Example 1, databases: The Cambridge Structural Database

- Databases are more or less cohesive sets of data
  - Different sizes

Example 1, databases: The Cambridge Structural Database

- Databases are more or less cohesive sets of data
  - Different sizes
  - Labeled, unlabeled or partialy labeled

Example 1, databases: The Cambridge Structural Database

- Databases are more or less cohesive sets of data
  - Different sizes
  - Labeled, unlabeled or partialy labeled
  - Data can be synthetic, experimental or a combination of both

Example 1, databases: The Cambridge Structural Database

- Databases are more or less cohesive sets of data
  - Different sizes
  - Labeled, unlabeled or partialy labeled
  - Data can be synthetic, experimental or a combination of both

Example 1, databases: The Cambridge Structural Database

- Databases are more or less cohesive sets of data
  - Different sizes
  - Labeled, unlabeled or partialy labeled
  - Data can be synthetic, experimental or a combination of both
- The Cambridge Structural Database was established in 1965 by Olga Kennard

Example 1, databases: The Cambridge Structural Database

- Databases are more or less cohesive sets of data
  - Different sizes
  - Labeled, unlabeled or partialy labeled
  - Data can be synthetic, experimental or a combination of both
- The Cambridge Structural Database was established in 1965 by Olga Kennard
- It contains 1.25 million **experimental** structures of small molecules

Example 1, databases: The Cambridge Structural Database

- Databases are more or less cohesive sets of data
  - Different sizes
  - Labeled, unlabeled or partialy labeled
  - Data can be synthetic, experimental or a combination of both
- The Cambridge Structural Database was established in 1965 by Olga Kennard
- It contains 1.25 million **experimental** structures of small molecules
- Curated, maintained, centralized, with added-value products on top (user interfaces, etc.)

Example 1, databases: The Cambridge Structural Database

- Databases are more or less cohesive sets of data
  - Different sizes
  - Labeled, unlabeled or partialy labeled
  - Data can be synthetic, experimental or a combination of both
- The Cambridge Structural Database was established in 1965 by Olga Kennard
- It contains 1.25 million **experimental** structures of small molecules
- Curated, maintained, centralized, with added-value products on top (user interfaces, etc.)
- Large database projects are important to share data and homogenize formats and quality (easier said than done...)

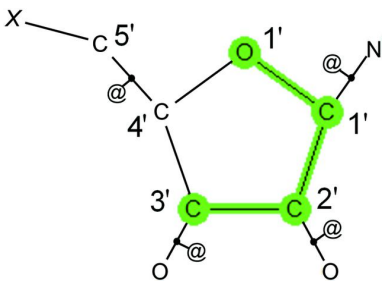
Example 1, databases: The Cambridge Structural Database (Sykes et al.)

Example 1, databases: The Cambridge Structural Database (Sykes et al.)

- Example from the literature: Sykes et al. (2011)

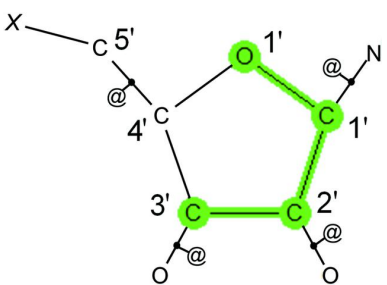
Example 1, databases: The Cambridge Structural Database (Sykes et al.)

- Example from the literature: Sykes et al. (2011)
- Mined structural info of aminofuranoside (!) containing molecules from the CSD



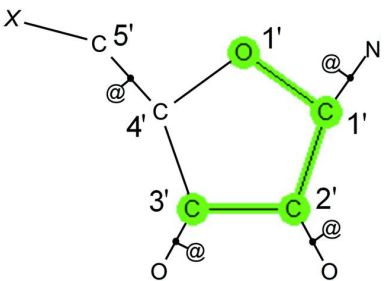
Example 1, databases: The Cambridge Structural Database (Sykes et al.)

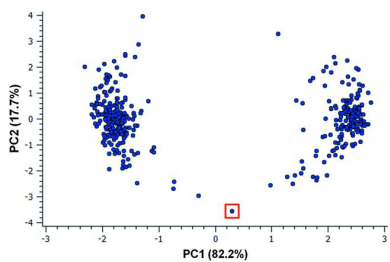
- Example from the literature: Sykes et al. (2011)
- Mined structural info of aminofuranoside (!) containing molecules from the CSD
- Can be characterized by 5 torsional angles (angle between two planes)



Example 1, databases: The Cambridge Structural Database (Sykes et al.)

- Example from the literature: Sykes et al. (2011)
- Mined structural info of aminofuranoside (!) containing molecules from the CSD
- Can be characterized by 5 torsional angles (angle between two planes)
- Performed principal component analysis (PCA) to reduce the dimensionality from 5 to 2 (which explain most of the variance in the data)





Example 1, databases: The Cambridge Structural Database (Sykes et al.)

Example 1, databases: The Cambridge Structural Database (Sykes et al.)

- What did we see in this example?

Example 1, databases: The Cambridge Structural Database (Sykes et al.)

- What did we see in this example?
  - A large database

Example 1, databases: The Cambridge Structural Database (Sykes et al.)

- What did we see in this example?
  - A large database
  - "Data mining": select only those molecules with aminofuranoside rings (very small subset of the whole database)

Example 1, databases: The Cambridge Structural Database (Sykes et al.)

- What did we see in this example?
  - A large database
  - "Data mining": select only those molecules with aminofuranoside rings (very small subset of the whole database)
  - A "dimensionality-reduction" or "low-dimensional embedding" technique, which performs a classification task on unlabeled data: inputs/descriptors (torsional angles) but the output/labels are not provided. The PCA algorithms works out the output (PC1 and PC2) based on the distribution of inputs

Example 1, databases: The Cambridge Structural Database (Sykes et al.)

- What did we see in this example?
  - A large database
  - "Data mining": select only those molecules with aminofuranoside rings (very small subset of the whole database)
  - A "dimensionality-reduction" or "low-dimensional embedding" technique, which performs a classification task on unlabeled data: inputs/descriptors (torsional angles) but the output/labels are not provided. The PCA algorithms works out the output (PC1 and PC2) based on the distribution of inputs
- Data representation (dimensionality reduction and embedding) is a rich field you can learn about by googling around (and it's easy to run examples on your own data!)

Example 2, active learning: Iterative training in MLPs

Example 2, active learning: Iterative training in MLPs

- What to do when you do not know which data you need (or is most useful) to train your model?

Example 2, active learning: Iterative training in MLPs

- What to do when you do not know which data you need (or is most useful) to train your model?
- In ML potentials (MLPs), DFT data is used; expensive to generate and the configuration space is huge (all the ways how the atoms can arrange themselves)

Example 2, active learning: Iterative training in MLPs

- What to do when you do not know which data you need (or is most useful) to train your model?
- In ML potentials (MLPs), DFT data is used; expensive to generate and the configuration space is huge (all the ways how the atoms can arrange themselves)
- We need an efficient way to obtain relevant inputs (structures) to compute synthetic data (outputs/labels/energies and forces in MLPs)

Example 2, active learning: Iterative training in MLPs

- What to do when you do not know which data you need (or is most useful) to train your model?
- In ML potentials (MLPs), DFT data is used; expensive to generate and the configuration space is huge (all the ways how the atoms can arrange themselves)
- We need an efficient way to obtain relevant inputs (structures) to compute synthetic data (outputs/labels/energies and forces in MLPs)
- (we will take a closer look at MLPs at a later session)

Example 2, active learning: Iterative training in MLPs

- What to do when you do not know which data you need (or is most useful) to train your model?
- In ML potentials (MLPs), DFT data is used; expensive to generate and the configuration space is huge (all the ways how the atoms can arrange themselves)
- We need an efficient way to obtain relevant inputs (structures) to compute synthetic data (outputs/labels/energies and forces in MLPs)
- (we will take a closer look at MLPs at a later session)
- Start with a low-quality interim MLP

Example 2, active learning: Iterative training in MLPs

- What to do when you do not know which data you need (or is most useful) to train your model?
- In ML potentials (MLPs), DFT data is used; expensive to generate and the configuration space is huge (all the ways how the atoms can arrange themselves)
- We need an efficient way to obtain relevant inputs (structures) to compute synthetic data (outputs/labels/energies and forces in MLPs)
- (we will take a closer look at MLPs at a later session)
- Start with a low-quality interim MLP
- Generate structures with your disposable MLP

Example 2, active learning: Iterative training in MLPs

- What to do when you do not know which data you need (or is most useful) to train your model?
- In ML potentials (MLPs), DFT data is used; expensive to generate and the configuration space is huge (all the ways how the atoms can arrange themselves)
- We need an efficient way to obtain relevant inputs (structures) to compute synthetic data (outputs/labels/energies and forces in MLPs)
- (we will take a closer look at MLPs at a later session)
- Start with a low-quality interim MLP
- Generate structures with your disposable MLP
- Collect snapshots (possibly along the whole path) for DFT calculations

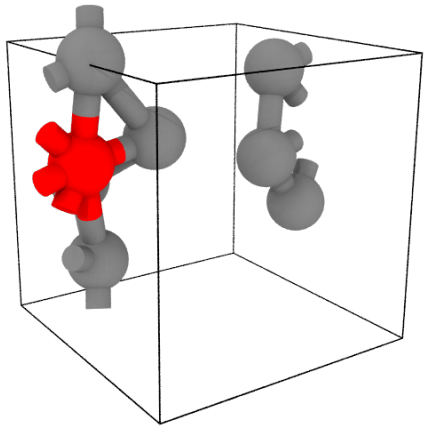
Example 2, active learning: Iterative training in MLPs

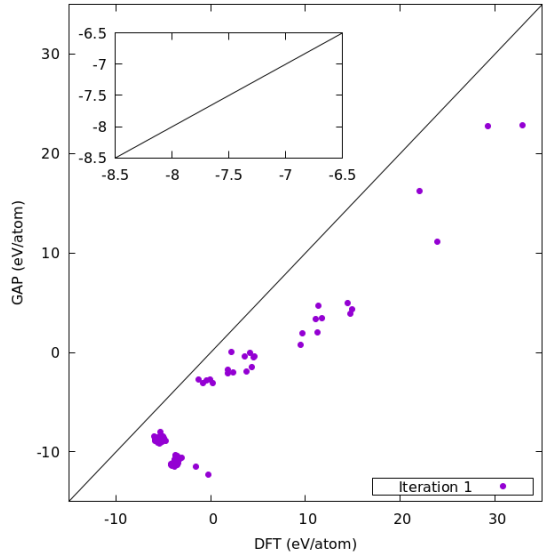
- What to do when you do not know which data you need (or is most useful) to train your model?
- In ML potentials (MLPs), DFT data is used; expensive to generate and the configuration space is huge (all the ways how the atoms can arrange themselves)
- We need an efficient way to obtain relevant inputs (structures) to compute synthetic data (outputs/labels/energies and forces in MLPs)
- (we will take a closer look at MLPs at a later session)
- Start with a low-quality interim MLP
- Generate structures with your disposable MLP
- Collect snapshots (possibly along the whole path) for DFT calculations
- Add to the training set and retrain the MLP

Example 2, active learning: Iterative training in MLPs

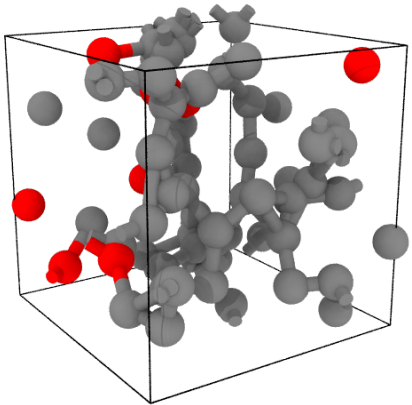
- What to do when you do not know which data you need (or is most useful) to train your model?
- In ML potentials (MLPs), DFT data is used; expensive to generate and the configuration space is huge (all the ways how the atoms can arrange themselves)
- We need an efficient way to obtain relevant inputs (structures) to compute synthetic data (outputs/labels/energies and forces in MLPs)
- (we will take a closer look at MLPs at a later session)
- Start with a low-quality interim MLP
- Generate structures with your disposable MLP
- Collect snapshots (possibly along the whole path) for DFT calculations
- Add to the training set and retrain the MLP
- Repeat until desired accuracy is reached

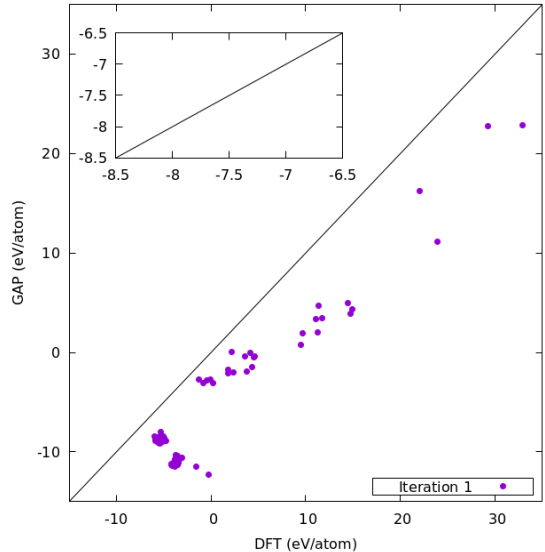
Example 2, active learning: Iterative training in MLPs



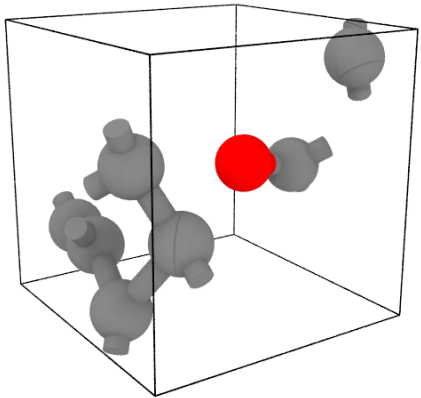


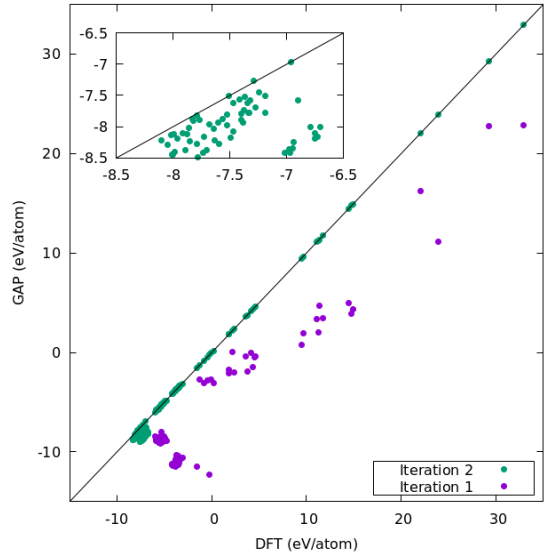
Example 2, active learning: Iterative training in MLPs



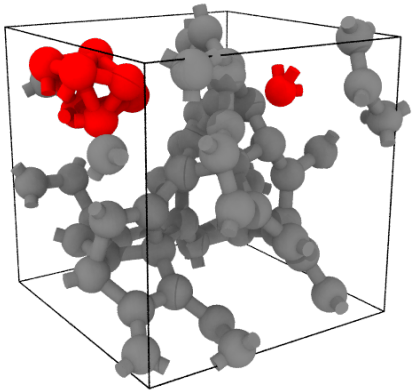


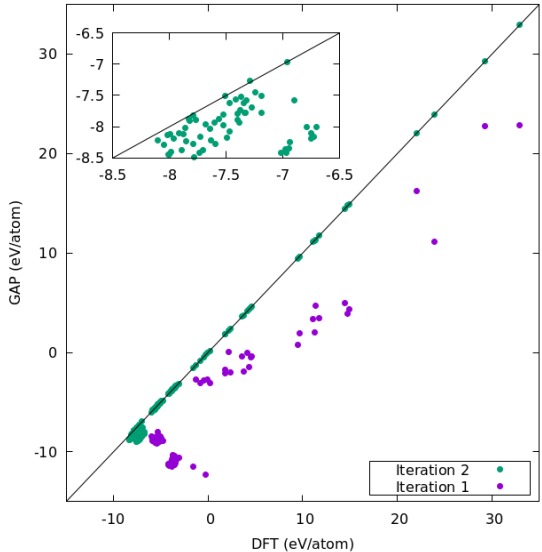
Example 2, active learning: Iterative training in MLPs



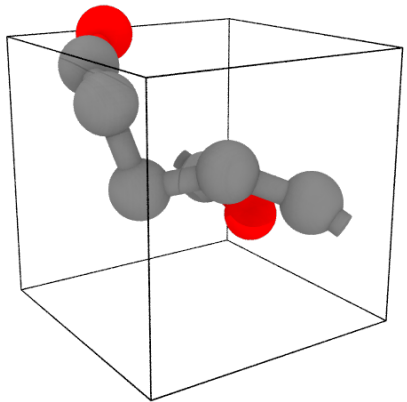


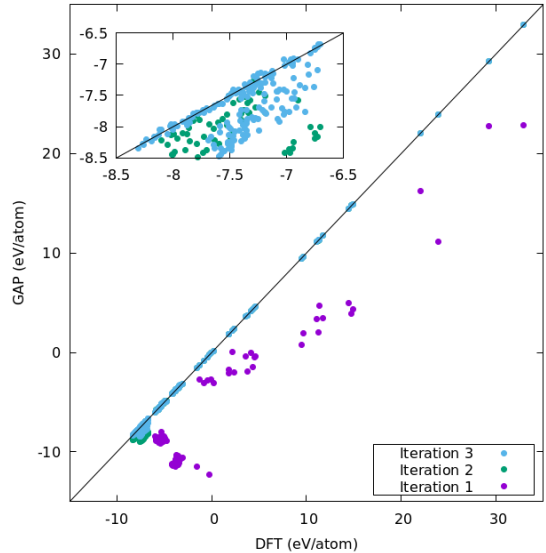
Example 2, active learning: Iterative training in MLPs



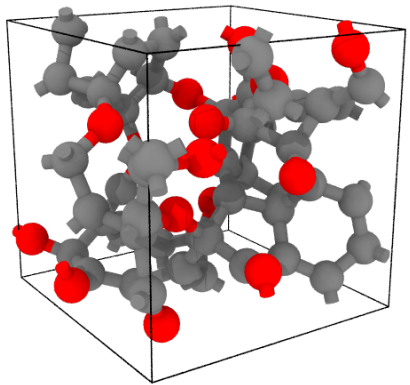


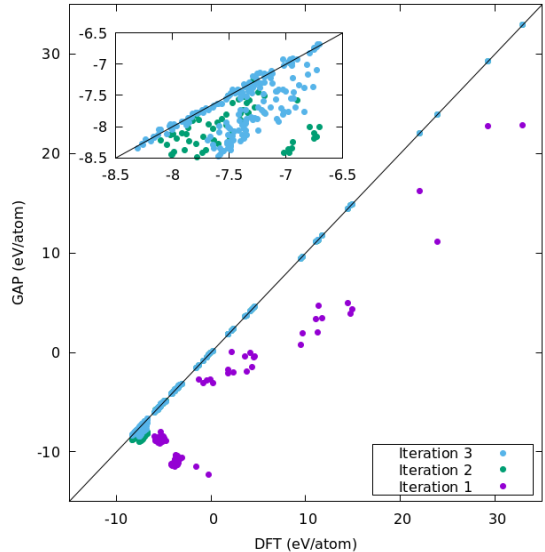
Example 2, active learning: Iterative training in MLPs



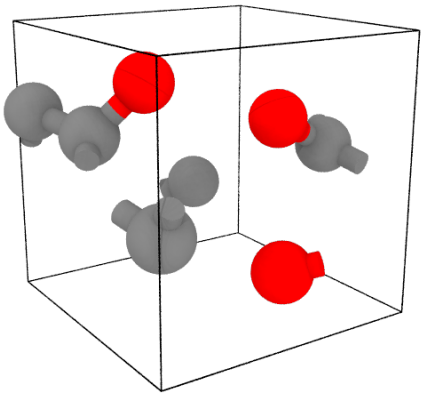


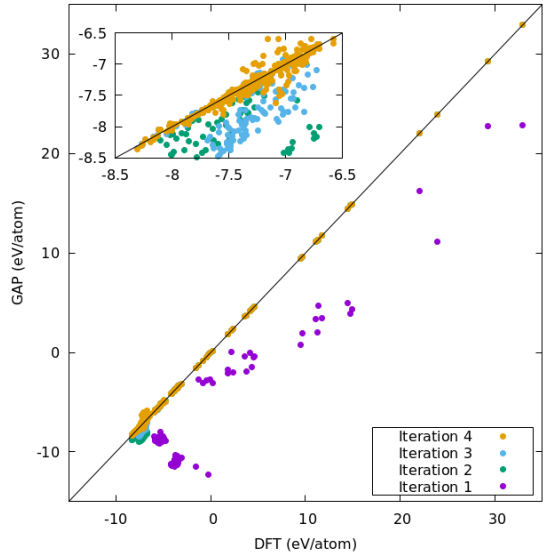
Example 2, active learning: Iterative training in MLPs



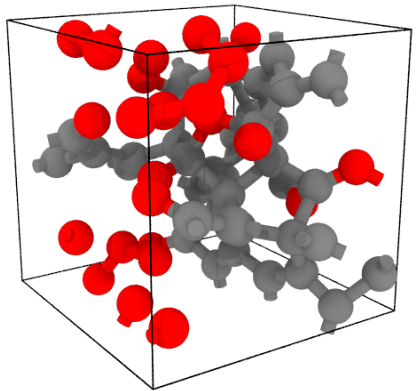


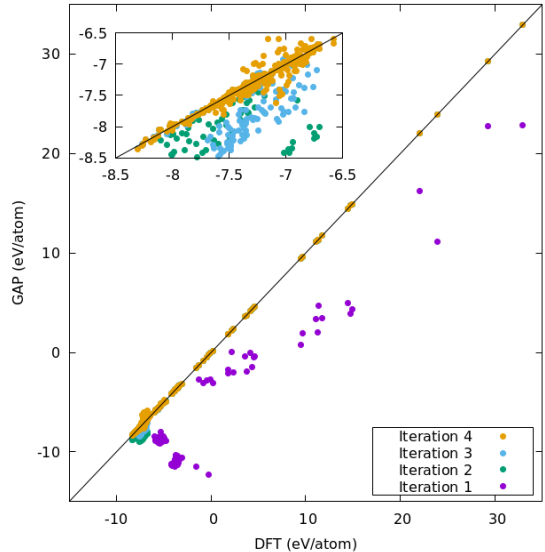
Example 2, active learning: Iterative training in MLPs



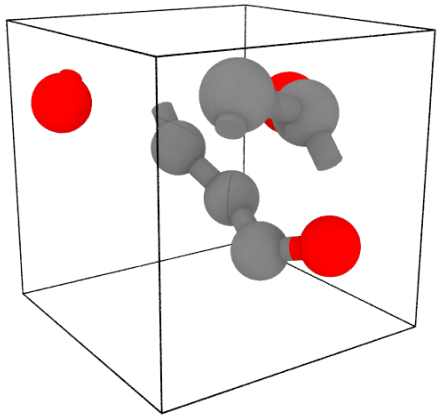


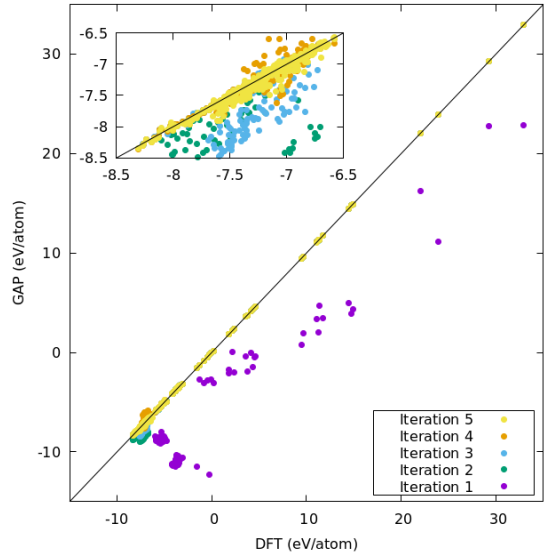
Example 2, active learning: Iterative training in MLPs



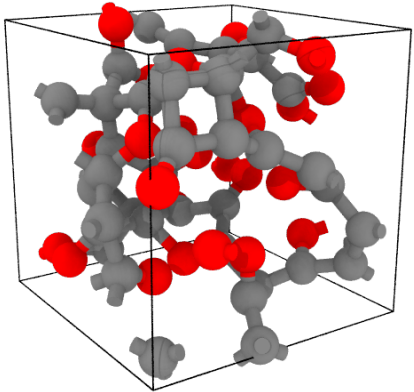


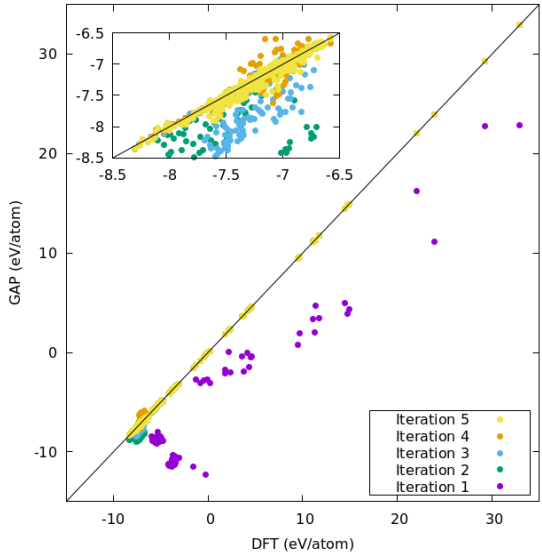
Example 2, active learning: Iterative training in MLPs



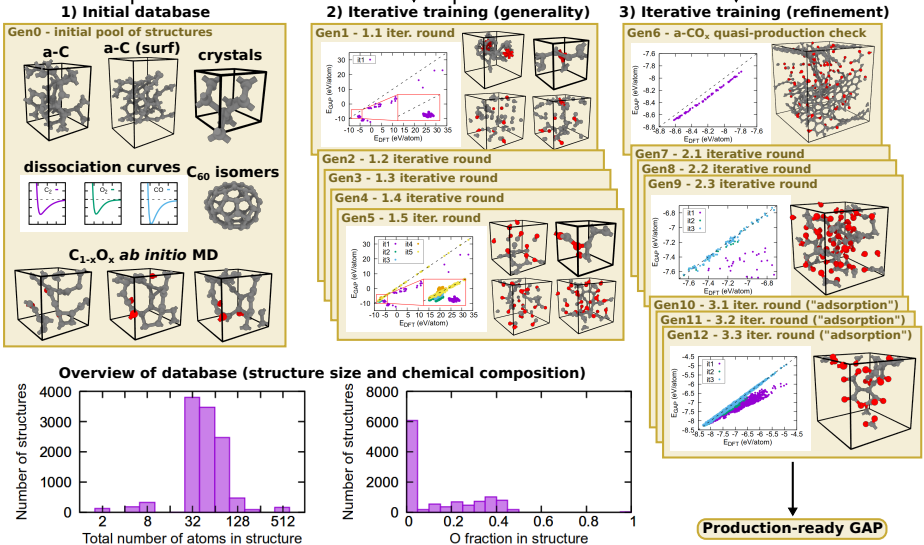


Example 2, active learning: Iterative training in MLPs





Example 2, active learning: Iterative training in MLPs



Example 3, surrogate model optimization: Finding new hydrocarbons

Example 3, surrogate model optimization: Finding new hydrocarbons

- Generative AI finds patterns in the training data and allows to generate similar patterns

Example 3, surrogate model optimization: Finding new hydrocarbons

- Generative AI finds patterns in the training data and allows to generate similar patterns
- This usually follow a fragment-based approach, e.g., put together with higher likelihood functional groups that tend to be nearby in the training data

Example 3, surrogate model optimization: Finding new hydrocarbons

- Generative AI finds patterns in the training data and allows to generate similar patterns
- This usually follow a fragment-based approach, e.g., put together with higher likelihood functional groups that tend to be nearby in the training data
- Another way to generate new molecules and materials is to replace expensive calculations (or even experiments) with surrogate models, for instance energy as a function of atomic coordinates $E(\{\mathbf{r}_i\})$

Example 3, surrogate model optimization: Finding new hydrocarbons

- Generative AI finds patterns in the training data and allows to generate similar patterns
- This usually follow a fragment-based approach, e.g., put together with higher likelihood functional groups that tend to be nearby in the training data
- Another way to generate new molecules and materials is to replace expensive calculations (or even experiments) with surrogate models, for instance energy as a function of atomic coordinates $E(\{\mathbf{r}_i\})$
- If we accelerate the calculations of the property to be optimized by a factor of n, then the optimization with the surrogate model can proceed n times faster too

Example 3, surrogate model optimization: Finding new hydrocarbons

- Generative AI finds patterns in the training data and allows to generate similar patterns
- This usually follow a fragment-based approach, e.g., put together with higher likelihood functional groups that tend to be nearby in the training data
- Another way to generate new molecules and materials is to replace expensive calculations (or even experiments) with surrogate models, for instance energy as a function of atomic coordinates $E(\{\mathbf{r}_i\})$
- If we accelerate the calculations of the property to be optimized by a factor of n, then the optimization with the surrogate model can proceed n times faster too
- An example is the search for new hydrocarbons using molecular dynamics that I have done in my group in collaboration with Rina Ibragimova (2025)

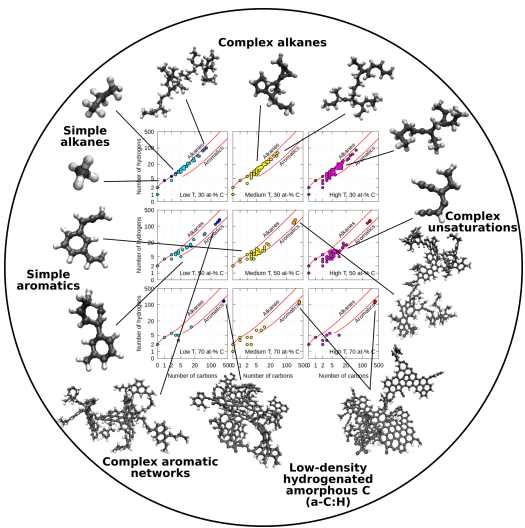
Example 3, surrogate model optimization: Finding new hydrocarbons

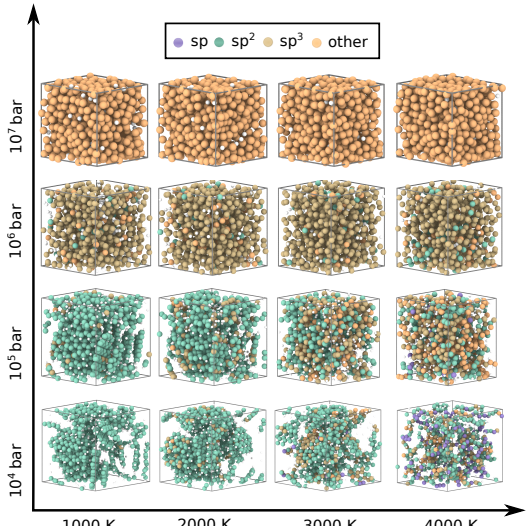
- Generative AI finds patterns in the training data and allows to generate similar patterns
- This usually follow a fragment-based approach, e.g., put together with higher likelihood functional groups that tend to be nearby in the training data
- Another way to generate new molecules and materials is to replace expensive calculations (or even experiments) with surrogate models, for instance energy as a function of atomic coordinates $E(\{\mathbf{r}_i\})$
- If we accelerate the calculations of the property to be optimized by a factor of n, then the optimization with the surrogate model can proceed n times faster too
- An example is the search for new hydrocarbons using molecular dynamics that I have done in my group in collaboration with Rina Ibragimova (2025)
- In this case the speed up of the surrogate model is $n \sim 10^4$ or more (it improves with system size)

Example 3, surrogate model optimization: Finding new hydrocarbons

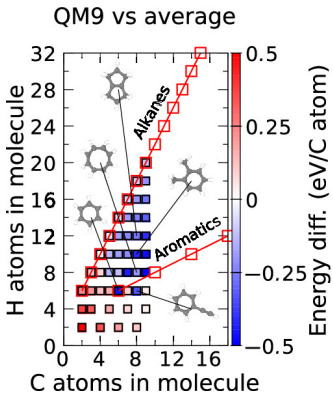
- Generative AI finds patterns in the training data and allows to generate similar patterns
- This usually follow a fragment-based approach, e.g., put together with higher likelihood functional groups that tend to be nearby in the training data
- Another way to generate new molecules and materials is to replace expensive calculations (or even experiments) with surrogate models, for instance energy as a function of atomic coordinates $E(\{\mathbf{r}_i\})$
- If we accelerate the calculations of the property to be optimized by a factor of n, then the optimization with the surrogate model can proceed n times faster too
- An example is the search for new hydrocarbons using molecular dynamics that I have done in my group in collaboration with Rina Ibragimova (2025)
- In this case the speed up of the surrogate model is $n \sim 10^4$ or more (it improves with system size)
- We can do very extreme simulations (high T, high P) and explore the configuration and chemical space

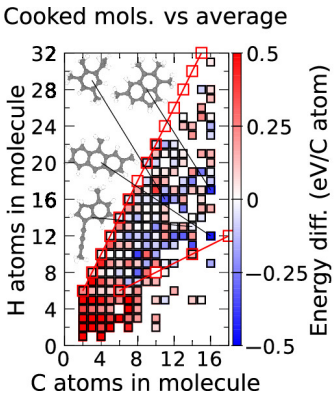
Example 3, surrogate model optimization: Finding new hydrocarbons

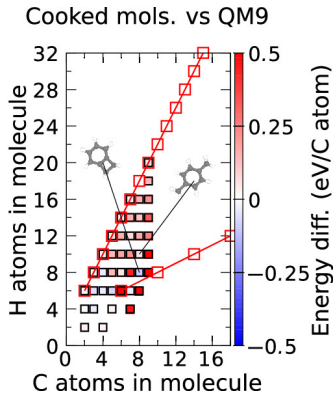




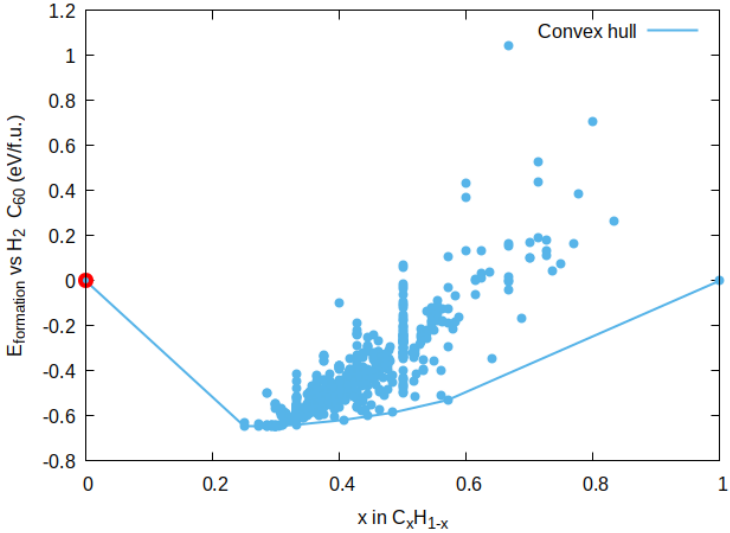
Example 3, surrogate model optimization: Finding new hydrocarbons



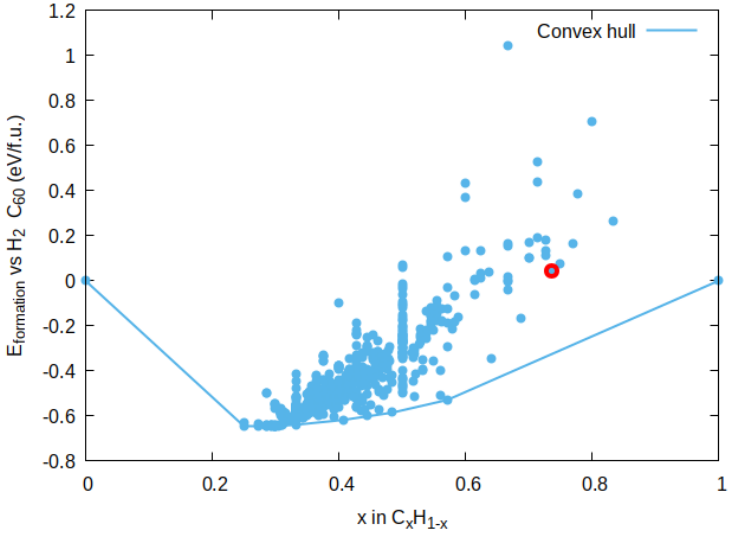




Example 3, surrogate model optimization: Finding new hydrocarbons

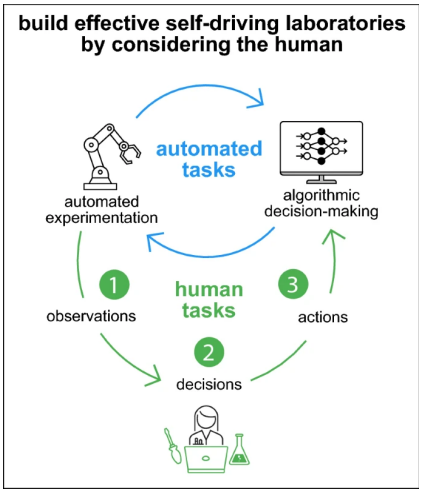


Example 3, surrogate model optimization: Finding new hydrocarbons



Final example: Self-driving labs (I)

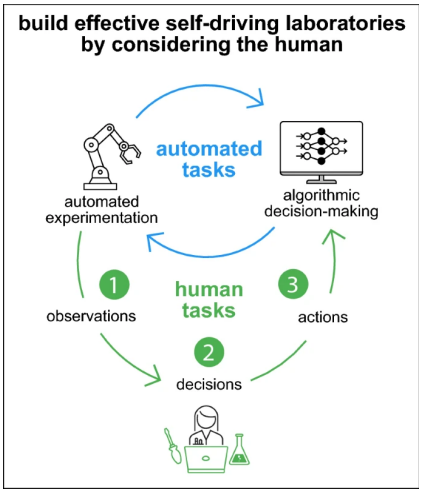
Final example: Self-driving labs (I)



B.P. MacLeod, F.G.L. Parlane, C.P. Berlinguette. MRS Bullet. 48, 173 (2023)

- Couples the machine learning model to the experimental data acquisition in "real time" (on demand)

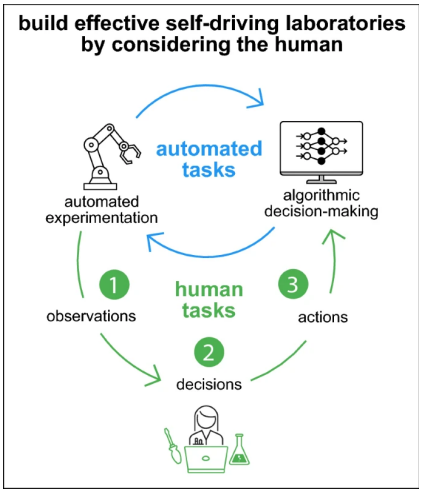
Final example: Self-driving labs (I)



B.P. MacLeod, F.G.L. Parlane, C.P. Berlinguette. MRS Bullet. 48, 173 (2023)

- Couples the machine learning model to the experimental data acquisition in "real time" (on demand)
- Pros:

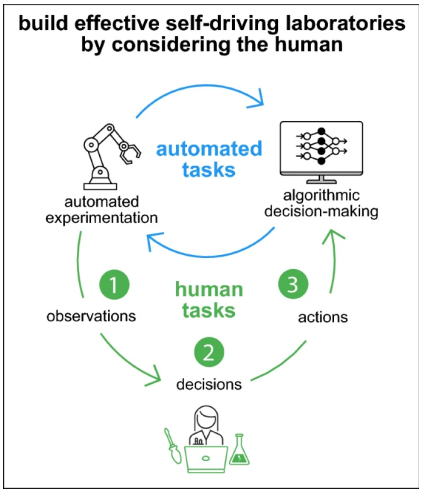
Final example: Self-driving labs (I)



B.P. MacLeod, F.G.L. Parlane, C.P. Berlinguette. MRS Bullet. 48, 173 (2023)

- Couples the machine learning model to the experimental data acquisition in "real time" (on demand)
- Pros:
  - Full integration of experiment and simulation ("on the fly" validation)

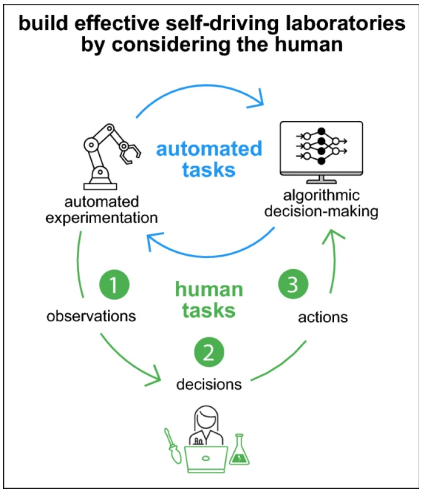
Final example: Self-driving labs (I)



B.P. MacLeod, F.G.L. Parlane, C.P. Berlinguette. MRS Bullet. 48, 173 (2023)

- Couples the machine learning model to the experimental data acquisition in "real time" (on demand)
- Pros:
  - Full integration of experiment and simulation ("on the fly" validation)
  - High-throughput experimentation: fast turnaround and discovery

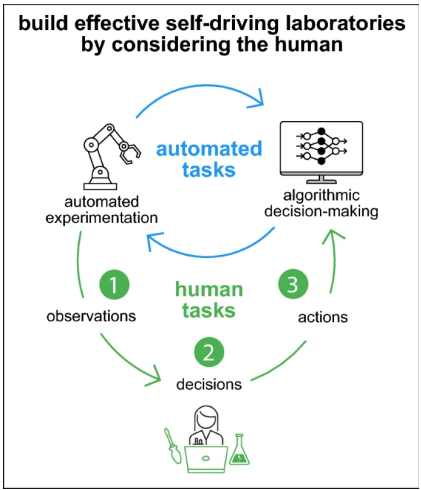
Final example: Self-driving labs (I)



B.P. MacLeod, F.G.L. Parlane, C.P. Berlinguette. MRS Bullet. 48, 173 (2023)

- Couples the machine learning model to the experimental data acquisition in "real time" (on demand)
- Pros:
  - Full integration of experiment and simulation ("on the fly" validation)
  - High-throughput experimentation: fast turnaround and discovery
- Cons:

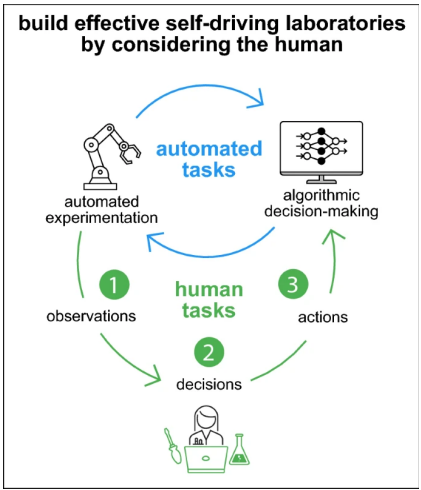
Final example: Self-driving labs (I)



B.P. MacLeod, F.G.L. Parlane, C.P. Berlinguette. MRS Bullet. 48, 173 (2023)

- Couples the machine learning model to the experimental data acquisition in "real time" (on demand)
- Pros:
  - Full integration of experiment and simulation ("on the fly" validation)
  - High-throughput experimentation: fast turnaround and discovery
- Cons:
  - High multidisciplinarity increases risks and cost (e.g., robotics might be involved)

Final example: Self-driving labs (I)



B.P. MacLeod, F.G.L. Parlane, C.P. Berlinguette. MRS Bullet. 48, 173 (2023)

- Couples the machine learning model to the experimental data acquisition in "real time" (on demand)
- Pros:
  - Full integration of experiment and simulation ("on the fly" validation)
  - High-throughput experimentation: fast turnaround and discovery
- Cons:
  - High multidisciplinarity increases risks and cost (e.g., robotics might be involved)
  - A custom automated setup will usually be limited to relatively simple experiments (for now!)

Self-driving labs basics (II)

Self-driving labs basics (II)

- Also known as “autononoums labs”

Self-driving labs basics (II)

- Also known as “autononoums labs”
- The key idea is the ability of the SDL to make *autonomous decisions* towards achieving a goal

Self-driving labs basics (II)

- Also known as “autononoums labs”
- The key idea is the ability of the SDL to make *autonomous decisions* towards achieving a goal
- The whole SDL pipeline is part of a *reinforcement learning* problem

Self-driving labs basics (II)

- Also known as “autononoums labs”
- The key idea is the ability of the SDL to make *autonomous decisions* towards achieving a goal
- The whole SDL pipeline is part of a *reinforcement learning* problem
  - A human states the *goal* of the experiment

Self-driving labs basics (II)

- Also known as "autononoums labs"
- The key idea is the ability of the SDL to make *autonomous decisions* towards achieving a goal
- The whole SDL pipeline is part of a *reinforcement learning* problem
  - A human states the *goal* of the experiment
  - The goal is often formulated in terms of maximizing or minimizing some objective/loss function (or figure of merit)

Self-driving labs basics (II)

- Also known as "autononoums labs"
- The key idea is the ability of the SDL to make *autonomous decisions* towards achieving a goal
- The whole SDL pipeline is part of a *reinforcement learning* problem
  - A human states the *goal* of the experiment
  - The goal is often formulated in terms of maximizing or minimizing some objective/loss function (or figure of merit)
  - The SDL has a recipe for how to autonomously collect data, e.g., automated synthesis of a compound plus analysis of the target properties

Self-driving labs basics (II)

- Also known as "autononoums labs"
- The key idea is the ability of the SDL to make *autonomous decisions* towards achieving a goal
- The whole SDL pipeline is part of a *reinforcement learning* problem
  - A human states the *goal* of the experiment
  - The goal is often formulated in terms of maximizing or minimizing some objective/loss function (or figure of merit)
  - The SDL has a recipe for how to autonomously collect data, e.g., automated synthesis of a compound plus analysis of the target properties
  - An ML model is trained with the available data that predicts the property of interest and gives error estimates

Self-driving labs basics (II)

- Also known as "autononoums labs"
- The key idea is the ability of the SDL to make *autonomous decisions* towards achieving a goal
- The whole SDL pipeline is part of a *reinforcement learning* problem
  - A human states the *goal* of the experiment
  - The goal is often formulated in terms of maximizing or minimizing some objective/loss function (or figure of merit)
  - The SDL has a recipe for how to autonomously collect data, e.g., automated synthesis of a compound plus analysis of the target properties
  - An ML model is trained with the available data that predicts the property of interest and gives error estimates
  - Bayesian optimization is used to select the new data points according to some (user-defined) acquisition function; the acquisition function establishes the tradeoff between exploitation (verify model predictions of optima) and exploration (improve model accuracy)

Self-driving labs basics (II)

- Also known as "autononoums labs"
- The key idea is the ability of the SDL to make *autonomous decisions* towards achieving a goal
- The whole SDL pipeline is part of a *reinforcement learning* problem
  - A human states the *goal* of the experiment
  - The goal is often formulated in terms of maximizing or minimizing some objective/loss function (or figure of merit)
  - The SDL has a recipe for how to autonomously collect data, e.g., automated synthesis of a compound plus analysis of the target properties
  - An ML model is trained with the available data that predicts the property of interest and gives error estimates
  - Bayesian optimization is used to select the new data points according to some (user-defined) acquisition function; the acquisition function establishes the tradeoff between exploitation (verify model predictions of optima) and exploration (improve model accuracy)
  - Continue until we achieve the objective or run out of time/funds

Self-driving labs basics (II)

- Also known as "autononoums labs"
- The key idea is the ability of the SDL to make *autonomous decisions* towards achieving a goal
- The whole SDL pipeline is part of a *reinforcement learning* problem
  - A human states the *goal* of the experiment
  - The goal is often formulated in terms of maximizing or minimizing some objective/loss function (or figure of merit)
  - The SDL has a recipe for how to autonomously collect data, e.g., automated synthesis of a compound plus analysis of the target properties
  - An ML model is trained with the available data that predicts the property of interest and gives error estimates
  - Bayesian optimization is used to select the new data points according to some (user-defined) acquisition function; the acquisition function establishes the tradeoff between exploitation (verify model predictions of optima) and exploration (improve model accuracy)
  - Continue until we achieve the objective or run out of time/funds
- In practice, not all steps are always automated

Self-driving labs basics (III)

Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)

Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out

Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation

Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation
- The expertise requirements are highly multidisciplinary (this can be a barrier to adoption):

Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation
- The expertise requirements are highly multidisciplinary (this can be a barrier to adoption):
  - You need to know chemistry

Self-driving labs basics (III)



Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation
- The expertise requirements are highly multidisciplinary (this can be a barrier to adoption):
  - You need to know chemistry
  - You need to know programming

Self-driving labs basics (III)

```
M107
;TYPE:Custom
G90 ; use absolute coordinates
M83 ; extruder relative mode
M104 S215 ; set extruder temp
M140 S60 ; set bed temp
G32 ; true bed leveling

M190 S60 ; wait for bed temp
M109 S215 ; wait for extruder

G29 ; mesh bed leveling
G1 X10 Y-3.0 Z0.5 F6000.0
G92 E0.0
```

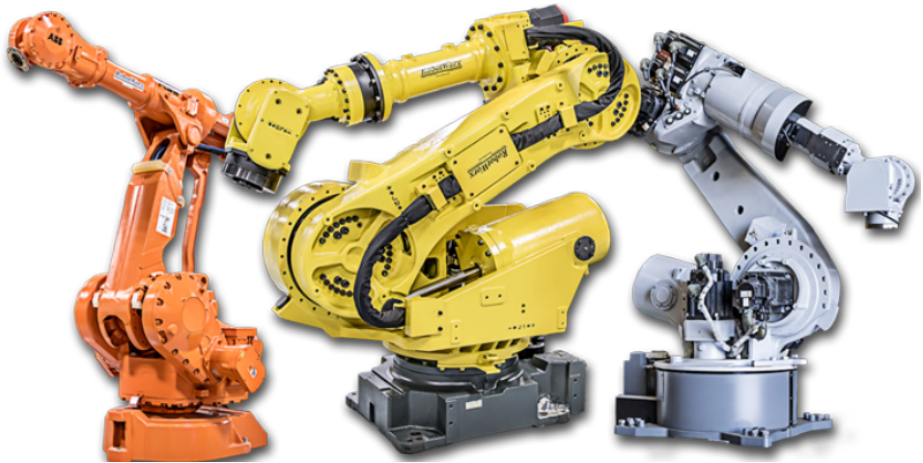
```
ok
ok
ok
```



Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation
- The expertise requirements are highly multidisciplinary (this can be a barrier to adoption):
  - You need to know chemistry
  - You need to know programming
  - You need to know robotics

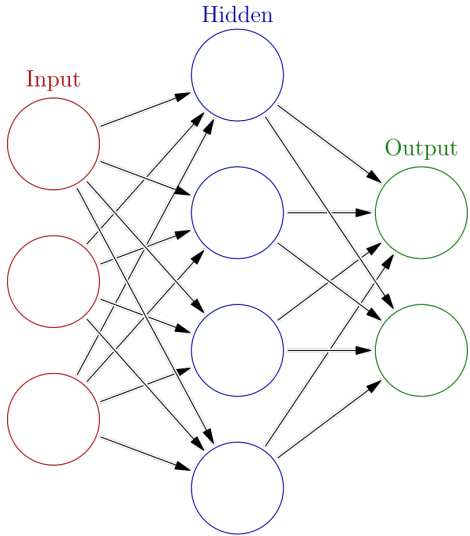
Self-driving labs basics (III)



Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation
- The expertise requirements are highly multidisciplinary (this can be a barrier to adoption):
  - You need to know chemistry
  - You need to know programming
  - You need to know robotics
  - You need to know ML

Self-driving labs basics (III)



Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation
- The expertise requirements are highly multidisciplinary (this can be a barrier to adoption):
  - You need to know chemistry
  - You need to know programming
  - You need to know robotics
  - You need to know ML
- Maybe you don't need to be a world-class expert in all, but you need familiarity with each

Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation
- The expertise requirements are highly multidisciplinary (this can be a barrier to adoption):
  - You need to know chemistry
  - You need to know programming
  - You need to know robotics
  - You need to know ML
- Maybe you don't need to be a world-class expert in all, but you need familiarity with each
- Industrial R&D may have access to these resources, but academic research needs to adapt

Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation
- The expertise requirements are highly multidisciplinary (this can be a barrier to adoption):
  - You need to know chemistry
  - You need to know programming
  - You need to know robotics
  - You need to know ML
- Maybe you don't need to be a world-class expert in all, but you need familiarity with each
- Industrial R&D may have access to these resources, but academic research needs to adapt
- I am planning a MSc-level course on lab automation at CHEM for the near future (probably based on cheap 3D printers)

Self-driving labs basics (III)

- Personal opinion: SDLs are the future of experimental R&D in chemistry and materials science (in some cases, they are the present)
- Currently, only relatively simple experiments can be carried out
- Those can be done much more efficiently than in regular (human-led) experimentation
- The expertise requirements are highly multidisciplinary (this can be a barrier to adoption):
  - You need to know chemistry
  - You need to know programming
  - You need to know robotics
  - You need to know ML
- Maybe you don't need to be a world-class expert in all, but you need familiarity with each
- Industrial R&D may have access to these resources, but academic research needs to adapt
- I am planning a MSc-level course on lab automation at CHEM for the near future (probably based on cheap 3D printers)
- Future of chemical education will likely involve more math, statistics and programming (possibly mechatronics too)

Examples of SDLs: the A-Lab

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

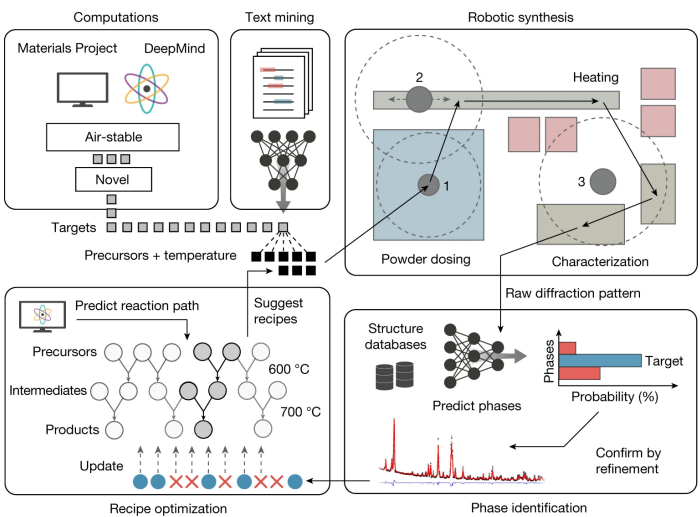
Open access

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, historical data from the literature, machine learning (ML) and active learning to plan and interpret the outcomes of experiments performed using robotics. Over 17 days of continuous operation, the A-Lab realized 41 novel compounds from a set of 58 targets including a variety of oxides and phosphates that were identified using large-scale ab initio phase-stability data from the Materials Project and Google DeepMind. Synthesis recipes were proposed by natural-language models trained on the literature and optimized using an active-learning approach grounded in thermodynamics. Analysis of the failed syntheses provides direct and actionable suggestions to improve current techniques for materials screening and synthesis design. The high success rate demonstrates the effectiveness of artificial-intelligence-driven platforms for autonomous materials discovery and motivates further integration of computations, historical knowledge and robotics.

Examples of SDLs: the A-Lab





Examples of SDLs: the A-Lab

https://www.youtube.com/watch?v=RLO4sfK37w4

Examples of SDLs: the A-Lab (criticism)

PRX ENERGY **3,** 011002 (2024)

Perspective

Challenges in High-Throughput Inorganic Materials Prediction and Autonomous Synthesis

Josh Leeman,[1] Yuhan Liu,[2] Joseph Stiles,[1,3] Scott B. Lee,[1] Prajna Bhatt,[2] Leslie M. Schoop,[1,3,*] and Robert G. Palgrave[2,†]

[1] *Department of Chemistry, Princeton University, Princeton, New Jersey 08540, USA*
[2] *Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, United Kingdom*
[3] *Princeton Materials Institute, Princeton University, Princeton, New Jersey 08540, USA*

(Received 30 January 2024; published 7 March 2024)

Materials discovery lays the foundation for many technological advancements. The prediction and discovery of new materials are not simple tasks. Here, we outline some basic principles of solid-state chemistry, which might help to advance both, and discuss pitfalls and challenges in materials discovery. Using the recent work of Szymanski *et al.* [Nature 624, 86 (2023)], which reported the autonomous discovery of 43 novel materials, as an example, we discuss problems that can arise in unsupervised materials discovery and hope that by addressing these, autonomous materials discovery can be brought closer to reality. We discuss all 43 synthetic products and point out four common shortfalls in the analysis. These errors unfortunately lead to the conclusion that no new materials have been discovered in that work. We conclude that there are two important points of improvement that require future work from the community, as follows. (i) Automated Rietveld analysis of powder x-ray diffraction data is not yet reliable. Future improvement of such, and the development of a reliable artificial-intelligence-based tool for Rietveld fitting, would be very helpful, not only for autonomous materials discovery but also for the community in general. (ii) We find that disorder in materials is often neglected in predictions. The predicted compounds investigated herein have all their elemental components located on distinct crystallographic positions but in reality, elements can share crystallographic sites, resulting in higher-symmetry space groups and—very often—known alloys or solid solutions. This error might be related to the difficulty of modeling disorder in a computationally economical way and needs to be addressed both by computational and experimental material scientists. We find that two thirds of the claimed successful materials in Szymanski *et al.* are likely to be known compositionally disordered versions of the predicted ordered compounds. We highlight important issues in materials discovery, computational chemistry, and autonomous interpretation of x-ray diffraction. We discuss concepts of materials discovery from an experimentalist point of view, which we hope will be helpful for the community to further advance this important new aspect of our field.

DOI: 10.1103/PRXEnergy.3.011002

Examples of SDLs: the A-Lab (some thoughts)

Examples of SDLs: the A-Lab (some thoughts)

- The complexity of realizing an autonomous lab for automated novel materials synthesis, characterization and refinement was enormous

Examples of SDLs: the A-Lab (some thoughts)

- The complexity of realizing an autonomous lab for automated novel materials synthesis, characterization and refinement was enormous
- The fact that it works at all almost seems like a myracle: impossible just a few years ago

Examples of SDLs: the A-Lab (some thoughts)

- The complexity of realizing an autonomous lab for automated novel materials synthesis, characterization and refinement was enormous
- The fact that it works at all almost seems like a myracle: impossible just a few years ago
- It required highly disciplinary team of experts

Examples of SDLs: the A-Lab (some thoughts)

- The complexity of realizing an autonomous lab for automated novel materials synthesis, characterization and refinement was enormous
- The fact that it works at all almost seems like a myracle: impossible just a few years ago
- It required highly disciplinary team of experts
- Paying more attention to the chemistry (domain knowledge requirement) would have prevented rediscovering existing compounds and the criticism

Examples of SDLs: the A-Lab (some thoughts)

- The complexity of realizing an autonomous lab for automated novel materials synthesis, characterization and refinement was enormous
- The fact that it works at all almost seems like a myracle: impossible just a few years ago
- It required highly disciplinary team of experts
- Paying more attention to the chemistry (domain knowledge requirement) would have prevented rediscovering existing compounds and the criticism
- Having domain experts (e.g., X-ray crystallographers) on board could have helped...

Examples of SDLs: the A-Lab (some thoughts)

- The complexity of realizing an autonomous lab for automated novel materials synthesis, characterization and refinement was enormous
- The fact that it works at all almost seems like a myracle: impossible just a few years ago
- It required highly disciplinary team of experts
- Paying more attention to the chemistry (domain knowledge requirement) would have prevented rediscovering existing compounds and the criticism
- Having domain experts (e.g., X-ray crystallographers) on board could have helped...
- ... but if it was for the traditionalists, nothing would have been done

Examples of SDLs: the A-Lab (some thoughts)

- The complexity of realizing an autonomous lab for automated novel materials synthesis, characterization and refinement was enormous
- The fact that it works at all almost seems like a myracle: impossible just a few years ago
- It required highly disciplinary team of experts
- Paying more attention to the chemistry (domain knowledge requirement) would have prevented rediscovering existing compounds and the criticism
- Having domain experts (e.g., X-ray crystallographers) on board could have helped...
- ... but if it was for the traditionalists, nothing would have been done
- My take: equip chemists and materials scientists with the tools to utilize state-of-the-art technology (programming, math, statistics, robotics); also a cultural shift is needed: the future will combine experiment and simulation inextricably

Examples of SDLs: the A-Lab (some thoughts)

- The complexity of realizing an autonomous lab for automated novel materials synthesis, characterization and refinement was enormous
- The fact that it works at all almost seems like a myracle: impossible just a few years ago
- It required highly disciplinary team of experts
- Paying more attention to the chemistry (domain knowledge requirement) would have prevented rediscovering existing compounds and the criticism



- Having domain experts (e.g., X-ray crystallographers) on board could have helped...
- ... but if it was for the traditionalists, nothing would have been done
- My take: equip chemists and materials scientists with the tools to utilize state-of-the-art technology (programming, math, statistics, robotics); also a cultural shift is needed: the future will combine experiment and simulation inextricably
- Whether the A-Lab failed or succeeded for this paper, it led to PROGRESS

Examples of SDLs: autonomous battery research

Accelerating Aqueous Electrolyte Design with Automated Full-Cell Battery Experimentation and Bayesian Optimization

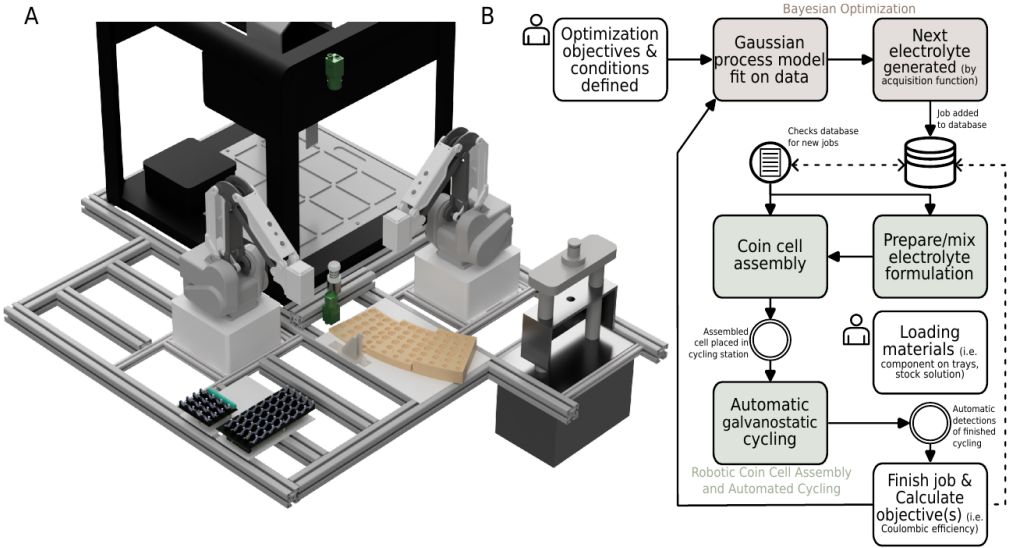
Jackie T. Yik[1], Carl Hvarfner[2], Jens Sjölund[3], Erik J. Berg[1,**], and Leiting Zhang[1,*]

[1]Department of Chemistry – [*]Angström Laboratory, Uppsala University, Uppsala, Sweden
[2]Department of Computer Science, Lund University, Lund, Sweden
[3]Department of Information Technology, Uppsala University, Uppsala, Sweden
[*]Correspondence: leiting.zhang@kemi.uu.se
[**]Correspondence: erik.berg@kemi.uu.se

SUMMARY

The integration of automation and data-driven methodologies offer a promising approach to accelerating materials discovery in energy storage research. Thus far, in battery research, coin-cell assembly has advanced to become near fully-automated but remains largely disconnected from data-driven methods. To bridge the disconnect, this work presents a self-driving laboratory framework to accelerate electrolyte discovery by integrating automated coin-cell assembly, galvanostatic cycling of $LiFePO_4||Li_4Ti_5O_{12}$ organic-aqueous full-cells, and Bayesian optimization for selecting subsequent experiments based on prior results. The study explored an organic-aqueous hybrid electrolyte system comprising four co-solvents and two lithium-conducting salts. Using this framework, cells with optimized electrolyte cycled with at least 94% Coulombic efficiency. Additionally, online electrochemical mass spectrometry revealed the optimized organic co-solvents successfully mitigated parasitic hydrogen evolution reaction. The results highlight the potential of combining Bayesian optimization with autonomous full-cell experimentation, while contributing new electrolyte design insights for next-generation aqueous batteries.

Examples of SDLs: autonomous battery research



Examples of SDLs: autonomous battery research (some thoughts)

Examples of SDLs: autonomous battery research (some thoughts)

- Huge resources (human and monetary) are needed to realize SDLs following the A-Lab approach

Examples of SDLs: autonomous battery research (some thoughts)

- Huge resources (human and monetary) are needed to realize SDLs following the A-Lab approach
- Autonomous labs can also be successfully realized with smaller resources (colleague Erik Berg told me 30k EUR for their robotic setup)

Examples of SDLs: autonomous battery research (some thoughts)

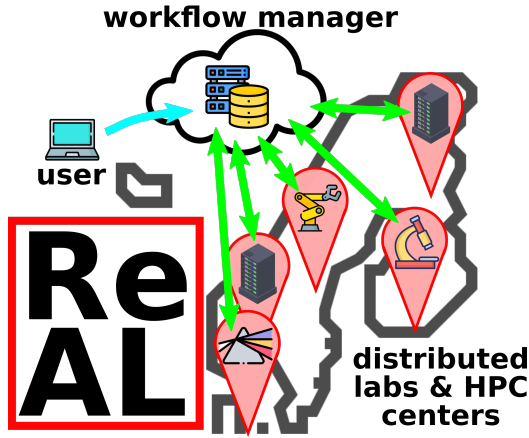
- Huge resources (human and monetary) are needed to realize SDLs following the A-Lab approach
- Autonomous labs can also be successfully realized with smaller resources (colleague Erik Berg told me 30k EUR for their robotic setup)
- Virtually every lab might be able to benefit from *some* degree of automation

Examples of SDLs: autonomous battery research (some thoughts)

- Huge resources (human and monetary) are needed to realize SDLs following the A-Lab approach
- Autonomous labs can also be successfully realized with smaller resources (colleague Erik Berg told me 30k EUR for their robotic setup)
- Virtually every lab might be able to benefit from *some* degree of automation
- The main barrier is access to human resources with sufficient expertise in automation-enabling technologies *and* domain knowlege

Examples of SDLs: autonomous battery research (some thoughts)

- Huge resources (human and monetary) are needed to realize SDLs following the A-Lab approach
- Autonomous labs can also be successfully realized with smaller resources (colleague Erik Berg told me 30k EUR for their robotic setup)
- Virtually every lab might be able to benefit from *some* degree of automation
- The main barrier is access to human resources with sufficient expertise in automation-enabling technologies *and* domain knowlege
- The future of chemistry and materials research is exciting!



Acknowledgements



Funded by
the European Union





• This work is based upon work from COST Action CA21126 - Carbon molecular nanostructures in space (NanoSpace), supported by COST (European Cooperation in Science & Technology).

• COST (European Cooperation in Science & Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

• www.cost.eu
• https://www.cost.eu/actions/CA21126/
• https://research.iac.es/proyecto/nanospace/