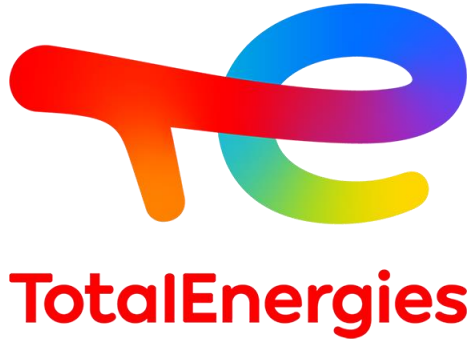


Numerical days: hands on AI for chemistry and materials

Cécile Pereira, Kamila Kazmierczak, Mauricio Araya

June 25th, 2024

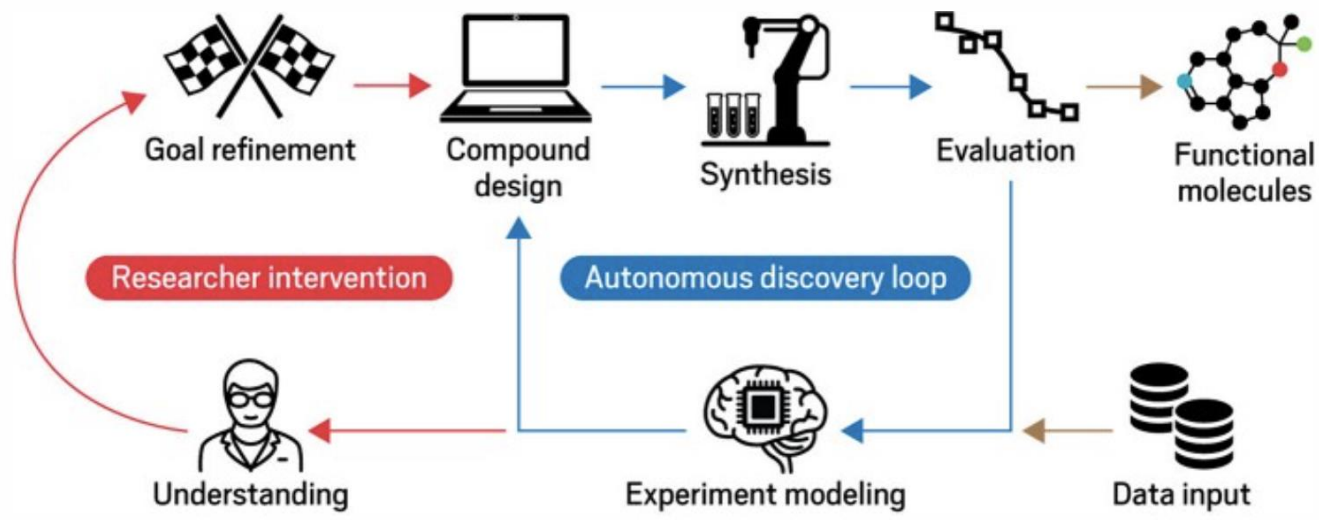
Example of projects at TotalEnergies

Prediction of properties
- Polymerization catalysts: performance
- CO2 capture
  - Partial atomic charges
  - CO2 uptake

Optimization
- Antioxydants

Prediction of chain of reactions
- Batteries

Surrogate models to replace simulations
- CFD
- Heat diffusion



Credit: Adapted from Connor W. Coley/Will Ludwig/C&EN

Source: Connor Coley/Massachusetts Institute of Technology

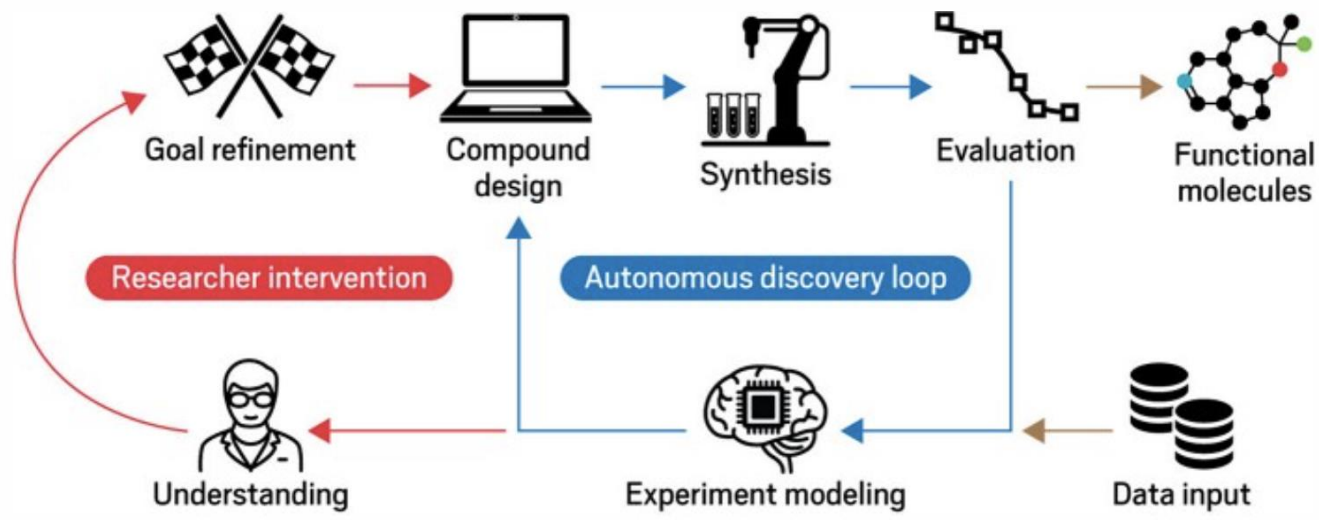
Example of projects at TotalEnergies

Prediction of properties
- **Polymerization catalysts**: performance
- CO2 capture
  - Partial atomic charges
  - CO2 uptake

Optimization
- Antioxydants

Prediction of chain of reactions
- Batteries

Surrogate models to replace simulations
- CFD
- Heat diffusion



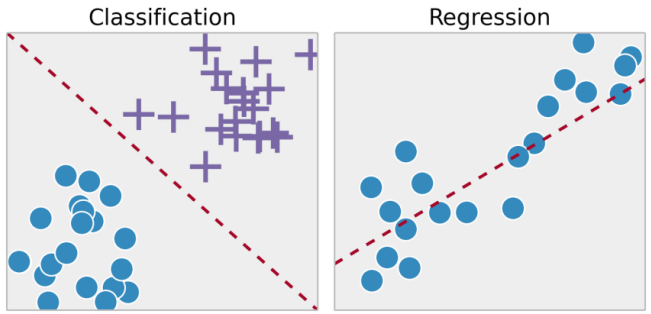
Credit: Adapted from Connor W. Coley/Will Ludwig/C&EN

Source: Connor Coley/Massachusetts Institute of Technology

Prediction of properties

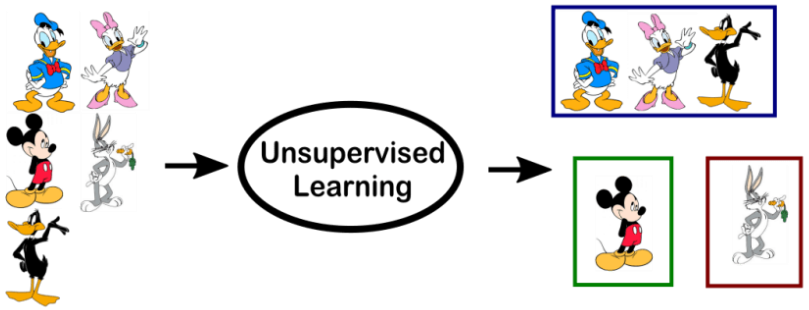
I have DATA

I have labels:
I want to predict something



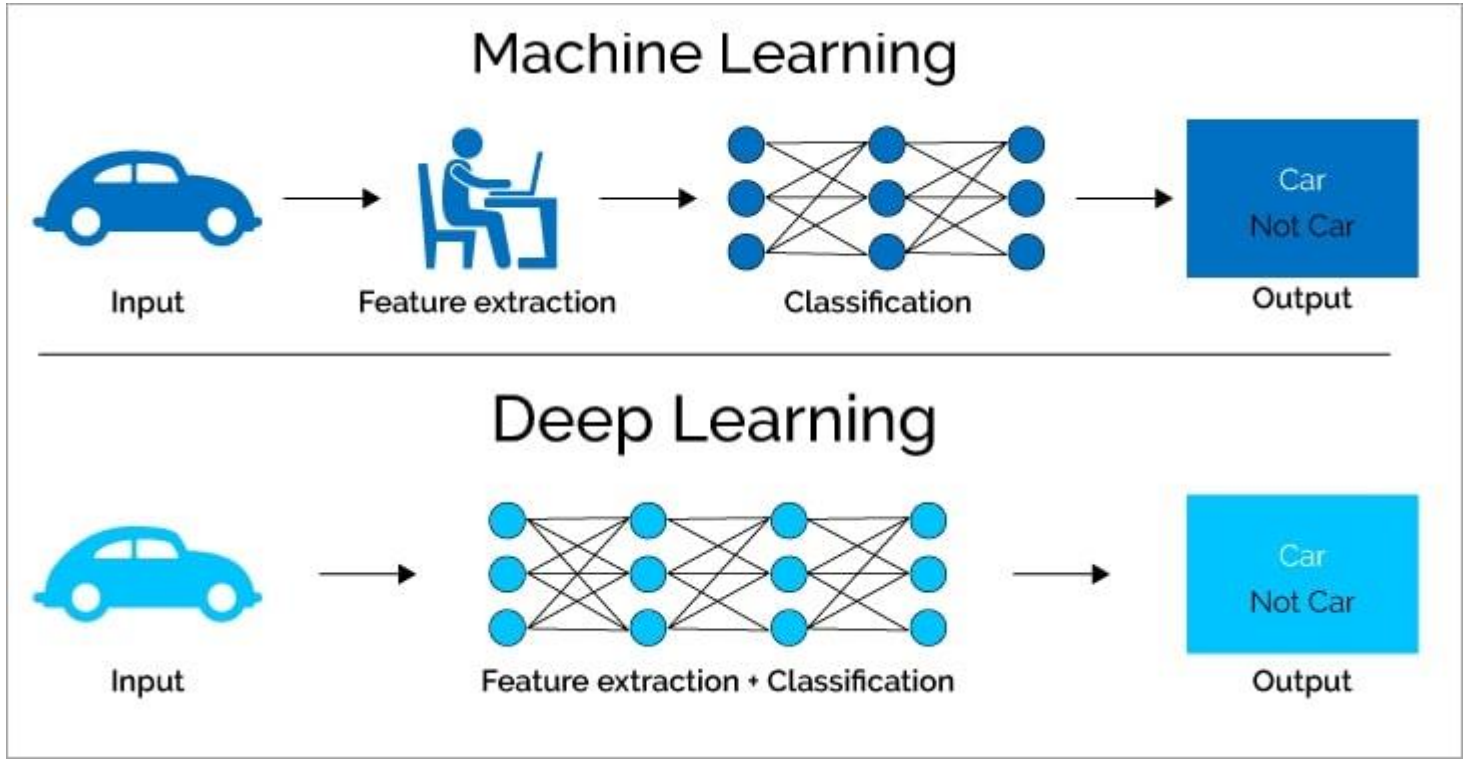
Supervised machine learning

I don't have labels:
I want to group by similarities



Unsupervised machine learning

Machine learning or deep learning



Classical machine Learning:

I provide the features (column)

DeepLearning: the model automatically extract the features from the dataset

Project 1: QSA/PR for POL

TotalEnergies

Billions of possible polymerization catalysts, impossible to synthesize all of them

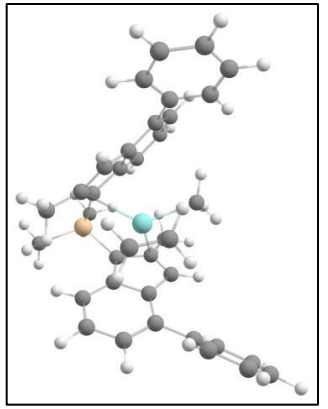
(synthesis and purification of 1 complex ~6-12 months(!))

Problem statement

Quantitative **S**tructure – **A**ctivity/**P**roperties **R**elationships models to predict performance of catalysts before their synthesis

QSA/PR

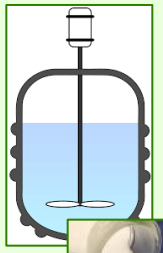
Focus only on the synthesis of the most promising catalysts to increase our chance of success



Project 1: QSA/PR for POL

Data collection

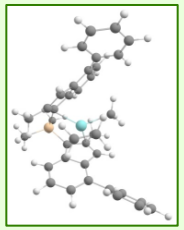
Data preparation



Experimental input

~50 experimental results

Small dataset!



~20 descriptors (features)

Numerical input

DFT calculations

Schrödinger

Work executed

Project 1: QSA/PR for POL

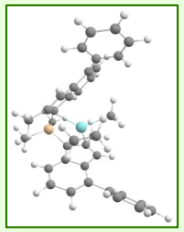

Data collection


Data preparation


Data modelling


Model validation

Experimental input

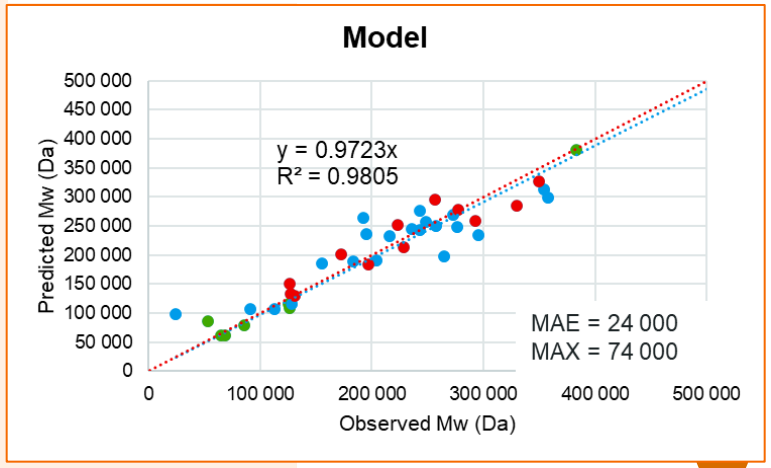


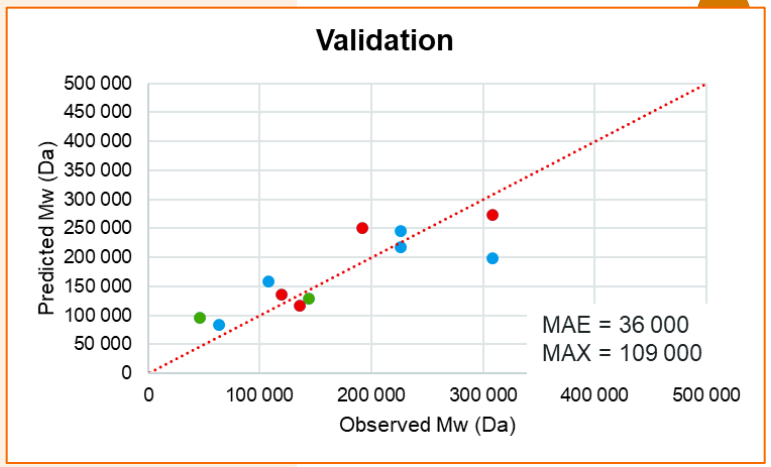
Numerical input

Supervised Machine Learning Regression models

Statistical model

Example of created model





Work executed

Work in progress

Project 1: QSA/PR for POL

Data collection

Data preparation

Data modelling

Model validation

User-friendly interface

Everyday utilization

TotalEnergies

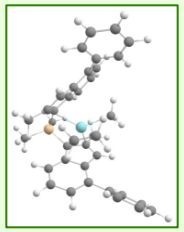
Experimental input

Supervised Machine Learning
Regression models

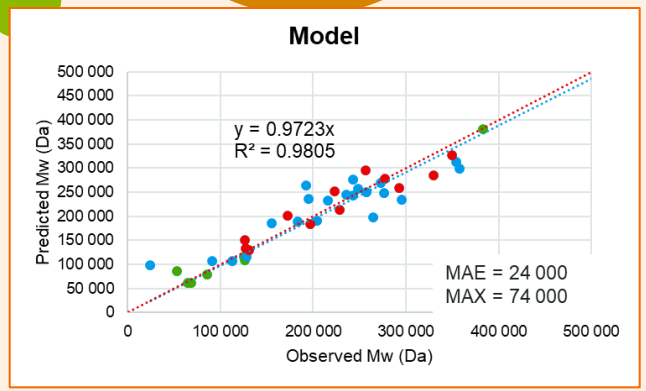
Statistical model

Predictions need to be **fast**, and executed in **user-friendly** manner

python™



Numerical input



in silico predictions

Predictions still need **numerical input data**

DFT or *(New tool) Neural-network Potential*

Schrödinger

MATLANTIS

Work executed

Work in progress

Work in progress (early stage)

Hands on: on google collab

Generate features

Rdkit
matminer
Deepchem

https://github.com/total-sa/CChemMLDays

AutoML

Pycaret
H2O

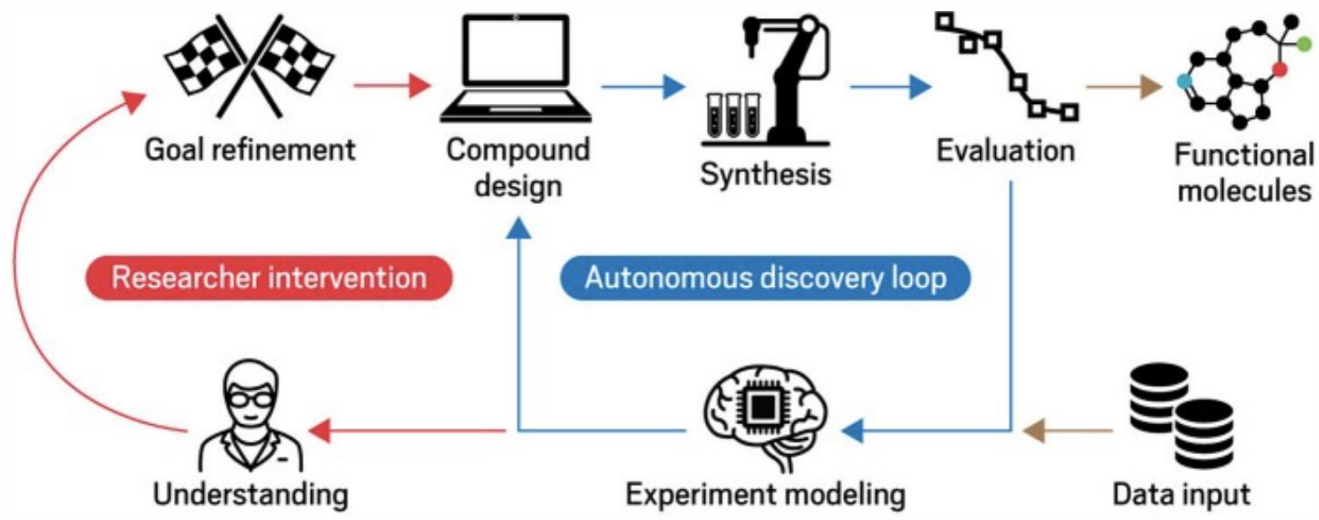
Example of projects at TotalEnergies

Prediction of properties
- Polymerization catalysts: performance
- CO2 capture
  - **Partial atomic charges**
  - CO2 uptake

Optimization of molecules or materials
- Antioxydants
- MOFs

Prediction of chain of reactions
- Batteries

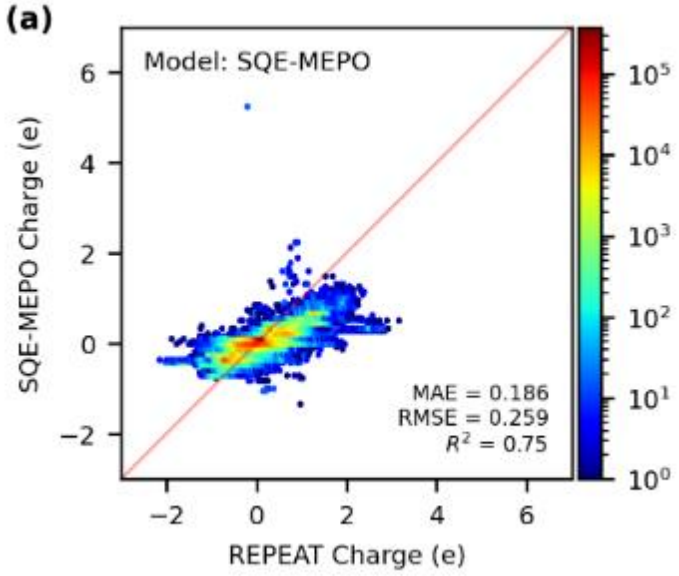
Surrogate models to replace simulations
- CFD
- Heat diffusion
- Process optimization for CO2 capture



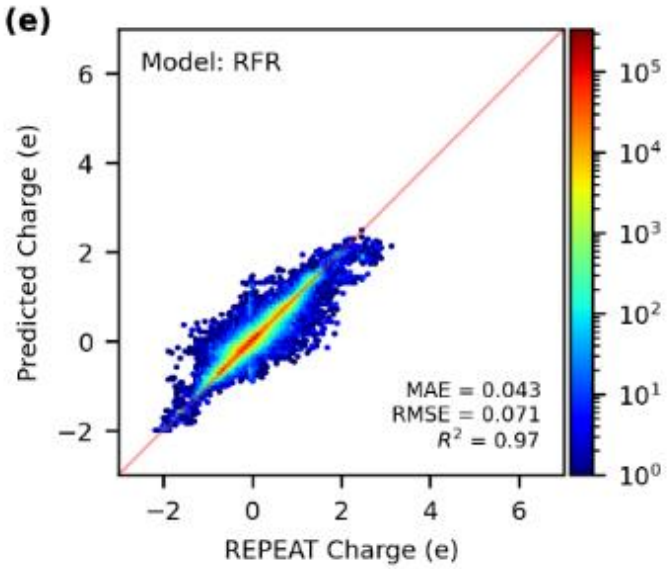
Credit: Adapted from Connor W. Coley/Will Ludwig/C&EN

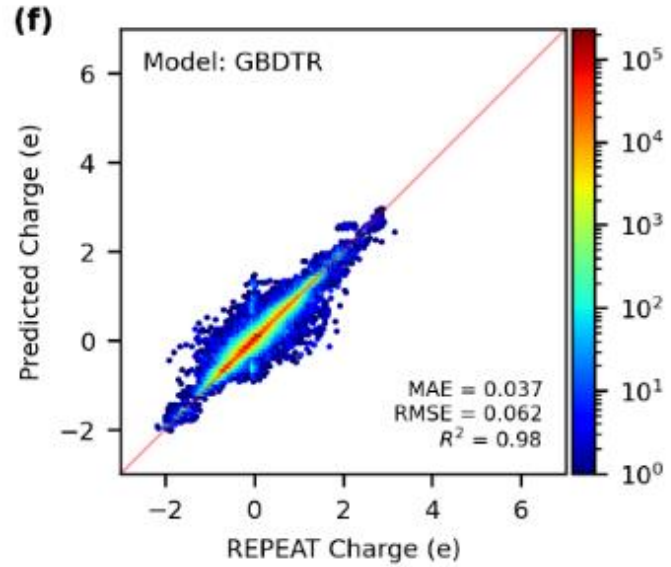
Source: Connor Coley/Massachusetts Institute of Technology

Machine learning to predict partial atomic charges

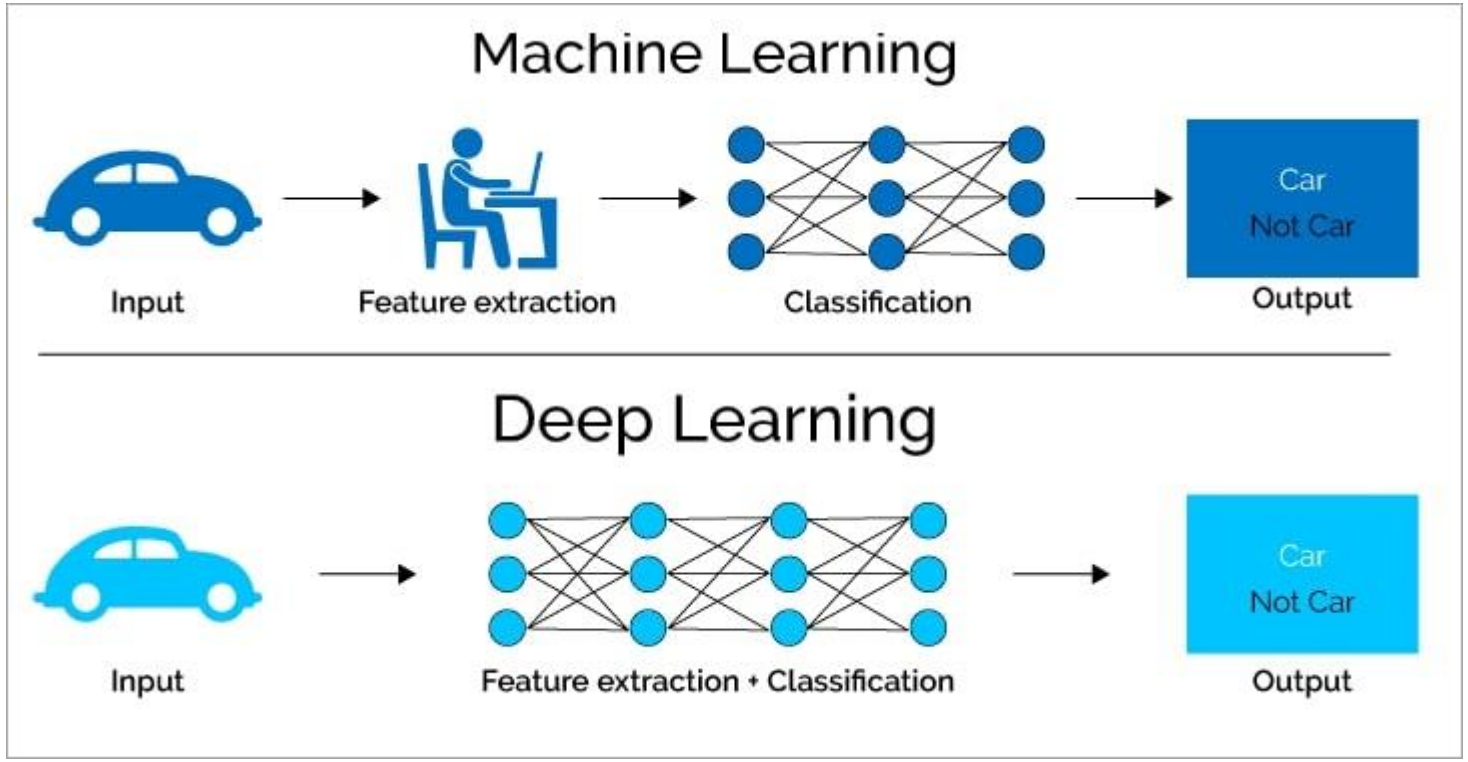


- Prepare the dataset
- Generate features
- Run autoML





Machine learning or deep learning

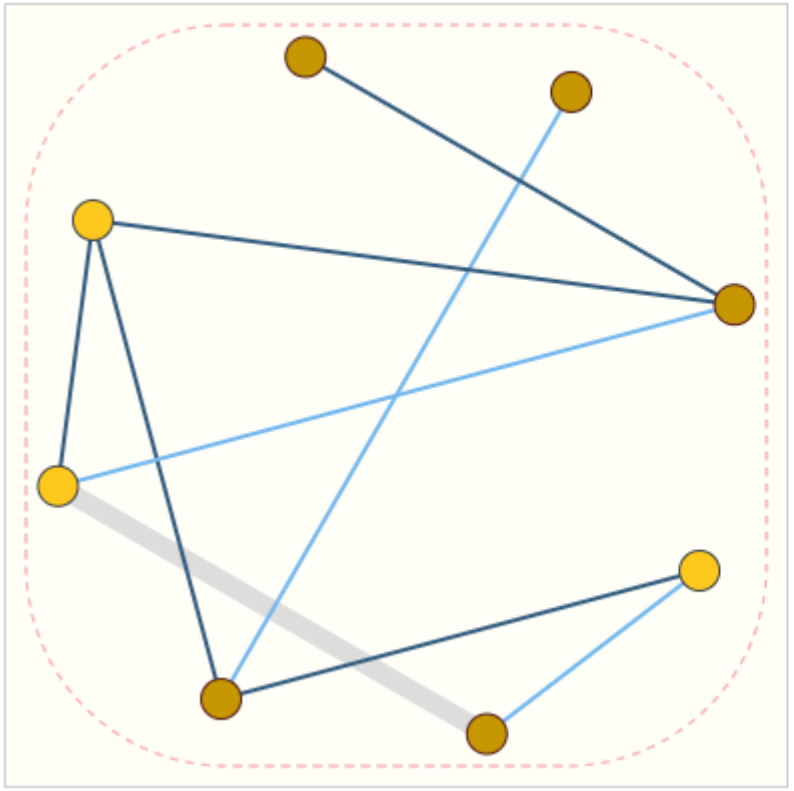


Classical machine Learning:

I provide the features (column)

DeepLearning: the model automatically extract the features from the dataset

Graph neural network



Nodes

```
[0, 1, 1, 0, 0, 1, 1, 1]
```

Edges

```
[2, 1, 1, 1, 2, 2, 1, 1]
```

Adjacency List

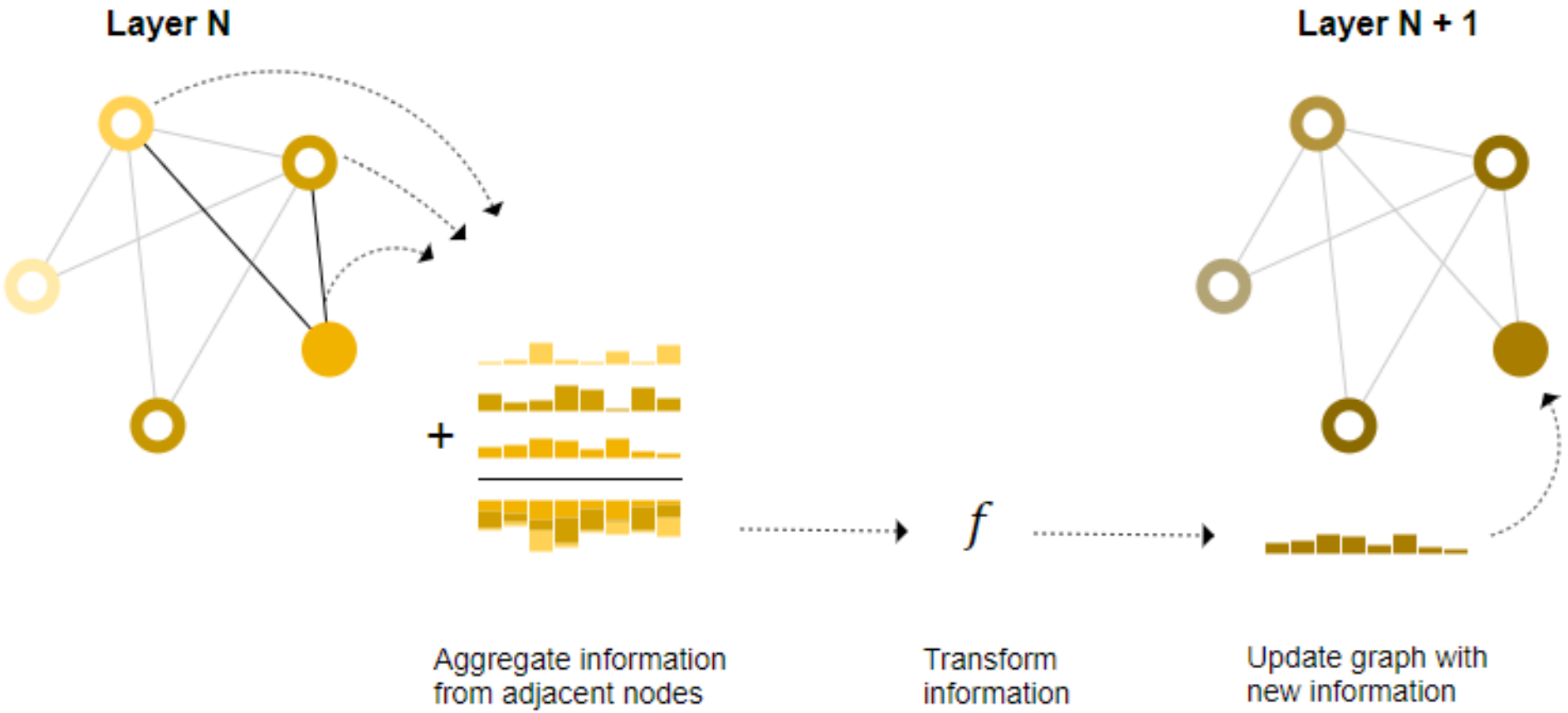
```
[[1, 0], [2, 0], [4, 2], [4, 3],
 [6, 2], [7, 3], [7, 4], [7, 5]]
```

Global

```
0
```

https://distill.pub/2021/gnn-intro/

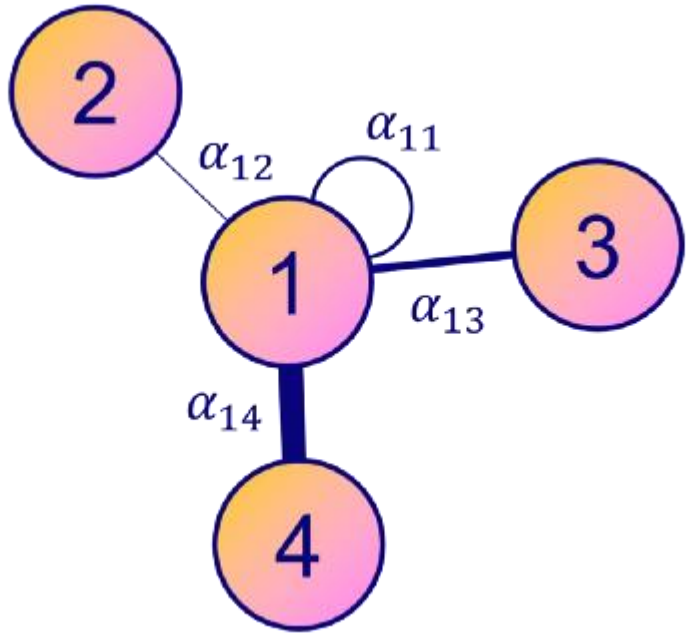
Graph neural network



Attention mechanism

GNN: every neighbor has the same importance

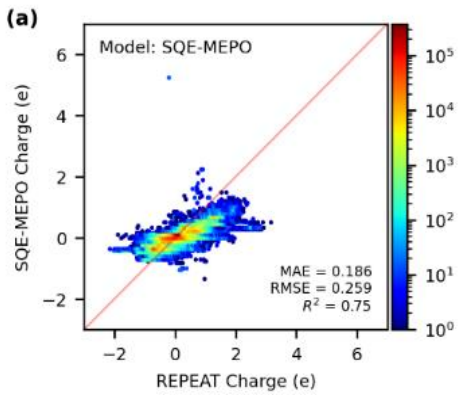
Graph attention network: the attention mechanism assigns a weighting factor to every connection
=> Value of the weights is learned with a neural network

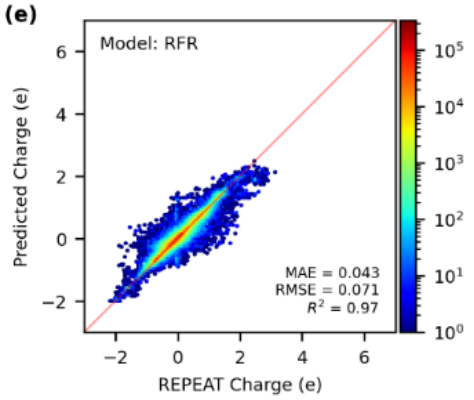


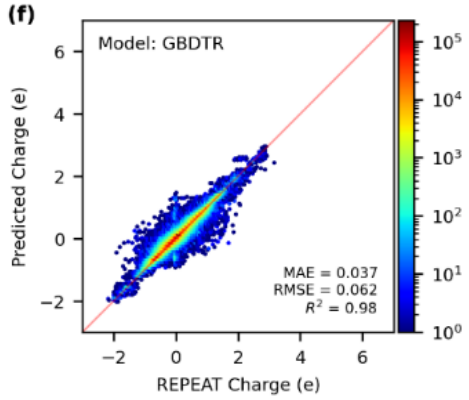
$$h_1 = \alpha_{11}\mathbf{W}x_1 + \alpha_{12}\mathbf{W}x_2 + \alpha_{13}\mathbf{W}x_3 + \alpha_{14}\mathbf{W}x_4$$

https://towardsdatascience.com/graph-attention-networks-in-python-975736ac5c0c

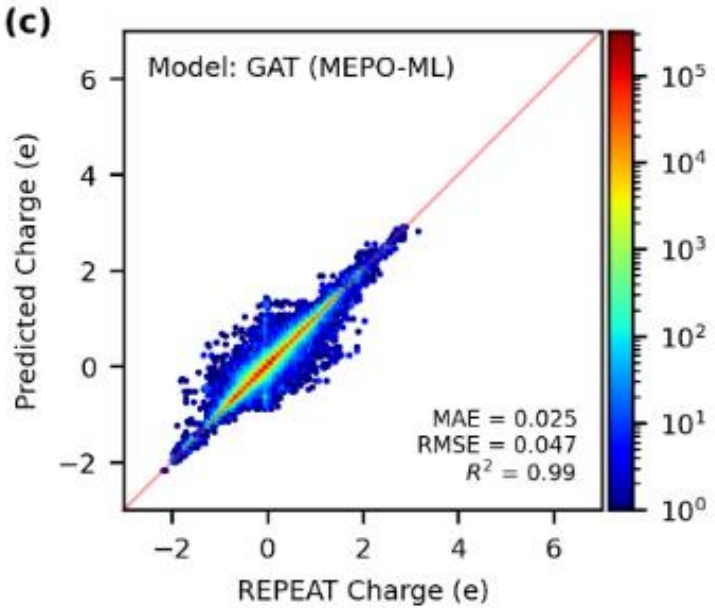
Machine learning to predict partial atomic charges

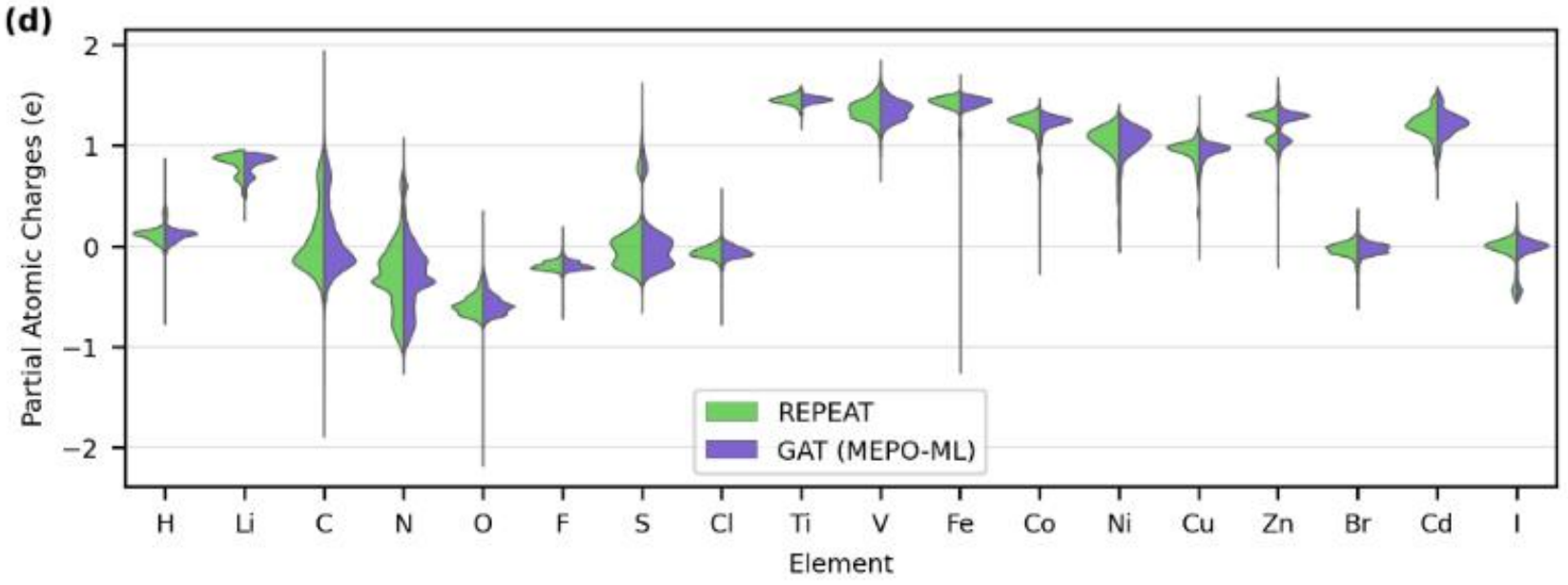






Take into account the structure (graph)





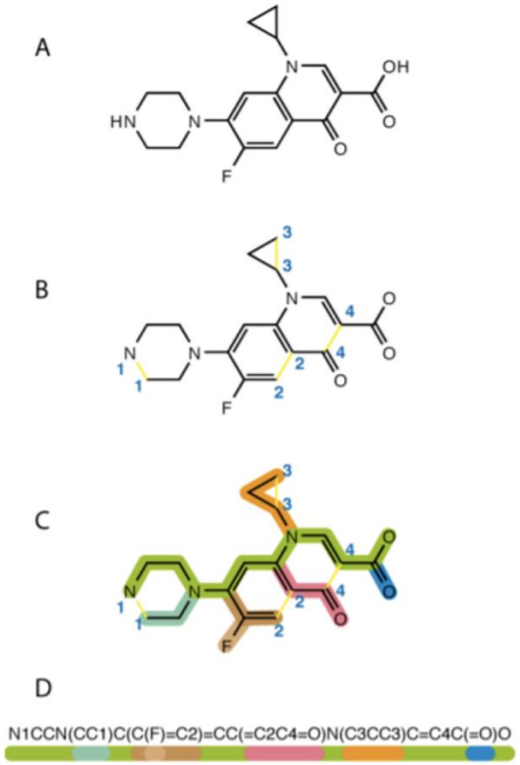
Hands on: look at the GAT model

Model and code are available on the github:
https://github.com/uowoolab/MEPO-ML

Optimization of antioxydants

- Two classes of deep learning models for molecular design:
  - Generative model
  - Reinforcement learning

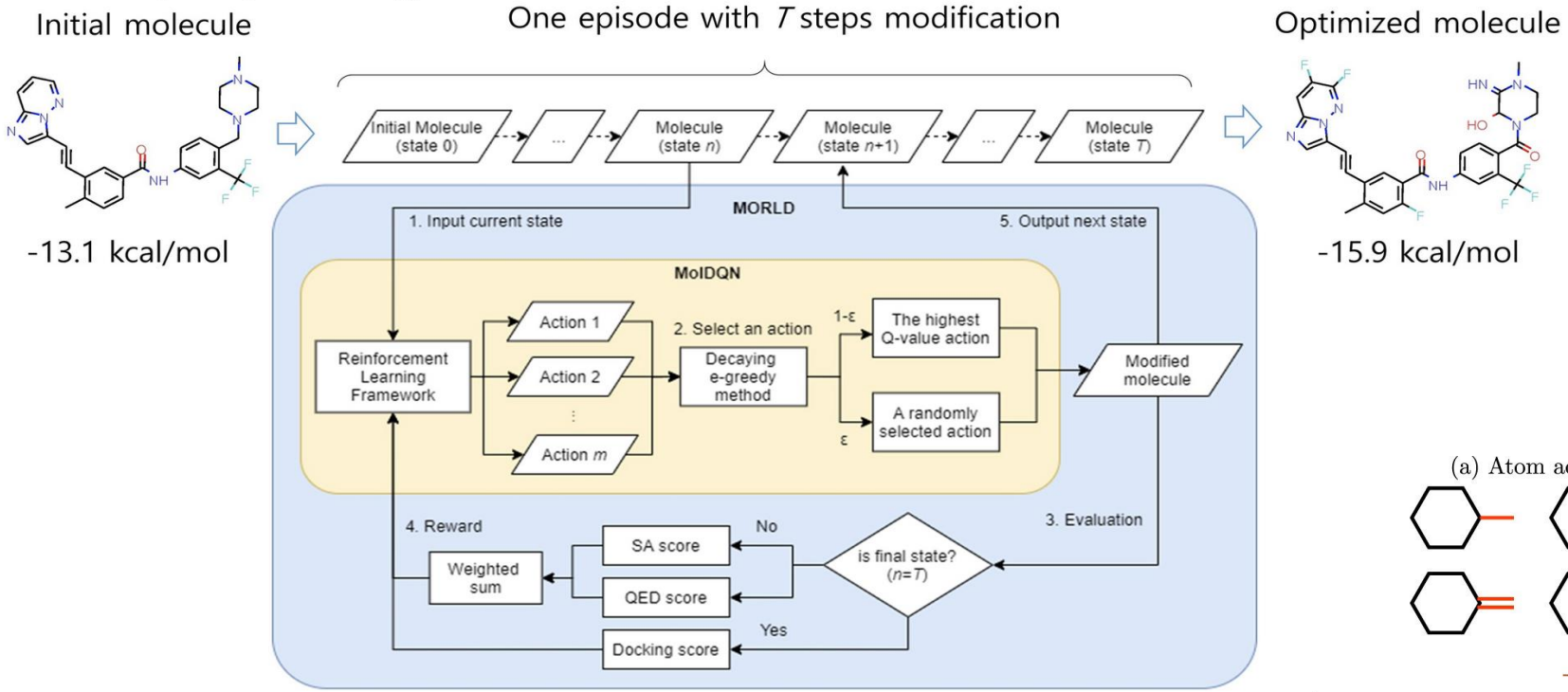
- How to let a machine recognize molecules?
  - Translate a molecule into a SMILES string (a line notation for describing the structure of chemical species using short ASCII strings)
  - Convert a SMILES string into a vector, i.e. Morgan fingerprint
  - Alternative approach: use a graph to represent a molecule

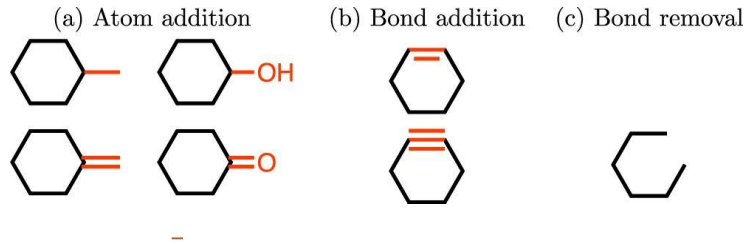


SMILES generation algorithm for Ciprofloxacin: break cycles, then write as branches off a main backbone

Optimization of antioxydants

- RL model learns actions to modify a molecule into an optimized molecule (with better property)
  - Game: modify an initial molecule into a better molecule
  - Typical example: MolDQN algorithm
- MolDQN learns T steps modifications and use a reward function to measure the quality of the generated molecules

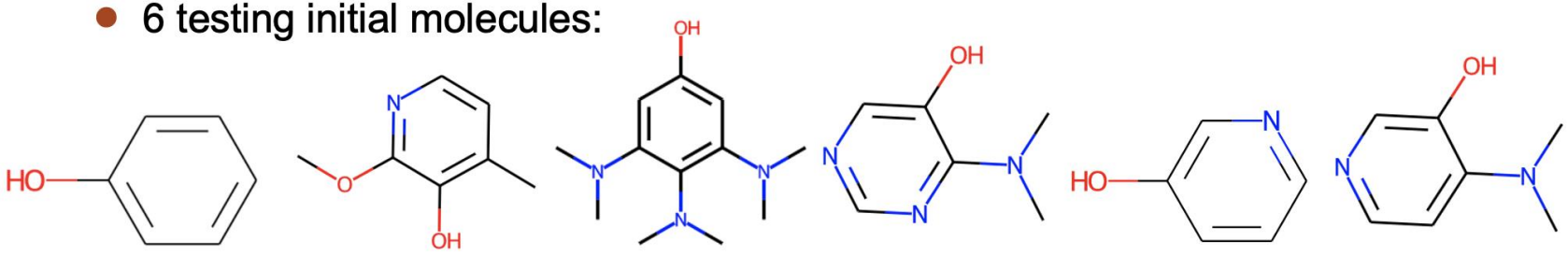




Patent filed, 2021

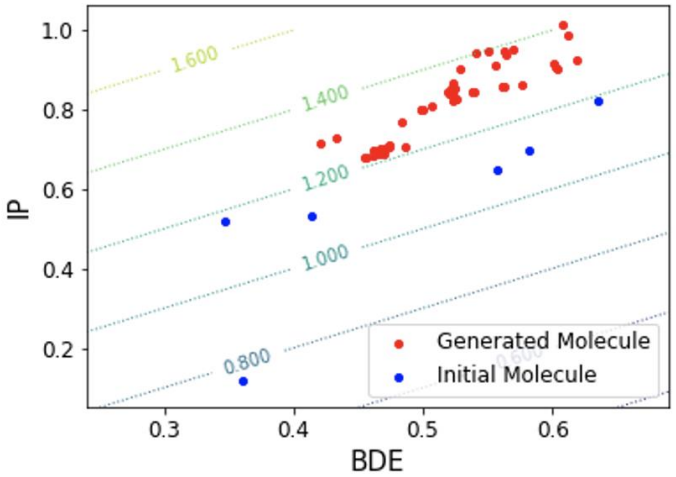
Optimization of antioxydants

- 6 testing initial molecules:

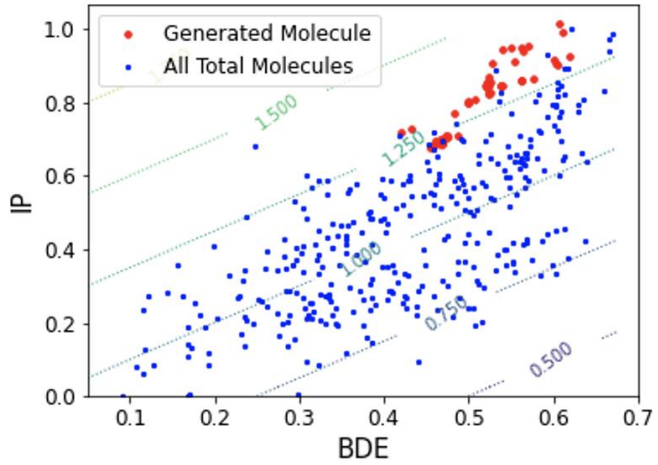


Molecule A Molecule B Molecule C Molecule D Molecule E Molecule F

- Quality of molecules generated by MT-MolDQN
  - Dash line: BDE_IP score = (1-normalized BDE) + normalized IP



Compare to initial molecules



Compare to all the TOTAL antioxidants

Also, given the parallelism introduced to the RL loop we reduced the standard time-to-solution by 3x

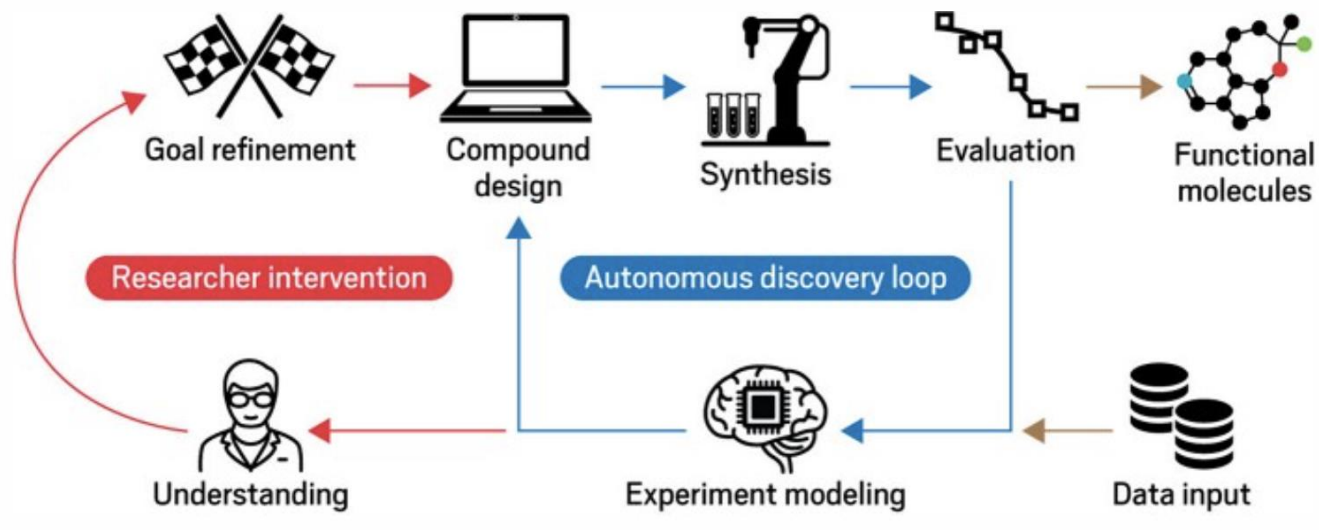
Example of projects at TotalEnergies

Prediction of properties
- Polymerization catalysts: performance
- CO2 capture
  - **Partial atomic charges**
  - CO2 uptake

Optimization of molecules or materials
- Antioxydants
- MOFs

Prediction of chain of reactions
- **Batteries**

Surrogate models to replace simulations
- CFD
- Heat diffusion
- Process optimization for CO2 capture

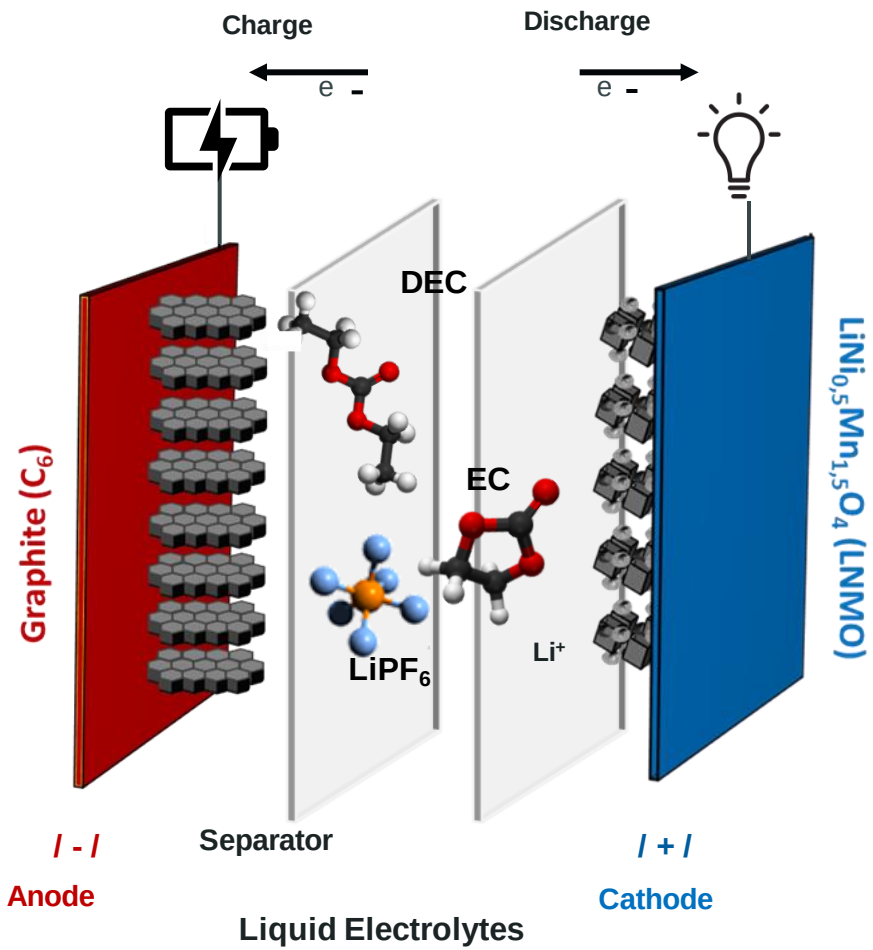


Credit: Adapted from Connor W. Coley/Will Ludwig/C&EN

Source: Connor Coley/Massachusetts Institute of Technology

ML-based screening of species on chain reaction networks

Complete module

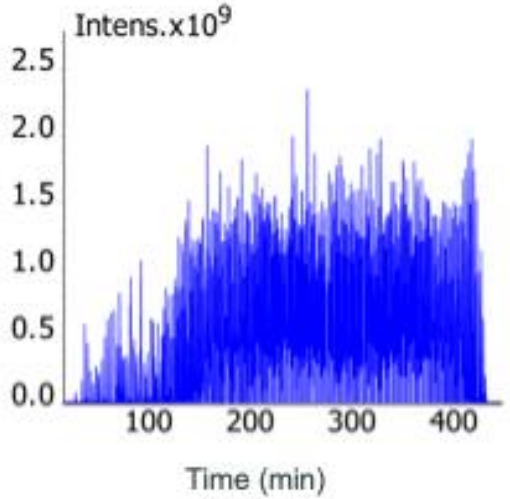


- **Two solid electrodes**
  - Negative electrode (Anode): Mainly graphite based
  - Positive electrode (Cathode): Transition metals oxides composed with multiple metals such as Cobalt, Nickel, Manganese
- **Liquid electrolytes**
  - Between 1 and 4 solvents
  - One or two salts
  - One to four additives
  - Example: Ethylene carbonate/Propylene carbonate + 1M LiPF6 + 2%(wt.) vinyl carbonate
- **Separators**
  - Polymers, coated membranes

FTICR-MS need CNR and CNR need FTICR-MS...

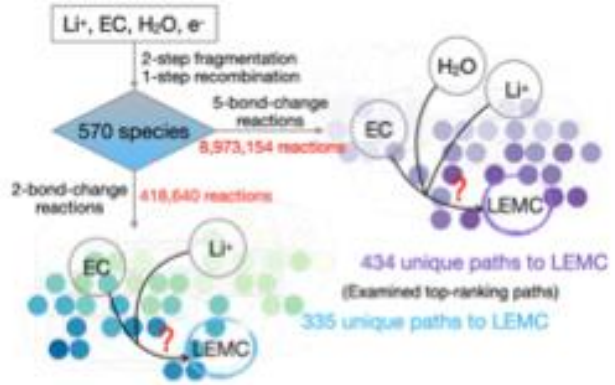
CRN coupled to FTICR-MS for a major advance in LIB interfaces understanding?

FTICR-MS



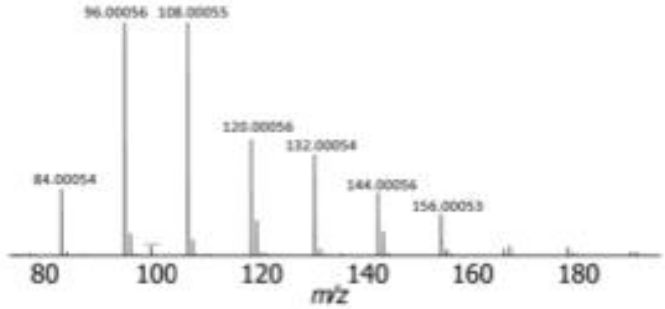
Thousands of MS signals

Use of CNR



CNR to select the targeted ions

Signal processing

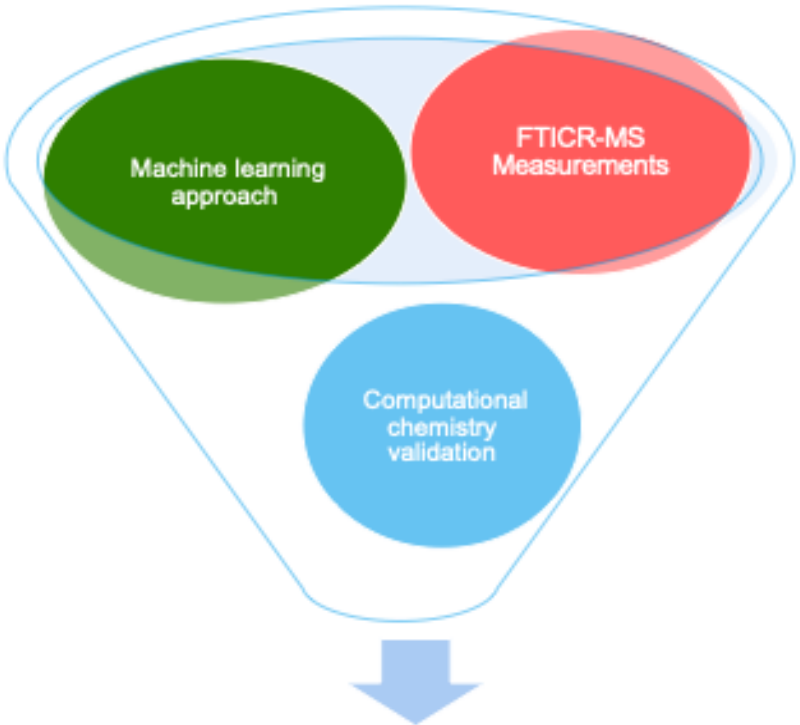


Check the presence of the targeted molecules in FTICR-MS data

CNR: chain reaction network

Proposal

Coupling of data-driven and experimental approaches to understand and improve electrode organic layer (SEI) composition



- Use machine learning algorithms* **patented** by Numerics Platform to generate and optimize millions of potential species in SEI layer
- Take advantage of our unique experimental capability of FTICR-MS measurements operated by DPP-ANA to filter real SEI species
- Validate species using computational chemistry performed by DPP-NUM

***Patent EP4152336A1**

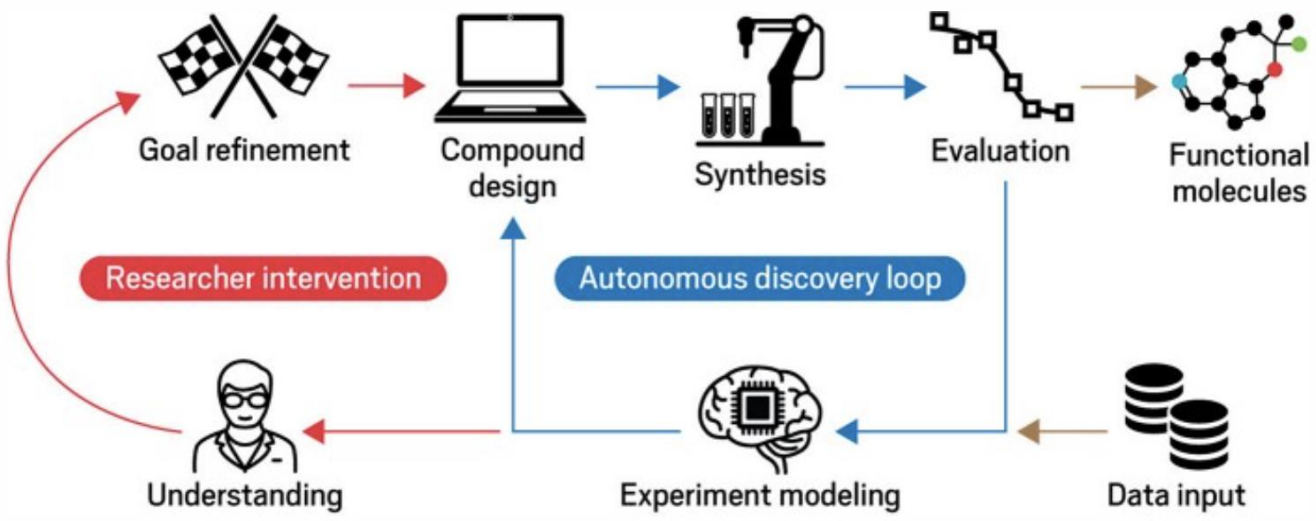
Conclusions

Prediction of properties
- Polymerization catalysts: performance
- CO2 capture
  - Partial atomic charges

Compound design
- Antioxydants
- MOFs

Prediction of chain of reactions
- Batteries

Surrogate models to replace simulations
- CFD
- Heat diffusion



Credit: Adapted from Connor W. Coley/Will Ludwig/C&EN

Source: Connor Coley/Massachusetts Institute of Technology

ML on chemistry and material
- Ask specialist to get idea of features you should work with
- Design: think also about synthesis and lifetime of the product
- Evaluation: surrogate model to replace simulations

Thanks

Kamila Kazmierczak

Mauricio Araya

Sophie Loehle

Cécile Pereira

Tom Woo

Jun Luo

& the computational chemistry community