ARISTOTLE UNIVERSITY OF THESSALONIKI

Machine Learning in chemistry for everyone: A practical guide for quick development of predictive models

George S. Fanourgakis

December, 9 2024

Faculty of Sciences, School of Chemistry,
Laboratory of Quantum and Computational Chemistry

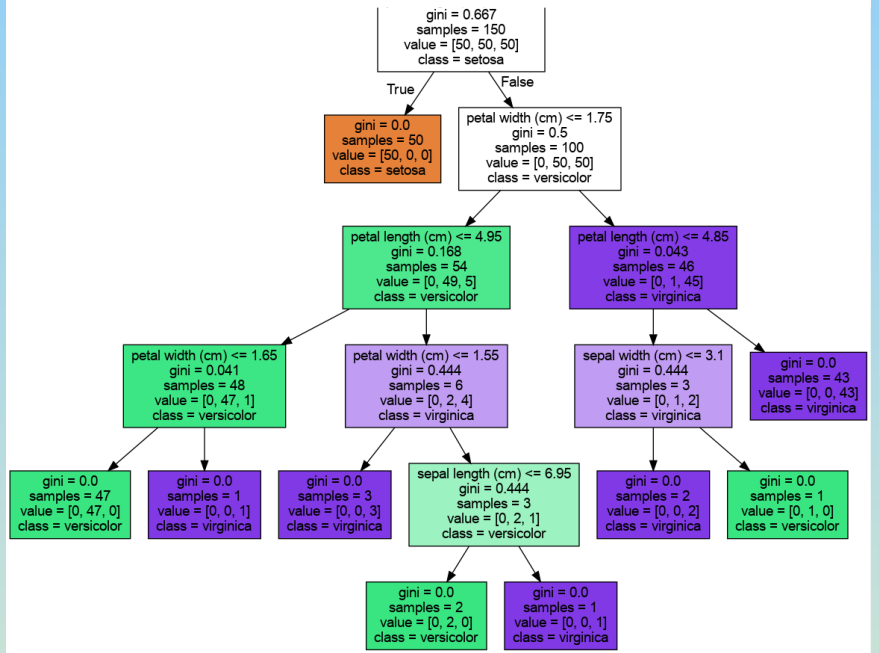
ML in Chemistry

1. Drug Discovery and Design: ML algorithms can predict properties of new compounds, identify potential drug candidates and optimize drug design processes. This accelerates the discovery of new medications reducing cost and time involved.
2. Material Science: ML helps in predicting the properties of new materials, optimizing their synthesis, and discovering novel materials with desired characteristics. This is particularly useful in developing advanced materials for various applications.
3. Chemical Synthesis: ML models assist in retrosynthesis, which involves predicting the sequence of chemical reactions needed to synthesize a target molecule. This helps chemists design efficient synthetic routes.
4. Catalysis: ML is used to design and optimize catalysts, which are substances that increase the rate of chemical reactions. This can lead to more efficient industrial processes and the development of greener chemical reactions.
5. Quantum Chemistry: ML techniques are applied to solve complex quantum mechanical problems, such as predicting molecular properties and simulating chemical reactions at the quantum level. This enhances our understanding of fundamental chemical processes.
6. Predictive Modeling: ML models can predict the outcomes of chemical reactions, the stability of compounds, and the behavior of chemical systems under different conditions. This helps in planning experiments and interpreting results.
7. Molecular Property Prediction: ML models can predict various properties of molecules, such as solubility, boiling points, and reactivity. This helps chemists understand molecules will behavior in different environments and conditions.
8. Environmental Chemistry: ML is used to model and predict the behavior of pollutants in the environment. This includes predicting the degradation pathways of chemicals and their impact on ecosystems.
9. Spectroscopy Analysis: ML algorithms can analyze spectroscopic data (e.g., NMR, IR, UV-Vis) to identify chemical compounds and understand their structures. This speeds up the process of analyzing complex mixtures.
10. Process Optimization: In industrial chemistry, ML is used to optimize chemical processes, improving efficiency and reducing waste.This includes optimizing reaction conditions and scaling up from lab to production scale.
11. Battery Research: ML aids in the development of new battery materials by predicting their performance and stability. This is essential for creating more efficient and longer-lasting batteries.

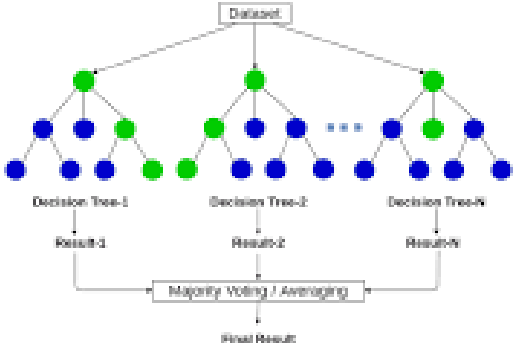
Machine Learning Algorithms

Tree based algorithms

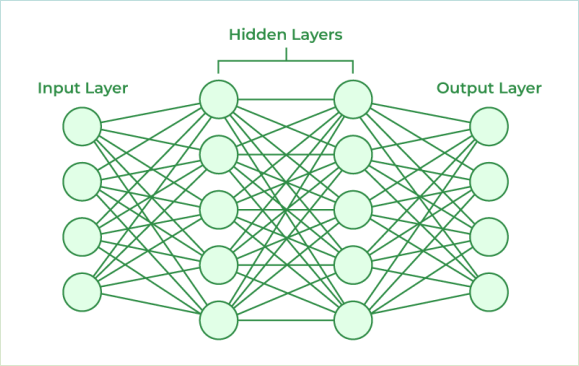
Decision Tree (DT)

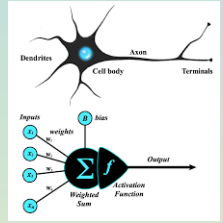


Random Forest (RF)
Extra Random Trees (ERT)



Artificial Neural Networks





Development of ML models in Python

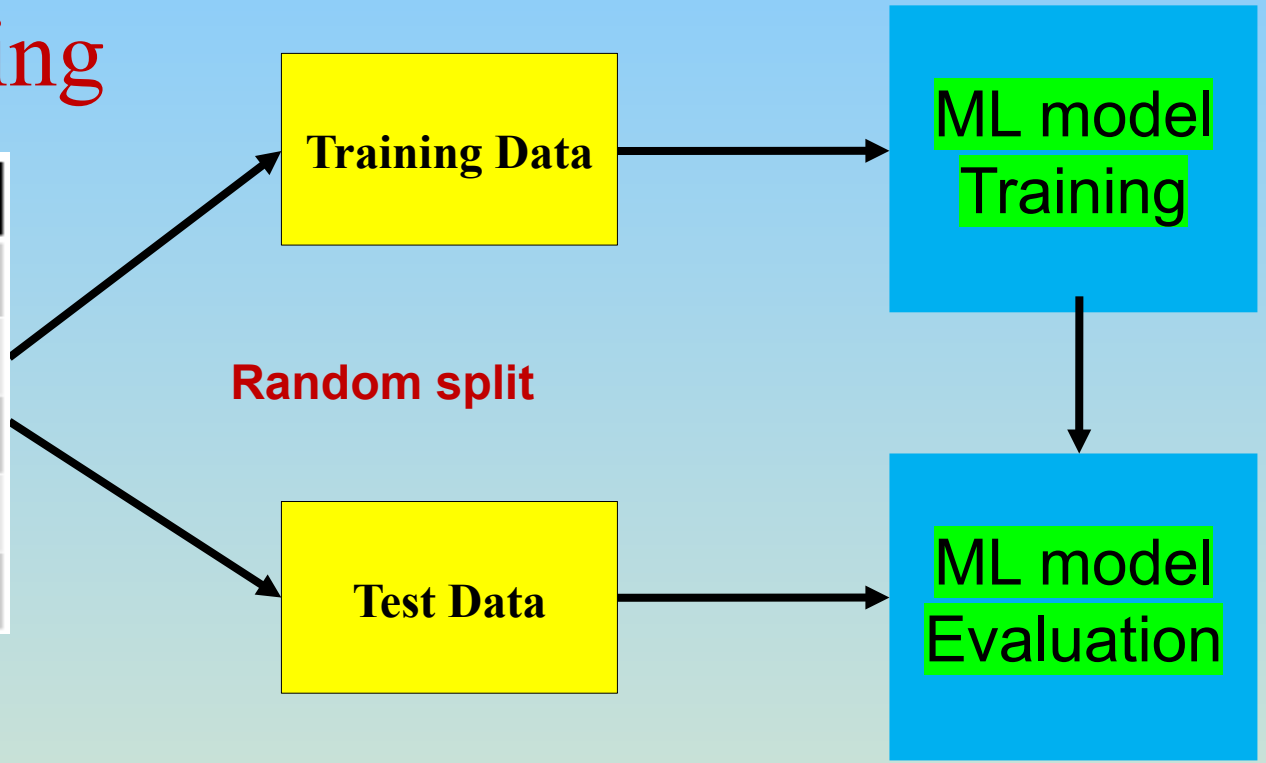
Main Steps

1. Data import
2. Data Cleaning
3. Data splitting into Training/Test Sets
4. Specification of a ML algorithm
5. Training of the ML model
6. Evaluation of the predictive model
7. ML model predictions
8. Understanding and improving model predictions

Development of a ML predictive mod

Supervised Learning

Available data	
Features (Descriptors)	Target
$x_1^1, x_1^2, \ldots, x_1^M$	y_1
$x_2^1, x_2^2, \ldots, x_2^M$	y_2
$x_N^1, x_N^2, \ldots, x_N^M$	y_N



Statistical Metrics

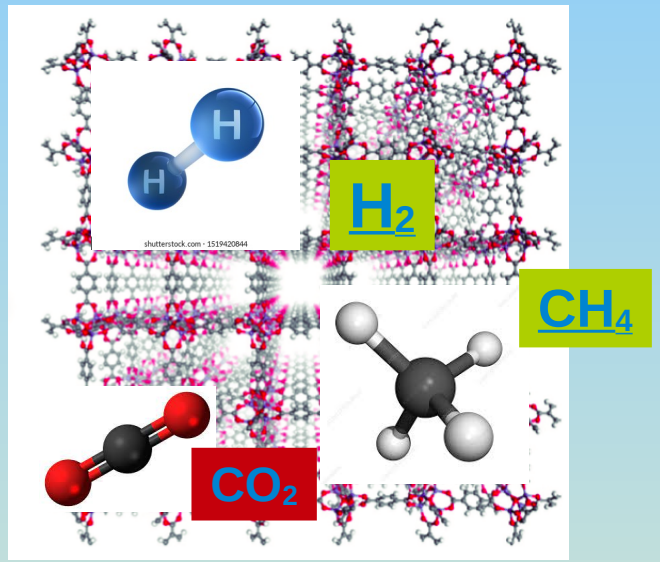
Mean absolute error (MAE) $\frac{1}{n}\sum|y_i - \hat{y}_i|$ y_i = observation $\hat{y}_i$ = prediction

Coefficient of determination $R^2 = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$ (unitless. =1 means perfect agreement)

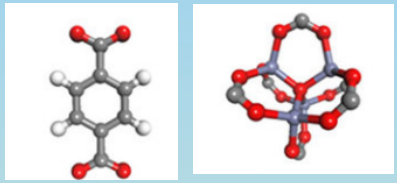
Problem under consideration

Metal Organic Frameworks (MOFs) for Energy and Environment

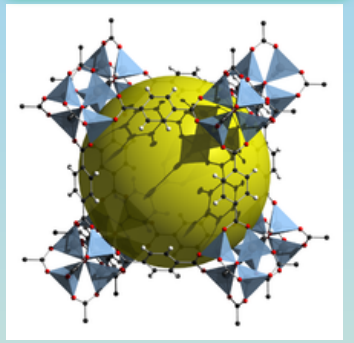
Theoretical & Computational Chemistry



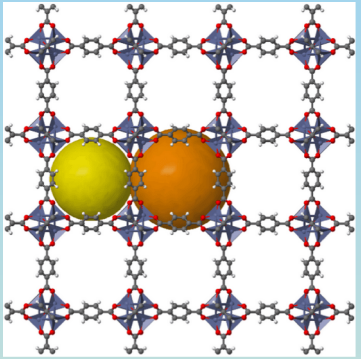
Ab initio

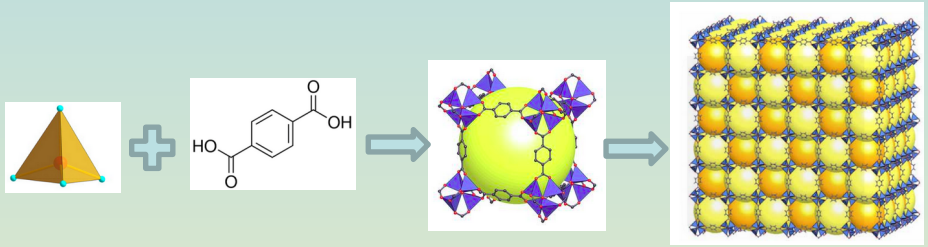


DFT - Empirical ff

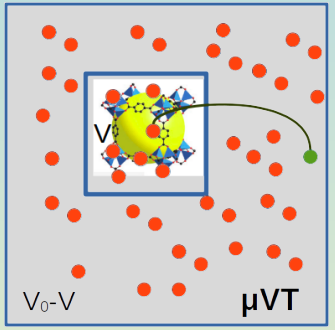


MD - GCMC





GCMC simulations



.

"Computational Screening of Trillions of MOFs for High-Performance Methane Storage" S. Lee, et al., *Appl. Mater. Interf.* 2021, 13, 20, 23647

Descriptors

Structural Descriptors (Experimental & Theoretical)

- **Density: mass density of MOF**
- **Pore limiting diameter (pld): the smallest diameter of the pores within the MOF structure**
- **Largest Cavity Diameter (lcd): the largest diameter of the cavities within the MOF structure**
- **pore volume (pv): the total volume of pores within the MOF structure.**
- **surface area (sa): the total area available on the internal and external surfaces of the MOF structure.**
- **Void fraction (vf): the ratio of the volume of the voids (empty spaces) within the MOF structure to the total volume of the MOF**

Probe atoms Descriptors (Theoretical)

- **Probes atoms (4 sizes): Probe atoms that account for the energetical features of the MOF**

G. S. Fanourgakis et al. J. Chem. Phys. A (2019) **123** 6080-6087

MOFs Dataset (methane P=1 atm, T=298 K)

CH4_P1_full.csv - LibreOffice Calc

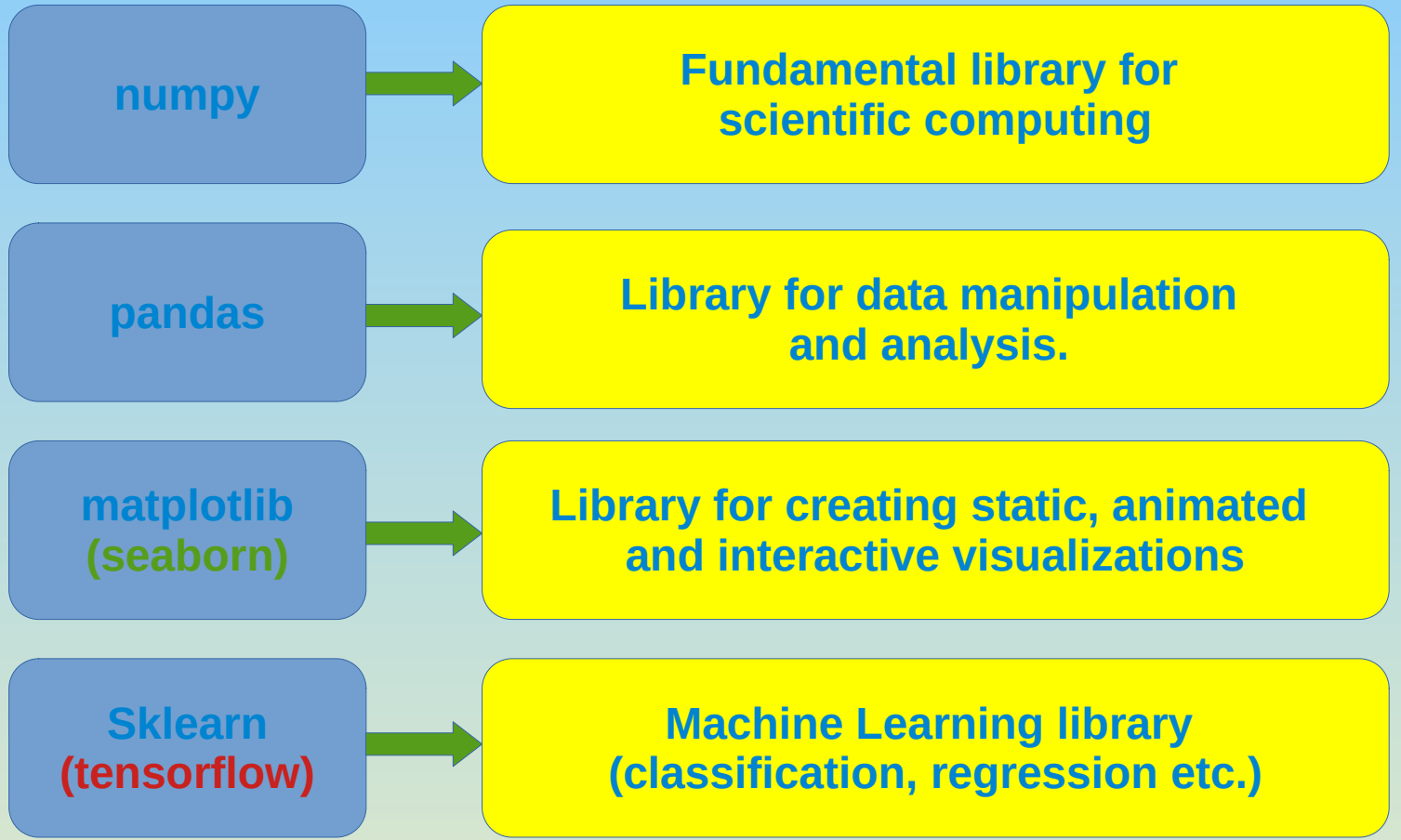
	name	density	pld	lcd	vsa	voidfrac	porevol	probe1	probe2	probe3	probe4	adsorption
2	MOF-00001	1.525	2.50	4.46	0.0	0.23	0.26	1.102	1.825	4.665	6.886	51.4
3	MOF-00002	1.566	2.44	3.54	0.0	0.17	0.24	0.896	1.256	0.754	0.013	3.9
4	MOF-00003	1.152	6.80	11.25	1185.3	0.69	0.56	1.459	2.049	3.193	5.755	34.1
5	MOF-00004	1.761	4.14	4.82	646.2	0.35	0.28	0.971	1.211	1.574	1.947	20.2
6	MOF-00005	1.788	4.08	4.81	591.7	0.34	0.27	0.994	1.181	1.487	2.086	20.1
7	MOF-00006	1.812	4.05	4.77	598.4	0.34	0.27	0.935	1.242	1.556	1.876	36.9
8	MOF-00007	1.848	4.02	4.74	547.5	0.34	0.26	0.925	1.183	1.579	1.921	20.0
9	MOF-00008	1.906	3.97	4.73	523.0	0.34	0.25	0.984	1.175	1.608	1.885	36.0
10	MOF-00009	1.929	3.94	4.70	486.1	0.33	0.24	0.965	1.240	1.525	1.845	20.7
11	MOF-00010	1.970	3.92	4.70	460.6	0.33	0.24	0.987	1.206	1.510	1.886	20.8
12	MOF-00011	1.918	4.05	6.23	1454.0	0.49	0.30	0.833	0.947	1.112	1.430	5.5
13	MOF-00012	1.347	3.82	4.33	329.4	0.20	0.33	0.633	1.039	1.708	1.286	28.7
14	MOF-00013	1.158	4.03	5.08	1007.6	0.55	0.53	1.455	2.076	2.992	3.766	43.3
15	MOF-00014	1.219	6.10	6.86	948.0	0.37	0.42	0.921	1.315	2.207	4.050	33.6
16	MOF-00015	1.590	3.01	4.47	0.0	0.26	0.27	0.688	0.906	1.371	1.821	22.4
17	MOF-00016	1.787	2.42	3.96	0.0	0.20	0.23	0.640	0.756	0.717	0.272	10.0
18	MOF-00017	2.602	2.95	3.60	0.0	0.10	0.09	0.448	0.687	0.711	0.857	17.0
19	MOF-00018	1.735	3.73	5.10	569.5	0.29	0.27	0.741	1.023	1.419	2.209	24.5

~5000 MOFs

- ID
- Structural Descriptors
- Probes Descriptors
- target

Development of ML models in Python

Python Libraries



Using Aristotle HPC

About Jupyter

https://hpc.it.auth.gr/applications/jupyter/

Setting up environment for ML

```
mkdir envs
cd envs
python -m venv myCustomEnv
source myCustomEnv/bin/activate
pip install --upgrade pip
pip install jupyter
python -m ipykernel install --user --name my-custom-env --display "My Custom Environment"
pip install pandas numpy matplotlib seaborn scikit-learn
```

About the problem under study