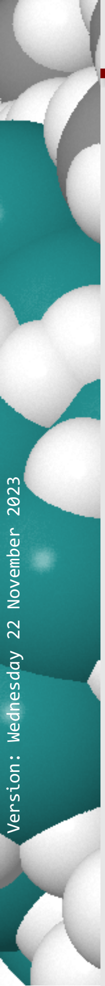


Version: Wednesday 22 November 2023

Python in the Physical Chemistry Lab

[pyPhysChem]

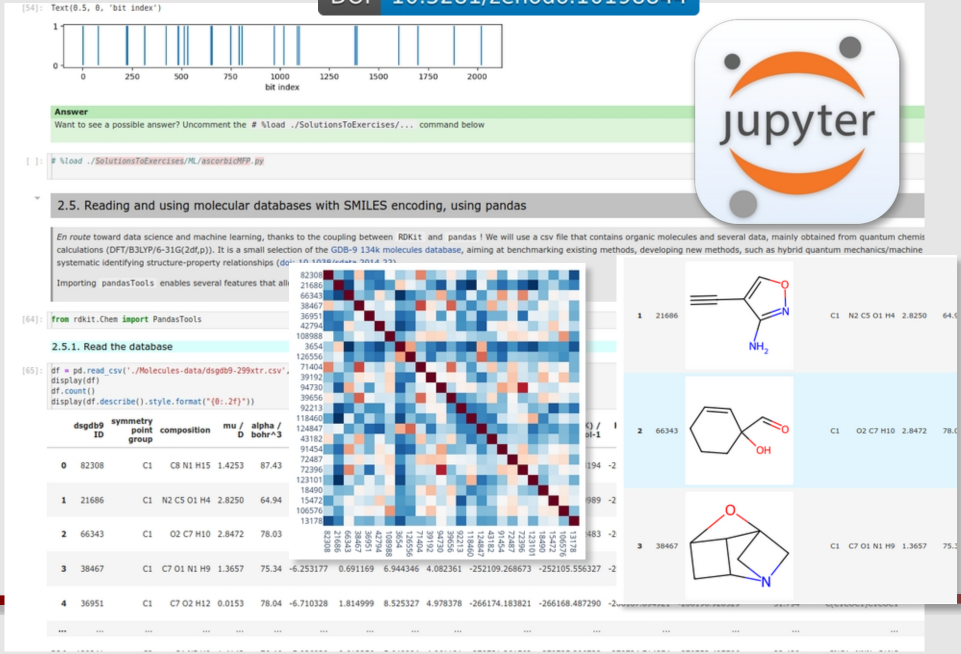
Université de Toulouse

"Talktorials" in physical chemistry and data science / machine learning

S. Christodoulou, Iann C. Gerber, F. Jolibois, R. Poteau

Python in the Physical Chemistry lab (pyPhysChem) github repository, release v. 1.8.0 (2023)

DOI 10.5281/zenodo.10198844



```
[54]: Text(0.5, 0, 'bit index')
```

Answer

Want to see a possible answer? Uncomment the `# %load ./SolutionsToExercises/...` command below

```
[ ]: # %load ./SolutionsToExercises/ML/ascorbicMFP.py
```

2.5. Reading and using molecular databases with SMILES encoding, using pandas

En route toward data science and machine learning, thanks to the coupling between `RDKit` and `pandas` ! We will use a csv file that contains organic molecules and several data, mainly obtained from quantum chemis calculations (DFT/B3LYP/6-31G(2df,p)). It is a small selection of the GDB-9 134k molecules database, aiming at benchmarking existing methods, developing new methods, such as hybrid quantum mechanics/machine systematic identifying structure-property relationships (doi: 10.1038/sdata.2014.22).

Importing `pandasTools` enables several features that all

```
[64]: from rdkit.Chem import PandasTools
```

2.5.1. Read the database

```
[65]: df = pd.read_csv('./Molecules-data/dsgdb9-299xtr.csv',
      display(df)
      df.count()
      display(df.describe().style.format("{0:.2f}"))
```

	dsgdb9 ID	symmetry point group	composition	mu / D	alpha / bohr^3							) / ol-1	I
0	82308	C1	C8 N1 H15	1.4253	87.43							194	-2
1	21686	C1	N2 C5 O1 H4	2.8250	64.94							989	-2
2	66343	C1	O2 C7 H10	2.8472	78.03							483	-2
3	38467	C1	C7 O1 N1 H9	1.3657	75.34	-6.253177	0.691169	6.944346	4.082361	-252109.268673	-252105.556327	-2	
4	36951	C1	C7 O2 H12	0.0153	78.04	-6.710328	1.814999	8.525327	4.978378	-266174.183821	-266168.487290	-2	
...	...	...	...	...	...	...	...	...	...	...	...	...	...

1	21686		C1	N2 C5 O1 H4	2.8250	64.9
2	66343		C1	O2 C7 H10	2.8472	78.0
3	38467		C1	C7 O1 N1 H9	1.3657	75.3

- interactive python
- images / videos
- mathematical equations
- enriched text (markdown)

Université de Toulouse

Motivation

integration of verbal explanations with numerical demonstrations or computer algebra system-based demonstrations proves to be an influential pedagogical tool

let’s call them “talktorials”

Université de Toulouse

Python in the Physical Chemistry Lab

[pyPhysChem]

python

talktorials specifically tailored for computational chemistry and data science/machine learning

such mixing is not new, but until recently it was restricted to rather simple applications that required a great deal of development effort

What is new is:

- the combination of Python's popularity and libraries

- the interactive nature of Jupyter Notebooks

- personal computers performance; the prevalence of real-world applications that can quite easily be adapted for students thanks to Python libraries available in a lot of domains

- the strong community support

- the ease of reproducibility that makes tutorials more effective, as learners can directly use the code provided to experiment and build upon it

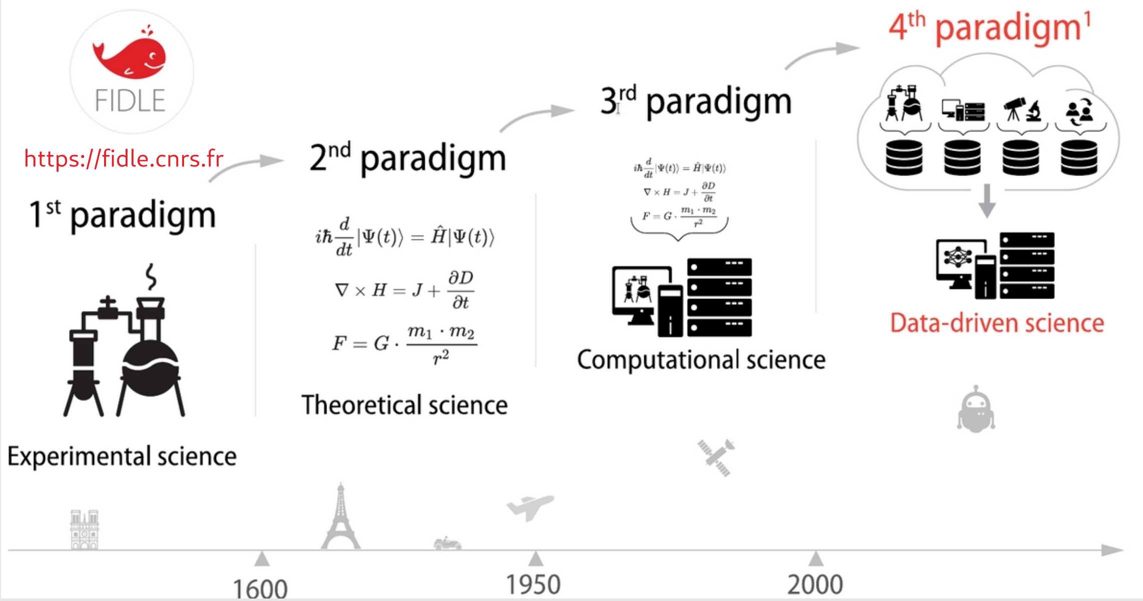
Version: Wednesday 22 November 2023

Motivation

The emergence of such innovative approaches in the realm of computational chemistry is truly encouraging

It not only enables learners to grasp theoretical concepts but also offers a practical perspective on their application

For students specialising in computational chemistry who develop their own scripts, they acquire a dual skill set that could be sought after in various areas of research and industry



Which students (so far...)?

masters' (graduate) degrees

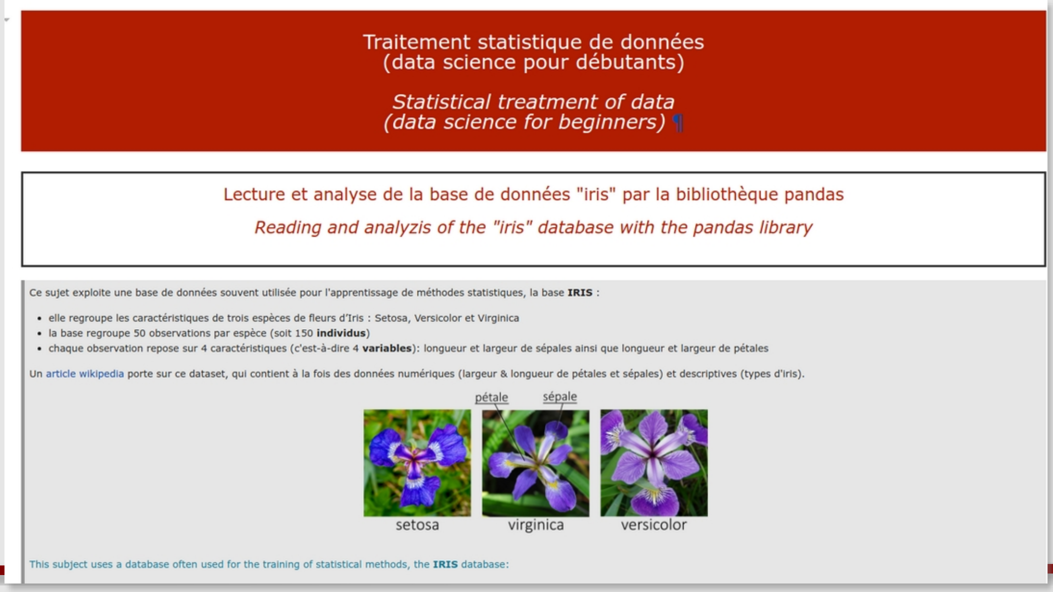
NANOSCALE SCIENCE AND ENGINEERING

Théories · Modélisations · Simulations · Atomistiques

Réseau Français de Chimie Théorique

bachelor degree

all chemistry students, 2nd year



Traitement statistique de données
(data science pour débutants)

Statistical treatment of data
(data science for beginners) ¶

Lecture et analyse de la base de données "iris" par la bibliothèque pandas

Reading and analyzis of the "iris" database with the pandas library

Ce sujet exploite une base de données souvent utilisée pour l'apprentissage de méthodes statistiques, la base **IRIS** :

- elle regroupe les caractéristiques de trois espèces de fleurs d'Iris : Setosa, Versicolor et Virginica
- la base regroupe 50 observations par espèce (soit 150 **individus**)
- chaque observation repose sur 4 caractéristiques (c'est-à-dire 4 **variables**): longueur et largeur de sépales ainsi que longueur et largeur de pétales

Un article wikipedia porte sur ce dataset, qui contient à la fois des données numériques (largeur & longueur de pétales et sépales) et descriptives (types d'iris).

This subject uses a database often used for the training of statistical methods, the **IRIS** database:

Version: Wednesday 22 November 2023

Summary & Outlook

Version: Wednesday 22 November 2023

Python in the Physical Chemistry Lab

[pyPhysChem]

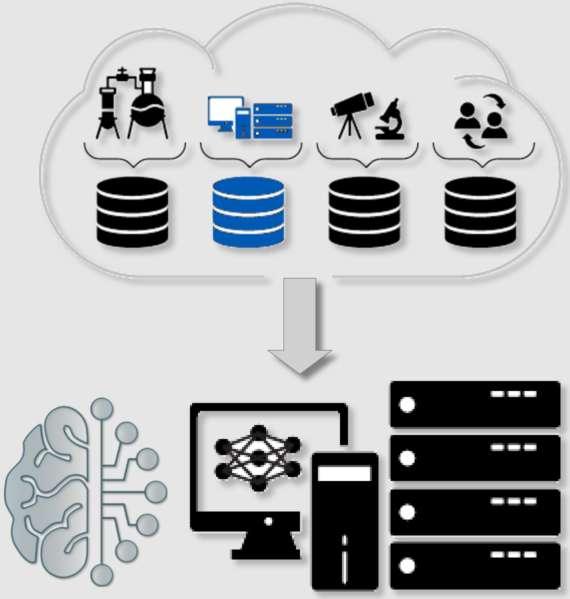
Université de Toulouse

https://github.com/rpoteau/pyPhysChem

DOI 10.5281/zenodo.10198844

- python for physicists and chemists in a nutshell
- Computer Algebra System
- Physical chemistry (incl. quantum chemistry)
- coding and use of representations of molecular structures and related data
- Data science and ML

can we expect a strong convergence between quantum and computational chemistry, data science and machine learning?



data-driven science

increase in the number of students in the master's programmes in theoretical and computational chemistry?

… unless we do not really give them a dual skill set that could be sought after in various areas of research and industry