Task 3: Representation Matters — Fingerprints, Descriptors, and Beyond.

Weeks: 6-7
Title: *"Do You Know What You're Feeding Your Model?"*

Objectives:

- Interpret molecular representations
- Assess feature relevance and usefulness for the task
- Explore alternatives to default representations

Deliverables:

- A report explaining what features exist in Mordred and MACCCS
    - Do you think they are relevant to your problem? (justify your answer)
    - Do you think there are other features that need to be considered? (provide examples)
    - **Push a pdf report to the GitHub repo by 6 PM the Wednesday before next session**
- Update your Jupyter notebook with a concise introduction to the representations you are using.
    - A general idea on what they try to describe
    - An assumption on how much you think they are relevant to your problem
    - A suggestion for few more features to consider
    - **Push your update to the GitHub repo by 6 PM the Wednesday before next session**
- A presentation where you explain your findings

Description

After thoroughly inspecting the problem and finding the holes in your understanding, you realized that it is not advisable anymore to take things for granted. Just because something has been done in a certain way for a long time does not mean it is correct. You understood that a dataset published in a reputable journal is not necessarily a good source. You

understood that data is way more than a simple column. And now, with your new mindset, and before jumping back to implementation, you keep the game of doubt, and you ask:

But what about the representations we used? Are they truly good, or will they be tricky just as the dataset was?

You decide to investigate the descriptors you used to understand what they really are. Each set of descriptors (Mordred or MACCS) is constituted of columns. Each column encodes a certain feature about the molecule. You decided to answer the following questions:

- What does a representation mean?
- What are these features in our representations?
- Are they what we need to describe a molecule to predict solubility?
- Are there irrelevant or redundant features?
- We now know a lot about solubility, are there features that we hoped to see, but we didn't?

Important notes for your deliverables

- You are welcome to use any language model for this task.
    - However, it is your full responsibility to make sure that the references used are accurate and reliable.