Task 4: Deep Dive — What's in Your Data?

Weeks: 8-9
Title: *"You Say Data, I Say Chemistry"*

Objectives:

- Critically assess when a curated dataset is good or not for a given purpose
- Improve data quality through curation

Deliverables:

- A report explaining the state of the training dataset you were handed
    - What types of source variations are there? Would you characterize them as inconsistencies?
    - Can you think of reasons for this variability? What's the "intention" or story behind the database itself?
        - Was it "greedily" compiled with as much data as possible?
        - Or was a created with a specific goal in mind?
        - Was any curation used at all for creating it?
    - How to account for (correct) these inconsistencies?
    - Now that you have the full table, what are OTHER viable ways to prune the dataset? (more on this below)
    - **Push a pdf report to the GitHub repo by 6 PM the Wednesday before next session**
- Create a new dataset with better curation
    - If not possible, push whatever you tried but failed.
    - **Push your new dataset to the GitHub repo by 6 PM the Wednesday before next session**

Description

You are still on the mission of helping your CEO harness the power of machine learning. At the beginning of your journey, machine learning was as simple as data → model → prediction score. But now you know it's much more...

The chemistry

You inspected the chemistry of the problem, and you realized it was a tough one.

Yes, we want to "replace" experimental setups with "efficient" machine learning. But if one doesn't know what they are replacing, one may never succeed.

And if one doesn't understand and respects the complexity of what is happening in the lab, one may also never succeed.

The representations

You then understood that a model does not consume chemistry the way humans do. It consumes it through the representation we use for a molecule. This representation can make or break our model. Or they can be bland, not harming but also not benefiting.

The answer to whether one representation is good or not will depend on the task one is performing. Therefore, knowing the chemistry of the problem is a prerequisite for knowing the quality of a representation.

The evaluation

You then understood that evaluation cannot be done properly through a single number. All single numbers try to summarize a big picture. It's helpful only when the big picture is nice and clean. But from all what you have understood so far, the big picture in our field is noisy and uncertain.

Therefore, proper model evaluation must go beyond single numbers and towards behavior assessment. Where does the model succeed and fail? And why?

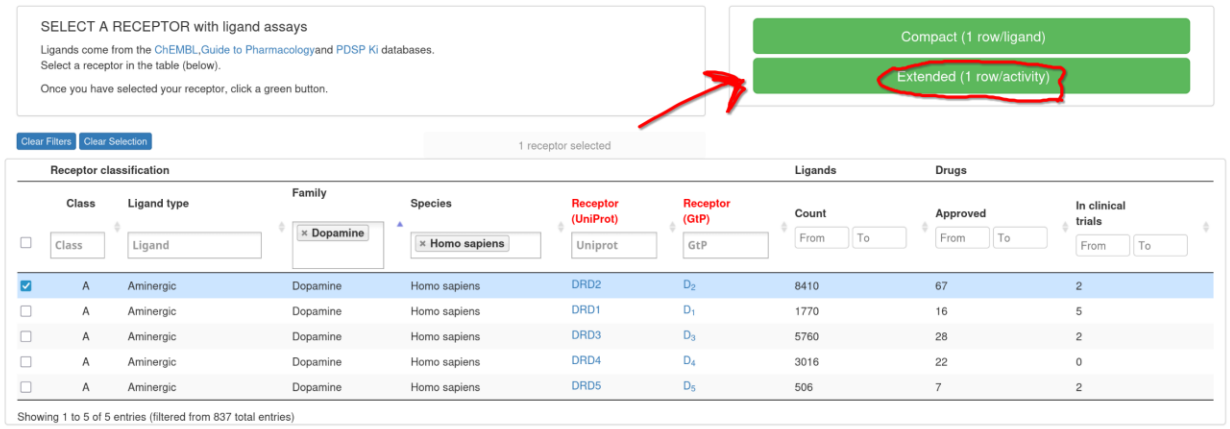
The data

The last piece of this big puzzle is the data. The data is what is given to the model and what is "thought of" as representing our problem

- Your first task would be to define what "data" is? How is it different from "the chemistry" and "the representation"? Include this in your report.

Then, you will do a deep dive in the data given to you to understand what it is, where it comes from, and whether different experiment types can be "mixed" for the task the CEO has given you (or not)!

For background, your data came from this portal and filtered down to the receptor you studied through these steps.



Instead of giving you the full dataset, we excluded some assays randomly so that you train your data on the majority of the filtered compounds and test it on a new one. We will attach the full datasets with their metadata for your reference.

You need to write a report explaining potential ways to group “similar” experiments. Your task is to come up with ways to group experiments by similarity, s.t. subsets of data “similar **enough**” can be used for the model. For this, you can look at other columns of the dataset and find out what they mean. Then, you could determine what is the experimental setup used by most experiments. You can be creative about how to cluster the data:

- Exact string-matching of the experiments’ description? Is this maybe too strict? Does the dataset break down into many subpopulations?
- Publication of origin? Is this robust with respect to publications which themselves aggregate other third-party sources?
- Fuzzy-string matching of experiment descriptions, like e.g. with https://github.com/rapidfuzz/RapidFuzz ? How would you cluster those?
- Throw everything into an LLM and see how it clusters experimental setups? There is nothing intrinsically wrong with this; you just need to convince your CEO not to substitute her workforce (=YOU) with the LLM. So you need to be able to explain to her what the LLM is doing.

One way or the other, you need to answer your CEO:

- Was the original database a good starting point for this task at all?
- Do you have tools in your toolbox to pre-process or prune the data to train a good-enough ML model?
- What things would you suggest to the GPCRdb in case they changed their approach on how they curate the data?

Finally, let's see if one can put these proposed solutions into action. Provide two different ways to group the input data into subsets with their respective rationalization.

You will not be judged on whether the model works well or not, but on your capacity to think critically.

Remember: your CEO would rather keep good employees capable of knowing when to stop and look for better data than having employees willing to keep a bad model because "it performs well".