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**
- □ A presentation where you explain your findings

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

You received three datasets; the first one is called AqSolDB, and it was obtained from this publication. After all the chemistry you have learned, you are now slightly equipped to look at this paper and judge its data curation quality. You can consult this paper for inspiration. But, if you apply your critical abilities, you can spot clearer problems with the curation of the AqSolDB dataset.

The other two datasets are coming from these two different sources. Challenge data 1 is coming from this entry on ChEMBL and the second challenge data is coming from this publication.

You need to write a report explaining

1. How the AqSolDB was curated?

2. Is there an apparent problem in the way it was curated? If yes, how will it affect the training of the machine learning model?
3. What are the challenge datasets? Are they different or similar from the AqSolDB in terms of experimental measures?

Once you identify the problems with the curation, this is where the solutions can kick in. They say the first step of solving a problem is identifying the problem, and this is not a cliché at all.

If you were the person to write the AqSolDB paper, and you now have some updated experimental knowledge, how would have you curate the data differently?

- What would you have accounted for when picking and collecting the molecules?
- What features or metadata will you include as a default important Info?
- How would this have helped you train the ML pipeline?
- Would this new curation ensure improved performance on the challenge dataset?
- Include a section in your report where you explore such questions.

Finally, let's see if one can put these proposed solutions into action. You will need to go through the different sources where the AqSolDB paper went through and see if you can retrieve the same molecules but with the enhanced curation you envisioned.

- If you managed to curate the whole (or part) of the dataset with your newer high quality curation scheme, push this dataset to your repo.
- If you couldn't manage, extend the report with the approaches you tried but failed you.