

**1. What is the main differences between ML and AI. In some paper they are used together.**

A: ML is subfield of AI where systems learn patterns from data. So ML is inside AI. Although they are kind of used interchangeably. (RLRC)

**2. As AI models become more capable and autonomous, how do organizations balance the need for speed and efficiency with the growing demand for transparency, accountability, and human oversight, especially in high-stakes decisions?**

As IBM famously said in 1979 “a computer can never be held accountable, therefore a computer must never make a management decision”. Decision-making and actual scientific insight are for humans, but we can use AI and ML to help us make better decisions by allowing us to get more insights across more data than a single human can manage. (NC)

**3. Thank you so much for your presentation, i have a question about what is a good k-fold value, is 5 a common one. and also, whether the number of random state affect strongly on the model**

A: It depends on the size of your dataset, but it is normally is between 5 to 10. 5 is normally for larger datasets. (RLRC)

**4. Is the data processing and curation done using any particular software.**

A: In practice, for most data processing and curation I use Python. you can check this as an example of notebooks for data curation: <https://github.com/openkinome/kinodata> (RLRC)

I also use Python. When dealing with lots of numbers, I would recommend using data visualization (Spotfire, Plotly, Power BI, or your favourite dataviz tool) as you go, because aggregate statistics can hide trends and errors. (NC)

**5. I am chemistry teacher in high school. I teach. Chemistry but I am interested in research of medicinal plant also. So, how AI and machine learning will help me to integrate the studying and working in it?**

I've personally found NotebookLM a great tool for helping me learn. Rather than rely on an LLM for learning, which can cause problems, you can use it to help you find relevant concepts in very large source material. I am using it alongside reading a textbook and training from an expert. (NC)

**6. Comparing R and python which one is more efficient for curation and which one is more reliable?**

Both are really great languages and I would not want to recommend one over the other! It's fair to say Python generally has more specialist libraries and is more widely used in ML, but R has a strong community and [reticulate](#) allows calling Python from R where need be. The best language is probably the one you're most comfortable with. (NC)

**7. I am a senior Biochemistry student and currently working as an R&D technician in a physical chemistry lab. I have been teaching myself Python to automate my lab data and simulations, but I lack a local mentor or a community in my country. From your perspective, what is the most critical skill I should focus on to bridge the gap between 'wet-lab' experience and AI-driven computational chemistry? Also, which open-source communities would you recommend for a self-taught student to find professional guidance?**

That's really great! There are thousands of Python tutorials and training sessions online. If you're interested in computational organic chemistry, I think the strongest skill is actually having a lot of knowledge and intuition about organic mechanisms, as expert domain knowledge is always in short supply. Chemistry social media has declined in recent years, but Bluesky has some great [computational chemistry starter packs](#). (NC)

**8. Ok Thank you, I also wish to know how you clean the data. The amount of data in your case is nearly 2 lack. I also wish to know is the data input given in excel**

I use the pandas and rdkit modules in Python, and often write my own chemistry cleaning functions, plus validate from multiple sources like ChemSpider and PubChem where possible. For specialist matters (metal catalysts, in my case) I also curate things manually in great detail. (NC)

**9. Thank you very much for your presentation! Is there a programm you use or suggest for ELN and and/or knowledge managment tool that could be implemented easier into ML?**

It very much depends on your use case, but TU Darmstadt's [ELN Finder](#) is a useful tool! Not all ELNs will provide level of the FAIR data you need. (NC)

**10. when we have limited dataset, what is good strategy to curate the dataset that would help to eliminate the noise in training, but also keep it to be sufficient, any advice. if you cannot throw any data point due to limitation, what is good strategy to minimise the noise ?**

There will always be noise. For reaction data, we can increase the input amount and also ensure it is high-quality (from reliable high-throughput experimentation, with lots of useful metadata). Even without this, it is good to "fill in" sparse data areas wherever possible; sometimes synthetic data will be sufficient. You might want to design a quality

score metric and try varying the cutoff to see where you get your best results, without overfitting. (NC)