Q&A
3 November 2021
Machine Learning and AI for Drug Design
Professor Ola Engkvist (AstraZeneca & Chalmers University)

Q1: Any ideas on how techniques used in small drug molecules can be scaled to the protein level?

I think there's quite a lot of things going on for antibody design. I think it's a much more difficult problem you have less data, and the algorithm probably needs to be very sophisticated. To move into peptide space, you can start to approach it in the same way There's a lot of research going on and I think it's fair to say they are a few years, behind the small molecules, but they are combining machine learning and AI with physics-based modelling, and it will ultimately have an impact.

Q2: I thought transformers need a huge amount of training data. What do you use for training data when you are using a transformer for molecular optimisation?

In the article (https://jcheminf.biomedcentral.com/articles/10.1186/s13321-021-00497-0), we looked to generate matched molecular pairs for the whole CHEMBL database and used it as training set. Clearly, it's a fair point that the transformer needs a lot of data, but you can also apply pre-training tricks to learn it on unlabelled data and then you only need to have a small set of labelled data, but it worked very well when we generated matched molecular pairs from the whole CHEMBL database, and we published the results earlier this year.

Q3: Great talk! Has this generative process described for creating structures from SMILES been tested for metal catalyst design? Would you expect this to perform with similar success?

We have not tested it, but I don't see any reason why it shouldn't work, if you have a good dataset to start with, and you know what you would like to achieve you can score the proposed catalyst and it might very well work. I have not done it myself, but I don't see any fundamental differences why you cannot train on completely different chemicals space. In drug like chemical space as well as for catalysts or inorganic compounds.

Q4: Great talk. Question about the MELLODDY project. Did they get it right that you used just very like a feedforward standard neural network but just very wide?

It's a feed forward type of network, yeah.

Why was the motivation instead of, for example is not graph convolution some, you know a little bit more modern?

It's multi-task learning so I would defend it, it's quite modern. We have done a lot of tests of the different algorithms, and I hope some will be published

Maybe the last, maybe a little bit controversial. Don't you think that the pay-off you know the delta AUC that you get out of the massive effort? Yes, it's statistically significant. But it's rather minor?

I agree for year two we just showed the principle that it works, that you have a statistical difference. Year three is about optimizing the difference between the multi-pharma and single pharma models. And I think it will need to be meaningful differences between a single and multi-pharma model. So, say if you use it internally, we'd use the models together with REINVENTt. I would like to see different molecules design with a multi-pharma model versus a single-pharma model to call it a success. So, I think it's fair to say we reached the year two milestone, but for year three to be a success we will need to see a significantly larger difference.

Q5: I know that your company is investing significantly in supporting universities in training students for the new world. What would be, in your view, of the ideal kind of training for the new version of a chemist that's coming forward?

I think that they need to be more fluent in automation and data analysis. I don't think everybody needs to be a fluent python programmer, but I think really to emphasise more automation and statistical analysis. I think that there's also other aspects, it's also important to understand that we also work with new drug modalities now in the pharma industry, including a lot of different variants of nucleotide therapy. So, I think it's also important to get a broad education and courses, so the students actually understand that, that it's not the only machine learning AI and Automation that is changing in the pharma at the moment it's a much versus a broader view of drug modalities now then we had maybe 10 years ago.

I think it's very interesting because we see this push for that the more computer friendly side, and I think the students appreciate that and see what that leads to. But that's not at the expense of learning that biochemistry side, which maybe not all chemistry courses can cover, and I think your point about automation is very interesting because yes, we're upscaling. Many university labs and teaching laboratories are beginning to use much more modern equipment, but actually it's about teaching the students how to automate things, which is a slightly different aspect as well. But of course the degree is limited in time, so we'll have to try and work out how to give them the principles

There's a lot of things happening at the same time and to be able to cover that in a good way and also keep the basic understanding it's definitely a challenge, I agree.

But one we need to face, I mean clearly people work in teams they don't have to be experts at everything, but they have to know enough to be able to work in the team. I think it's that that's what we should be aiming for.

Absolutely.

Q6: As the reproducibility is sometimes a problem with already existing data, training neural network with such defective data may lead to a bad prediction? How do you think it can be addressed?

I think uncertainty quantification is very important. We need to apply it much more consistently so it's not only about the accuracy, it's also about the quantifying the answer alternatives. Also, of course, interpretability can help that the model prediction is for the right reason. But, of course in the end it is data from experiments. It is the experimental settings, and we have to explore method that take the experimental uncertainty into account. Like a version of probabilistic random forest. We tried actually to model the probability distribution from the experiment which you can do with the method. I think in the end we are still in the in a lucky situation. If we do a wrong prediction, we might synthesize the wrong molecule even though you want to synthesize the right molecule, but if it happens that you synthesized the wrong molecule the damage is limited. It is much different if you work in the clinical setting where you need to really be 100% that you make the right decision. We can do the wrong prediction once in a while for the next molecule to make, but we should of course do much more right predictions than wrong predictions.