Q&A
17 November 2021
What a Medicinal Chemist Needs to Know about Explainable Artificial Intelligence
Dr Alexander (Al) G. Dossetter (Medchemica Ltd)

Q1: We see non-additivity in many cases that limit the applicability of this approach. How can you explain that? Can you tell which cases this approach will work and where should we avoid it?

For our particular technology, we supply the stats data going forward and you have the ability to drill back to the original data, particularly for the matched pairs. This enables you to see the compounds that are most similar to what you're working on. So, we argue further by saying "which of the applicable pairs that are nearer to the chemical matter that I'm working on?" That helps us, kind of, addressed that problem. But overall, we present the data, "this is what's happened globally". A lot of what we focus on, during our training, is to be able to go into that data and understand it, so users are able to make that judgment. To give you a sense of scale when I first started using matched pairs, we spent about three days pulling all the data together, to then be able to make a decision about how applicable the pairs are. So now the first thing we're doing is organising all that data to be able to make the better decision. But overall, you do touch upon quite an important challenge in this world.

Q2: Do you work only with Pharmaceuticals? What chemical descriptions do you use for the compounds?

Our client base is large pharma all the way down to individuals in universities. Our tools are available online and you only need a web browser to use them. Optimized for Google Chrome, and it does work in Safari and Firefox and Edge. So, descriptors? On the slide I had earlier on, we actually encode the fragments of molecules with all other hydrogens in place as absolute structures for matched pair analysis. That is the only way it works; you need that level of precision. As a result, our databases are 500 gigabytes in size, basically there are 40 million chemical transformations encoded on there and that's the way to make sure the computers aren't biased. For the machine learning methods, and I briefly touched on it, the descriptors we first came up with produced a brilliant system, amazing. We could take fragments with the linker to another fragment and all the hydrogens described. Fantastic, brilliant models, and very accurate predictions, but very easy to fall out of domain with a different chemical series. So, we changed the descriptors, and made them simpler, to the likes of hydrogen bond donor, hydrogen bond acceptor, aliphatic group, aromatic group and atom linkers. And by doing two descriptors with the linker chain between them, we encode much more accurately, but we have softened the accuracy of the model so that it means when a new molecule comes in that it's never seen before it can provide some sort of prediction, and more importantly, the basis of how those predictions come about. So have explainable models. Just to build on this further, sometimes people go "do you not include molecular weight and do you not include lipophilicity, and types of descriptions like these". No, we don't, and that's very deliberate, because what we found is you're double counting the chemical groups. This leads to predictions with higher errors. Because we have been medicinal chemists in large pharma and still are working on live projects and have been through years of QSAR and understand the domain and so be able to organize and think about descriptors very carefully before building models.

Feel free to drop us an email anytime and we can have a chat further on zoom or find me on LinkedIn!