Q&A
1 December 2021
Chemical Space Exploration
Professor Jan Jansen (University of Copenhagen)

Q1: Did you have any problems with maintaining diversity in your populations? And how did you deal with it?

Yes, we do. Each individual genetic algorithm search tends to lose diversity a lot. You're zeroing in on one particular target, but if you run different genetic algorithm searches, it zeroes in on another molecule. So, if you combine the final populations of several runs, then you have fairly good diversity and so the question is now, are there ways to increase or force diversity? Basically, by changing the including diversity in the scoring function. The question is, is it better to run fewer genetic algorithm searches, but with this diversity parameter? Or is it better to run many genetic algorithm searches without the diversity criterion and then combining the final populations. So that's one of the things we have to look at. If you only can afford to run one genetic algorithm search, then yes, you need to somehow ensure that genetic population is diverse, so that can be done, just by insisting that they have, for example, molecular weight, different molecular weights, or different properties that are fast to compute.

Q2: In GA, how do you determine the population size?

That's trial and error and also depends on how expensive your scoring function is. So, if you're scoring function is very expensive, you typically have to keep your population size low and also the number of generations low and just hope for the best. So basically, it's an empirical parameter. You can help ensure that small populations work by doing some pre-screening. So instead of justice randomly populating your initial population, you could do some pre-screening of the property or interest to them and then only include high scoring molecules that are available or known in your initial population. And then you could probably get away with a smaller one, but there's a lot of empiricism in these genetic algorithm runs.

Q3: How often do the TS's found by xTB give more than 1 negative frequencies with DFT?

So, we haven't kept track of that, how often does the transition state search succeeds? Usually when it doesn't succeed it's often because you have two or more imaginary frequencies. We haven't really addressed that problem specifically, but usually the success rate is fairly good - better than 50%. Basically, it depends on how much time and effort you want to spend on finding that transition state. So, if it looks like it has a very good score at the semi empirical level, there are many things you can try to get your transition state search to work. But if you just do a first pass and give it 50 molecules, then maybe 25 will work at the first shot, and then the rest you have to go in and deal with a manually. That is by far where we spend most of our computational efforts. It's not really in the genetic algorithm, it's doing the DFT refinement for the final population. We've actually gone in now, and also with the catalyst search, we also have a synthetic accessibility measure in there. Before we start looking at transition states we really do as much sort of synthetic accessibility analysis, so if you can't make the molecule anyway, there's no point in trying to find a transition state. But if everything lines up, you have a molecule, at the semi empirical level, that's very promising. It looks like it's easy to make then we would spend a lot of effort trying to find the DFT year transition states to verify it.

Q4: Are there any methods which you applied to your GA to accelerate the evolution?

No. So far, a straight 'plain vanilla' genetic algorithm has worked. The one thing we have messed around with a bit is the selection criteria. So, when I say we pick according to score, there's actually many ways of doing that. There's where the probability is directly proportional to the score. You can also just rank them, and then pick according to the rank in the population and things like that. And for some scoring functions that has some effect. But some of the other accelerants that's out there, we haven't really had to use it yet because when we run this sort of very simple genetic algorithm code, we get enough promising candidates to try to refine that the DFT level, and that's where we

end up spending most of our time, not sort of generating more things to try right, but actually trying them.

Q5: For the total chemical space of 10^60, does that include molecules of all size? About what is the chemical space size for reasonable drug sized molecules (with say less than 50 carbons)?

This is for that kind of space, this estimate was basically made for drug-like molecules, so these are small molecule organic with not too many different atoms, so there's no high molecular weight in there and the number is not large because you have high molecular weights or because you have strange atoms like silicon or boron, these are just plain organic molecule. But there are so many of them because there are so many different combinations of functional groups that can be combined in so many different ways. So, personally if it's just possible molecules, I actually believe the spaces is even bigger. It's not clear what percentage would be judged synthetically accessible? That's the question, and that's why you often see lower estimates of this space. I think that's actually when people start trying to estimate the percentage of synthetically accessible molecules. But of course, that will change as the synthetic machinery gets better.

Q6: Do you use xTB's optimizer or ORCA's optimizer?

xTB. I think the xTB optimizer is fine, but the real issue here is the speed. And by speed I also mean the set-up getting the program loaded into memory - getting it started. If you're doing thousands of very fast xTB calculations, these things start to matter and then these big quantum programs that are huge, and you're loading a lot of stuff into memory that you actually don't need for the xTB calculations, those things that are not really time consuming, but they are if you're doing them, thousands of times.

Q7: I missed how he got the SA values, mind showing it again?

there will be a reference in this paper (Steinmann, Casper, and Jan H. Jensen. 2021. "Using a Genetic Algorithm to Find Molecules with Good Docking Scores." Peer J Physical Chemistry 3 (May): e18.). But it's basically just an analysis of molecules that have been made, that you can be made, but fragments are in there, and if you have a high percentage of those fragments in your molecule, then it'll judge it to be synthetically accessible. But it's an old method, at least 10 years old or something like that. And the reference will be in here (Ertl P, Schuffenhauer A. 2009. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics1(1):8.).