Q&A
8 December 2021
Audacity of huge: Machine Learning for the discovery of transition metal catalysts and materials
Professor Heather Kulik (MIT)

Q1: Is the machine learning model you used with DFT suitable for other quantum chemistry methods like MP2 and CCSD(T)? Please explain why or why not.

So, in a different part of my research program, we spend a lot of time asking ourselves where do errors come from if our materials are strongly correlated, where should we have been using DFT models? The reason that everything I show is with DFT is because the computational cost is so high to generate a few thousand points with coupled cluster, and in transition metal chemistry there's a less obvious hierarchy of correlated wavefunction methods sitting above DFT. Some of the things that we've done, it absolutely is amenable to training on those datasets. Some of the things we've done is we've predicted strong correlation with multireference diagnostics over these spaces, so we've trained models and then looked at hotspots of strong correlation, that's one thing, because then you can identify that you definitely need to go beyond DFT. Implicit in the double hybrids is MP2 correlation. So obviously MP2 is feasible, but it tends not to perform better necessarily on its own for transition metal chemistry, and then one of the things we've done is we've used the same approach to learn the difference between the DFT answer and the coupled cluster answer and identify when that difference is going to be large. And that's a paper that's coming out shortly and in that work what we've done is we've focused on identifying places where there's an imbalance in the degree of strong correlation between the points being compared. That's when going beyond DFT is most essential. And naturally, we've also shown some machine learning of experimental data. We've also done that in the context of spin crossover and predicting properties that are less sensitive to the DFT functional. So, there's no inherent limitation except that we have to have a bunch of high-quality data that we really trust, and historically that hasn't been something that we've been able to get on transition metal complexes that have 50 atoms or more.

Q2: How many electrons are in the DFT calculations you are doing and how much time do the DFT calculations take?

I guess I'm not going to be able to convert it to electrons. The biggest systems we were studying were 200 atoms, at least 100-something of which were heavy atoms. The smallest complexes we ever study are about 25 atoms, so they're all bigger than the small molecule sets that people typically study. In terms of valence electrons, you multiply all the organic/the heavy organic by at least six and then add with an ECP add about 24 for the metal centre, so a decent number of electrons. I should mention we use GPU accelerated DFT with the with the code TeraChem which we develop which helps accelerate these larger DFT calculations, certainly at the cost of making some other shortcuts on the basis set, but in general, these are big calculations. We're able to complete most of them in five days or less, but most of them take considerably more than a couple hours. Even with a fast DFT code.

Q3: What hardware and software are you using for the DFT calculations?

The first bit, actually there is one subtlety. The first bit I show we run everything through with a modest double zeta basis set and we use TeraChem, which is a GPU accelerated quantum chemistry code developed at Stanford and that we use our code molSimplify to run in an automated fashion. We write our own codes to automate that process, and that allows us to get a large number of calculations to converge. When I was showing the 23 different density functionals in that case, what we do is we actually take that calculation out of TeraChem and we pass it to PSI 4, mainly because we could write an interface to PSI 4, but we've also done this with Orca and other codes, and the main idea is that we translate and we keep the wavefunction as close to the original result as possible, and then we can carry out other density functionals that maybe aren't implemented in TeraChem. So, for instance, the Minnesota functionals are not in TeraChem and some of those double hybrids are not in TeraChem. So implicitly our approach is agnostic to the code that we use, we try to use whatever code is going to have the methodology we want, and we write our own. Our own molSimplify automatic design ensures

that it's pretty seamless to switch between these codes, but you know TeraChem to generate the initial geometry and stuff is chosen for speed. And we use a range of graphics cards so it is GPU accelerated and you can get better performance out of the high end of GPU cards, but we also we tend to use the low-cost gamer cards that do just as well. The benefit increases slightly if you do the higher end cards, but the cost increases dramatically, so we tend to do this on one or two GPUs. The bigger molecules may be on two GPUs. And just pretty low-cost standard cards, both locally on our own cluster as well as in in supercomputing resources.

Q4: What type of water or solvent model do you use with DFT?

We use implicit solvent models for everything I talked about including things like logP. We use a polarisable continuum model and that's, again, for speed. You could imagine certain properties if you care about explicit hydrogen bonding interactions. You might want to do more of a QM/MM approach. In other parts of my group, we worry about that, but for here even the logP I was showing for the RFBs that was the difference in the solvation free energy in a polarisable continuum model between water and octanol.

Q5: Do you favour the consensus vs delta learning approach for DFT data?

I think they both have promise, we've had trouble in the delta learning specifically figuring out what we put at the top of the ground truth. We have some small transition metal complex data that I mentioned we're working on putting together. It's relatively small, and it looks promising, but we have to treat coupled cluster as the ground truth or DLPNO coupled cluster as the ground truth, and if you read into literature, that's not necessarily something people consistently believe is going to be more robust than the consensus approach. So, I think the consensus approaches adds value and it's also cheaper.

Q6: I thought the low-cost gamer GPUs do not have FP64. Are you doing your computations in FP64?

We use mixed precision that's built into TeraChem, which takes advantage of the fact that most pieces of the quantum chemistry calculation do not require double precision all the time. But, as newer cards have come out, the difference between the double and single precision, speed, the differential has been less and less, so that has been less of a problem, but it was originally the problem you would think about with these cards. It's not something that comes up that often in quantum chemistry, and it comes up even less in molecular dynamics, so a lot of things are good enough in single precision.