Q&A
3 November 2021
Accelerating design of organic materials with Machine Learning and AI
Professor Olexandr Isayev (Carnegie Mellon University)

Q1: How do you treat long range interaction within the descriptor? Or those are neglected entirely?

So there are a few approximations in the original ANI neural network, long range interactions are truncated to six Angstroms. In new AIMNet architecture, it's all data driven, i.e. no physical descriptor and we do not have explicit equation. It's all learning implicitly by the neural network, by the design because essentially one little atomic environment passes message from another, and therefore you can feel the presence of another environment in a certain distance. Therefore, it's implicit, we don't have a specific operation, neural network essentially learns that by itself, end to end.

Q2: Are the torsion energies exploring an "extrapolation" behaviour of the network? As I understand, the sampling is done close to eq. geometries.

All examples I showed you is a here or there, they all extrapolation. So, we train a set of small organic molecules and up to 15 heavy atoms. These examples were specifically selected were molecules not present in the training data, and therefore it's all extrapolation and you can think about that building this map you can use it as a DFT or you can use neural network and obtain very similar answer.

Q3: For the 19F MRI agent project, did you predict water solubility with free-energy perturbation or something else?

In the first project the solubility was predicted through a straightforward ML model. It was a binary classification and the reason we use it as the binary was because it was difficult to measure solubility under pandemic conditions. What we did is essentially qualitative experiment, if a specified quantity of polymer diluted in the solvent, and solution was clear, we call it soluble. If it was cloudy or polymer precipitated, we called it insoluble. And hence we predicted solubility qualitatively, there was no simulation. It's trained to just a simple experiment.

Q4: How well do the models work on simulating enzyme catalysis? Have you explored such avenue?

Out of the box, it probably will not work for catalysis, because the neural network was trained only on ligands i.e., small molecules, and at this point there's no coupling to the protein. But what we are working on right now extends our small molecule forcefield to include proteins. And then you can do the simulation. The force field is reactive in principle, you can describe reaction if It's been trained to see chemical reactions, but at this point it totally fails because it's never seen a catalysis and reaction, and how it happens.

Q5: Can it also be applied to bond breaking?

Yes, given the proper training. We're working on fully reactive forcefield. But out of the box, for example learning model will not work because the training data does not have any reactive data. But this is more of a data problem rather than methodological problem. The neural network can describe, given the proper data, a chemical reactivity.

Q6: Can this approach be applied to proteins? Remodel flexible chains or even fold proteins?

Yes, again, it is a work in progress, and we plan to release the counterpart of little network work projects. Stay tuned!