Scaling law for machine learning of chemically functionalized metal organic frameworks

Yuta Aoki, Schrödinger, K.K.

Functionalization of linkers in metal organic frameworks (MOFs) can greatly increase the structural diversity of MOFs. Considering that the structural diversity of MOFs is originally so high, it is almost impossible to investigate the properties of all the possible structural patterns for functionalized MOFs only with the physics-based simulation. Machine learning is one of the solutions to deal with such a huge search space. The question is how many structural patterns we should sample as the training dataset to achieve an enough prediction accuracy.

I will clarify the scaling law of the machine learning prediction accuracy with respect to the proportion of the sample size in the total search space. To demonstrate it, I use the dataset of the CO2 capturing capability for 560 parent (unfunctionalized) MOF structures and 10995 structures of their functionalized counterparts. All the parent structures and some of the functionalized structures are used as the training data. I first hold out 20% of the functionalized structures as the test data and change the number of the functionalized structures in the training dataset. The result shows a clear scaling law and using 5-10% of the total structures for the training is enough to achieve R2 of 0.8-0.9.