From 1 to 200 features in model regression – what's the gain? Transforming materials characterization signals to structural information

Kersti Hermansson[1], Shokirbek Shermukhamedov[1] and Jolla Kullgren[1]

[1] Department of Chemistry−Ångström, Uppsala University, Box 531, S-75121 Uppsala, Sweden
kersti@kemi.uu.se

Key Words: ***Descriptors, NMR chemical shifts, Vibrational frequencies, Machine-learning prediction***

Abstract

Advances in machine-learning (ML) have transformed computational chemistry. ***One*** such leap forward is the advancement of machine-learning interatomic potentials or force-fields. Unlike traditional classical potentials, which rely on predefined functional forms, machine-learned force-fields learn directly from quantum-chemical data, typically DFT.

However, the work presented here is not about energies, forces and interatomic potentials. It is about the fitting of spectroscopic properties to mathematical models using ML. We have chosen two of the most widely used materials characterisation methods in academia and industry: IR and Raman vibrational spectroscopy [1] and NMR chemical shifts.

We present a "progression of regressions" where we monitor the predictive performance of the resulting ML models as a function of (i) the complexity of the descriptors used, and (ii) the physics contents.

For adequately homogeneous datasets, we reach RMSE values indicative of high predictive precision: < 0.1 ppm for proton chemical shifts and about 10 cm^{-1} for OH vibrational frequencies. [Unpublished work]

References

[1] Kullgren, J., Röckert, A., & Hermansson, K. Water in Crystals: A Database for ML and a Knowledge Base for Vibrational Prediction. J. Phys. Chem. 127, 13740-13750 (2023).