Predictions and/or insight? - ML and physics-based NMR and IR spectroscopy for water in, and on, crystals

Shokirbek Shermukhamedov[1], Kersti Hermansson[1] and **Jolla Kullgren**
[1] Dept of Chemistry-Ångström, Uppsala University, Uppsala, Sweden

jolla@kemi.uu.se

Background and aims. This project aims to discover structure–property relations for H_2O in and on solids, using ML with some physics-based flavour. The properties are spectral signatures (IR, NMR and a little bit of XPS). In this presentation, we evaluate several key structural descriptors (features) based on their ability to capture the spectral signatures.

Our first goal is to generate predictive models for spectral signatures from quantum-mechanical materials data.

Under favorable conditions, high-quality experimental OH vibrational frequencies and ^{1}H chemical shift measurements for well-behaved ices (for example), may be as accurate as whitin 5−10 cm^{-1} and 0.02−0.05 ppm. *Can our computational-based predictions match this quality? Does the predictability depend strongly on the nature of the spectroscopy? Does it depend strongly on the "nature" of the water*? The systems examined constitutes a rich plethora spanning from ice polymorphs (**cf. Fig. 1**) → crystalline bulk hydrates → water molecules at surfaces.

Our second goal is to find out how far we can reach in predictability if we trade some predictive power for model evaluation speed or model interpretability.

Methodology. Our descriptors were selected to represent a progression from a geometric machine-learning (ML) descriptor towards those with gradually more physics (chemistry) contents, and with varying short- and long range effects.

Some key methodological aspects of our study are:

(i) Our descriptors are either only intermolecular (we mask out intramolecular effects) in order to emphasize the perturbing effects of the local environment in our prediction models, or include all geometrical information available. (The effect can be drastic!)

(ii) To achieve as consistent comparisons between descriptors as possible, we streamline the training and use the same DFT functional, consistent hyperparameters, and the same measures of quality in all comparisons.

(iii) We also use a second training philosophy for comparison, namely one where we fine-tune the hyperparameters to achieve the "best possible" prediction with each descriptor.

(iv) OH vibrations are strongly anharmonic and for the training set generation we solve a quantum vibrational Schrödinger equation. Also the NMR calculations are state-of-the-art.

(v) The effect of adding physics-based descriptors, such as electric field from the surroundings, is also examined.

Results. The results include data from Refs. [1-4] as well as new and unpublished data. **Fig. 1** shows the capability of our descriptors in predicting the DFT reference values for vibrational frequencies and ^{1}H NMR chemical shifts. The descriptors in the figure are ordered in terms of their complexity and reveals that the simple geometric hydrogen-bond descriptor, R(H···O), is remarkably accurate. For ^{1}H NMR chemical shifts, a higher degree of descriptor complexity is required to achieve comparable accuracy. With the more elaborate machine learning (ML) descriptor encoding more contributions from the external environment of the water molecule, we can achieve a near-perfect correlation. This comes at the cost of reduced interpretability and naturally leads to the question: *Can we reach a good balance between prediction and insight*? (**Fig. 2**)

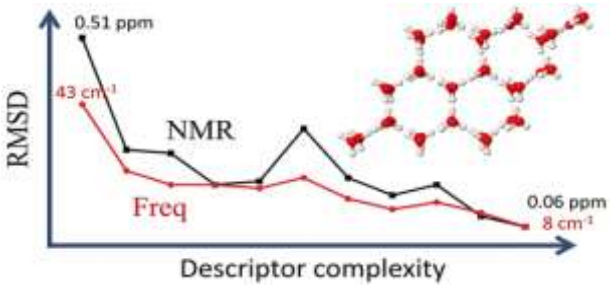

Figure 1. Root mean square deviations (RMSD) in the prediction of DFT computed vibrational frequencies and ^{1}H NMR chemical shifts using descriptors of increasing complexity. The descriptor at the right-hand end is MACE, and to the left it is a single H-bond distance only.

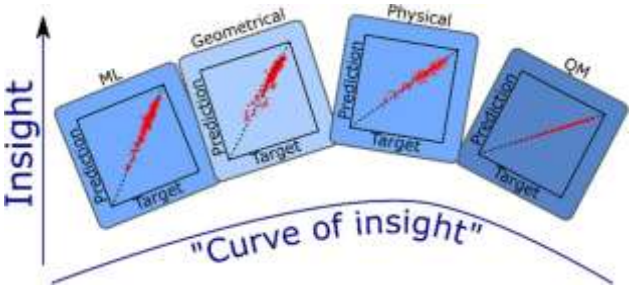

Figure 2. A schematic overview of descriptors with varying degrees of complexity, physical insight and level of predictive power.

References

[1] Andreas Röckert, Jolla Kullgren, Kersti Hermansson, J. Chem. Theory Comput., 12 (2022), 7683–769.
[2] Andreas Röckert, Jolla Kullgren, Daniel Sethio, Lorenzo Agosta Kersti Hermansson, J. Chem. Phys., 159 (2023), 044705.
[3] Jolla Kullgren, Andreas Röckert, Kersti Hermansson, Phys. Chem. C, 127 (2023), 13740–13750.
[4] Shokirbek Shermukhamel, Jolla Kullgren, Daniel Sethio and Kersti Hermansson, J. Chem. Theory Comput. In press (2026)