Modelling functional materials

Data-driven materials modelling – is domain-knowledge needed?

Jolla Kullgren, Andreas Röckert, Kersti Hermansson
Department of Chemistry-Ångström, Uppsala University, Sweden (*kersti@kemi.uu.se*)

1. Materials modelling: Is materials knowledge needed?

Modelling can mean different things. Typically one of the following:

(A) **Computer simulations** that generate data and phenomena based on scientific/engineering **EQUATIONS** and materials relations.

(B) Statistical **data-driven modelling** (≈ machine-learning≈ "AI") that generates models based entirely on **DATA.**

(C) Mixes thereof, i.e. (1) + (2)

In data-driven modelling (**B**), **domain knowledge** enters via the selection of variables (features, descriptors) and via constraints. If very many features are used, most **insight is lost**, but the **prediction capability** may be large.
In this poster we will explore descriptors that can predict vibrational spectra.

2. Regression philosophy

To make the comparison between descriptors as unified as possible, . . .

- We use the same data points in the regression.
- We use the same regression method (Gaussian process regression)
- We use the same measures of quality for all descriptors.

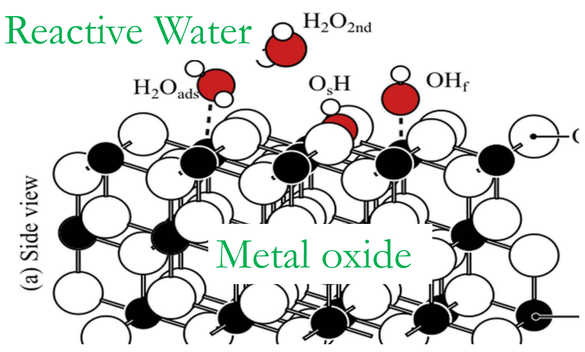
3. Objective: Predict spectra

Objective: Predict IR/Raman spectra for water and OH groups on functinal surfaces. Goal: good predictive model + insight (= physics/chemistry). We will evaluate the performance of *"typical ML" descriptors => to "descriptors with more physics".* [1, 2]

4. Metal oxide surfaces – in industry & society

The use of metal oxide materials

- pollution control
- energy generation and storage
- water splitting → H_2 → fuel cells
- microeletronics, catalysis, self-healing coatings, paint, gas-sensing, ceramics, biomedicine, ...



Surface OH groups from water are central in most of these applications.

5. Results: Validation of training data vs experiment

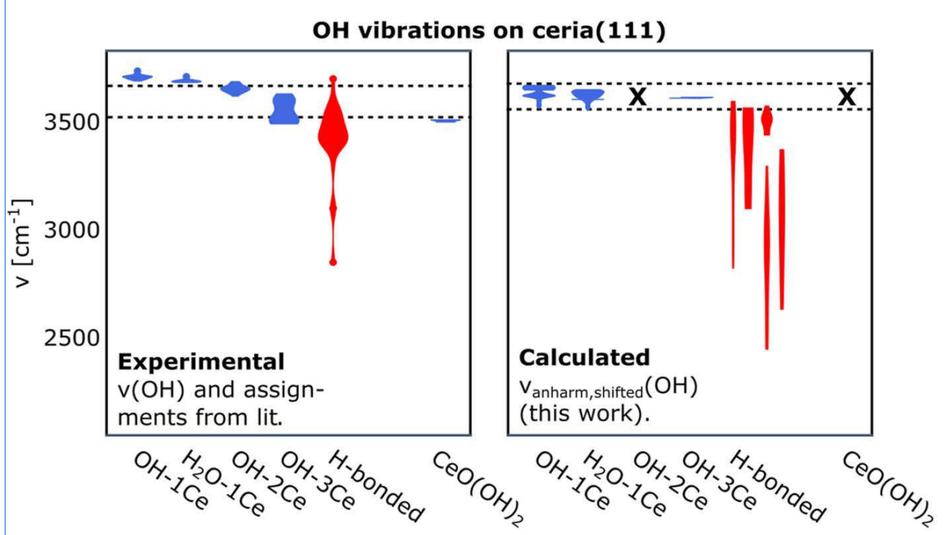


Fig. Comparison of experimental spectral data (literature, left) and computed data (right, generated from DFT calc.). The good agreement validates that the DFT method gene-rated good training data. The data are shown as violin plots.

Fig. We created two data-bases of water and OH (i) in bulk and (ii) on surfaces [1,2].

6. Results: Found "best" descriptor

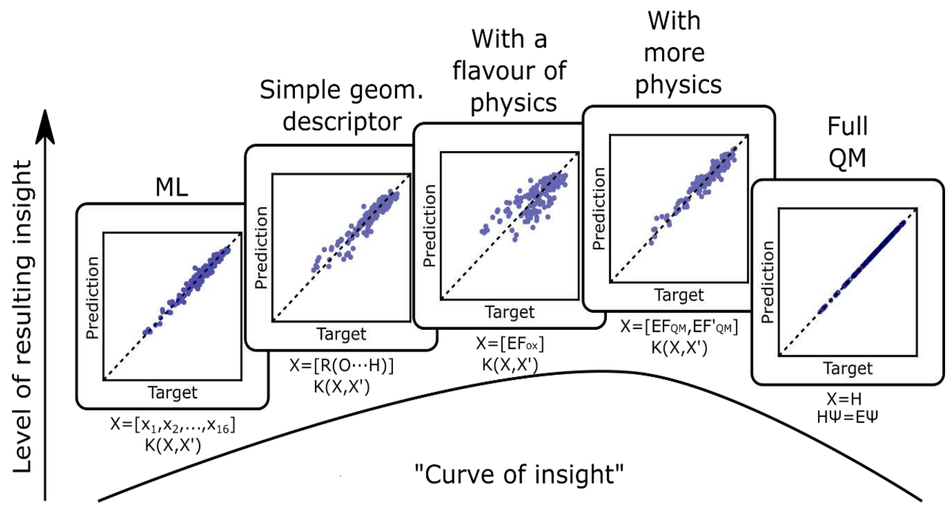


Fig. Illustration of the progression of our descriptors in terms of the "amount" of physics coded into them and their respective level of insight. The scatter plots show the agreement between predictions and reference values.

7. Conclusion: **Yes. Crucial!**

EMMC Focus Area 1:
"Model development and validation"
"... stands for everything that has to do with the **capabilities** of materials models and modelling workflows, and validation of them. Application to challenging problems of industrial relevance also belong here.

UPPSALA UNIVERSITET
Vetenskapsrådet
eSENCE THE E-SCIENCE COLLABORATION

References

[1] Water in crystals – a database for ML and a knowledge-base for vibrational prediction , J Kullgren, A Röckert, K Hermansson
J Phys Chem C, Accepted for publication 12 April (2023)

[2] Predicting Frequency from the External Chemical Environment: OH Vibrations on Hydrated and Hydroxylated Surfaces
A Röckert, J Kullgren, K Hermansson
J. Chem. Theory Comput. 18, 7683−7694 (2022)