Contribution ID: **30** Type: **Poster**

Exploring the limits of the random forest algorithm for the classification of X-ray absorption spectra

Tuesday 9 April 2024 18:30 (20 minutes)

X-ray absorption spectroscopy gives access to a wealth of information regarding the local structure and electronic properties of materials. However, data analysis is significantly more time-consuming than the acquisition and initial data reduction. Decoding the information relies on comparison with similar compounds for which the spectrum–property mapping is already established, a task that is very often performed by visual inspection.

Machine learning (ML) is revolutionizing many fields with its ability to extract and learn patterns in big data without having to provide additional prior information other than the data itself. ML models give access to instantaneous predictions of properties and observables, which makes them particularly attractive for performing autonomous experimental acquisitions or real-time analysis of the measured data.

Current ML applications for X-ray spectroscopy mainly use neural networks (NN), which require extensive computational datasets as training data.[1,2] These are very time-consuming to build and are often linked to large-scale computing facilities access. Alternatively, one can use tailor-made ML models that are less data-hungry and can be trained significantly faster. One such ML algorithm is the random forest, which has already shown promising applications for analyzing visible and infrared spectra.[3]

In the present contribution, we will present the application of the random forest algorithm to identify the coordination environment of iron in a given compound from the corresponding K-edge X-ray absorption spectrum. As we train the model using theoretical data, but we use it to infer properties on measured spectra, we analyze the different sources of errors that limit the quality of the prediction, such as spectral shift, normalization, and noise level. In addition, we explore the use of SMOTE to tackle the class imbalance, a common issue in such datasets as most materials in nature tend to adopt a small set of specific coordination environments.

[1] Timoshenko, J.; Anspoks, A.; Cintins, A.; Kuzmin, A.; Purans, J.; Frenkel, A. I. Phys. Rev. Lett. 2018, 120 (22), 225502. https://doi.org/10.1103/PhysRevLett.120.225502.
[2] Rankine, C. D.; Madkhali, M. M. M.; Penfold, T. J. J. Phys. Chem. A 2020, 124 (21), 4263–4270. https://doi.org/10.1021/acs.jpca.0c03723.
[3] Lee, S.; Choi, H.; Cha, K.; Chung, H. Microchem. J. 2013, 110, 739–748. https://doi.org/10.1016/j.microc.2013.08.007.

Primary author: RETEGAN, Marius (ESRF)

Co-authors: Mr ZECCHI, Federico (ESRF); Dr GLATZEL, Pieter (ESRF)

Presenter: RETEGAN, Marius (ESRF)

Session Classification: Posters

Track Classification: MLC