Pretraining Foundation Models: Unleashing the Power of Forgotten Spectra for Advanced Geological Applications

An-Sheng Lee¹, Hsuan-Tien Lin², and Sofia Ya Hsuan Liou¹

¹Department of Geosciences and Research Center for Future Earth, National Taiwan University, Taipei, Taiwan
²Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

X-ray fluorescence (XRF) core scanning, renowned for its high-resolution, non-destructive, and user-friendly operation, is pivotal in geological research for analyzing chemical, physical, and biological signals. Despite the extensive applications of XRF data for various research purposes, the quantification of this data into specific geological proxies remains challenging due to the inherent non-linearity caused by simple sample pretreatment during core scanning. Leveraging advancements in deep learning, computing power and large-scale scientific drilling programs, our study aims to address this non-linearity by harnessing the often-overlooked raw XRF spectra stored in laboratory databases. We introduce an approach involving self-supervised pretraining on 54,643 spectra from marine sediments in the high-latitude sectors of the Pacific Ocean (cruises SO178, SO264, PS97, PS75, LV29). Our model, underpinned by a deep bidirectional image transformer (ViT-base), is trained to reconstruct heavily masked spectra (75%) with an $R^2$ accuracy of 0.996, demonstrating its proficiency in feature extraction from limited data portions. This foundational model is anticipated to serve as a versatile tool for various downstream geological applications after finetuning with specific labeled data, such as quantifying high-resolution calcium carbonate ($\text{CaCO}_3$) and detecting machinery anomalies. Future work includes expanding the spectrum database with diverse materials and machine settings to enhance the model's generalizability, ultimately broadening its applicability beyond core scanning for geological applications to encompass all XRF measurement techniques.