Goldschmidt 2024 Abstract
https://doi.org/10.46247/gold2024.24121

Machine Learning Models for Evaluating Biological Reactivity in Dissolved Organic Matter over time: A Case Study in the Three Gorges Reservoir, China

CHEN ZHAO[1], KAI WANG[2], QIANJI JIAO[3], XINYUE XU[4], DR. YUANBI YI[5], PENGHUI LI[6], JULIAN MERDER[7] AND DING HE[1]

[1]The Hong Kong University of Science and Technology
[2]Southern University of Science and Technology
[3]Xidian University
[4]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[5]Department of Ocean Science and the Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), The Hong Kong University of Science and Technology, Hong Kong, China
[6]School of Marine Sciences, Sun Yat-sen University, Zhuhai, China
[7]Department of Global Ecology, Carnegie Institution for Science, Stanford

Presenting Author: czhaobk@connect.ust.hk

Reservoirs exert a profound influence on the cycling of dissolved organic matter (DOM) in inland waters by altering flow regimes. Biological incubations can help to disentangle the role that microbial processing plays in the DOM cycling within reservoirs. However, as the complex DOM composition poses a great challenge to the analysis of such data, it remains elusive how DOM molecules respond to microbial processing over different time scales. Here we tested if the emergence of interpretable machine learning (ML) methodologies can contribute to capturing the relationships between molecular reactivity and composition. We developed time-specific ML models based on 7-day and 30-day incubations simulating the biogeochemical processes in the world's largest reservoir, the Three Gorges Reservoir in China, over shorter and longer water retention periods, respectively. Besides the well-recognized predictive power of ML methodologies, we delved into the processes of tuning the ML models to acquire additional interpretability. We used an under-sampling strategy to improve model performance and simultaneously observed the variations in model performance metrics for different biological reactivity pools over incubations with different durations. Results showed that that bio-produced pools exhibited higher heterogeneity than the others, likely resulting from the complicated reaction paths to generate products. Moreover, it is revealed that shorter incubation periods resulted in a broader range of molecules disappearing, with a greater contribution of stochasticity, while the longer incubation allowed the successive biodegradation of oxygen-poor compounds, with a greater contribution of directed degradation. Employing a classical ML algorithm combined with a straightforward under-sampling technique, and Shapley values to identify feature importance, we uncovered a novel perspective in understanding the dynamics of DOM over time.