Goldschmidt 2023 Abstract
https://doi.org/10.7185/gold2023.16031

Improving Understanding of Dissolved Organic Matter by Using Machine Learning to Predict Stable Carbon Isotope based on Molecular Abundances

DR. YUANBI YI[1], TONGCUN LIU[2], JULIAN MERDER[3], CHEN HE[4], HONGYAN BAO[5], PENGHUI LI[6], SI-LIANG LI[7], QUAN SHI[4] AND DING HE[1]

[1]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

[2]Zhejiang A&F University

[3]Department of Global Ecology, Carnegie Institution for Science, Stanford

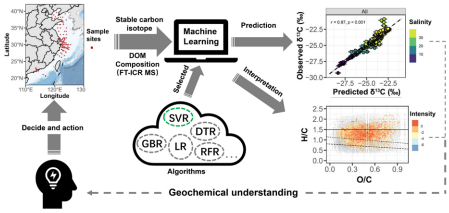
[4]State Key Laboratory of Heavy Oil Processing, China University of Petroleum, Beijing, China

[5]Xiamen University

[6]Sun Yat-Sen University

[7]Institute of Surface-Earth System Science, School of Earth System Science, Tianjin University, Tianjin, China

Presenting Author: yuanbiyi@ust.hk



Dissolved organic matter (DOM) consists of lots of molecular formulas (MFs), playing a unique function in aquatic systems. However, it remains unclear how MFs determine bulk DOM properties. Here, we use Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR MS) to characterize DOM molecular composition from 510 samples, distributed in the China Coastal Sea, of which more than 320 samples contain accompanying stable carbon isotope ($\delta^{13}C$) measurements. A machine learning model (ML) was created based on $\delta^{13}C$ contained samples, resulting in a mean absolute error (MAE) of 0.30‰ on the test data. The developed model was used to predict in remaining samples without $\delta^{13}C$ and other published datasets. The result shows that ML selected 5199 MFs to predict $\delta^{13}C$, and the observed $\delta^{13}C$ values were more effectively tracked than traditional approaches. Besides, the ML revealed that the elemental composition would be primarily impacted when DOM was transferred from land to ocean. This work suggests that with increasing molecular research and larger learning datasets, ML could be considered a new paradigm for studying the non-linear and non-monotonic connections behind the complex DOM molecular composition.