Goldschmidt 2026 Abstract
https://doi.org/10.46427/gold2026.38034

Leveraging Interpretable Machine Learning with Small Geochemical Datasets to Predict Trace Element Speciation in Groundwater

LOGAN GOULETTE[1] AND ATHENA A. NGHIEM[2]

[1]University of Wisconsin - Madison
[2]University of Wisconsin-Madison

Arsenic (As) is a carcinogenic metalloid found in drinking water throughout the world. Within a groundwater system, the partitioning of As is a biogeochemical process affected by local geology, hydrology, and redox. Because of its detrimental effect on human health, many studies have sought to understand As mobilization in groundwater, resulting in abundant geochemical data. In this big data era, machine learning (ML) has been a powerful tool to model the environmental processes which drive As concentrations above the World Health Organization's limit of 10 g/L. However, the speciation of As remains an overlooked mechanism of As mobilization which has not been explored, likely because there are limited geochemical data available for modeling. While the widely recognized method for As mobilization is the release of sorbed As via reductive dissolution of iron oxides, in reducing environments with abundant sulfur, poorly sorbing thiolated As (thioarsenic) species may form. Though these species can accelerate groundwater contamination, they are difficult to preserve and process-based geochemical models remain unable to resolve their behavior. This presents a unique opportunity to utilize interpretable ML techniques to analyze these limited datasets to understand and predict the geochemical conditions which create thioarsenic species. This study utilizes a small dataset of 288 observations from Bangladesh, Vietnam, and China to create and compare decision-tree based models that predict the presence of monothioarsenate (MTA), a marker of As thiolation. Our study utilizes a maximum of 8 typically measured geochemical features such as pH, total iron and total arsenic to classify MTA-containing groundwaters. This small data and limited feature approach is not common in geochemistry ML due to the potential for unstable, or overfit models. Thus we test model stability by comparing test/train splits, utilizing nested cross validation, and by eliminating overfit models using techniques such as partial dependence plot visualization. This effort produced models with 82 - 90% accuracy on resampled and hold out data. The uptake in use of interpretable ML techniques allows geochemists to analyze signals even in small data scenarios, pushing forward our understanding of the formation and stability of thioarsenic species in different hydrological environments.