Goldschmidt 2025 Abstract
https://doi.org/10.7185/gold2025.30849

Unidentified dataset dependencies may inflate machine learning performance metrics in geochemistry studies

TARYN SCHARF[1], MATTHEW DAGGITT[2], LUC DOUCET[3] AND CHRISTOPHER L. KIRKLAND[1]

[1]Timescales of Mineral Systems Group, Curtin Frontier Institute for Geoscience Solutions, School of Earth and Planetary Sciences, Curtin University, Perth, WA 6103, Australia
[2]University of Western Australia
[3]Curtin University

Machine learning algorithms have been widely used to develop predictive models on *big data* in the geosciences. The reliability of reported model performance metrics is important if models are to inform decision making. However, model performance risks being overstated when unidentified data dependencies are present in geological datasets. Data dependencies can result in information leakage between training and testing subsets, causing a model to have artificially higher performance on the testing subset and reduced generalisability. Overstated model performance may contribute to the scientific 'reproducibility crisis' and risks decreasing confidence in machine learning applications.

We review recent peer-reviewed geochemical publications, assessing what proportion of the described methodologies explicitly consider data-leakage and what proportion may be vulnerable to overly optimistic results. Geochemical datasets often comprise multiple analyses drawn from a single sample, which imposes a hierarchical structure on the dataset. We demonstrate the effects of hierarchical data structures on algorithm performance by simulating a dataset of samples comprising multiple analyses that are related through latent variables (e.g., individual analyses linked to a common compositional sample). Simulation studies provide insights into the conditions that exacerbate the impact of data leakage in hierarchical datasets. Finally, we move beyond simulation studies by accessing the magnitude of the effect of information leakage within a subset of published datasets.

This work highlights that dependencies are often present in geological datasets. If these dependencies are unidentified and untreated, they may result in over-optimistic estimates of model performance.