geoconvention
Calgary • Canada • May 12-14 2025

Beyond Black Box: Integrating Domain Knowledge with AI/ML for Geochemistry Applications

Jagoš R. Radović[1], Jeffrey F. Van Humbeck[2], Ana Vielma[1], Stephen R. Larter[3]

[1]Center for Petroleum Geochemistry (UH-CPG), Dept. of Earth and Atmospheric Sciences, University of Houston

[2]Department of Chemistry, University of Calgary

[3]Department of Earth, Energy, and Environment, University of Calgary

Summary

The integration of artificial intelligence (AI) and machine learning (ML) in geoscience is transforming how complex datasets are analyzed, uncovering hidden geochemical patterns and improving analytical workflows (Larter et al., 2019). This study presents two case studies showcasing approaches that integrate domain-specific geoscientific knowledge to enhance AI/ML model performance.

The first case study focuses on asphaltene analysis, a critical area due to asphaltenes' complex molecular structures and their significant industrial implications. Asphaltenes are not only relevant to understanding petroleum systems, but they also have promising applications in advanced materials, particularly in the production of carbon fibers (Saad et al., 2022). By incorporating molecular relationships into ML models, our approach significantly improves predictive performance, reducing mean-squared error (MSE) and enhancing the interpretability of complex asphaltene mixtures.

The second case study applies ML-driven feature selection to refine lithofacies classification within Monterey Formation. Differentiating lithofacies transitions in high-dimensional geochemical datasets poses a challenge for traditional methods like Principal Component Analysis (PCA) and Hierarchical Clustering Analysis (HCA). By leveraging Random Forest algorithms combined with t-distributed stochastic neighbor embedding (t-SNE), we successfully identified specific geochemical markers that differentiate geochemical transitions, offering new insights into complex sedimentary processes.

In both cases, domain expertise was essential in guiding AI/ML workflows, ensuring scientifically meaningful results. These findings reinforce the necessity of adapting AI/ML methodologies to geological data complexities rather than applying generic analytical frameworks, not informed by relevant geoscientific context.

Workflow

Case Study 1: Machine Learning for Molecular Pattern Recognition

- Dataset: 35 high-resolution mass spectra of asphaltene extractions.
- Approach: Incorporating molecular relationships to constrain ML predictions.
- Key Innovation: Use of Formula Information Gain (FIG) to rank predictive molecular relationships, optimizing ML feature selection.

Case Study 2: ML-Driven Lithofacies Differentiation in the Monterey Formation

- Dataset: 17 bitumen samples, extracted from a core taken in the Monterey Formation within the Santa Maria Basin, were analyzed using GC-MS/MS.
- Approach: Comparing PCA, HCA, and Random Forest + t-SNE, revealing monoaromatic steranes as potential markers of organic input shifts.

- Robust scaler preprocessing + supervised feature extraction (Random Forest) + t-SNE clustering allows to interpret a distinctive geochemical signature for each lithofacies (UL1 and UL2).

Results

The first case study explores the role of domain-informed ML in analyzing complex organic mixtures, specifically petroleum-derived asphaltenes. Using Physics-Informed Neural Networks (PINNs) and molecular relationship constraints, we developed an optimized ML model that achieves a 40% reduction in mean-squared error (MSE) compared to unconstrained approaches (Le et al., 2024). The inclusion of domain knowledge enhances model interpretability and predictive accuracy, as shown in Figure 1, which compares performance with increasing domain knowledge incorporated into the model.

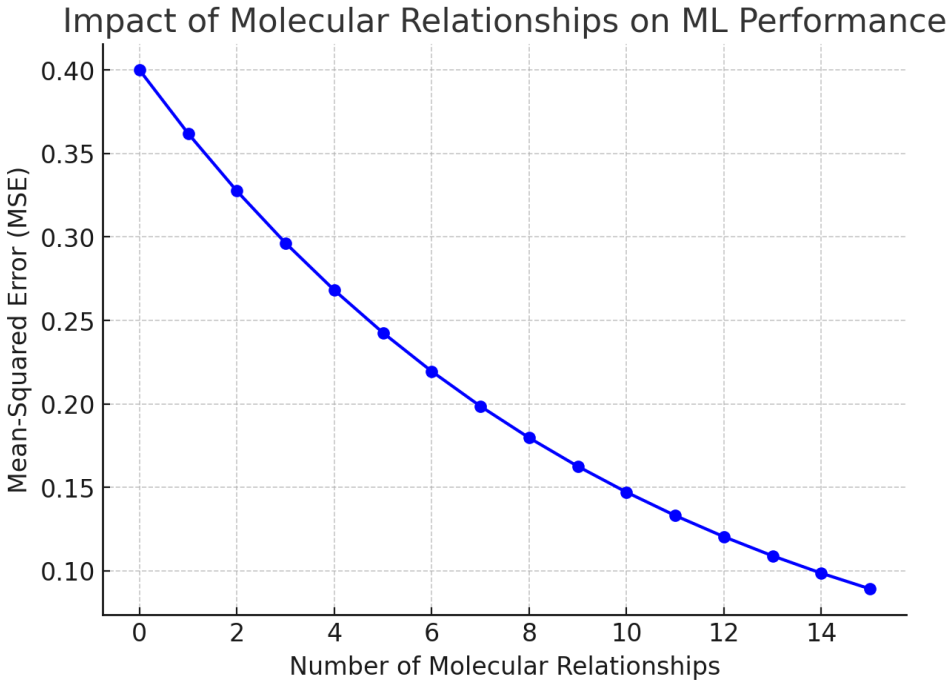

Figure 1. Impact of Molecular Relationships on ML Performance. This generalized illustrative figure, derived from observations in the Le et al. (2024), demonstrates how incorporating domain knowledge into ML models reduces mean-squared error (MSE). As more molecular relationships are added, MSE decreases significantly, with the greatest improvement occurring within the first 10–11 relationships. Beyond this point, the performance gains plateau, indicating diminishing returns. This trend aligns with findings from Le et al. (2024), where the optimal model used 11 molecular relationships, achieving a 40% reduction in MSE compared to unconstrained ML.

The second case study applies ML to explore potential drivers behind the unique lithofacies transition within the Monterey Formation. This challenge is amplified by the high-dimensional, multivariate nature of the dataset, where distinguishing between correlation and causation in multivariate geochemical analyses is challenging (Vielma et al., 2024). Traditional approaches, such as Principal Component Analysis (PCA) and Hierarchical Clustering Analysis (HCA), fail to

capture the full complexity of the dataset. Instead, a Random Forest (RF) and t-distributed stochastic neighbor embedding (t-SNE) workflow was implemented to improve classification accuracy. This approach successfully identifies unique chemical markers supporting the hypothesis that a shift in organic input is a primary driver of lithofacies change.

Novel/Additive Information

This work demonstrates two illustrative AI/ML approaches in geoscience: (1) constraining models with domain knowledge for complex mixture analysis and (2) AI-driven feature selection for geochemical classification. These case studies reinforce the necessity of adapting AI/ML workflows to geological data complexities, rather than forcing datasets into predefined analytical frameworks.

Acknowledgements

Case Study 1 was supported by Alberta Innovates through the Carbon Fiber Grand Challenge (Phase I) and by allowing access to the Asphaltene Sample Bank. It was also supported by the Canada First Research Excellence Fund—Global Research Initiative. J.L.M. is supported by the Canada Research Chairs program. Katelyn Lee and Justin L. MacCallum are acknowledged for their contributions to Case Study 1.

References

Larter, S., Radovic, J., Silva, R. and Huang, H., 2019. The Evolution of Petroleum Systems Analysis: A Future For Petroleum Geochemistry?. In *AAPG Hedberg Conference, The Evolution of Petroleum Systems Analysis*.

Le, K., Radović, J.R., MacCallum, J.L., Larter, S.R. and Van Humbeck, J.F., 2024. Machine Learning in Complex Organic Mixtures: Applying Domain Knowledge Allows for Meaningful Performance with Small Data Sets. *Journal of the American Chemical Society*, 146(32), pp.22563-22569.

Saad, S., Zeraati, A.S., Roy, S., Saadi, M.A.S.R., Radović, J.R., Rajeev, A., Miller, K.A., Bhattacharyya, S., Larter, S.R., Natale, G. and Sundararaj, U., 2022. Transformation of petroleum asphaltenes to carbon fibers. *Carbon*, *190*, pp.92-103.

Vielma, A., Curiale, J.A., Carvajal-Ortiz, H., Radović, J.R., Fu, Q., Malloy, T.B. and Bissada, K.A., 2024. Paleoredox and lithofacies assessments in deepwater intervals of the Monterey Formation, Santa Maria Basin, California: Insights from organic sulfur geochemistry. *International Journal of Coal Geology*, *294*, p.104606.