Title: Coupling AI-based Modeling and Molecular Soil Organic Matter at Regional-scale

Authors/Affiliations: Satish Karra, Emily Graham, Odeta Qafoku, John R. Bargar (Environmental Molecular Sciences Laboratory/Pacific Northwest National Laboratory), and the Environmental Molecular Sciences Laboratory team

Focal Area(s): Our white paper is framed around the focal area of *the importance of high-potential datasets and how combining multiple datasets leads to scientific insights into the methane cycle. Our approaches in this white paper are also aligned with improving the measurement coverage toward reducing uncertainty in mechanistic models.*

Science or Technological Challenge: The Environmental Molecular Sciences Laboratory (EMSL) spearheads a 10-year National Molecular Observations Network (MONet)[1] initiative for the BER, with the objective to build a national network of environmental sampling and sensing sites along with methods to provide molecular-level and microstructural information on soil, water, resident microbial communities, and biogenic emissions. For instance, in the current phase of the MONet initiative, data types, including metagenomics, respiration, mineral organic matter, hydraulic properties, and geochemistry, are being collected from core samples from a wide range of ecoregions within the US[3]. In coordination and partnership with other observational networks (ARM, AmeriFlux, NEON), the objective is to make these molecular observations and the data from field-deployed sensors, available to the BER community. To make these multi-modal data streams accessible to domain scientists, modelers, and data scientists, who study the methane cycle, we aim to build a suite of data and modeling products and avail them to the BER community via the MONet portal. *EMSL is strongly positioned to bridge fundamental ModEx gaps by building key products for the BER community. We envision that AI-based methods will be central to these products and are key to accelerating BER community science toward eliciting the mechanics of the methane cycle.*

Rationale: Several hydro-bio-geochemical natural and anthropogenic processes in the soil, water, and atmosphere, and their complex interactions, contribute to methane fluxes. Characterization of the underlying fundamental molecular-scale and microstructural processes (e.g., geochemistry, omics, etc.) is needed to parameterize and validate the individual process models and their coupling. One of the major contributors to an increase in uncertainty in models is the lack of such data. The MONet initiative at EMSL aims to facilitate the availability of such data to advance Model-Experiment integration and to enhance the predictive power of the multiscale models for carbon and nitrogen fluxes including the methane cycle. Specifically, the key gaps that we will address are:

- Lack of multi-modal molecular and microstructural data with metadata capture that follow FAIR principles for soils across the US and the resident microbes and their availability to the BER community.
- Availability of molecular and microstructural data (e.g., analysis, integration, and visualization) and modeling (e.g., pore models for transport) tools, along with the tools that integrate data and models (e.g., parametrization, sensitivity analysis, uncertainty quantification).

- AI methods can potentially play a major role in these tools and workflows. However, AI methods need data[4] across plot, ecosystem, and regional scales, and the collection of multi-modal molecular and microstructural data is thus needed.

The EMSL MONet soil characterization program, which began user operations in Feb 2023, provides such molecular data at regional and CONUS scales. MONet is collecting and analyzing soil cores using standardized workflows that can be optimized to provide data critical to AI-informed studies of the methane cycle.

Narrative: Our overall approach is to build a web-based data platform to make the MONet observational data, along with AI-based data and modeling tools, available to the BER community. We briefly detail our vision for the role of AI within data and modeling software products on this platform:

- AI and graph-based methods for data analysis and visualization: Classical unsupervised methods, such as principle component analysis[6], and non-negative matrix factorization[7], have proven to be powerful ways of identifying patterns and dominant features, and in correlating multi-modal datasets. They can be used to identify key signatures in multi-dimensional datasets and reduce dimensionality to visualize data effectively. For instance, our preliminary non-negative matrix factorization analysis on soil biogeochemical and microbial data from EMSL's '1000 Soil Pilot' project (a pilot program to MONet), showed clear correlations between dissolved organic matter and environmental stresses, such as flow, pH, and wildfire occurrence. In addition to making unsupervised ML-based tools available, we will build visualization tools based on network theory and graph-based methods for clustering and finding similarities in multi-modal data streams[8].
- AI for multiscale modeling: To enable the transfer of information (or upscale) from molecular- and microstructural- (pore-) scale to the site, regional, and eventual global Earth system models, AI-based methods can play a significant role. For example, we will provide users with pore-scale and models to perform flow and reactive transport simulations that utilize the MONet data, which will then inform averaged parameters, such as reaction rates or permeability, needed in site/regional scale models. AI methods such as deep learning[8,9] can be used to train on data from such simulations and to build surrogate models for upscaling information. These surrogate models will represent the relationships between molecular and microstructural information of interest to the user. Akin to constitutive models or equations of states, the AI-based surrogate models can be used in larger-scale simulations.
- AI for data-model integration: Recently, AI-based models based on deep learning, including approaches that constrain balance laws[10] or mimic balance laws[11], have become popular. These AI-based models are much faster to run and have shown to be effective for parametrization[12], and towards quantifying uncertainty.[13] We will provide users with workflow components that will enable these analyses.

References:

1. EMSL Five-Year Strategic Plan, 2021: https://content-qa.emsl.pnl.gov/sites/default/files/2021-07/EMSLStrategicPlanFY2021_0.pdf
2. MONet data being collected: https://content-qa.emsl.pnl.gov/sites/default/files/2023-02/EMSL0419_MonetFlyer.pdf
3. Molecular Observation Network – ecoregion table: https://content-qa.emsl.pnl.gov/sites/default/files/2023-02/MONet_%20Ecoregion%20Table.pdf
4. There is no AI without data. Communications of the ACM, November 2021, Vol. 64 No. 11, Pages 98-108 doi:10.1145/3448247
5. Taguchi, Y. H. Unsupervised feature extraction applied to bioinformatics: A PCA based and TD based approach. Springer Nature, 2019.
6. Wang, Yu-Xiong, and Yu-Jin Zhang. "Nonnegative matrix factorization: A comprehensive review." IEEE Transactions on knowledge and data engineering 25.6 (2012): 1336-1353.
7. Aittokallio, Tero, and Benno Schwikowski. "Graph-based methods for analysing networks in cell biology." Briefings in bioinformatics 7.3 (2006): 243-255.
8. Tang, Meng, Yimin Liu, and Louis J. Durlofsky. "Deep-learning-based surrogate flow modeling and geological parameterization for data assimilation in 3D subsurface flow." Computer Methods in Applied Mechanics and Engineering 376 (2021): 113636.
9. Ahmmed, B., Mudunuru, M. K., Karra, S., James, S. C., & Vesselinov, V. V. (2021). A comparative study of machine learning models for predicting the state of reactive mixing. Journal of Computational Physics, 432, 110147.
10. Karra, S., Ahmmed, B., & Mudunuru, M. K. (2021). AdjointNet: Constraining machine learning models with physics-based codes. arXiv preprint arXiv:2109.03956.
11. Haghighat, E., Raissi, M., Moure, A., Gomez, H., & Juanes, R. (2021). A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. Computer Methods in Applied Mechanics and Engineering, 379, 113741.
12. Raissi, M. (2018). Deep hidden physics models: Deep learning of nonlinear partial differential equations. The Journal of Machine Learning Research, 19(1), 932-955.
13. Gasmi, C. F., & Tchelepi, H. (2022). Uncertainty Quantification for Transport in Porous media using Parameterized Physics Informed neural Networks. arXiv preprint arXiv:2205.12730.