Shifting the Paradigm of Characterization and Remedy Decisions With Application of Machine-Learning Algorithms and Molecular Biological Tools

Andrew Madison and Skyler Sorsby (WSP USA)
Scott Porman (Mobil Oil Australia Pty Ltd)

June 5, 2024

Introduction

Biogeochemical CSM Development

Multiple Lines of Evidence (MLOE) Framework

Primary LOE – COCs

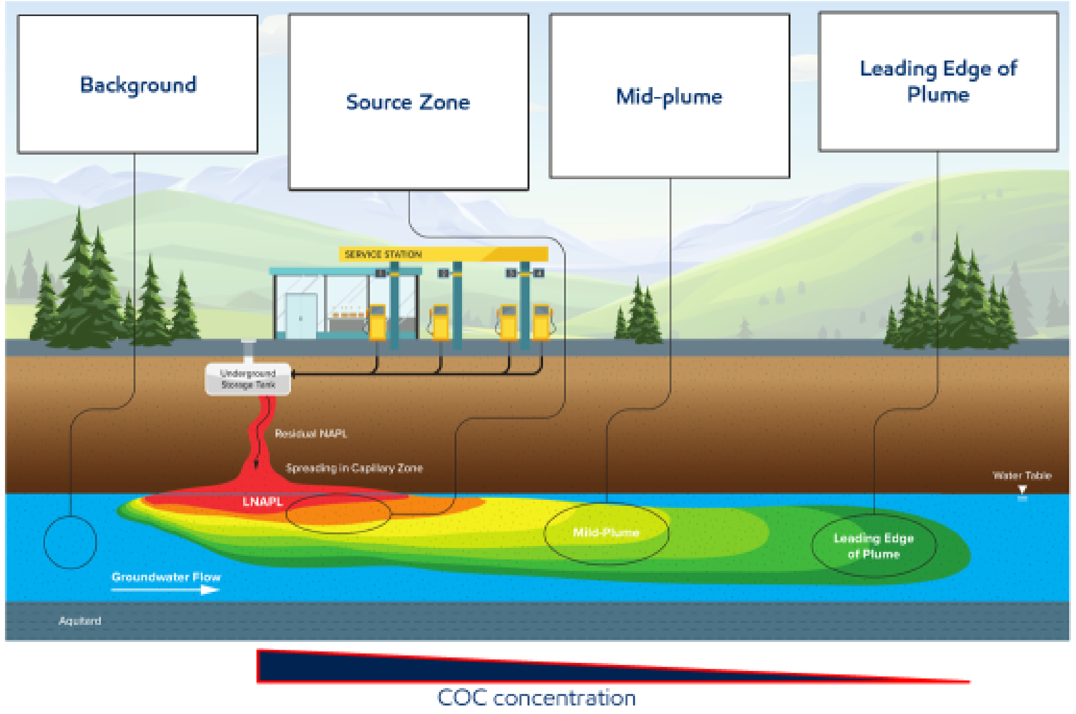
- Spatiotemporal trends in contaminant conditions
    - Contaminant fate and transport

Secondary LOE – Geochemistry

- Spatiotemporal trends in contaminant geochemical conditions
    - Key biodegradation/biotransformation processes

Tertiary LOE – MBTs

- Perform microbiological sampling at focused locations within area of interest
    - Microbial community composition
    - Quantify abundance of key functional genes and associated microorganisms



Key, Sorsby, Wang & Madison 2022 – Frontiers in Microbiology

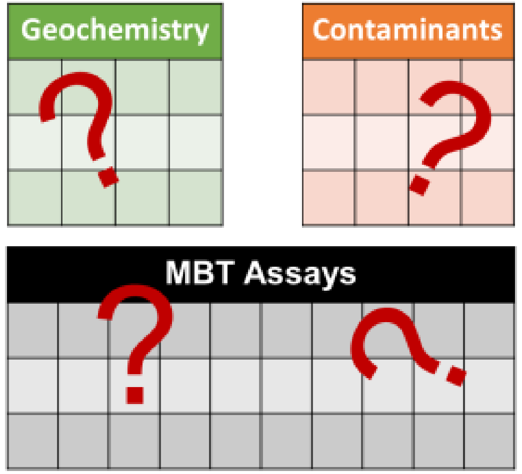
CSM is utilized to inform and guide contaminated site management including potential bioremediation application

MBT Data

Overcoming Interpretation Challenges to Increase Uptake & Site Understanding

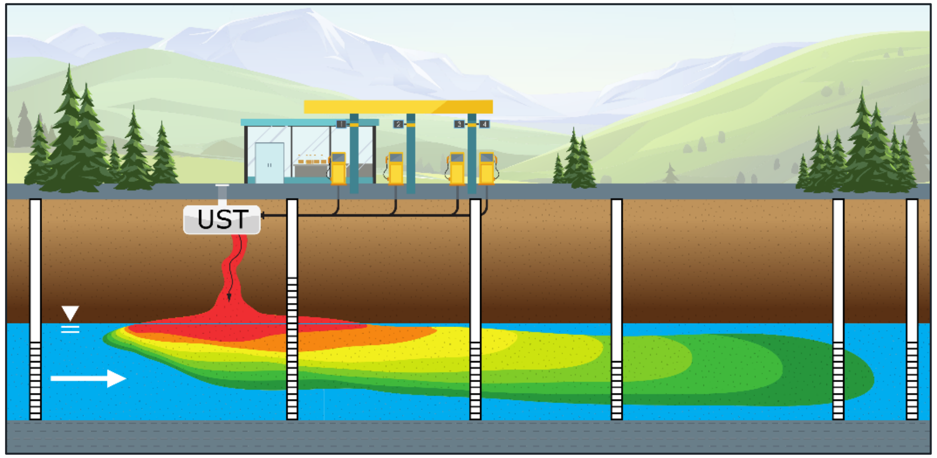
Potential Challenges:

- Integration into CSM requires microbiology or related subject matter expertise
- Large datasets for few sampling locations
- Common data reduction include rules-of-thumb (e.g., filtering data by % abundance) and laboratory/vendor prescriptions that can eliminate subtle trends



Suggested Solution:

- **Leverage machine learning to help jump-start and guide site-specific MBT interpretation**

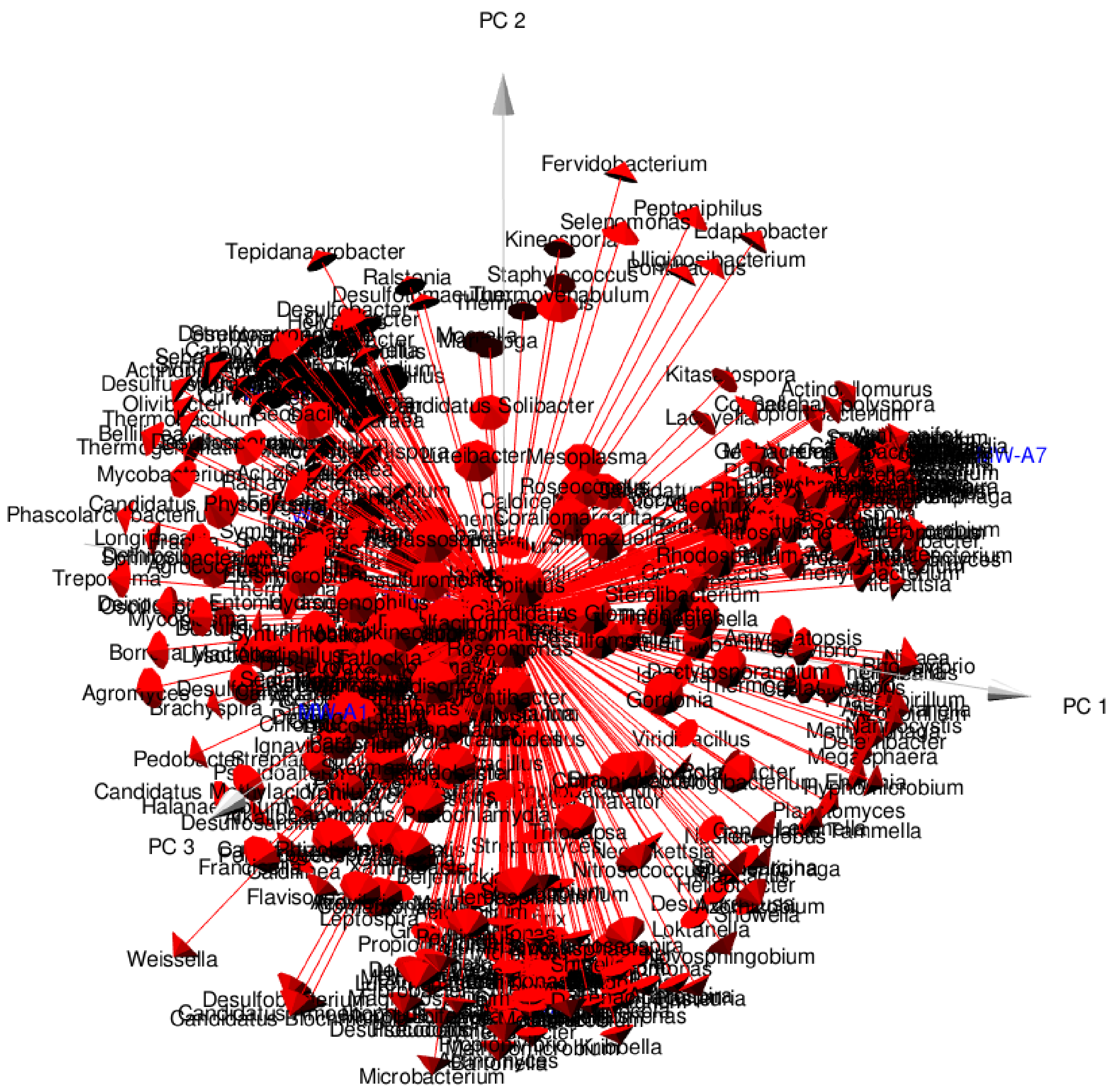


"Curse of Dimensionality"

- Inherent to contaminated site dataset where parameters (COCs, geochemistry, MBTs) outnumber the number of samples (P >> N)

- Classical multivariate exploratory analyses (e.g. principal components analysis – PCA) can produce cluttered/unclear results

- "Unconstrained" unsupervised learning algorithms (like PCA) "overfit" the data (e.g., noise is mistaken for signal)



PCA Biplot of 16S rRNA Sequencing Data

A New Approach – Machine Learning Applications

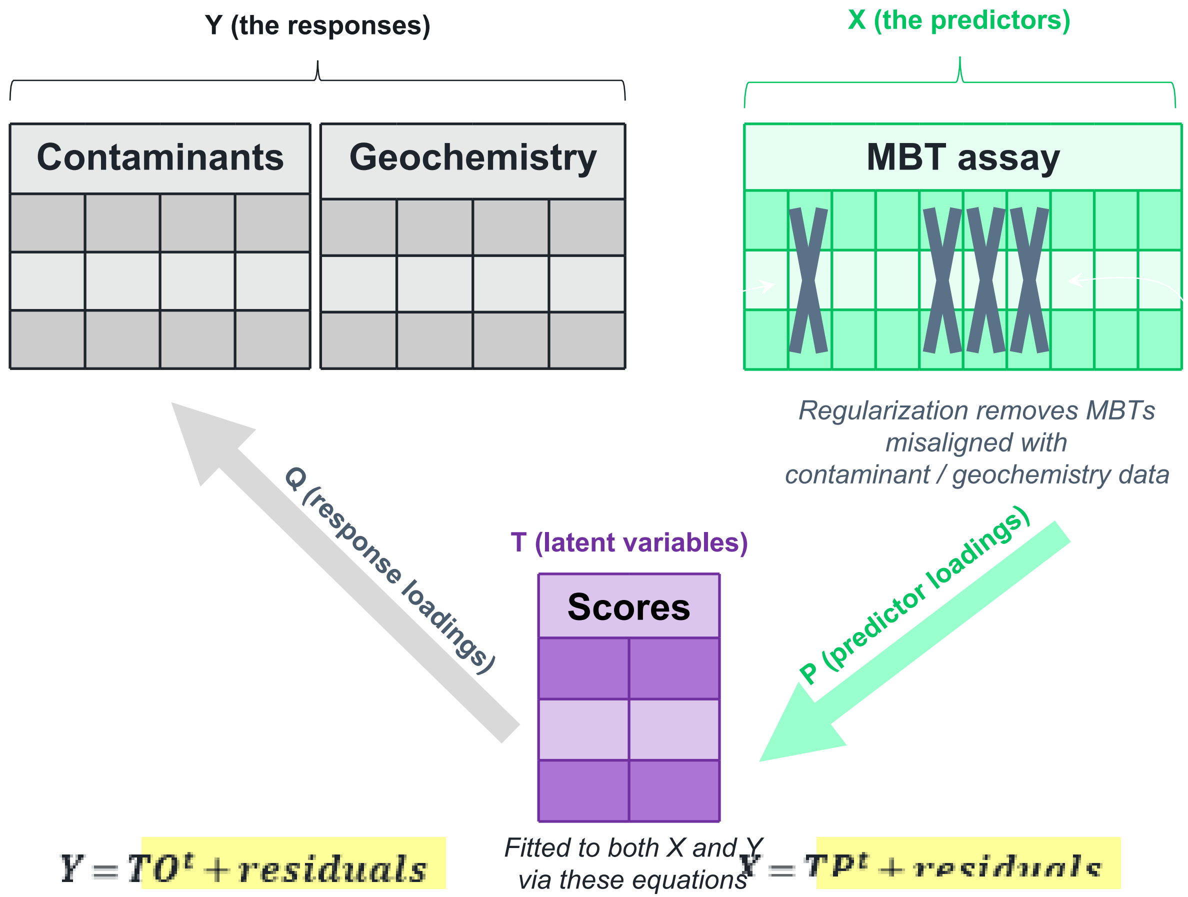
Lessons from Bioinformatics and Ecological Modeling

Constrained Ordination (Supervised Machine Learning):

- Learns relationship between predictor (e.g., MBT) and response variables (e.g., COCs & geochemistry)
- Order samples along (biogeochemical) gradients

Regularization:

- Removes predictors that do not align with responses



Applicable Algorithms

Sparse Partial Least Squares (sPLS) and **Sparse Redundancy Analysis (sRDA)** preserve biogeochemical structure & perform feature selection simultaneously

Field-Scale Application

Case Study

Site Background

- Former petroleum hydrocarbon (PHC) bulk storage terminal

COCs

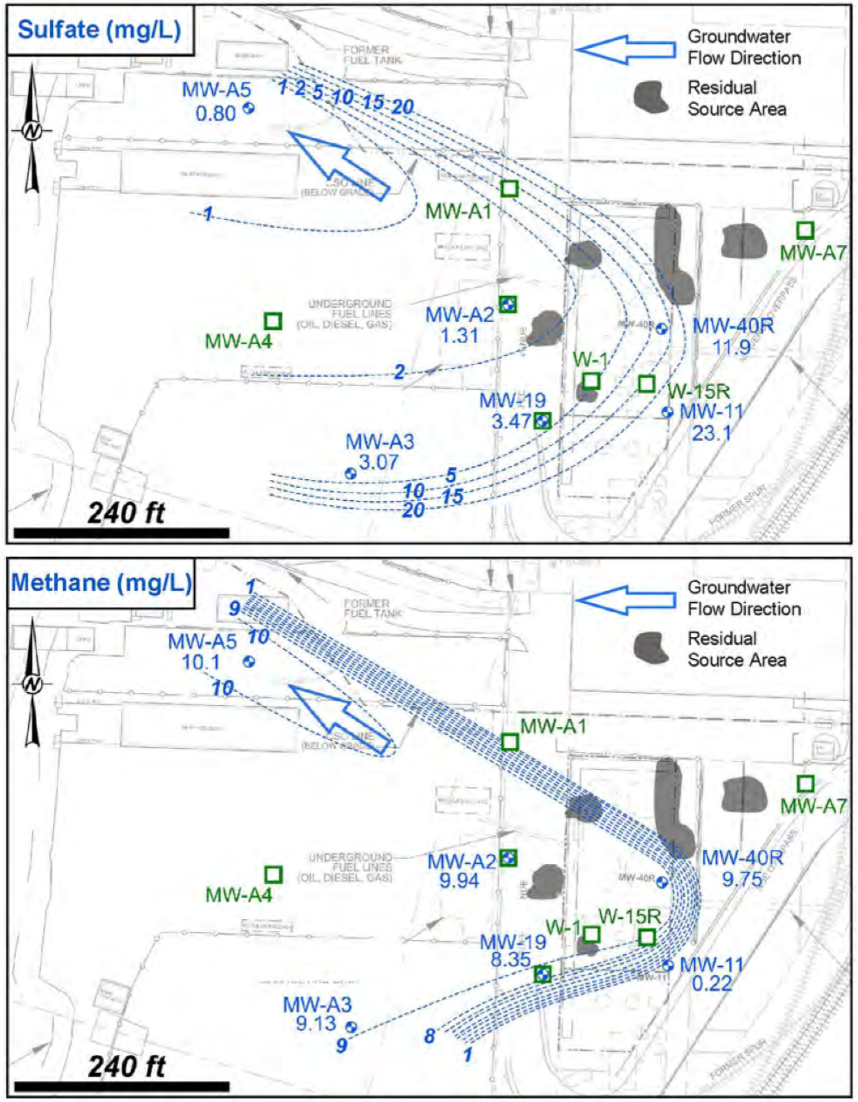
- Dissolved-phase diesel and gasoline groundwater plume emanating from residual NAPL source area

Geochemistry

- Sulfate depletion in source and downgradient areas
- Increased dissolved methane levels in source and downgradient plume

KEY SITE QUESTIONs:

- Is intrinsic biodegradation occurring?
- What are key organisms involved
- Can their contaminant-degrading activities be enhanced?



Madison et al. 2023 – Frontiers in Microbiology

Field-Scale Application

Case Study

Application of Machine Learning

Dataset:

- Groundwater samples collected for COCs and geochemical parameters from upgradient, source, downgradient, and sidegradient monitoring wells
  - Capture spatial extent of COCs and geochemical changes

- 16S rRNA analysis performed for 7 locations (upgradient, source, cross-gradient, and downgradient) to assess microbial community composition
  - Sample locations informed by CSM

Applied **Constrained Ordination** algorithms to assess trends / relationships

Field-Scale Application

Constrained Ordination

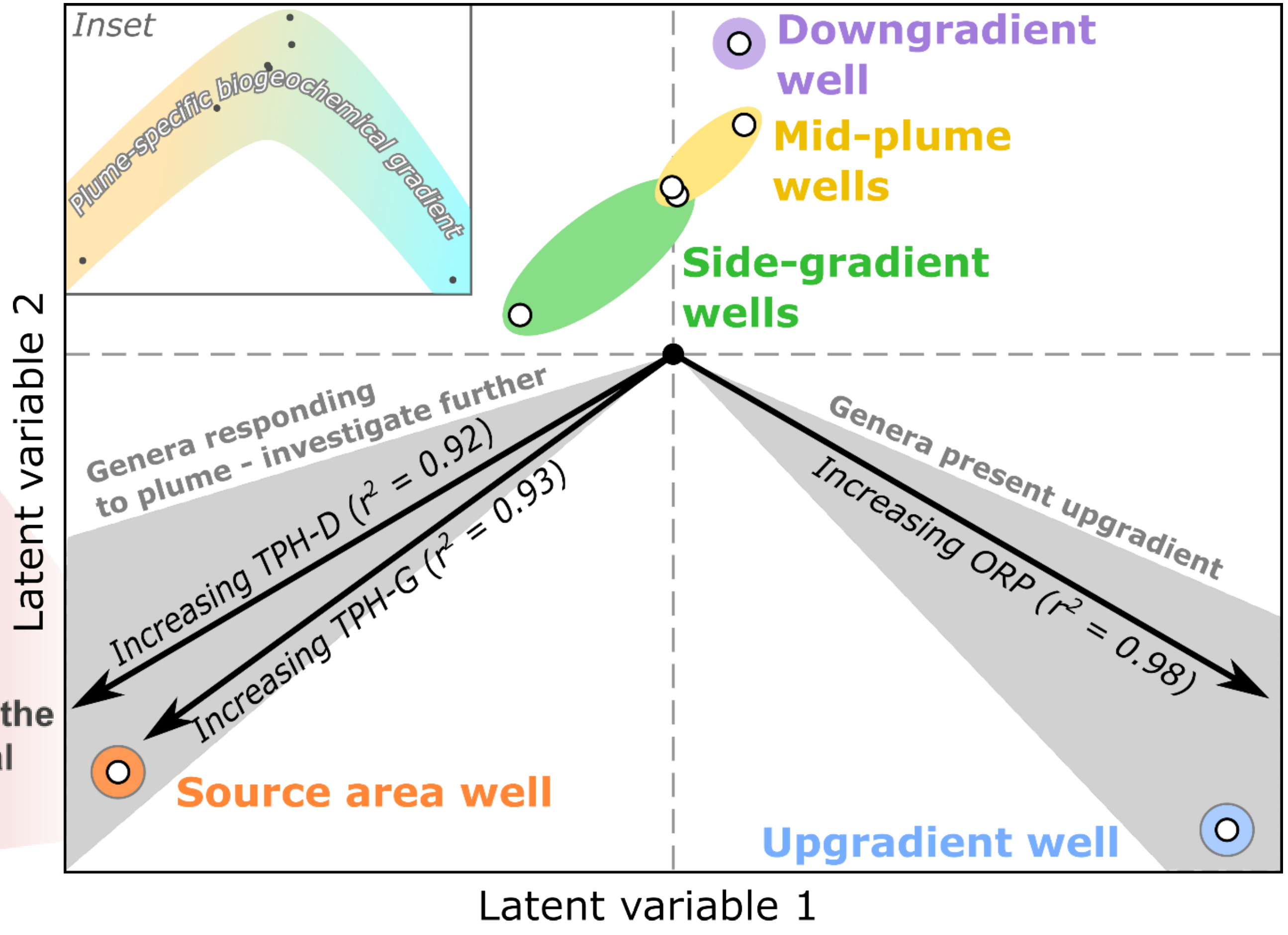
Sparse Partial Least Squares Results (sPLS)

Output: Relevant microbial genera

Acholeplasma
Acidovorax
Actinopolyspora
Anaerobranca
Azospira
Blautia
Caldisericum
Caloramator
Campylobacter
Carboxydocella
Cystobacter
Dehalococcoides
Emticicia
Enterobacter
Heliorestis
Negativicoccus
Petrotoga
Sebaldella
Streptosporangium
Sulfuricurvum
Syntrophus
Thermogemmatispora
Trabulsiella

Feature selection returns a subset of genera that best align with the biogeochemical gradient

sPLS Output



Field-Scale Application

Interpreting Outputs

Correlation of Selected Genera

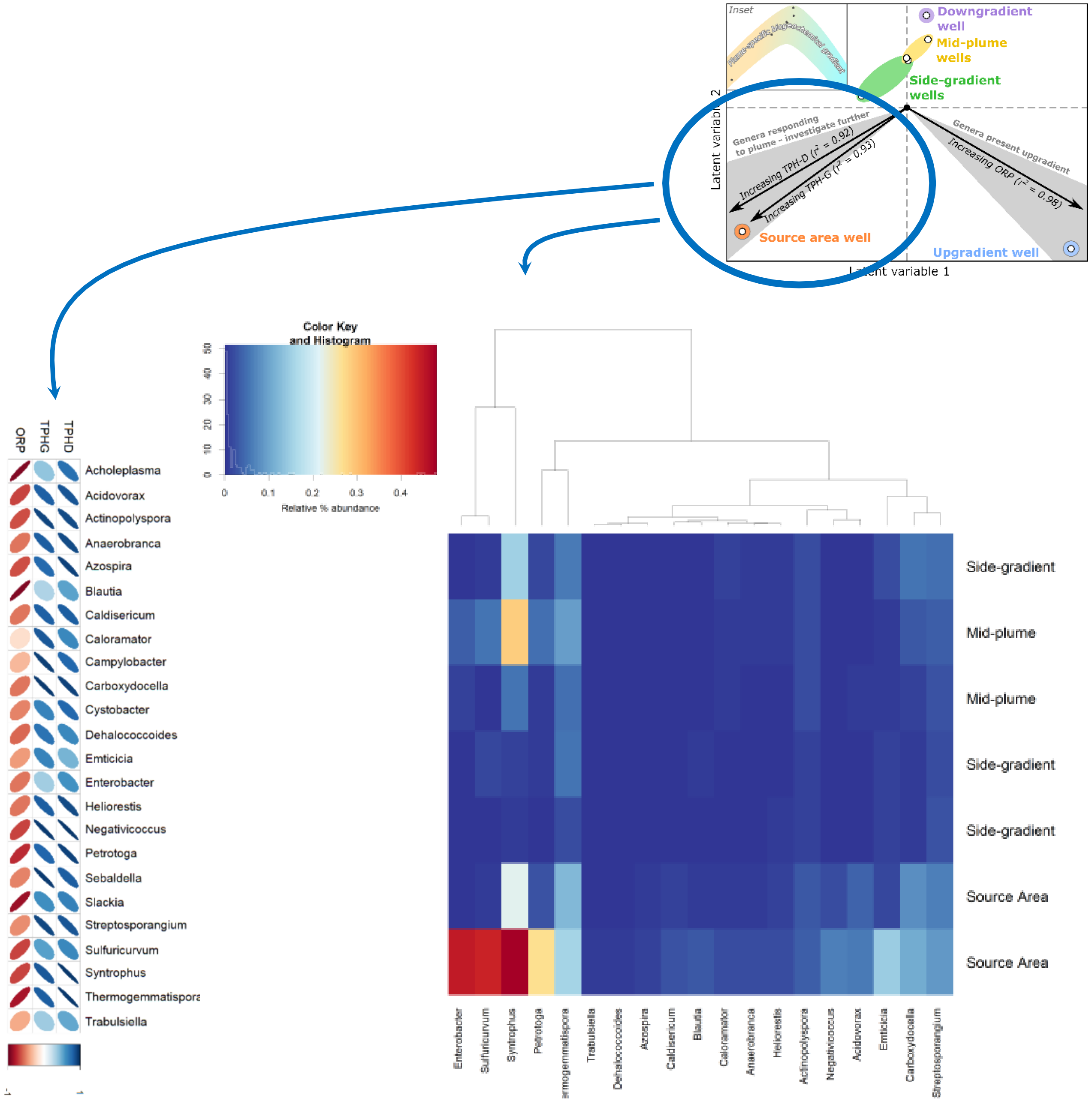
- Mathematical regularization “weights out” most genera and keeps only those [mathematically] responding to the plume
- Bivariate correlation calculated for selected genera
    - Correlation strength increases with TPH and decrease with ORP – aligns with CSM

Selected Genera Abundance

- Constrained ordination tells us which MBTs co-vary with the plume
- Heatmap compares relative abundances to see which genera are both responding *and* potentially thriving

Implications for bioremediation design:

Anaerobes present, including genera that degrade PHCs with sulfate as electron acceptor

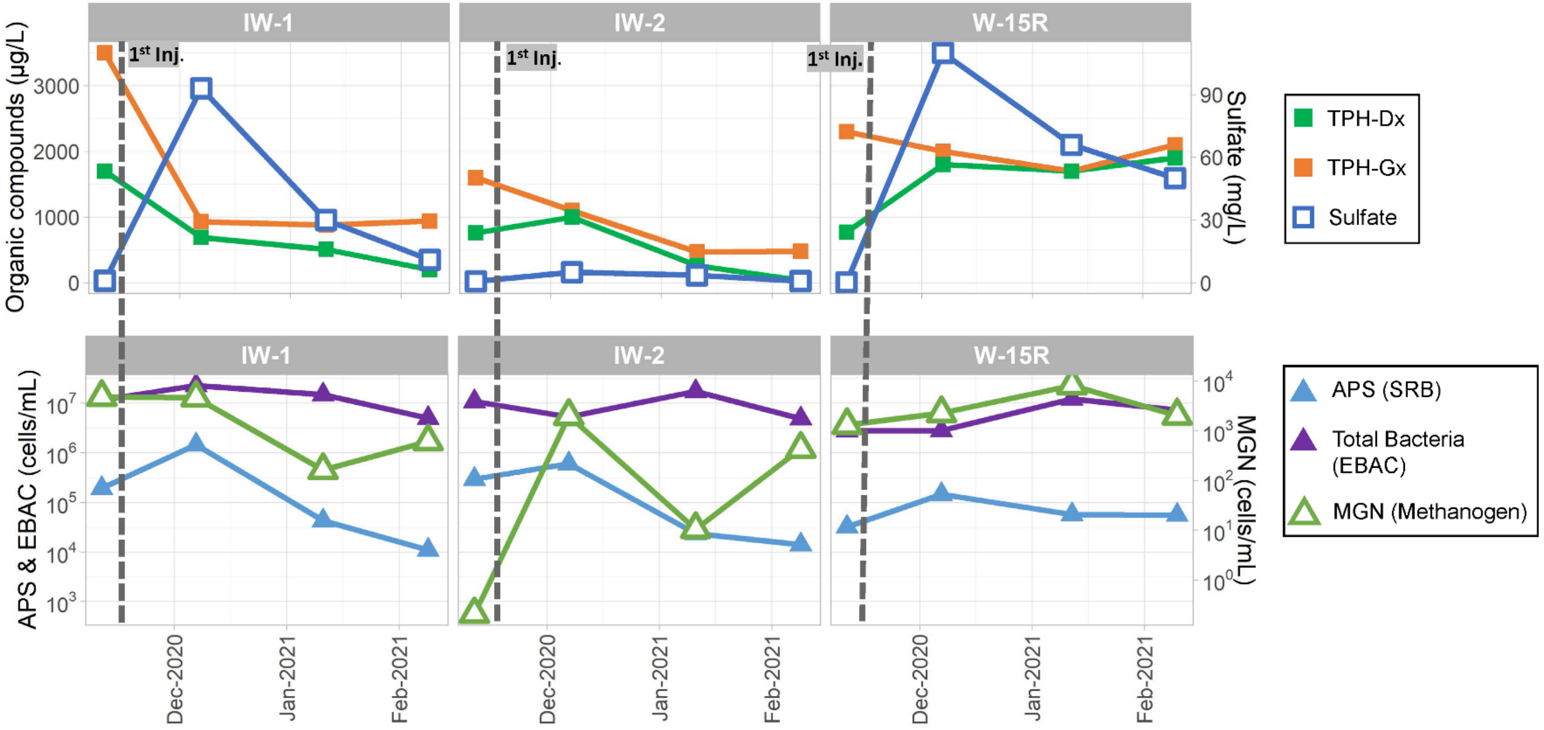


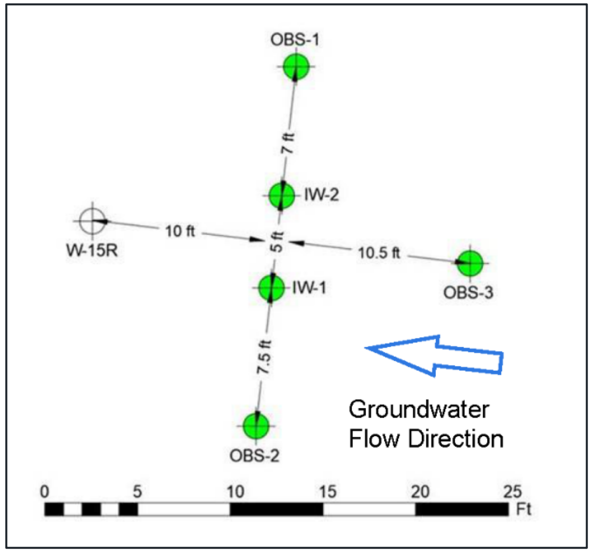
Field-Scale Application

Verifying / Extending the Machine Learning Results

Enhanced In Situ Biodegradation Pilot Study

- Addition of sulfate (magnesium sulfate) to subsurface within source area to stimulate TPH-degrading activities of indigenous sulfate-reducing bacteria






frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 10 February 2023
DOI 10.3389/fmicb.2022.1005871

Increasing *in situ* bioremediation effectiveness through field-scale application of molecular biological tools

Andrew S. Madison[1*], Skyler J. Sorsby[1], Yingnan Wang[2] and Trent A. Key[3]

Next Steps

Extending Machine Learning Applications

- Extending application beyond “simple” sites
- Automated approach to distill multivariate biogeochemical patterns and guide data interpretation

Potential Applications:

- Identification of key microorganisms that potential degrade or transform emerging contaminants
- Application to portfolios to identify sites with potential for bioremediation applications
- Simple visuals for regulator and stakeholder engagement

Key Takeaways

- **Algorithms can help with the design process, but need context in order to be effective**
- Machine learning tools can distill multivariate biogeochemical patterns difficult or infeasible to pick out manually
- "Big data" is not needed – but sufficient biogeochemical data (i.e., gradient) to represent the area of interest is required
- Parameters used to adjust data models require subject matter expertise
- Assess outputs alongside other lines of evidence (e.g., maps, trends, distributions and key MBTs)

Thank you

Andrew Madison, PhD

Vice President and Technical Fellow

andrew.madison@wsp.com

Ph: (609) 450-1325

wsp.com