DOE BSSD Metrics Progress Report Q4: 09/06/2022

SFA Laboratory Research Manager: Paul D. Adams[1], PDAdams@lbl.gov
SFA Technical Co-Manager: Adam P. Arkin[1], APArkin@lbl.gov
[1]**Lawrence Berkeley National Laboratory,** Berkeley CA 94720

Investigators: Paul D. Adams[1,2], Adam P. Arkin[1,2], Nitin S. Baliga[3], Romy Chakraborty[1], Adam M. Deutschbauer[1], Matthew W. Fields[4], Terry C. Hazen[5,6], Trent R. Northen[1], Michael W.W. Adams[7], Eric J. Alm[8], John-Marc Chandonia[1], Aindrila Mukhopadhyay[1], Gary E. Siuzdak[9], David A. Stahl[10], Peter J. Walian[1], Jizhong Zhou[11]

Participating Institutions: [1]**Lawrence Berkeley National Laboratory,** Berkeley CA 94720; [2]University of California at Berkeley, CA 94704; [3]Institute for Systems Biology, Seattle, WA 98109; [4]Montana State University, Bozeman, Mt 59717; [5]University of Tennessee, Knoxville, TN 37916; [6]Oak Ridge National Laboratory, Oak Ridge, TN 37831; [7]University of Georgia, Athens, GA 30602; [8]Massachusetts Institute of Technology, Cambridge, MA 02139; [9]Scripps Research Institute, La Jolla, CA 92037; [10]University of Washington, Seattle, WA 98105; [11]University of Oklahoma, Norman, OK 73072

Q4 Target: Report on capabilities to model the activities of microbial communities in environmental samples

The terrestrial, shallow subsurface is a complex and microbially active habitat located beneath the top-most surface soil layers, comprised of sediments (inorganic or organic unconsolidated material that can originate from the weathering of rock transported by wind, water or ice), rocks, gas, porewater and groundwater [1,2]. In terms of DOE research, subsurface environments contain a large diversity of microorganisms under low nutrient conditions that significantly impacts the carbon, nitrogen, phosphorus, sulfur, and mineral cycles. For example, up to 40% of the microbial biomass and 10^{16}–10^{17} g C on Earth resides within the terrestrial subsurface [3–5]. Typically, subsurface environments contain less labile organic matter (OM) compared to surface soils, and the degree of connectivity to surface waters (*e.g.*, rivers, streams, precipitation) can vary drastically. Although water covers 70% of the Earth's surface, roughly 1% is readily available for human use, and a vast majority (~95%) of the Earth's consumable and available freshwater is groundwater [6–8]. Despite the importance of groundwater for global consumption, agriculture, and industry, the role of microbial communities in the maintenance of groundwater ecosystems is not well understood, particularly for sites impacted by human activity. Understanding microbial community structure and function within the subsurface is critical to assessing overall quality and maintenance of groundwater. A central goal of ENIGMA is to use high-resolution observation of subsurface microbial community dynamics in order to extract critical principles that can explain assembly and activity as well as develop the methods to translate the critical principles to create generalizable models that can predict future microbial community composition and function given environmental constraints.

On an ecosystem scale, there is limited information regarding the exact relationship between community composition, function, and environmental constraints between groundwater and subsurface porous media that can help explain the distribution of microbial biomass and activity that ultimately impacts the fate and transport of nutrients and contaminants of interest. This report and summary of hydrogeological modeling as well as modeling for the associated microbial communities will focus on aspects of chosen sites at the Oak Ridge-Field Research Center (OR-FRC) at the Y-12 Complex. The OR-FRC contains 'shallow' freshwater subsurface environments (mainly porous/granular) that can have a high degree of connectedness with the surface and are impacted by mixed wastes (*e.g.*, organic and inorganic including radionuclides). These shallow, subsurface environments are common across DOE sites that are impacted by a wide array of contaminants that have detrimental impacts on human and environmental health. For example, DOE spends ~$6B/year managing and treating DOE superfund sites, yet the roles of microbes in the subsurface at these sites are still poorly understood and underexploited.

Traditionally, the shallow subsurface can be separated into three distinct zones based on moisture content in relation to the water table configuration termed the vadose, capillary fringe and saturated zones (Figure1). The vadose zone represents the upper most boundary of the subsurface, and following precipitation events, the vadose zone can experience high saturation levels as vertical infiltration proceeds downward to the water table, yet residual pore water can persist creating varying levels of water and gas saturation [9]. The capillary fringe exists at the interface of the saturated and vadose zone, is dynamic, and is highly dependent upon fluctuations of the local water table [4]. This fluctuating interface has been shown to be a 'hotspot' of biogeochemical activity [10,11]. The saturated zone (*i.e.*, at/below water table) of most aquifers consists of

porous material and voids are filled with water. Generally, the direction of water flow in the saturated zone can be in any direction (horizontal and/or vertical).

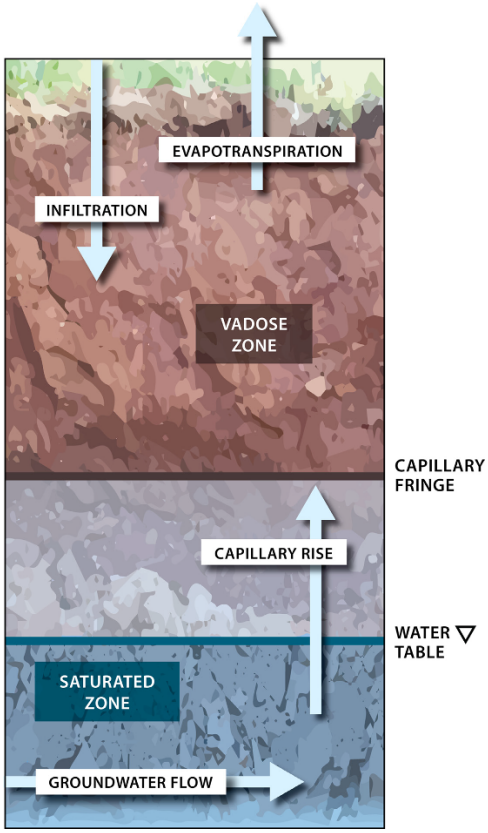

Figure 1. Conceptual illustration of representative shallow subsurface environment that includes the vadose, capillary fringe and saturated zones. Arrows depict the movement of water within and between these zones that creates dynamic conditions

With respect to the subsurface environments dictated by water content and the impact on microbial communities, much attention has been given to the water table position and sediments in the saturated and capillary fringe zones. At least for the OR-FRC, nutrients and contaminants of interest tend to be highest at these boundaries, and therefore, the associated constraints and the impacts on the microbial communities are of high interest. The transitional boundary between the vadose and saturated zones can experience drastic changes in geochemical parameters (*e.g.*, pH and dissolved oxygen), particularly during rain events, and thus, the impacts on microbial activity in terms of geochemical cycling is a target for modeling to enable prediction for the fate and transport of nutrients and contaminants.

Field site. Sites at the OR-FRC are being used to develop and apply our integrated approach that includes microbiologists (ecology, molecular, physiology), biochemists, hydrologists, environmental engineers, and modeling. The FRC contains highly contaminated DOE legacy regions along with excellent field infrastructure available for research instrumentation and 30+ years of biogeochemical data collected by other scientists in previous work. It has well-mapped hydrology and geology and complex gradients of nutrients, stressors and contaminants, making the FRC an excellent site to study the reciprocal interactions of environmental factors on microbial ecology and activity that in turn impact the distribution of nutrients and contaminants. A large number of active wells enable efficient groundwater sampling, and new well installation for depth-indexed sediment cores is relatively cost-effective when needed.

It has become increasingly apparent that microbial populations have distinct physiologies and functions in the shallow subsurface but the potential relationships between biotic and abiotic parameters of the ecosystem are not well understood [2,12,13], particularly hydrogeochemical parameters across the different zones at fine enough resolution. Many questions remain regarding cell interactions with sediments, the distribution and rate of microbial activity, cooperative/competitive microbial interactions, and mechanisms of distribution in the shallow subsurface mixing zones. Due to sampling challenges and the complexity of the heterogeneous subsurface matrix that ranges across the vadose, capillary fringe and saturated zones, few field sites have been comprehensively described and studied despite the important ecosystem functions associated with shallow subsurface systems. The shallow subsurface has historically been considered a stable environment, but it is now clear that temporal and seasonal dynamics influence hydrological mixing

and thus microbial populations, and we have made similar observations at the OR-FRC [14]. Aquifer recharge and fluctuating water table can occur via seasonal patterns, and not surprisingly, the transition zones between the variably saturated and saturated zones have been shown to be an important habitat for microbial diversity and activity. Technological advances for sample retrieval and fine-scale analyses (spatial, temporal, cellular) of subsurface samples are needed, including samplers that can retrieve intact porous media (*i.e.*, sediments) to enable better maintenance of U.S. water sources.

The ENIGMA science focus area (SFA) is a multi-disciplinary, multi-institutional research effort focused on addressing these foundational knowledge gaps by studying groundwater and sediment microbiomes in the shallow subsurface at the contaminated OR-FRC. We seek to discover and characterize the reciprocal interactions between the microbial communities and the geochemical and geophysical parameters of the shallow subsurface within the contamination plume. Our ambition is to do so at sufficient resolution to causally predict the active biotic and abiotic mechanisms mediating key processes such as denitrification, and ultimately predict the future changes in contaminant fate that possibly arise from natural and anthropogenic perturbations. Outcomes are significant both in the fundamental science of community ecology and in gaining an applied understanding of biologically-mediated subsurface processes in contaminated sediments. Moreover, we aim to extract the fundamental principles in order to generalize models to new sites. This report and summary of hydrogeological modeling as well as modeling for the associated microbial communities will focus on our developing framework based upon chosen sites at the OR-FRC.

We have recently organized field information through an integrative model-driven framework inspired by microbial ecology and reactive transport modeling termed **F**ramework for **I**ntegrated, **C**onceptual, and **S**ystematic **M**icrobial **E**cology (FICSME) [15]. The FICSME conceptual model is intended to provide a framework toward mechanistic models of subsurface microbial communities that are genome-informed. For earth system models or models at the scale of a watershed, the contributions and dynamics of microbial communities may be coarsely represented, often simply via microbial biomass. However, at the scale of millimeters to meters at which ENIGMA works, it is much more important to understand how microbes (with potentially different functions) disperse at the site and where they will become abundant. It remains to be understood how these microscale interactions may propagate to the watershed scale and beyond (if at all). The FICSME model is a way to coordinate field work with laboratory experiments (as described in our Q1, Q2, and Q3 reports) necessary for parameterization. For example, laboratory experiments can provide data on how microbial members respond to different concentrations of metals [16], but field measurements provide information about the resident microbes at the site, dispersal paths, biogeochemical ranges, *etc*. The FICSME model is initially a means to formally enumerate the individual components, and the variables in the conceptual equation can be mapped to the taxonomic, chemical and physical entities identified as important during our research and the equation terms track how they interact to affect spatiotemporal abundance. In particular for this report, groundwater flow and geochemical dynamics have been observed in relation to microbial populations for the developing FICSME model (*e.g.*, how water flow contributes to biological and chemical dispersal).

The ENIGMA program is designed to allow increasingly high-resolution and multimodal characterization of subsurface biogeochemical dynamics in situ and controlled perturbation of those dynamics for training and testing models designed to identify critical environmental constraints on microbial activity. Based on

field data, parameters that constrain the FICSME model in terms of water flow, geochemical dynamics, and subsequent changes in microbial populations are being estimated. Ultimately, the model is being built to estimate and predict the fate of nutrients and contaminants. At its highest resolution these measures will be integrated into genome-informed reactive transport models, but it is an active area of research to start with smaller heterogeneous statistical, mechanistic, and increasingly causally informed machine learning models to build towards a robust integrated community model. The below sections provide examples of field work that provides the means to sample parameters of interest as well as the incorporation of field data into hydrogeochemical, metabolic, and ecological models to determine the relationships between water flow and geochemistry to the distribution of microbial cells and activities important to contaminants of interest (*e.g.*, nitrates).

Statistical modeling for microbial populations

There is an overwhelming number of variables that impact the dynamics of any environmental process. Modeling everything, even everything measurable, is neither an option nor desirable. It is better to derive the smallest set of variables that have the maximal predictive power so that attention can be focused on core principles and actionable information. It is also desirable that the predictions are driven by causal explanations because these are more generalizable to situations not yet observed and are more likely to allow specific intervention to change outcomes. Therefore, a primary task in ENIGMA wherein we are measuring nearly 100 chemical species, tens of other environmental variables, tens of thousands of microbial species and millions of genes is to determine which of these are most proximally responsible for the observed critical processes and their persistence over time and multiple events.

The first approaches tend to use fairly simple statistical approaches in which various forms of statistical hypothesis testing (e.g. Mantel tests), static regressions (*e.g.*, ANOVA, ordinations), or temporal statistics (*e.g.*, time-lagged correlations or auto-regressive moving average models) are used to identify specific observed variables which correlatively predict potential outcomes. These can become quite sophisticated in some cases with structured relation models, for example, and corrections for the compositional nature of especially the metagenomic data. ENIGMA has at one time or another used most of these to analyze data while moving towards the FICSME model.

Detection of microbial interactions in any environment let alone a largely inaccessible subsurface environment can be challenging. There are theoretical and practical reasons why this may be even more difficult than hoped due to the dynamic nature of microbial communities. But even deriving an average sense of the interactions among microbial taxa from, for example, cross-correlation of temporal or spatial co-occurrence presents challenges. First, correlation is not causation so without understanding the mechanisms by which organisms could interact we are left with weak inference at best even when controlling for indirect effects. Second, the nature of microbiome measures tends to permit only relative abundances to be measured. That is, the absolute amount of a species is not being measured, but rather a relative percentage within some error that can lead to errors in estimation of relation. Early on, ENIGMA developed some of the first methods to control for this sort of error and produced a popular tool called SPARCC [17]. In this work we showed how not properly controlling for these effects could lead to completely erroneous inference of correlations among organisms. It also began to inform exactly how

much data would be necessary to estimate these interactions. These considerations have fed into the design of the ecological modeling frameworks and associated designs of experiments below.

One of the first follow-on models we attempted brought in an interpretable machine learning method called random forests [18]. Our goal was to predict the distribution of key environmental variables across the Y-12 site from the composition of the microbial community observed. The idea being that the most predictive community members were likely causally linked to the chemical process being predicted and the combinations of members necessary for accurate prediction were candidates for performing complementary functions. To power this model, we needed to sample enough locations with variation in pH, uranium and nitrate to cover the relevant ranges and in enough different 'contexts' to render the predictions robust to environmental variation. While it is still a challenge to rationally design these sorts of experiments, based on prior field surveys we were able to pick ~100 groundwater wells that had been previously operating and we measured these environmental variables along with amplicon sequencing of the communities. From this, we were able to show using random-forest regression that there were critical taxa, slightly different sets for each of over 26 environmental variables, that could fairly accurately forecast the abundance of each chemical and found that many of these had mechanisms expected to be related to the forecasted variable. Further, the same modeling framework was successfully applied to hydrocarbons released during the Exxon Valdez oil spill showing its generalizability. Variations of this framework have, since then, become increasingly sophisticated and allow the integration of more types of data and better representation of causal priors [19,20]. However, we had little sense how these predictions would vary over time and why the relationships can change.

In a more recent study [14], three shallow wells (FW301, FW303, FW305) in a non-contaminated shallow aquifer in the OR-FRC were sampled approximately 3 times a week over a period of three months to measure changes in groundwater geochemistry and microbial diversity. The wells displayed some degree of hydrochemical variability over time unique to each well, and microbial community composition of a given well was on average > 50% dissimilar to any other well at a given time (days). Yet, functional gene diversity remained relatively constant. These results indicated that up to half of a microbial community could change within a couple of days, likely related to hydrogeochemical changes at the local scale. In addition, despite high turnover in microbial populations, the overall metabolic functions of the entire community would not change significantly which indicated some degree of functional redundancy. Similarity percentage (SIMPER) analysis revealed that variability in the wells differed in the impacts between low abundance, highly-transient populations and more highly abundant and frequently present taxa. These results suggested local scale effects between wells likely related to a combination of heterogeneous flow and geochemistry. Despite these differences, time-series analysis of all three wells using vector auto-regressive models and Granger causality showed unique relationships between species number and geochemistry over time in each well, highlighting local-scale effects that can be tested in the laboratory and the field. The results indicated temporally dynamic microbial communities over short time scales, with day-to-day population shifts in local community structure influenced by available source community diversity and local groundwater hydrochemistry.

The results above demonstrated the need to conduct *in situ* measurements with increased temporal resolution for both hydrogeochemical parameters as well as microbial populations that perform different

functions in order to capture local scale changes that were potentially related to the distribution of different microbial populations associated with activity of interest (*i.e.*, denitrification).

Finer Resolution of Microbial Populations and Associated Metabolic Potential
Ecological modeling to dissect ecological processes

The diversity, structure, and succession of biological communities are governed by complex ecological processes, such as selection, dispersal (immigration and emigration), diversification (speciation and extinction), and drift (random birth and death)[21–23]. These processes can be deterministic (selection) or include obvious stochasticity (drift, dispersal and diversification). Disentangling these ecological processes is crucial, but challenging, in ecology, especially for microorganisms, and various qualitative or quantitative approaches have been developed based on multivariate analyses, neutral theory models, or ecological null models [24].

Considering the limitations of current methods, we developed two new approaches to quantitatively assess different processes based on ecological null models [25,26]. One is a general framework to estimate the relative importance of stochastic ecological processes in shaping community structure, with a new index, normalized stochasticity ratio (NST) [25]. The other is a quantitative framework (iCAMP) to estimate the relative influence of selection, dispersal, and drift in different phylogenetic groups [26]. We tested the new approaches with simulated communities and demonstrated substantial improvement in quantitative and qualitative performance. The two methods have been widely applied to microbial communities in various natural, engineered, and host-associated systems (animal and plants).

Enabled by the new approaches, we further explored our newly proposed framework about the general relationship between environmental stress and ecological processes. Environmental stresses are major drivers of community variation, but little is known about effects on microbial assembly mechanisms. We proposed a framework with four major schemas about how stresses affect ecological stochasticity, selection, dispersal, and drift in microbial community assembly. Our field site has extremely large ranges of various geochemical properties, and we ranked the stress levels of samples from our site with a proposed comprehensive evaluation index based on >30 measured factors. The relative importance of each ecological process was assessed by our new approaches, and the results clearly supported our hypotheses, showing an increase in selection and decreases in stochasticity, dispersal limitation, and drift as stress increased. We further investigated the associations between environmental factors and ecological processes with improved statistical methods. The results indicated the heterogeneous selection (i.e., selection leading to more dissimilar communities) was primarily related to abnormal pH, deficiency of cobalt and molybdenum, and high concentrations of some heavy metals. Moreover, the spatial distribution patterns of heterogeneous selection and dispersal limitation corresponded fairly well with major contaminants and geo-hydrological characteristics.

Revealing different mechanisms in different phylogenetic groups (bins) is a major advantage of our new approach (iCAMP). In the over 260 microbial groups observed at our site, 60%, 31%, and 6.5% were dominated by dispersal limitation, heterogeneous selection, and drift, respectively (Figure X). The most abundant three groups governed by heterogeneous selection were from as yet uncultivated phyla. Interestingly, the ENIGMA team has obtained quite a few isolates that are very similar (>97%) to the

dominant taxa in the top dispersal- or drift-governed groups, but much fewer isolates similar to those in top selection-controlled bins, probably due to narrow niche preference to the specific *in situ* environment.

As field work continues, we are working on the further development of the ecological modeling approach to include trait-based information (e.g., functional genes from metagenomic analysis) and temporal ecological null models, as well as a new framework leveraging ecological modeling results to improve our reactive transport modeling.

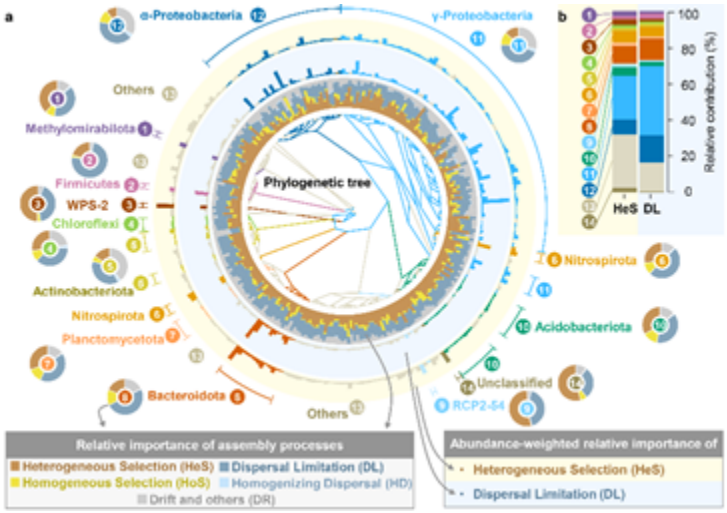

Figure 2. Ecological processes in different phylogenetic groups (bins). a. Phylogenetic tree (center), relative importance of different processes (inner annulus), and abundance weighted importance of dispersal limitation (DL, middle annulus) and heterogeneous selection (HeS, outer annulus) in each bin, as well as the relative importance of different processes in each phylum and each class of proteobacteria (donut plots by the phylum/class names). b. Relative abundance of different phyla/classes governed by HeS or DL.

The ecological modeling results demonstrated that different microbial groups are impacted by different stressors, and this in turn, can cause different rates of stochasticity, selection, dispersal, and drift that results in different microbial community structures. The different stressors can be ranked, and results highlighted the importance of pH extremes as well as the presence/absence of particular heavy metals. The identified populations/consortia become targets for laboratory work to further inform genetic and physiological responses to hydrogeochemcial parameters for the FICSME model.

In an effort to integrate microbial functional data (*i.e.*, biochemical catalysis) with field-relevant constraints, models intermediate between highly detailed reactive transport models and more statistical models can be constructed. In these cases, microbes are often 'models' as classes of biocatalyst based upon genomic traits which are then converted to chemical reaction sets and thermodynamic constraints that impact processes of interest. Members of ENIGMA have used models such as the microbial enzyme-mediated decomposition (MEND, [27]) to carry out these analyses. While these models simplify the incorporation of biological information and provide a natural means for integrating genomic information with field scale models, the experimental designs necessary to calibrate and test these models are relatively sophisticated, time-consuming and expensive. In general, statistical models as above are used to identify the critical environmental, taxonomic and functional relationships that must be considered from high-resolution time-space-condition field data. Functional assignments are made for observed taxa (inferred from amplicon sequence-based classification into a microbial functional class or through metagenomic gene functional analysis). If quality annotations can be achieved, biochemical pathways can be estimated and fit into standard models of geochemical transformations.

To fit model kinetics, ample time-series and perturbation data must be present to allow both training and testing of the models. The MEND model is based upon the consumption of soluble organic matter along

with the incorporation of fundamental variables that impact the reactions. These include pH, soil water potential, temperature and direct action of biomass. There have now been a number of successful uses of such models for carbon cycle processes at large scales in grasslands over a 12 year period [28,29] and to predict the effects of warming. While these models can be powerful in predicting the overall environmental processes at scale, the opposite question, why specific microbes are present or persistent or together at specific locations over time is not directly answered. To connect the mechanisms that adapt an organism and its community to a particular location for carrying out a particular process may require a finer grained modeling framework. This may also be the level required if organisms (or communities) are ever to be engineered effectively for operation in open environments.

Finer Resolution of Hydrogeochemical Parameters Over a Designated Space and Time

The reactive transport model (RTM) in groundwater is aimed to quantitatively describe and predict the distribution of chemicals accounting for the transport and transformations, both abiotic and biotic, integrating various active hydrological, geochemical, and microbial processes [30,31]. Different processes can be represented in a RTM at multi-spatial scales, from nanometers to hundreds of meters and beyond. For example, flow and advection of solutes can be simulated at pore scale (as fine as nanometer) when the individual pores are represented explicitly and can also be simulated at larger scales (up to field scale) when the porous medium is treated as a continuum. Models for reactive microbial processes can be developed for individual isolates in the laboratory and larger spatial scales when specific microbial impacts on environments are considered. The increasing capability of sampling and chemical and microbial meta-omics analysis allows us to capture the subsurface heterogeneity at a fine scale. However, the computational cost of fine-scale simulations is very high if such models are applied to a larger scale site. Therefore, developing multiscale models makes it feasible to incorporate the fine-scale information in larger-scale applications with acceptable computational costs.

Interests in microbial communities in subsurface systems are growing as contributors to global carbon and nitrogen cycling in relationship to hydrogeochemical parameters, but as noted above, the exact relationships between these parameters and the distribution of microbial communities and activities is still poorly understood and is difficult to predict. This is because microbial community composition and its activity depend on site-specific environmental conditions, such as rainfall perturbations and heterogeneous hydraulic conductivity that can be linked to dissolved oxygen dynamics. In the modeling of N turnover in the shallow subsurface, it is necessary to incorporate the changes of the microbial community in response to hydrological perturbations and the subsequent biogeochemical processes. In order to achieve this type of integrated modeling, we will develop a reactive transport model of nitrogen in the subsurface system for the ENIGMA Subsurface Observatory (SSO) site. The model utilizes the FICSME framework [15] as the foundation for developing an effective reactive transport code in conjunction with a module utilizing omics and community data. The model is parameterized and calibrated by many different types of data collected by the ENIGMA SFA (e.g., meteorological, hydrological, microbial, geophysical, and geochemical datasets). As a novel capability to the modeling community, our work incorporates omics data into the model, connecting the microbial community to nitrogen cycling in the meso-field scale subsurface system, and allows the microbial community to evolve in response to geochemical and flow conditions.

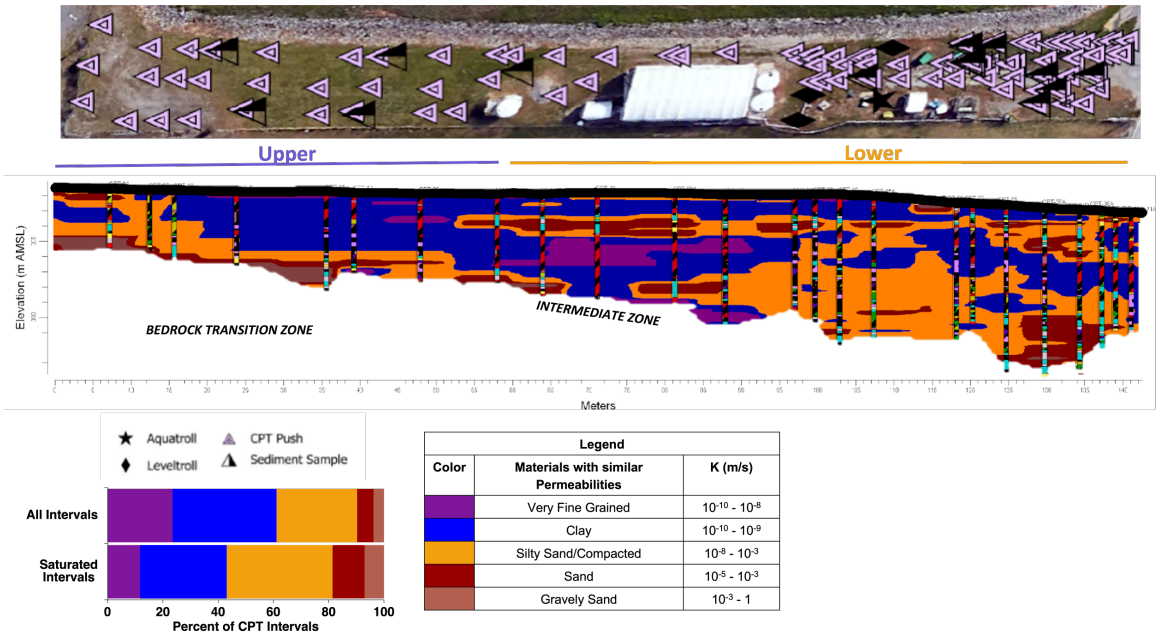

Figure 3. Cone Penetrometer Testing (CPT) identified sediment behavior types to refusal bat the intermediate zone overlying the weathered bedrock. A zone that has long been identified as a major zone of flow, recharge, and transport. The cross section of a three-dimensional volumetric model was constructed to quantify permeability characteristics of Area 3. The volumetric model used five classifications with a blending method to solve for subsurface volumes between CPT pushes where direct linear mapping was insufficient. Multiple iterations of a random model generated probable solutions and identified silty sand/compacted material as the most abundant soil type. Image: Andrew Putt UTK

High-resolution three-dimensional models of the shallow subsurface had not been previously done at the Y-12 site [32]. The team used for the first time a cone penetrometer (CPT) to effectively identify areas of lesser heterogeneity where sample composition was largely homogeneous at a ninety-five percent confidence interval. 130 pushes were made in 16 days from the surface to 6-9 m depths. The penetration testing identified a majority of the similar soil composition profiles to be located within the saturated material of lower area 3 and identified the abundant soil types as silty and compacted sediments at the potential SSO site (near wells of interest FW106, FW112, and the newly drilled tracer testing flow path well EFPW01). During CPT activities, a significantly higher and direct response in local well geochemistry and water level was measured in the tested wells, suggesting decreased connectivity in certain zones between the wells. Activities with the greatest response have been identified near a former buried stream channel and in material where advanced weathering and dissolution is likely occurring. This unequal and directional response may be related to soil properties (*e.g.*, macropores in fine grained materials) and the preferential flow identified in this study site rather than a direct relationship to soil behavior type. A subsequent analysis [33] suggests that pore pressures measured by the CPT penetration were highest in sediments where penetration activities were not identifiable. These results further suggest that the CPT technique was able to identify and distinguish transport pathways circumventing the intact, low-flow clay matrix extremely common in engineered and backfilled sites. These data have been used to construct estimates of permeability which are being tested in laboratory columns in combination with the FICSME modeling objective in an effort to model carbon and nitrogen cycling under environmentally relevant perturbations. These studies were critical for establishing new wells that will constitute the SSO.

In a more microbial/functional group centric approach, we have developed a microbial enzyme functional group-based RTM at the field-scale, in which the microbial-mediated biogeochemical reactions are presented through enzyme groups, representing the overall microbial function dynamics at the modeling region. The meta-omics measurements will be incorporated into the model calibration and validation.

We also built a primary flow field and solute transport model using the commonly applied MODFLOW and MT3D-USGS models. In addition, we modified MT3D-USGS to incorporate multiple microbial functional groups, which mediate the critical metabolic processes in groundwater, such as organic matter fermentation, methanogenesis, denitrification, nitrification, dissimilatory sulfate reduction/oxidation, and uranium (VI) reduction. Since the groundwater reactive transport process combines complex hydrological, geochemical, and microbial processes, it is challenging to estimate various types of parameters. Therefore, we developed a parameter optimization procedure for field-scale groundwater RTM based on the Shuffled Complex Evolution (SCE) algorithm. To enhance the computation efficiency, we developed the parallel computation program with the OpenMPI interface on a supercomputer.

The new model was applied to a historical time-series data (13 months) of emulsified vegetable oil injection activity at the OR-FRC site as an example to show the development and performance of the model. The total area of the modeling zone was around 3,700 m^2, covering 21 observation wells. The modeling area was discretized into grids with 0.5 m × 0.5 m × 2 m for the finite-difference calculation. Firstly, we used the MODFLOW and water level observations to estimate the spatial distribution of hydraulic parameters, i.e., hydraulic conductivity and recharge rate, and simulate the transient flow field after EVO injection. The mean residual of the water heads in the optimization was <0.01 m. Then the MT3D-USGS and the developed SCE optimization algorithm were applied to calibrate the baseline model and simulated the chemical transport processes. Two recommended statistical measures, correlation coefficient and PBIAS [34,35], were adopted to evaluate the RTM performance. The results showed that the model simulations for tracer test and chemical transport were all at the level of acceptable performance (correlation coefficient ≥0.42, PBIAS <±70%) or better for most observation wells and all observed chemicals (bromide, acetate, nitrate, sulfate, and uranium). Next, we will integrate the omics data into the model calibration and validation. By comparing with the baseline RTM, we will evaluate the omics-informed RTM to see if we can improve modeling performance or reduce the uncertainty by introducing necessary functional enzymes/genes.

Summary

In the initial stages of characterizing the OR-FRC, we started with modeling marker genes (both general and functional) representing microbial communities in relation to expanding geochemical parameters. Along these lines, tools were also developed to minimize cross-correlations of temporal and/or spatial data that typically co-occurs within field data. This has been an issue that challenges microbial ecology of any environment (from soil/water to human health), and work from the ENIGMA group helped account for these challenges. Subsequent model development incorporated machine-learning techniques to gain predictive power between identified combinations of microbial groups that aligned with environmental variation (*e.g.*, occurence of heavy metals or radionuclides). Variations of this framework have been further adapted to also be predictive at different field sites with different conditions. However, as we worked to include and sample sediments and sediment zones along with groundwater, we realized the underappreciated variability in microbial communities along with the dynamic nature of hydrogeochemical

parameters in the context of both low-time frame and high-time frame ecological processes that were important in shaping the *in situ* microbial communities. Not only do these processes need to be measured at increasing spatial and temporal resolution, but microbial traits need to be more integrated within reactive transport models that represent important hydrogeochemical parameters that dictate macro-scale processes. In essence, how do we inform landscape models with meaningful microbial signatures and behaviors that function at the micro-scale but impact processes of interest (*e.g.*, N-cycling) at the macro-scale? Therefore, we are now co-designing the field observatories to help calibrate and test genome-informed reactive transport models to capture microscale flow dynamics that represent field parameters and that can incorporate microbial genetic and phenotypic signatures and predict the fate and transport of contaminants and processes of interest. Other parts of ENIGMA focus on identifying the key signatures (microbial groups and their activities) and the environmental constraints that play significant roles in the processes of interest, and these parameters are then used in the developing models. Looking to the future, we aim to use the FICSME framework to demonstrate the ability to predict how perturbations to bio-systems (e.g., soil, sediments, groundwater) impact ecosystem services and vice versa and sites within and beyond the Y-12 site.

Bibliography

1 Atekwana EA, Werkema DD & Atekwana EA (2006) Biogeophysics: the effects of microbial processes on geophysical properties of the shallow subsurface. In *Applied Hydrogeophysics* (Vereecken H, Binley A, Cassiani G, Revil A, & Titov K, eds), pp. 161–193. Springer Netherlands, Dordrecht.

2 Smith HJ, Zelaya AJ, De León KB, Chakraborty R, Elias DA, Hazen TC, Arkin AP, Cunningham AB & Fields MW (2018) Impact of hydrologic boundaries on microbial planktonic and biofilm communities in shallow terrestrial subsurface environments. *FEMS Microbiol Ecol* **94**.

3 Whitman WB, Coleman DC & Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95**, 6578–6583.

4 Griebler C & Lueders T (2009) Microbial biodiversity in groundwater ecosystems. *Freshw Biol* **54**, 649–677.

5 McMahon S & Parnell J (2014) Weighing the deep continental biosphere. *FEMS Microbiol Ecol* **87**, 113–120.

6 Danielopol DL & Griebler C (2008) Changing Paradigms in Groundwater Ecology - from the 'Living Fossils' Tradition to the 'New Groundwater Ecology.' *Int Rev Hydrobiol* **93**, 565–577.

7 Griebler C, Malard F & Lefébure T (2014) Current developments in groundwater ecology--from biodiversity to ecosystem function and services. *Curr Opin Biotechnol* **27**, 159–167.

8 Dennehy KF, Reilly TE & Cunningham WL (2015) Groundwater availability in the United States: the value of quantitative regional assessments. *Hydrogeol J* **23**, 1629–1632.

9 Jones AA & Bennett PC (2014) Mineral microniches control the diversity of subsurface microbial populations. *Geomicrobiol J* **31**, 246–261.

10 Silliman BR & Bertness MD (2002) A trophic cascade regulates salt marsh primary production. *Proc Natl Acad Sci USA* **99**, 10500–10505.

11 Berkowitz B, Silliman SE & Dunn AM (2004) Impact of the capillary fringe on local flow, chemical migration, and microbiology. *Vadose zone j* **3**, 534–548.

12 Hall-Stoodley L, Costerton JW & Stoodley P (2004) Bacterial biofilms: from the natural environment to infectious diseases. *Nat Rev Microbiol* **2**, 95–108.

13 Anneser B, Pilloni G, Bayer A, Lueders T, Griebler C, Einsiedl F & Richters L (2010) High resolution analysis of contaminated aquifer sediments and groundwater—what can be learned in terms of natural attenuation? *Geomicrobiol J* **27**, 130–142.
14 Zelaya AJ, Parker AE, Bailey KL, Zhang P, Van Nostrand J, Ning D, Elias DA, Zhou J, Hazen TC, Arkin AP & Fields MW (2019) High spatiotemporal variability of bacterial diversity over short time scales with unique hydrochemical associations within a shallow aquifer. *Water Res* **164**, 114917.
15 Lui LM, Majumder EL-W, Smith HJ, Carlson HK, von Netzer F, Fields MW, Stahl DA, Zhou J, Hazen TC, Baliga NS, Adams PD & Arkin AP (2021) Mechanism across scales: A holistic modeling framework integrating laboratory and field studies for microbial ecology. *Front Microbiol* **12**, 642422.
16 Goff JL, Szink EG, Thorgersen MP, Putt AD, Fan Y, Lui LM, Nielsen TN, Hunt KA, Michael JP, Wang Y, Ning D, Fu Y, Van Nostrand JD, Poole FL, Chandonia J, Hazen TC, Stahl DA, Zhou J, Arkin AP & Adams MWW (2022) Ecophysiological and genomic analyses of a representative isolate of highly abundant *Bacillus cereus* strains in contaminated subsurface sediments. *Environ Microbiol*.
17 Friedman J & Alm EJ (2012) Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**, e1002687.
18 Smith MB, Rocha AM, Smillie CS, Olesen SW, Paradis C, Wu L, Campbell JH, Fortney JL, Mehlhorn TL, Lowe KA, Earles JE, Phillips J, Techtmann SM, Joyner DC, Elias DA, Bailey KL, Hurt RA, Preheim SP, Sanders MC, Yang J & Hazen TC (2015) Natural bacterial communities serve as quantitative geochemical biosensors. *MBio* **6**, e00326-15.
19 Lagergren J, Cashman M, Melesse Vergara VG, Eller PR, Gazolla JGFM, Chhetri HB, Streich J, Climer S, Thornton P, Joubert W & Jacobson D (2021) Climatic clustering and longitudinal analysis with impacts on food, bioenergy, and pandemics. *BioRxiv*.
20 Walker AM, Cliff A, Romero J, Shah MB, Jones P, Felipe Machado Gazolla JG, Jacobson DA & Kainer D (2022) Evaluating the performance of random forest and iterative random forest based methods when applied to gene expression data. *Comput Struct Biotechnol J* **20**, 3372–3386.
21 Hanson CA, Fuhrman JA, Horner-Devine MC & Martiny JBH (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* **10**, 497–506.
22 Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P & Ferrenberg S (2013) Patterns and processes of microbial community assembly. *Microbiol Mol Biol Rev* **77**, 342–356.
23 Vellend M (2010) Conceptual synthesis in community ecology. *Q Rev Biol* **85**, 183–206.
24 Zhou J & Ning D (2017) Stochastic community assembly: does it matter in microbial ecology? *Microbiol Mol Biol Rev* **81**.
25 Ning D, Deng Y, Tiedje JM & Zhou J (2019) A general framework for quantitatively assessing ecological stochasticity. *Proc Natl Acad Sci USA* **116**, 16892–16898.
26 Ning D, Yuan M, Wu L, Zhang Y, Guo X, Zhou X, Yang Y, Arkin AP, Firestone MK & Zhou J (2020) A quantitative framework reveals ecological drivers of grassland microbial community assembly in response to warming. *Nat Commun* **11**, 4717.
27 Wang G, Jagadamma S, Mayes MA, Schadt CW, Steinweg JM, Gu L & Post WM (2015) Microbial dormancy improves development and experimental validation of ecosystem model. *ISME J* **9**, 226–237.
28 Wang G, Gao Q, Yang Y, Hobbie SE, Reich PB & Zhou J (2022) Soil enzymes as indicators of soil

function: A step toward greater realism in microbial ecological modeling. *Glob Chang Biol* **28**, 1935–1950.

29 Guo X, Gao Q, Yuan M, Wang G, Zhou X, Feng J, Shi Z, Hale L, Wu L, Zhou A, Tian R, Liu F, Wu B, Chen L, Jung CG, Niu S, Li D, Xu X, Jiang L, Escalas A & Zhou J (2020) Gene-informed decomposition model predicts lower soil carbon loss due to persistent microbial adaptation to warming. *Nat Commun* **11**, 4897.

30 Bedekar V, Morway ED, Langevin CD & Tonkin MJ (2016) MT3D-USGS version 1: A U.S. Geological Survey release of MT3DMS updated with new and expanded transport capabilities for use with MODFLOW. *Techniques and Methods*.

31 Meile C & Scheibe TD (2019) Reactive transport modeling of microbial dynamics. *Elements* **15**, 111–116.

32 Putt AD, Kelly ER, Lowe KA, Rodriguez M & Hazen TC (2022) Effects of cone penetrometer testing on shallow hydrogeology at a contaminated site. *Front Environ Sci* **9**.

33 Kelly E (2021) Influence of Physical Variability of Highly Weathered Sedimentary Rock on Nitrate in Area 3 of the ENIGMA Field Research Site at Y-12. .

34 Moriasi DN, Gitau MW, Pai N & Daggupati P (2015) Hydrologic and water quality models: performance measures and evaluation criteria. *TransASABE* **58**, 1763–1785.

35 Moriasi DN, Arnold JG, M. W. Van Liew, Bingner RL, Harmel RD & T. L. Veith (2007) Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *TransASABE* **50**, 885–900.