ENIGMA DOE BSSD Metric Progress Report Summary: 09/29/2022

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

BSSD 2022 Performance Metric Summary

Goal: Develop new technologies for assessing microbial ecological processes in environmental samples.

Introduction

The ENIGMA (ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies) SFA is a consortium of 16 investigators at eleven institutions across the country led by Lawrence Berkeley National Laboratory. Established in 2009, ENIGMA researchers collaborate to create increasing multiscale, causal and predictive models of the reciprocal impacts of microbial communities on critical processes (e.g., Carbon and Nitrogen-cycling) within open ecosystems. We are specifically focused on addressing foundational knowledge gaps in groundwater and sediment microbiomes in the shallow subsurface at the contaminated Oak Ridge Field Research Site (OR-FRC) as a testbed for our approaches. Waste generated by research and production of nuclear materials at this site, including nitrate, acidity, radionuclides, and volatile organic carbons, have contaminated the watershed that passes across the site. The persistence and flow of these contaminants is mediated by the complex interplay of hydrogeochemical forces and is mediated by the activity of complex subsurface microbial communities.

Subsurface environments such as this contain a large diversity of microorganisms under low nutrient conditions that significantly impact 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[1–3]. Although water covers 70% of the Earth's surface, only ~1% is readily available for human use, and a vast majority (~95%) of the Earth's consumable and available freshwater is groundwater[4–6]. 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. 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. 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.

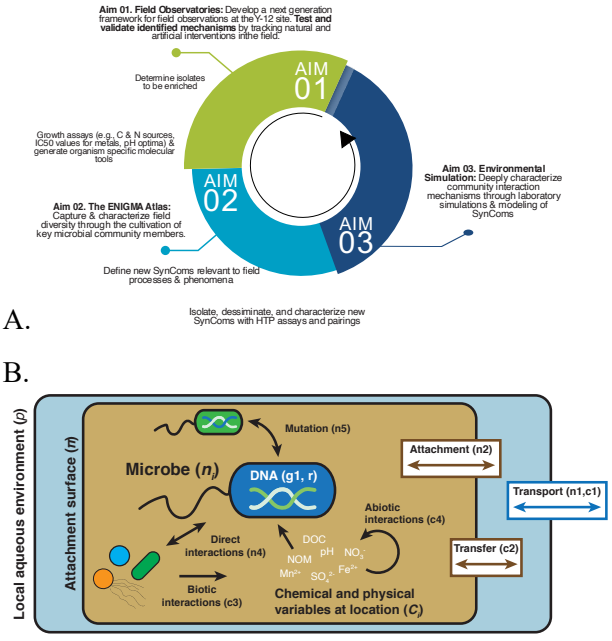

***Figure 1 A**. The iterative and overlapping structure of aims driving ENIGMA activity. **B**. A biology-centric cartoon of the process types captured in the FICSME framework.*

The ENIGMA workflow uses sophisticated, increasingly model-driven, field experiments to discover the biotic and abiotic components of the geochemical and biological processes (See Figure 1A). From continuous weather and hydrological measurements to periodic measurements of hundreds of chemicals, biomass, metagenomes, and activity assays, we track the spatiotemporal dynamics of biogeochemical

processes from the local centimeter scale to kilometers. From these we infer the chemical, physical and microbial interactions predictive of these dynamics and estimate the ecological forces, both stochastic and deterministic, that shape community function. We then deploy a unique array of culturing, genetic, physiological and imaging technologies to capture this diversity in the laboratory and map the genetic basis for observed behaviors. Laboratory consortia are used to map gene function and investigate material flow within and among cells in conditions that simulate relevant field processes. These then are all used to 'fill' in an increasingly detailed model of microbial community dynamics (see Figure 1B).

Our ambition is to achieve sufficient resolution to causally predict the active biotic and abiotic mechanisms mediating key processes (*e.g.*, denitrification); dissect the dispersing and persistent microbial community components critical in space and time during these processes; and ultimately predict the future changes in contaminant fate from current observations and possibly arising from natural and anthropogenic perturbations. Outcomes are significant both in the fundamental science of community ecology and in the applied understanding of biologically mediated processes in contaminated sediments.

This report highlights the progress made by the ENIGMA SFA over the last few years. The topics covered presuppose a critical capability that has been developed and refined by ENIGMA during this and previous periods which is the deep functional observation of microbiome composition and dynamics in a complex geochemical environment which enable biological prediction critical to subsurface activities (*e.g.*, nitrate and metal reduction, carbon utilization, and greenhouse gas emission) under different conditions of pH, heavy metals, nutrients, and oxygen. The reports summarized herein describe how ENIGMA captures relevant field and laboratory data to understand microbial function and add processes and parameters to our evolving model. The four previous reports reported on the following topics:

1. **Report on the latest techniques for recovering microbial isolates from environmental samples.** [Q1 Performance Metric] One of the largest challenges in interpreting field data on microbiome dynamics is the dearth of knowledge on both the capabilities of the organisms observed therein even if full genomes can be recovered from measurements. It is critical to isolate representative and relevant (preferably active and persistent) diversity from the varied conditions at the target location so physiology and causative gene functions can be assessed (See Q2 report). Traditionally, it has been a challenge to cultivate the majority of organisms in any complex environmental sample, however, our SFA has developed enrichment conditions that better represent field conditions/nutrients to achieve a far higher coverage of the observed community. We now have thousands of isolates in our collection covering 5 phyla, 9 orders, 35 families and around 160 genera which map at the genus level to over 80% of the most abundant and ubiquitous genera observed in our field data. With these approaches, we are better able to represent the real-world communities in the laboratory and have a vastly increased ability to characterize the genetically encoded capabilities of these microbes.
2. **Describe progress on new techniques for characterizing microbial isolates in the laboratory.** [Q2 Performance Metric] Once obtained in the laboratory, the next challenge is to understand the biochemical capabilities of isolates. ENIGMA has developed several critical approaches that enable the scalable and quantitative assessment of phenotypes from growth to nutrient consumption and secretion, adhesion, and interspecific warfare. The genome is the fundamental code for possible heritable behaviors of isolates so obtaining the highest quality genomes is key. ENIGMA has developed a new hybrid assembly workflow (now in the DOE Systems Biology KnowledgeBase, KBase[7]) using short and long read sequencing which can lead to finished quality genomes that practically lead to higher gene numbers and better circularization and strain definition than short read assemblies alone. Using new computational tools such as FastTree 2[8], PaperBlast[9], GapMind[10,11], and SitesOnATree (some incorporated into KBase) we can accurately place these genomes in phylogenetic context and infer metabolic capabilities. Our KBase Organization allows us to organize and further annotate and model these genomes with flux-balance models which we can share with the community upon publication (e.g., This KBase static narrative[12]) and we are working with the KBase team on improving formal model-based prediction of growth and other

phenotypes using our other characterization data. Importantly, we have developed highly efficient exo-metabolomic workflows which map nutrient consumption/production conditions and can be used to constrain metabolic models and predict microbial interactions. Further, ENIGMA has pioneered novel genome-scale bar-coded genetic disruption, knock-down and over-expression technologies (RB-TNSEQ[13], CRISPRi[14], and DUB-Seq[15]) that have enabled ENIGMA to experimentally annotate/reannotate tens of thousands of genes which help us both understand individual microbes and to propagate this information across the tree-of-life. We have built computational systems that organize these analyses and information together for Findable, Accessible, Interoperable and Reusable operation in the form of a set of web-accessible resources that also prepare our data for integration into KBase.

3. **How laboratory simulations using microbial isolates informs the understanding of the microbial ecology in environmental samples.** [Q3 Performance Metric] Analysis of environmental, hydrogeochemical and biological data from the field, in combination with the high-resolution, genome-backed, physiological data from our scalable microbial isolation and characterization workflows (see 1 and 2 above), with, in part, the modeling frameworks discussed in our fourth report (described below), ENIGMA is inferring the critical taxa, biotic and abiotic functions that mediate key field processes including carbon utilization, nitrate reduction, sulfate reduction, methane and nitrous production, and metal reduction. To test the hypotheses that arise from this, ENIGMA chooses specific organisms for deeper characterization with more extensive analyses (transcriptomics, proteomics, metabolomics, and imaging) under increasingly realistic field conditions with attached and planktonic phases both in upflow and packed-bed reactors. These enable a more dynamic picture of response to environmental variation of growth, size-control, and adhesion systems; the ability to disperse and persist in different niches, and to evolve and adapt within these niches. Since many of the inferences from field data suggest that there are functional associations among multiple microbes, ENIGMA has developed the means of prioritizing and characterizing synthetic community compositions predicted to implement mechanisms explanatory of field observations. Specific gene- and protein-level mechanisms discovered in this way become targets for measurement in the field. To make this effective, ENIGMA has developed new methods in proteomics, metabolomics, and computation tools for inference of cellular regulatory networks.
4. **Report on capabilities to model the activities of microbial communities in environmental samples.** [Q4 Performance Metric] The central goal of ENIGMA is to learn how to create data-driven but increasingly causal and mechanistic models of how microbial communities assemble, grow, transform and adapt in complex open-biomes such as the terrestrial shallow subsurface. The standard ecological forces of dispersal, drift, and selection should be in the context of the physical processes that can drive these effects. Specialized experimental designs are required to train and test these models so they are predictive both within the site being studied and generalizable to sites with similar properties. To meet this challenge, ENIGMA has developed a suite of modeling tools ranging from strict machine-learning based methods such as random forests to identify the taxa and/or biological functions which together are the most predictive of environmental parameters[16], to neutral-model based ecological frameworks that predict whether observed abundances of biological elements are subject to different aspects of dispersal, drift or selection[17], to more mechanistic frameworks that are beginning to build-up genome-informed reactive-transport-like models that capture both abiotic and biotic mechanisms that interact with hydrological, geophysical, and chemical transformations[18]. To ensure these models have sufficient and well-structured data, ENIGMA has designed field experiments across the site to provide spatio-temporal data enabling prediction of the species whose presence and activity are most related to these processes and are likely persistent and/or under selection. We are now in the phase of developing a Subsurface Observatory (SSO) to vastly increase the resolution at which we can track microbial processes in time and space in geological context to build a multiscale model from microns-to-meters-to-kilometers of how biological processes impact processes across a contaminant plume within a watershed.

The latest techniques for recovering microbial isolates from environmental samples

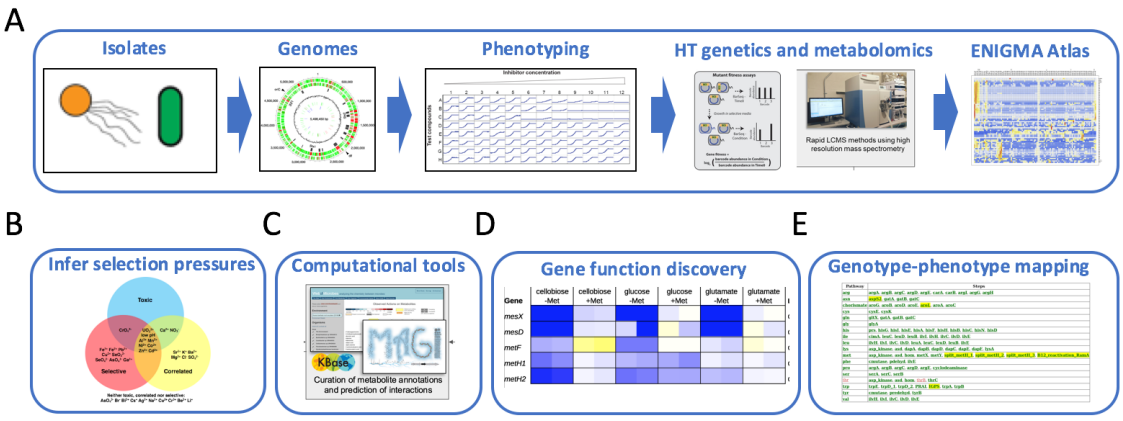

Figure 3. Overview and applications of ENIGMA characterization pipeline. A) Standard pipeline from isolates to rich experimental data including metabolomics and genetics data feeds into ENIGMA Atlas. (B-E) Applications of ENIGMA Atlas include correlating laboratory growth measurements to field data (microbial abundance and geochemistry) to infer key selective pressures in the environment [17], new computational tools for analyzing and visualizing high-throughput phenotyping data [13], systematic discovery of new gene functions [15], and tools for improved inference of metabolic pathways from genome sequence alone [9].

While the technologies for determining 'who is there and active' in microbial communities are rapidly improving, interpretation of this data remains difficult due to a lack of direct experimental knowledge of the genotypes, phenotypes, and physiology of microbial constituents. Even the baseline annotation of most observed genes is exceptionally poor, and it remains challenging to isolate all of the diversity. However, ENIGMA continues to create media and growth formats/conditions that better simulate the environments such that we are able to capture a greater diversity from the *in-situ* community. The inventions and results from our work are summarized in the Q1 report and below:

1. In groundwater, dissolved organic matter (DOM) derived from the sediment contributes to the available C source for microorganisms. Our previous study shows that sediment DOM contains a myriad of heterogenous organic compounds—mostly recalcitrant C such as lignin-like compounds and a small portion of relatively labile C such as carbohydrate- and protein-like compounds[19]. Other natural C sources available for microorganisms in groundwater can derive from dead microbial biomass turnover. We found cultures amended with complex natural organic C rapidly diversify at early stages of enrichment compared to those grown on simple, small organic C compounds. We obtained a total of 228 bacterial isolates representing 5 phyla, 17 orders, and 56 distinct species with these approaches. Our isolates represent both abundant (3-5% relative abundance at OTU level) and rare (< 0.01%) species from the initial groundwater sample. Of the species isolated, nine belong to candidate novel

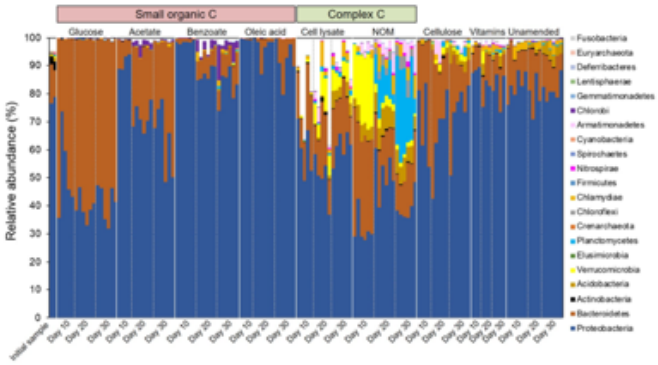

Figure 2 *Temporal community structures of the initial groundwater sample and each C-amended group reported as relative abundance of taxonomic phyla over 3 timepoints (Days 10, 20, and 30).*

species and three belong to candidate novel genera. Our cultivation strategy will benefit future development of effective and ecologically relevant cultivation/isolation strategies.

2. Metals are critical stressors and enzymatic co-factors in the terrestrial subsurface and this particular site has combinations that can carry out complex chemical transformations, can become scarce due to formation of insoluble minerals, and/or can be extremely toxic to most bacteria. Elevated metals include Al, Mn, U, Fe, Co, Cu, Ni, and Cd. Using the environmentally informed enrichment strategy, 22 different metal tolerant microbial strains were isolated from ORR groundwater and 88 different strains from ORR sediment. Using the technologies described in our Q2 and Q3 reports, we discovered new mechanisms for scavenging and tolerance to beneficial and toxic metals, respectively.
3. Chemical and physical parameters show distinct selection on communities in the planktonic and sediment attached fractions. The OR-FRC site is a complex geological location where the groundwater flows through complex sediment and rock formations and we hypothesize that isolation of critical strains of microbes observed at key locations arise in part from physical interactions with solution and surface chemistries. ENIGMA has designed up-flow packed bed reactors (PBR) to simulate select field conditions (*i.e.*, flow rate and particle size) observed at ORR-FRC to observe how environmental factors, including hydraulic variables such as average pore velocity, influences metabolic activity, community establishment, and cell distribution in a micropore environment. In recent work with these reactors, we demonstrate the strong differential selection of attached growth at low pH compared to either planktonic or attached at higher pH. Our low pH reactors resembled the low diversity found in the attached communities within our contaminant plume and were populated with specific genera reminiscent of *in situ* observations.
4. Field data indicates the presence of specific genera or species whose predicted metabolisms indicate the need for specific conditions for enrichment (*e.g.*, levels of electron donors/acceptors). For example, metagenomic depth profiling at the ORR-FRC showed that the Clade I vs Clade II nitrous oxide reducers are differentially distributed with depth, with Clade II variants enriched closer to the surface, thus being implicated in N_2O consumption. To obtain these, multiple sets of liquid enrichments with each containing a different electron donor for N_2O reduction were initiated in sealed Balch tubes as well as agar surfaces grown in anoxic steel pressure vessels. The combination of both approaches has helped to increase the diversity of N_2O reducing microbes isolated. ENIGMA is creating specific protocols for a number of our target genera informed by field measurements.
5. Finally, as isolates are characterized in an environmental context, a growth medium (soil-defined medium, SDM) based upon water-soluble soil exo-metabolites was formulated[12]. Although SDM was successfully used for exometabolomic profiling, only half (15/30) of the screened ORR-FRC isolates grew. This led to the construction of a new defined medium (NLDM) that supports the growth of a wider range of diverse soil bacteria. Additional compounds included were selected on the basis of their presence in R2A media, their presence in soil, and their usage across a diverse set of bacteria based on existing exometabolomic data[20]. Out of 110 of our isolates tested, 108 grew on NLDM, and preliminary results show that all 64 metabolites from the NLDM medium were utilized by at least 1 isolate. This study also revealed a high degree of phylogenetic niche conservatism for substrate use. This standard defined media sets the stage for standardized growth/phenotype comparisons of diverse microbes and provides data for models to predict utilization/production of metabolites during processes of interest.
6. All together ENIGMA has collected ~2500 diverse aerobic and anaerobic microbes from groundwater and sediment to date, this collection is continuously expanding, and we call this strain collection and its taxonomic and functional mapping the ENIGMA Environmental Atlas. Through this Atlas, ENIGMA is poised to provide a foundational dataset of genotype-phenotype relationships that greatly enhance our ability to interpret metagenomic data and infer the functions of thousands of uncharacterized genes – a key contribution to our understanding of subsurface microbial physiology.

Techniques for characterizing isolates in the laboratory

The fundamental goal for isolating the microbes above is to be able to characterize them at sufficient resolution, so that we can predict how they and their relatives will perform in the field. Ideally, with sufficient genotype-phenotype-ecotype information from field and lab studies, the new, high-quality genetic data from the field would be sufficient to predict how complex communities would function *in situ*. The ENIGMA Environmental Atlas seeks to collect enough diversity to cover not only large-scale phenotypic diversity conserved within genus level and above, but also how small variations in gene function may play a role in the adaptation and persistence of specific strains in the environment. To meet this challenge, ENIGMA has developed a suite of experimental and computational tools to characterize non-model bacterial isolates in the laboratory[21]. These experimental tools are designed to be flexible, such that the techniques can be applied to a wide diversity of bacteria, and inexpensive, to enable the multi-omic characterization of hundreds to thousands of strains per year. These measurements contribute to the functionalization of the ENIGMA Environmental Atlas and include the accurate mapping of genotype to phenotypes of field relevant microbes, discovery of thousands of new gene functions[22], and an experimental-computational framework for others to leverage for the systematic characterization of bacterial isolates and their gene products from other environments. Furthermore, the characterization data is being used to prioritize the development and investigation of synthetic microbial communities (SynComs) that recreate key phenomena observed in the field (See Q3 summary below). We summarize these results in our Q2 report. Key inventions and findings are summarized below:

1. Quality genome sequences are a prerequisite for any predictive mapping of genotype-phenotype relationships. Traditionally, most bacterial genomes have been assembled solely using short-read Illumina data, which has prevented the assembly of complete genome sequences. We have developed a hybrid assembly approach that uses both short and long read sequencing (backed by a highly effective high-molecular weight DNA purification protocol) that enables high quality, circularized bacterial genomes (with no estimated errors) which we can now scale to the thousands in our collection. We have built a computational tool called Jorg which has been incorporated in the KBase so all users have access, and all our genomes are uploaded into the ENIGMA Organization in KBase[23].
2. We have derived custom panels of field-observed carbon and nitrogen sources for high-throughput growth assays[24,25], and custom panels of metals and other inorganic ions[26] for screening the phenotypes of isolates across hundreds of conditions. For example, this high-throughput phenotyping data has been used to identify stressors that selectively impact the fitness of isolates in the laboratory and are explanatory of their distributions with the contaminant plume in the field. We have automated a large fraction of the growth assay and data analyses, and these are being added to the Atlas and being used to constrain metabolic and population growth models of these microorganisms.
3. ENIGMA has also developed a highly efficient exo-metabolomics workflow that measures the ability of a bacterium to consume and transform environmental compounds[19]. ENIGMA has pioneered the development of exometabolomics as a powerful laboratory tool for characterizing, among other things, the varying nutrient preferences of our isolates and how differences in substrate preferences impact microbial communities[27].
4. ENIGMA has a long history of developing genetic tools for non-model bacteria, including early efforts on targeted chromosomal engineering of sulfate-reducing bacteria[28]. Over the last few years, we have developed a number of sequencing based approaches to interrogate the functions of bacterial genes and uncover mechanisms of regulation that can be applied across a large diversity of microorganism including loss-of-function methods such as (RB-TnSeq supported by magic pools vectors[29] and CRISPRi and gain-of-function techniques such Dub-Seq[15] to systematically uncover phenotypes and functions for thousands of bacterial genes that previously were poorly understood[13]. To date, we have generated RB-TnSeq libraries in 66 diverse bacteria, performed nearly 14,000 genome-wide fitness assays, generated over 50 million gene fitness measurements, and

identified functions for thousands of previously uncharacterized bacterial genes. We have disseminated these reagents and created others on demand for well over 100 external groups to date. We have used all these methods to identify new metabolic pathways, transporter and other systems that provide tolerance to metal stressors, and host determinants of phage susceptibility or defense. For example, we identified a conditional phenotype for 1,137 non-essential protein coding genes including hundreds of genes with no previously known functionin one of ENIGMA's legacy model sulfate reducers *Desulfovibrio vulgaris* Hildenborough (DvH)[30].

5. To characterize gene regulation and signal transduction at scale, we have pioneered DAP-seq to identify binding motifs for transcriptional regulators in ORNL FRC isolates, and have used these tools to dissect the complex regulatory mechanisms these bacteria use to respond to environmental stresses present at the ORNL FRC[31].
6. To aid in the analysis and interpretation, ENIGMA has been developing new analytical tools and data management systems (for modeling see the Q4 summary below), both stand-alone and as part of the DOE KBase. These include: (i) CORAL as a central platform for data storage and data integration; (ii) a flexible comparative genome browser that can be rapidly generated using only genome .gbk files as the starting point; (iii) the fitness browser for RB-TnSeq data; (iv) isolate browser for interactive analysis of growth data, (v) Web of Microbes for exometabolomics data[13], and (vi) METLIN/XCMS for cloud-based analysis of metabolomics data[32]. We have historically also built tools to characterize the phylogeny of organisms and genes (FASTTREE)[8] that maps genes in a new genome to function through linkage to experimentally characterized genes from the literature (PaperBlast)[9]. Both are in KBase. We have also recently developed an integrated tool-suite for sophisticated accurate annotation of metabolic pathways in diverse bacteria and archaea using our data and other community collected experimentally validated protein annotations called GapMind [10,11]. Using these tools, we have been able predict carbon source preferences for isolates and MAGs (metagenome assembled genomes).
7. All these approaches help elucidate genotypes and phenotypes in isolates important to field processes and supporting the synthetic community work described in our Q3 report. For example, the genus *Rhodanobacter* has been identified as an abundant and ubiquitous microbe at our site, strains of which differentially seem to locate and grow within critical regions of the contaminant plume where denitrification processes are under extreme selection from metals and low pH. We have used the methods above to identify critical efflux pump and regulatory systems that allow a specific one of these strains to survive under these conditions.

Laboratory simulations with isolates informs understanding of the microbial ecology in environmental samples.

Field observed correlations across biotic and abiotic factors have generated hypotheses regarding how specific environmental parameters such as pH, metals, and carbon/oxygen availability might constrain microbial communities and associated processes. Development of synthetic communities and environmental simulations allows ENIGMA to more deeply dissect the molecular mechanisms that are specific to the observed field processes and might arise due to interaction among microbes rather than their individual activities. This entails focused and higher-resolution application of these genomic technologies along with deep transcriptomics, proteomics, and careful physiological measurements. Among the specific processes we are interested in understanding are the factors that modulate the reduction of nitrate since this is one of the most abundant and oddly persistent contaminants at the site. We have observed strong, variable modulation in various metals, pH, O_2, and carbon and that sometimes, microbes with complementary pathway configuration may co-occur and provide resilience across temporally varying conditions. We summarize some of our results from our Q3 report below:

1. Two important modes of respiration that are commonly observed in subsurface environments are nitrate and sulfate respiration, which are facilitated by Nitrate Reducing Bacteria (NRB) and Sulfate

Reducing Bacteria (SRB), respectively. SRB and NRB activity were observed to be stratified along the vertical transect of the sediment cores from the site analyzed, which was not surprising. ***What was surprising was the association of dissimilatory nitrate reduction to ammonia (DNRA) with SRB activity***. This suggested SRB activity, and potentially hydrogen sulfide, could impact NRB activity and promote DNRA over denitrification. This has been observed in other contexts, but the mechanism(s) are not well understood[33]. We hypothesized that hydrogen sulfide could impact NRB activity and potentially influence N_2O emissions. We showed this to be the case in the field isolate *Intrasporangium calvum C5* (*I. calvum*), a NRB with the ability to reduce nitrate to ammonia via DNRA or N_2O via partial denitrification. Hydrogen sulfide had a strong dose dependent inhibitory effect on the nitrate respiratory growth of *I. calvum*. Systems analysis, using time-resolved transcriptomics and metabolomics, revealed that inhibition was driven in part by dysregulation of branch chain amino acid (BCAA) biosynthesis and carbon uptake[34].

2. Using data from a deep survey of soils across the contamination plume, we identified an isolate with a 16S RNA identical to a strain from amplicon data that was ubiquitous and abundant at locations where nitrate and heavy metal were highest. The genome of this *Bacillus cereus* species designated "CPTF" suggested that this strain is capable of DNRA via NarGHI and NasDE[35,36]. Laboratory experimentation found that nitrite and ammonium are the two major end-products of nitrate-respiration, and further analysis showed a complex pattern of regulation to respond to different metal combinations. While these seem to enable the organism to continue to grow under these mixed stress conditions, there were combinations in which the second set in DNRA was inhibited leading to nitrite accumulation which then leads to toxicity and high pH sensitivity to growth. This accumulated nitrite *could* possibly enhance chemodenitrification.
3. One of the compelling observations from our metagenomics and isolate collection is that most organisms possessing denitrification pathway genes lack the complete pathway. The communities are a complex mixture of species that encode partial denitrification pathways that likely complete the process together. The environmental significance of pathway partitioning is mostly unexplored, but it is presumed partial denitrifiers in the subsurface complement one another via exchange of soluble nitrogen intermediates (i.e., nitrite, nitric oxide, nitrous oxide). We have shown that a pair of isolates related to those known to co-occur at our site with different partial denitrification pathways, *Rhodanobacter sp.* FW510-R12 (R12) and *Acidovorax sp.* GW101-3H11 (3H11), can perform complete denitrification only when co-cultured. 3H11 and R12 complement one another to completely reduce nitrate to nitrogen gas. We are now exploring how environmental factors such as metals, oxygen and carbon mediate these effects.
4. To aid in dissecting these processes we also invent new supportive technologies such as:
   a. New types of fluidized bed reactors (FBRs) and packed-bed reactors (PBRs) that simulate anaerobic conditions in the subsurface and allow the microbes to attach to particles or live as free-floating organisms (planktonic). We can couple systems-level approaches to the environmental parameters to differentiate the molecular function of each isolate in the community from distinctive growth modes (i.e., attached, or planktonic) within the same environment.
   b. An LC-MS/MS-based metaproteomic analysis to assess the structure of microbial communities based on species proteinaceous biomass contribution. This method provides accurate biomass composition of a given microbial community and insights into community expressed proteomes and metabolic activities. We have applied the technology to communities comprised of over 10 isolates.

Capabilities to model the activities of microbial communities in environmental samples

There are two complementary questions in tracking the microbial contributions to any environmental process: first, what fraction of an observed activity in space and time may be traced to reactions/actions of

biological origin, and second, what factors lead to the specific composition of microbial strains and their individual states in that same interval? The first of these may not require a detailed accounting of all the biological mechanisms that lead to the establishment, growth, and persistence of microbes at a given field site- instead requiring that more general qualities (such as the ability to reduce nitrate) are maintained). The latter question requires a far more detailed accounting of specific constraints on specific microbial species and their mechanisms. Linking these two is the critical question to address how phenomenon at the microscopic level (microns to millimeters) propagates up to meters and beyond. The OR-FRC is an exceptional complex environment at all these scales and the forces that act on the microbes that contribute to nitrogen and carbon cycle process and metal respiration are subject to strong and variable processes affecting their dispersal, catalytic activity and growth as cells are transported and establish across the three dimensions of the site and conditions change seasonally and over longer times. Modeling the activity of microbial communities over such a complex site to answer the questions above requires tight co-design of field experiments so that the specific parameters of the models may essentially be fit in such a way as to allow generalization and prediction outside the training data. ENIGMA has 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, Figure 4)[18]. The FICSME conceptual model is intended to provide a framework toward mechanistic models of subsurface microbial communities that are genome-informed into a reactive transport framework. We have been developing a suite of modeling tools to aid in estimating the critical microbial taxa, chemical species, and biological and abiotic mechanisms to incorporate into such a detailed model based on data from our laboratory work and a set of increasingly focused field experiments culminating in the design of our Subsurface Observatory. We outline our approaches in our Q4 report and summarize here.

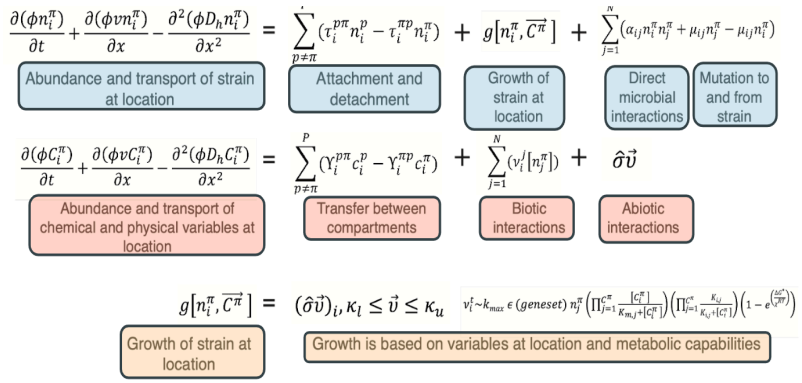

Figure 4. *The conceptual FICSME model organizes information about the population dynamics of microbes and chemical in both planktonic and attached phases driven by hydrological forces, phase partitioning, interactions with each other, and in the case of microbes their ability to grow given the local conditions. Experiments in the field and lab are directed towards elucidating the underlying mechanisms (reactions) within these equations or estimating the kinetic, mechanical, and thermodynamic parameters.*

1. ENIGMA has developed a set of effective and sometime novel computational approaches to identifying taxa or underlying functions and their possible interactions which are statistically predictive of geochemical conditions. This started with tools that recognized the composite nature of most metagenomic data from the field- that is- that the data was not a measure of absolute but rather relative abundances. Improper consideration of this leads to spurious estimates of correlation among taxa and our method, SparCC[37], has become one of the most popular tools that more correctly finds these correlations. We also pioneered the use of machine learning tools such as Random Forests[16] to find how well these taxa could predict environmental features- that is- could, for example, predict nitrate levels from microbial community composition. We took single time points from many locations across the OR-FRC as input data for training and testing, and we found that mean pH, nitrate, and Uranium levels could be predicted dependent upon taxonomic groups for each of these. These microbes then became the target for our isolation and characterization efforts.
2. These more static analyses, however, do not account for processes that may force change in the microbial community or local chemical/physical variables over time. We performed time-series

measurement to understand these dynamics more deeply. Using new workflows based on vector-autoregressive models and Granger causality we were able to discover that while there was a great variability and turn-over in the identity of microorganisms in the groundwater there was a general conserved set of metabolic functions and that the diversity of species and functions was linked strongly to the variations in geochemistry that occur over time[38]. This highlights the need to dissect the dispersal, drift and selection forces on the community.

3. To approach this, we developed an ecological null-model informed framework called iCAMP that predicts types and extent of ecological force that may likely explain the distribution of taxa driven by specific environmental parameters[17]. 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. 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. This provides information to the isolation approaches above.
4. Finally, as we recognize the limitation of the data used to parameterize the models above and come to understand what data will be necessary to scale-up to more mechanistic reactive transport models, we have launched an effort to characterize a specific area of our study site at much higher resolution and establish a Subsurface Observatory- a defined volume across the contamination plume, instrumented to allow nearly continual sampling of hydrological, chemical and biological elements across its three dimensions. As a baseline, we performed a dense matrix of cone penetrometry measurements across an area just downstream of a major contaminant catchment. This yielded high-resolution three-dimensional models of the shallow subsurface that estimated water paths, mineral and rock types, and configuration and permeability[39]. Associated biological and chemical measurements are giving us resolution on the gradients of these elements as they enter and cross the plume. This information is now being used to refine the specific region in which the SSO will be built.

Conclusions and Future Steps

Microbial communities are a powerful engine of environmental change for better and worse. Gaining an actionable understanding of how they form, what mediates their activities, and how they adapt over time as conditions change is critical to our ability to predict the effects of climate change, the impact of anthropogenic activities, and our ability to mitigate against the detrimental effects. However, the challenges in tracing the flow of matter and energy through the complex assemblages of abiotic and abiotic elements- which is likely necessary to create that understanding are manifold. The technologies that allow us to identify ‘who’ is present through improved metagenomic sequencing and assembly and even to infer which of the observed microbes may be active are improving rapidly although the spatial and temporal resolution at which we need to know this information is not entirely clear. Also improving is the ability to track the molecular abundances in time and space and in some cases, through isotope measurements, it is possible to infer the reactions (biotic or abiotic) that are likely impacting overall processes. However, more direct measurement of complex reaction fluxes, cellular growth rates, and dispersal rates are still difficult. ENIGMA has been sharpening approaches through co-design of experiments, measurement technologies, data analyses and models that enable more precise tracking and estimation of these features. Our increasingly sophisticated and controlled field observation systems are allowing more precise representation of the environments in which our microbial communities are evolving and acting with increasing high-resolution data on hydrology, chemistry, and biology. Our isolation and characterization systems are allowing us to better interpret that information by enabling deep biomolecular measurement of the genetic determinants of the metabolism, stress response and growth dynamics of a large diversity of relevant microorganisms though challenges remain in both accruing that relevant diversity from the field and being able to genetically manipulate it. Our environmental simulation platforms allow us to subject microbes to more realistic environments in the laboratory which permit a

much more realistic assessment of the mechanistic determinants of community function that can be mapped back to specific field locations. Albeit there are still challenges in knowing the exact features of the environment to simulate and how to construct relevant synthetic communities from isolate collections. Nonetheless, this deeply characterized site; isolate collection and characterization alone and within environmental simulations is becoming a unique integrated long-term resource for environmental systems biology The integration of our data and tools in our own framework as well as dissemination through the DOE KBase enables an increasing impact on science beyond our project. In addition, the isolate collection, and the biological reagents (*e.g.*, RB-TNSEQ) have also become community resources.

In our future work, we are greatly increasing our capabilities to track environmental and biological processes at fine spatiotemporal scales in well-defined and instrumented regions of the shallow subsurface. This entails building up a critical field infrastructure for continuous monitoring of water, weather, and critical chemical/physical variables and for rapid sampling for deep chemical and biological analyses. Better methods for identification of biological entities- cells, phage and other mobile elements are coming online along with better tools for measuring their viability and activity. These in turn will help better focus and parameterize the models that identify the microbes and mechanisms most likely causal for nitrate reduction, nitrous production, sulfate reduction, methane production, carbon utilization, and metal reduction at the site. The increased resolution of our environmental measurements should also guide better enrichment of the relevant, selected fraction of microbes from the site as we determine better the conditions in which they are surviving and thriving in the field. Improvements in our mining of gene transfer elements in the field and our increasing sophistication in genome analysis for exogeneous DNA defense systems should help us improve our ability to gain genetic control of these microbes and thus better assign functions to their genes and build models of their physiology. As these more relevant organisms enter the lab and our environments are better characterized, our ability to create controllable laboratory simulations of the field communities for deep mechanistic assessment improves. These data together will enable the true realization of a model in the FICSME framework that links relevant mechanisms and environmental outcomes like nitrate reduction and carbon use efficiency and provide us possible control. We believe then the underlying observations will be ready for generalization to other sites where mineral cycle processes are impacted under similar conditions.

References

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci USA. 1998;95: 6578–6583. doi:10.1073/pnas.95.12.6578

2. Griebler C, Lueders T. Microbial biodiversity in groundwater ecosystems. Freshw Biol. 2009;54: 649–677. doi:10.1111/j.1365-2427.2008.02013.x

3. McMahon S, Parnell J. Weighing the deep continental biosphere. FEMS Microbiol Ecol. 2014;87: 113–120. doi:10.1111/1574-6941.12196

4. Danielopol DL, Griebler C. Changing Paradigms in Groundwater Ecology - from the ‘Living Fossils’ Tradition to the ‘New Groundwater Ecology.’ Int Rev Hydrobiol. 2008;93: 565–577. doi:10.1002/iroh.200711045

5. Griebler C, Malard F, Lefébure T. Current developments in groundwater ecology--from biodiversity to ecosystem function and services. Curr Opin Biotechnol. 2014;27: 159–167. doi:10.1016/j.copbio.2014.01.018

6. Dennehy KF, Reilly TE, Cunningham WL. Groundwater availability in the United States: the value of quantitative regional assessments. Hydrogeol J. 2015;23: 1629–1632. doi:10.1007/s10040-015-1307-5

7. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase: the United States Department of Energy Systems Biology Knowledgebase. Nat Biotechnol. 2018;36: 566–569. doi:10.1038/nbt.4163

8. Price MN, Dehal PS, Arkin AP. FastTree 2 — approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5: e9490. doi:10.1371/journal.pone.0009490

9. Price MN, Arkin AP. PaperBLAST: Text Mining Papers for Information about Homologs. mSystems. 2017;2. doi:10.1128/mSystems.00039-17

10. Price MN, Deutschbauer AM, Arkin AP. Gapmind: automated annotation of amino acid biosynthesis. mSystems. 2020;5. doi:10.1128/mSystems.00291-20

11. Price MN, Deutschbauer AM, Arkin AP. Filling gaps in bacterial catabolic pathways with computation and high-throughput genetics. PLoS Genet. 2022;18: e1010156. doi:10.1371/journal.pgen.1010156

12. Goff JL, Szink EG, Thorgersen MP, Putt AD, Fan Y, Lui LM, et al. Ecophysiological and genomic analyses of a representative isolate of highly abundant Bacillus cereus strains in contaminated subsurface sediments. Environ Microbiol. 2022. doi:10.1111/1462-2920.16173

13. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, et al. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. MBio. 2015;6: e00306-15. doi:10.1128/mBio.00306-15

14. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013;152: 1173–1183. doi:10.1016/j.cell.2013.02.022

15. Mutalik VK, Novichkov PS, Price MN, Owens TK, Callaghan M, Carim S, et al. Dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits in bacteria. Nat Commun. 2019;10: 308. doi:10.1038/s41467-018-08177-8

16. Smith MB, Rocha AM, Smillie CS, Olesen SW, Paradis C, Wu L, et al. Natural bacterial communities serve as quantitative geochemical biosensors. MBio. 2015;6: e00326-15. doi:10.1128/mBio.00326-15

17. Ning D, Yuan M, Wu L, Zhang Y, Guo X, Zhou X, et al. A quantitative framework reveals ecological drivers of grassland microbial community assembly in response to warming. Nat Commun. 2020;11: 4717. doi:10.1038/s41467-020-18560-z

18. Lui LM, Majumder EL-W, Smith HJ, Carlson HK, von Netzer F, Fields MW, et al. Mechanism across scales: A holistic modeling framework integrating laboratory and field studies for microbial ecology. Front Microbiol. 2021;12: 642422. doi:10.3389/fmicb.2021.642422

19. Wu X, Wu L, Liu Y, Zhang P, Li Q, Zhou J, et al. Microbial interactions with dissolved organic matter drive carbon dynamics and community succession. Front Microbiol. 2018;9: 1234. doi:10.3389/fmicb.2018.01234

20. Kosina SM, Greiner AM, Lau RK, Jenkins S, Baran R, Bowen BP, et al. Web of microbes (WoM): a curated microbial exometabolomics database for linking chemistry and microbes. BMC Microbiol. 2018;18: 115. doi:10.1186/s12866-018-1256-y

21. Liu H, Deutschbauer AM. Rapidly moving new bacteria to model-organism status. Curr Opin Biotechnol. 2018;51: 116–122. doi:10.1016/j.copbio.2017.12.006

22. Price MN, Deutschbauer AM, Arkin AP. Four families of folate-independent methionine synthases. PLoS Genet. 2021;17: e1009342. doi:10.1371/journal.pgen.1009342

23. Lui LM, Nielsen TN, Arkin AP. A method for achieving complete microbial genomes and improving bins from metagenomics data. PLoS Comput Biol. 2021;17: e1008972. doi:10.1371/journal.pcbi.1008972

24. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature. 2018;557: 503–509. doi:10.1038/s41586-018-0124-0

25. Carlson HK, Lui LM, Price MN, Kazakov AE, Carr AV, Kuehl JV, et al. Selective carbon sources influence the end products of microbial nitrate respiration. ISME J. 2020;14: 2034–2045. doi:10.1038/s41396-020-0666-7

26. Carlson HK, Price MN, Callaghan M, Aaring A, Chakraborty R, Liu H, et al. The selective pressures on the microbial community in a metal-contaminated aquifer. ISME J. 2019;13: 937–949. doi:10.1038/s41396-018-0328-1

27. Erbilgin O, Bowen BP, Kosina SM, Jenkins S, Lau RK, Northen TR. Dynamic substrate preferences predict metabolic properties of a simple microbial consortium. BMC Bioinformatics. 2017;18: 57. doi:10.1186/s12859-017-1478-2

28. Keller KL, Bender KS, Wall JD. Development of a markerless genetic exchange system for Desulfovibrio vulgaris Hildenborough and its use in generating a strain with increased transformation efficiency. Appl Environ Microbiol. 2009;75: 7682–7691. doi:10.1128/AEM.01839-09

29. Liu H, Price MN, Waters RJ, Ray J, Carlson HK, Lamson JS, et al. Magic pools: parallel assessment of transposon delivery vectors in bacteria. mSystems. 2018;3. doi:10.1128/mSystems.00143-17

30. Trotter VV, Shatsky M, Price MN, Juba TR, Zane GM, De Leon KP, et al. Large-scale Genetic Characterization of a Model Sulfate Reducing Bacterium. BioRxiv. 2021. doi:10.1101/2021.01.13.426591

31. Garber ME, Rajeev L, Kazakov AE, Trinh J, Masuno D, Thompson MG, et al. Multiple signaling systems target a core set of transition metal homeostasis genes using similar binding motifs. Mol Microbiol. 2018;107: 704–717. doi:10.1111/mmi.13909

32. Majumder EL-W, Billings EM, Benton HP, Martin RL, Palermo A, Guijas C, et al. Cognitive analysis of metabolomics data for systems biology. Nat Protoc. 2021;16: 1376–1418. doi:10.1038/s41596-020-00455-4

33. Brunet RC, Garcia-Gil LJ. Sulfide-induced dissimilatory nitrate reduction to ammonia in anaerobic freshwater sediments. FEMS Microbiol Ecol. 1996;21: 131–138. doi:10.1111/j.1574-6941.1996.tb00340.x

34. Otwell AE, Carr AV, Majumder ELW, Ruiz MK, Wilpiszeski RL, Hoang LT, et al. Sulfur Metabolites Play Key System-Level Roles in Modulating Denitrification. mSystems. 2021;6. doi:10.1128/mSystems.01025-20

35. Kraft B, Strous M, Tegetmeyer HE. Microbial nitrate respiration--genes, enzymes and environmental distribution. J Biotechnol. 2011;155: 104–117. doi:10.1016/j.jbiotec.2010.12.025

36. Goff JL, Lui LM, Nielsen TN, Thorgersen MP, Szink EG, Chandonia J-M, et al. Complete Genome Sequence of Bacillus cereus Strain CPT56D-587-MTF, Isolated from a Nitrate- and Metal-Contaminated Subsurface Environment. Microbiol Resour Announc. 2022; e0014522. doi:10.1128/mra.00145-22

37. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8: e1002687. doi:10.1371/journal.pcbi.1002687

38. Zelaya AJ, Parker AE, Bailey KL, Zhang P, Van Nostrand J, Ning D, et al. High spatiotemporal variability of bacterial diversity over short time scales with unique hydrochemical associations within a shallow aquifer. Water Res. 2019;164: 114917. doi:10.1016/j.watres.2019.114917

39. Putt AD, Kelly ER, Lowe KA, Rodriguez M, Hazen TC. Effects of cone penetrometer testing on shallow hydrogeology at a contaminated site. Front Environ Sci. 2022;9. doi:10.3389/fenvs.2021.821882