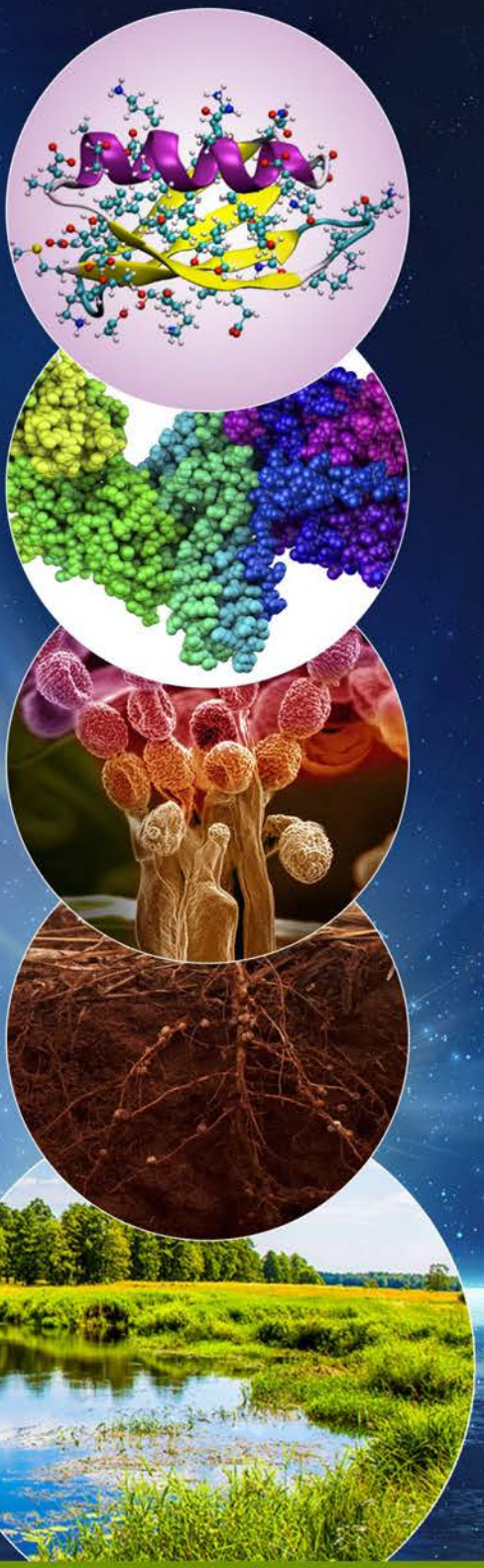
EMSL

2018

Modeling Microbial Dynamics and Processes from Cells to Ecosystems



Workshop WS30 held at the American Geophysical Union (AGU) Fall Meeting

December 9, 2018, 8:00 AM – 4:00 PM

Grand Hyatt Hotel, Washington, DC

Co-Organizers: Nancy Hess (PNNL/EMSL), Tim Scheibe (PNNL/EMSL) and Chris Henry (ANL/KBase)

Invited Presenters: Romy Chakraborty (LBNL), Hyun-Seob Song (PNNL), Pamela Weisenhorn (ANL), Chris Henry (ANL), Lee Ann McCue (PNNL), Roelof Versteeg (Subsurface Insights), Xuehang Song (PNNL), David Moulton (LANL).

U.S. DEPARTMENT OF ENERGY

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm


This document was printed on recycled paper.
(9/2003)

PNNL-28458

Modeling Microbial Dynamics and Processes from Cells to Ecosystems

Workshop WS30 held at the American Geophysical Union (AGU) Fall Meeting
December 9, 2018, 8:00 AM–4:00 PM
Grand Hyatt Hotel, Washington, DC

Co-Organizers: Nancy Hess (PNNL/EMSL), Tim Scheibe (PNNL/EMSL) and Chris Henry (ANL/KBase)

Invited Presenters: Romy Chakraborty (LBNL), Hyun-Seob Song (PNNL), Pamela Weisenhorn (ANL), Chris Henry (ANL), Lee Ann McCue (PNNL), Roelof Versteeg (Subsurface Insights), Xuehang Song (PNNL), David Moulton (LANL).

February 2019

Prepared for the U.S. Department of Energy's Office of Biological and Environmental Research under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Acknowledgments

The workshop was organized by staff from the Environmental Molecular Sciences Laboratory (EMSL) and the U.S. Department of Energy (DOE) Systems Biology Knowledge Base (KBase), both of which are supported by the DOE, Office of Biological and Environmental Research. We thank the American Geophysical Union (AGU) for coordinating workshop registration, venue, and other logistics and providing the opportunity for this workshop to take place. Special thanks to PFLOTRAN developer Glenn Hammond, who was not able to attend but nevertheless provided extensive support to the presenters in the PFLOTRAN segment of the workshop.

Acronyms and Abbreviations

AGU	American Geophysical Union
ANL	Argonne National Laboratory
API	Application Programming Interface
BER	Office of Biological and Environmental Research
DOE	U.S. Department of Energy
DOM	dissolved organic matter
EMSL	Environmental Molecular Sciences Laboratory
ENIGMA	Ecosystems and Networks Integrated with Genes and Molecular Assemblies
ESS-DIVE	Environmental Systems Science Data Infrastructure for a Virtual Ecosystem
FREDA	FTICR R Exploratory Data Analysis
FTICR	Fourier-transform ion cyclotron resonance
FTICR-MS	Fourier-transform ion cyclotron resonance mass spectrometry
FTIR	Fourier-transform infrared spectroscopy
IDEAS	Interoperable Design of Extreme-scale Application Software
JGI	Joint Genome Institute
KBase	Systems Biology Knowledge Base
KEGG	Kyoto Encyclopedia of Genes and Genomes
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
NOM	natural organic matter
NOSC	nominal oxidation state of carbon
PAF	Predictive Assimilation Framework
PCA	Principal Components Analysis
PNNL	Pacific Northwest National Laboratory
RTM	reactive transport model
SBIR	Small Business Innovation Research
SBR	Subsurface Biogeochemical Research
SFA	Scientific Focus Area
SOM	sediment organic matter
SRS	Savannah River Site
s-XAS	Soft X-ray absorption spectroscopy
WHONDRS	Worldwide Hydrobiogeochemical Observation Network for Dynamic River Systems

Contents

Figures

1.0 Introduction

The U.S. Department of Energy (DOE) Office of Biological and Environmental Research (BER) has made major investments in scientific user facilities and computational and data management tools, thereby creating a significant opportunity to integrate and leverage these resources to address the challenge of modeling microbial dynamics and processes across a range of temporal and spatial scales of interest to BER scientists. These resources include the following:

- Environmental Molecular Sciences Laboratory (EMSL, https://www.emsl.pnnl.gov/)
- Joint Genome Institute (JGI, https://jgi.doe.gov/)
- Systems Biology Knowledge Base (KBase, http://www.kbase.us/)
- High-performance reactive transport simulators such as PFLOTRAN (https://www.pflotran.org/)
- Community science networks for data generation such as WHONDRS (https://sbrsfa.pnnl.gov/whondrs.stm)
- Data management and public access platforms such as ESS-DIVE (https://ess-dive.lbl.gov/)
- Resources for high-quality community scientific software development such as IDEAS (https://ideas-productivity.org/)
- Toolkits for analysis and visualization of complex datasets such as EMSL's FTICR R Exploratory Data Analysis (FREDA) tool.

As part of the 2018 American Geophysical Union (AGU) Fall Meeting (https://fallmeeting.agu.org/2018/), AGU coordinated a number of scientific workshops (https://fallmeeting.agu.org/2018/scientific-workshops/), many of which were held on the Sunday preceding the main week of the conference. EMSL and KBase staff jointly organized this workshop to provide an opportunity for the community of interest to come together, learn more about these resources and how they are being (and could be) used by BER scientists, and explore opportunities for enhanced integration to solve critical scientific problems of interest to BER and the broader scientific community. This full-day workshop (attended by approximately 50 participants) explored recent advances, remaining gaps and challenges, and opportunities in current methods for modeling microbial metabolism. The workshop started with introductory research presentations from leading pioneers in this field, which demonstrated some of the cutting-edge research currently being undertaken using these resources. The presentations were followed by a series of tutorials and hands-on demonstrations of specific resources. The workshop culminated in a group brainstorming session in which participants discussed challenges and opportunities for using various combinations of DOE-funded facilities and platforms presented in the workshop to study environmental biogeochemistry at scales ranging from individual cells to communities to ecosystems. The agenda for the workshop is provided as Appendix A.

2.0 Scientific Presentations

Incorporation of microbial processes into quantitative numerical models at various levels of complexity has been approached using a variety of methods including genome-informed, trait-based methods, enzyme-based reaction networks, and flux balance analysis. The past two decades have seen an explosion of environmental microbiology data resulting from the development and maturation of a number of high throughput -omics technologies (e.g., genomics, proteomics, metabolomics, transcriptomics) and corresponding bioinformatics methods. While the scientific community

has widely embraced the use of -omics approaches to elucidate the roles of microbial community processes in ecosystem function, their systematic incorporation into quantitative models of microbial dynamics and processes remains a significant challenge. This challenge is particularly exacerbated by problems associated with variability (heterogeneity) across a range of temporal and spatial scales and associated difficulties in developing well-posed numerical models and parameterizations. For this workshop, we invited presentations from scientists working at three different DOE laboratories (Pacific Northwest National Laboratory [PNNL], Lawrence Berkeley National Laboratory [LBNL], and Argonne National Laboratory [ANL]) to describe cutting-edge research currently being conducted to address the integration of –omics and other environmental information to understand and quantify microbial processes in environmental systems.

2.1 Romy Chakraborty (LBNL) – Microbial Interactions with Dissolved Organic Matter

Traditional understanding of soil and sediment organic matter (SOM) revolved around the concept of humic substances—stable and chemically unique organic molecules that were commonly observed in soil extractions. However, recent evidence suggests that these humic substances are largely an artifact of the acid or base extraction methods used to isolate organic matter from mineral soils. Lehmann and Kleber (2015) argue that SOM is in fact a progression of decomposing organic compounds rather than large and persistent humic substances. Recent advances in methods for characterizing dissolved organic matter (DOM) have provided new insights into the complexity of environmental organic compounds and their interactions with microbial communities. Multidisciplinary experimental and computational tools are needed to fully understand the controls exerted by (and feedbacks to) microbial structure and function, and ultimately to predict carbon fluxes and transformations under various environmental conditions (Schmidt et al. 2011).

As an example of these challenges, results of a field study at the DOE Oak Ridge Reservation (Oak Ridge, Tennessee, USA) were presented (Wu et al. 2018). This research was performed under the Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA) project (http://enigma.lbl.gov), a Scientific Focus Area (SFA) at LBNL. Dissolved natural organic matter was fed to indigenous soil microbial communities from the capillary fringe of the shallow surficial aquifer at the Bear Creek Field Research Center (Vishnivetskaya et al. 2010). A number of different analytical techniques were applied to quantify microbial process dynamics and associated carbon transformations. Fourier-transform infrared spectroscopy (FTIR) did not provide any clear insights due to inherent limitations of the method. Soft X-ray absorption spectroscopy (s-XAS) and Fourier-transform ion cyclotron resonance mass spectrometry (FTICR-MS) provided a number of preliminary observations: (1) initially aryl functional groups (aromatics) were present in very low amounts, but microbial activity led to increased concentrations; (2) within 1.5 days, large molecules were being produced, likely derived from microbial biomass; however, the majority of molecules were still low molecular weight (100–500 Da) and the proportion of large molecules decreased over time; (3) carbon compounds containing nitrogen were preferentially degraded; and (4) protein and carbohydrate groups decreased over time while the proportion of lignin increased suggesting that recalcitrant carbon was persisting or being used at lower rates than other carbon. In terms of microbial responses, an initial increase in biomass growing on the labile fraction of added carbon was observed in conjunction with a large increase in CO_2 production. Using 16S rDNA and Geochip methods, it was demonstrated that the microbial community shifted as the carbon pool transitioned toward more recalcitrant compounds. Specifically, gene expression reflected a progression from dominance by copiotrophs (fast growth using labile carbon) to dominance by oligotrophs (slower growth but with greater diversity of carbon degradation potential). From these observations, it is concluded that complex and dynamic feedbacks are critical to system function: microbes play key roles in the transformation and generation of natural organic matter (NOM), which in turn directs microbial community evolution and succession in ecosystems. Potential approaches for using –omics integration tools and workflows (e.g., Figure 1) embodied in KBase were identified:

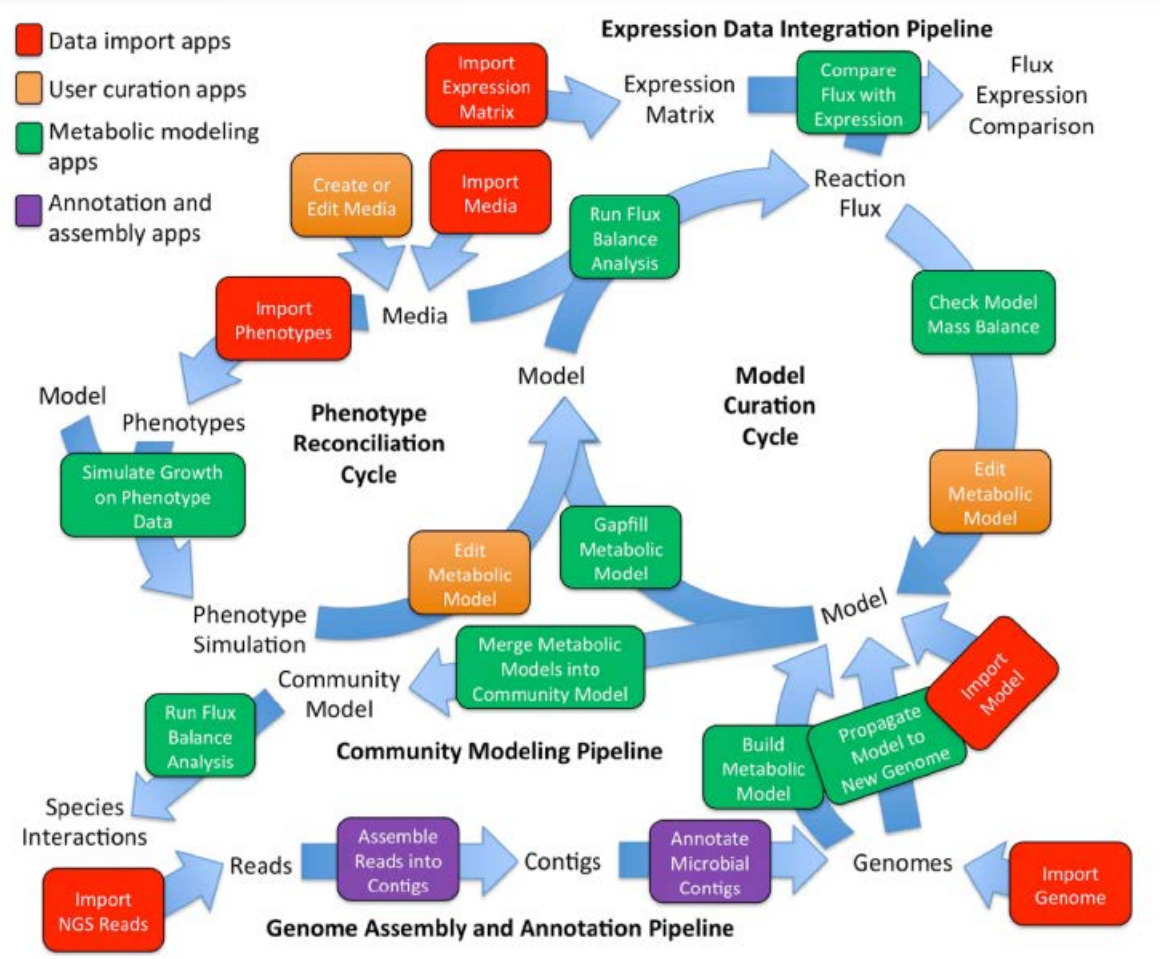


Figure 1 Genome assembly and annotation pipeline

- Metagenomic Sequence Analysis & Assembly:
  - Perform quality control and analysis with *FastQC* and *CheckM*
  - Trim reads to remove adaptors with *Trimmomatic*
  - Assemble binned contigs using *MEGAHIT*, *metaSPAdes*, and *IDBA-UD*
  - Bin contigs using *MaxBin2* or *metaBAT2*
  - Import, extract, and edit bins with various apps
- Community Exploration:
  - Annotate binned contigs with the *RAST* pipeline
  - Explore taxonomic abundance with *Kaiju*
  - Generate and merge metabolic models with *ModelSEED*
  - Produce phylogenetic trees from annotated genomes
  - Use metabolic modeling tools to analyze community metabolism

2.2 Hyun-Seob Song (PNNL) – A Metabolic Network-Based Approach to Biogeochemical Reaction Modeling

As –omics methods for characterizing microbial communities and environments have become more commonplace, we are increasingly recognizing that we need a multi-dimensional view of environmental metabolites. Microbial processes use many different substrates, including very diverse sources of carbon and nitrogen; and it has become evident that metabolite composition is a critical driver of microbial community dynamics (Graham et al. 2018). Metabolic network modeling (genome-informed biogeochemical pathway identification) is posed as an important approach that is beginning to allow us to translate the molecular-level understanding embodied in –omics data into models of ecosystem function.

Three examples of this approach were presented: (1) Prediction of autotrophic/heterotrophic interactions in a binary community; (2) comparative metabolic pathway analysis of two related microbial communities; and (3) integration of FTICR-MS data into metabolic pathway construction workflows. These examples are taken from research being performed by the PNNL Subsurface Biogeochemical Research (SBR) SFA (https://sbrsfa.pnnl.gov/) and the PNNL Genomic Sciences Program SFA (https://genomicscience.energy.gov/research/sfas/pnnl.shtml). In the first example, a KBase pipeline was used to develop metabolic network models for a microbial consortium of two metabolically dependent organisms: a cyanobacterium supports growth of an obligate aerobic heterotroph by providing organic carbon, O_2, and reduced nitrogen, as well as key B vitamins (McClure et al. 2018). Network analysis facilitated identification of the specific genes and metabolites associated with the interactions between these two organisms. The narratives developed for this example (represented as Jupyter notebooks, https://jupyter.org/) are publicly available in KBase for community reproduction and reuse (https://narrative.kbase.us/narrative/ws.13838.obj.1). The second example illustrated how a similar approach could be used in more complex communities, using functional guilds or superorganism representations to make the problem tractable. Comparisons were drawn between networks derived for microbial communities sampled from riverbank sediments in nearby sites that were vegetated (high C inputs) and non-vegetated (low C inputs) (Graham et al. 2017). A non-compartmentalized superorganism approach was used, in which the genomic potential of each community was treated as being embodied in a single organism without considering species partitioning. Metabolic networks were built for each site, and metabolic pathway analysis revealed that the two sites had a large number (hundreds) of pathways and compounds in common, but there were also a small number (~20) of unique pathways and compounds at each site. This example points out the potential value of incorporating FTICR-MS data into metabolic network construction processes and motivates the third example presented. While FTICR data cannot be directly incorporated into a genome-derived metabolic network due to the lack of overlapping compounds, this final example provided an idea of how KBase chemoinformatics tools can identify a scaffold of small molecules (representing putative substrates) implied by differences in masses between pairs of FTICR-measured carbon compounds. These can be mapped to known metabolic reactions (elements of a reaction pathway) through available databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg/kegg1.html). Preliminary networks generated from metagenomics data were supplemented by these reactions derived from metabolomics observations, leading to expanded metabolic reaction networks. This presentation concluded with a proposed pipeline (Figure 2), implemented using KBase tools, for integrating multiple types of –omics data into metabolic reaction models and ultimately into reaction networks that can be used in reactive transport models such as PFLOTRAN.

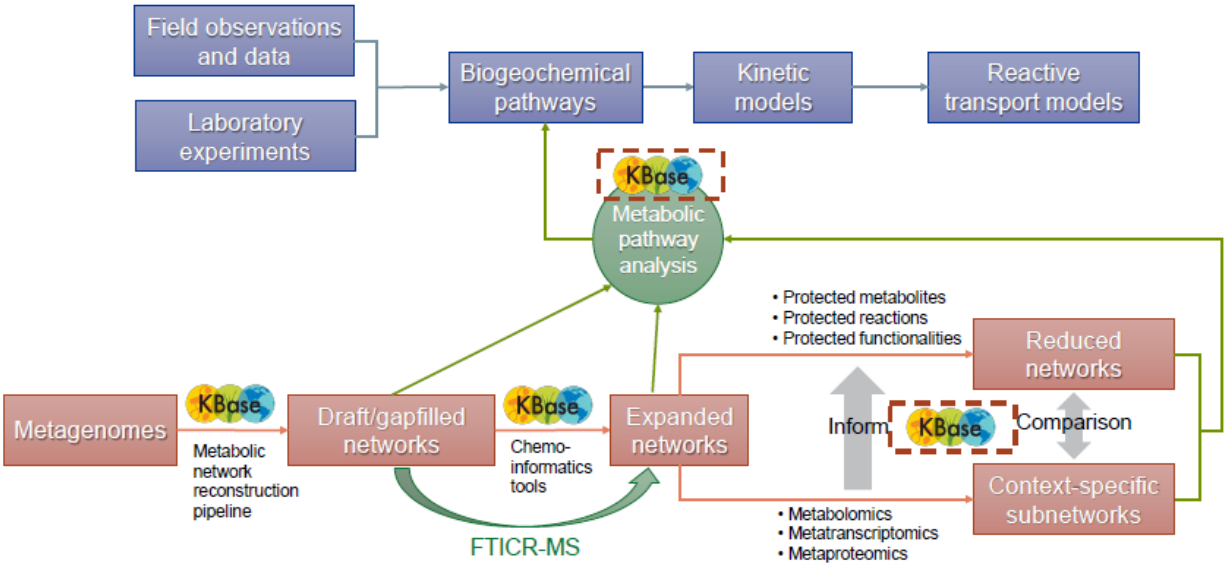


Figure 2 Workflow for integrating multiple –omics data to generate microbial reaction networks for use in reactive transport modeling

2.3 Pamela Weisenhorn (ANL) – Microbiome Heterogeneity Across the Redox Dynamic Zone

Sediment properties, hydrologic processes, and microbial community dynamics often interact to set up dynamic redox zonation in subsurface environments. These redox conditions strongly control the mobility of many redox-sensitive contaminants such as radionuclides and metals (Figure 3). A KBase microbiome/amplicon pipeline was presented in the context of redox dynamic zones being studied by the ANL SBR SFA project (https://doesbr.org/research/sfa/sfa_anl.shtml) at the Savannah River Site (SRS). The pipeline consists of six steps.

1. Creation of a taxon table: This step is currently performed outside the KBase system, but the KBase functionality is in development. The inputs to this step are raw sequence reads, which are used to make taxonomic assignments.
2. Data upload: The taxon table is imported into KBase in this step. First, the metadata is uploaded as a TSV or Excel file, then the taxon table is uploaded and attributes are assigned. The table upload requires a consensus sequence for each taxon, which provides a link between genome and community approaches.

Steps 3–5 all act on the matrix object uploaded in Step 2 and provide alternative approaches for exploratory analysis of the community.

3. Hierarchical cluster analysis: KBase provides a simple interface in which some of the parameters of the clustering algorithm can be specified by the user.

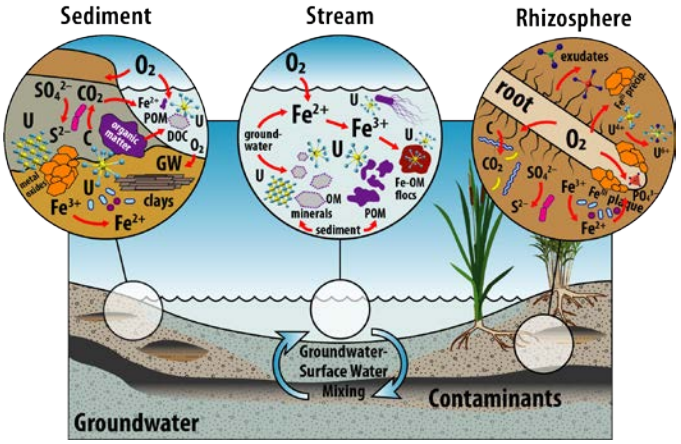


Figure 3 Schematic diagram showing critical interactions among hydrologic, geochemical and microbial processes in the groundwater-surface water mixing zone

4. Principal Components Analysis (PCA): Again, a simple interface allows specification of algorithm controls. A multi-dimensional PCA is performed on either the rows or columns of the uploaded matrix object as specified in the control interface.
5. Interaction Networks: First, a correlation matrix is computed from the uploaded matrix object, then the correlation matrix is used to construct a correlation network. Optionally, a correlation heat map can be plotted which indicates correlation between organisms; however, this step may be difficult if there are a large number of operational taxonomic units.

Finally, the outputs of steps 3–5 are used to guide selection of organisms and interactions for use in the metabolic models:

6. Metabolic modeling: Once the species and/or interactions of interest are identified, the consensus sequence links are used to identify the closest available genomes. These genomes are used to generate metabolic models that can be executed within the KBase framework to, for example, examine auxotrophies within the microbial community. The genome-scale metabolic models can in principle be linked to reactive transport models (Scheibe et al. 2009), although this remains an active area for development.

The pipeline was illustrated by application to wetland hydrobiogeochemical processes at the SRS, with emphasis on understanding linkages between microbial community dynamics and water quality (contaminant transport). Sediment cores were collected along a profile moving away from a stream. In the core nearest the stream, the sediments exhibited a distinct vertical redox profile: (1) shallow sediments are organic-rich and oxidized; (2) a deeper layer is organic-rich but reduced; and (3) the deepest layer is mineral sediment (organic-poor) and reduced. Sediment samples were taken from each of these redox zones to construct microcosm experiments. Sediments were mixed with creek water and sealed in a glovebox. Four treatments were applied (oxic [air-sparged] or anoxic [Ar-sparged], with or without 1mM glucose amendment), and microcosms were monitored over 32 days (5 sampling periods). Analyses were performed for pH, methane, hydrogen, CO_2, glucose, Fe(II), and microbial community composition. The 16S sequences obtained were fed into the pipeline described above, leading to the following conclusions:

- Post-incubation communities were dependent on the treatment but also on the initial community composition.
- Microbes responded to the treatment in consistent ways in the organic-rich sections. Legacy effects (initial conditions) exerted greatest control, then redox state (oxic-anoxic), and glucose addition had only secondary impact.
- Responses were distinct in the mineral layer, which showed effectively no clustering based on the various treatments.
- Interaction networks showed strong structure (a single large cluster) in organic sediments, especially in the oxic organic samples. However, there were only weak interactions in the mineral layer comprising small clusters of 7–8 organisms.
- After incubation, uranium was in a reduced condition under anoxic treatments (with or without glucose amendment).

3.0 Resource Overviews and Tutorials

Workshop participants were asked to bring their laptops, as the second portion of the workshop focused on tutorial-style introductions to a number of relevant computational resources. These were intended to introduce the participants to a range of available resources and give them an opportunity in some cases to try them out in real time. This portion of the agenda had three major components as described in the following sections.

3.1 KBase (DOE's Systems Biology Knowledge Base): A platform for integrating, analyzing, and modeling genomic data in a Jupyter-notebook-like interface

Chris Henry (ANL/KBase) presented an overview of the KBase analysis and modeling platform, with emphasis on a proposed roadmap for integrating systems biology with ecosystem modeling. KBase is a suite of shared open data management, analysis, and modeling capabilities. It includes 160 apps, each of which has a graphical user interface. Workflows can be constructed in the form of Jupyter notebooks, which provide reproducible records of computational processes for provenance capture and knowledge transfer. Existing workflows are provided for comparative genomics, metabolic model reconstruction, and many other elements; KBase also provides a software development kit, which provides flexibility for users to add their own tools and integrate them with Jupyter notebooks.

A live demonstration of KBase capabilities was provided, with participants encouraged to follow along on their own laptops and execute KBase workflows in real time. The pipeline focus for this demonstration was on the use of shotgun metagenomes to create functional microbial community models as input for Flux Balance Analysis simulations of species interactions. Participants that were new to KBase were encouraged to create a new account, which can be easily done by linking to the user's ORCID identification account (other options such as Google accounts are also provided). Upon signing into KBase, a dashboard (Figure 4) is presented containing available "narratives"—analysis workflows based on Jupyter notebooks. These include a set of available tutorial narratives as well as any narratives the user may have created and saved previously.

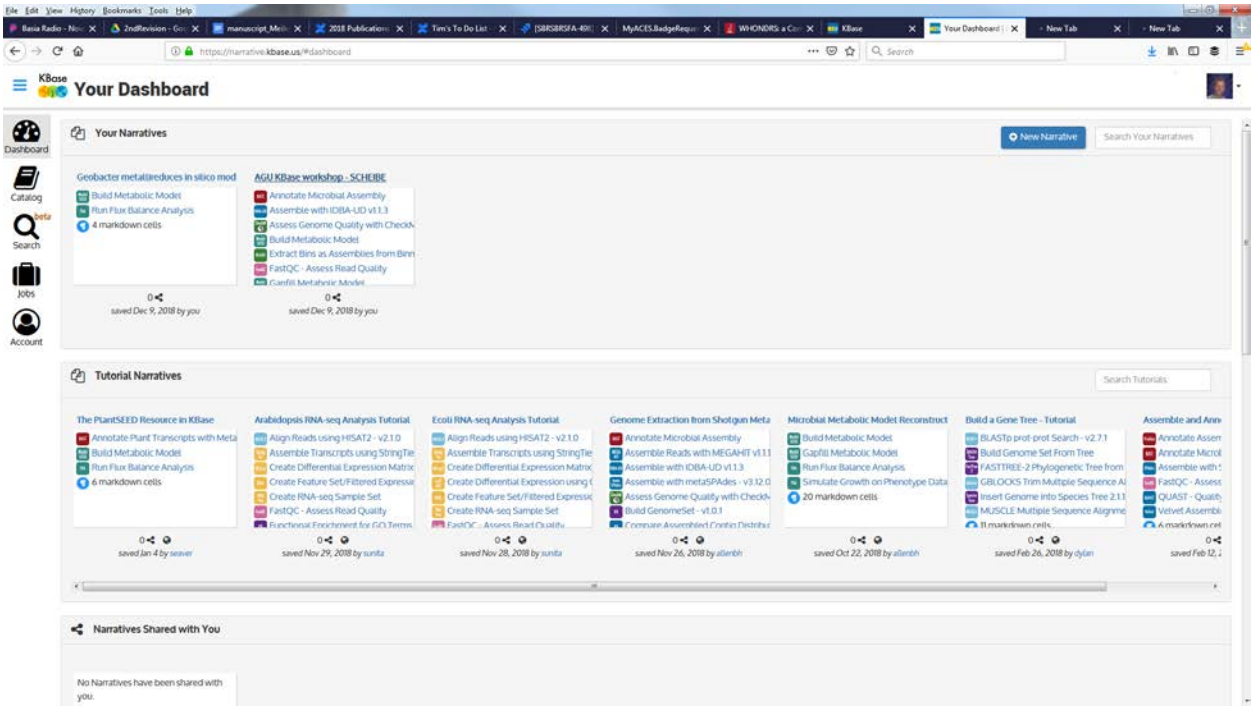


Figure 4 Screenshot of an example user dashboard in KBase

A tutorial had been created specifically for the workshop (see "AGU KBase Workshop" narrative in Figure 4), which the presenter walked through with the participants. A number of KBase capabilities were illustrated by example, including:

- Shotgun metagenomics information was assembled to form metagenome contigs, which were then binned into species contigs and annotated to form genomes

- Genome information was used to construct, gap-fill, and reduce metabolic network models, which were then executed using Flux Balance Analysis
- An illustration of how metabolic data could be used to support model expansion and novel pathway discovery

3.2 FREDA: Tools developed by experts at EMSL and Pacific Northwest National Laboratory for FTICR-MS data integration, visualization and modeling.

Lee Ann McCue (PNNL/EMSL) presented an overview and tutorial of the FREDA (FTICR R Exploratory Data Analysis) tool, an interactive interface for exploratory analysis of FTICR-MS DOM data. Such data are a form of metabolomics information, as organic carbon molecules in the environment are byproducts of metabolic processes. As illustrated in the presentation by Hyun-Seob Song described above, these data can potentially be integrated with other -omics data to generate metabolic reaction networks that can in turn be incorporated into reactive transport modeling codes such as PFLOTRAN.

This presentation took the form of a real-time demonstration of the web-based FREDA system, available at https://msc-viz.emsl.pnnl.gov/FREDA/. A similar demonstration can be viewed on YouTube at https://youtu.be/V7HWplbtT2A. FREDA analyses are initiated by uploading processed results of FTICR-MS measurements, which must be in a flexible prescribed format. Note that raw FTICR-MS data comprise mass spectra, which must be pre-processed separately (outside FREDA) to get them into the form that is ingested by FREDA (this step is typically performed by EMSL staff before data are provided to users). A FREDA dataset consists of two files in.csv format: (1) a data file that contains quantified data for each mass observed (rows) in each sample analyzed (columns); and (2) a molecular identification file that contains either the inferred molecular formula for each mass observed, or the corresponding counts of C, H, O, N, S, and P atoms in the molecular formula. An optional indicator can be provided if isotope peak identification was performed. A link to an example dataset is provided on the FREDA "Data Requirements" page. When uploading the dataset, an interactive menu asks the user for necessary information needed to interpret the data files. Once uploaded, the dataset can be pre-processed to calculate various metrics (values) as directed by the user. These include molecular ratios (O:C, H:C, N:C, P:C, and N:P, used to generate van Krevelen plots), the Kendrik mass and defect (used to generate Kendrick plots), the nominal oxidation state of carbon (NOSC, a measure of thermodynamic state), and several other measures. The preprocessing interface also provides some initial statistical and visual summaries of the selected metrics. The next step in the analysis is application of filters that allow the user to specify which mass peaks should be retained in the dataset for subsequent visualization. For example, a mass filter allows specification of a range of masses of interest. Interactive graphs show the impact of filtering on the number of data retained. Once filtering is complete, more complex visualizations can be generated. Currently, FREDA provides options to generate the following plots: (1) van Krevelen diagrams (Figure 5), (2) Kendrick plots, (3) density plots, and (4) custom scatter plots. These plots are highly interactive and can be customized by the user to a large degree (for example, selection of subgroups of samples for visualization or coloring graph points by different variable values). Graphs can be saved

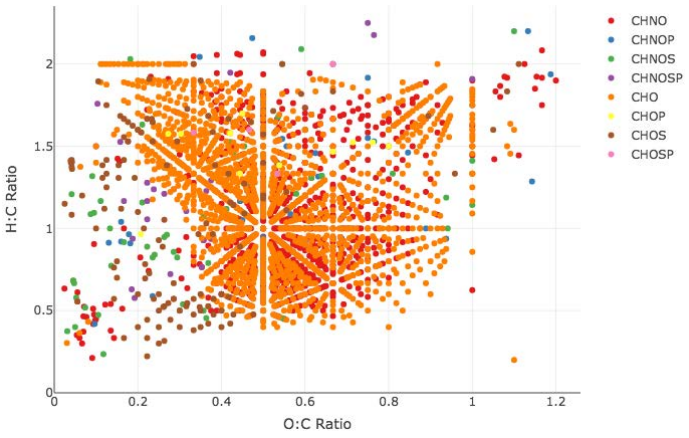


Figure 5. A van Krevelen plot generated by FREDA. This plot uses molecular ratios to group organic carbon molecules into general classes (e.g., lipid, carbohydrate, protein, etc.).

in.png format, and/or the plot parameters can be saved allowing for downloads of the processed data or images in other formats (pdf or tiff).

3.3 PFLOTRAN—A high-performance subsurface flow and biogeochemical reactive transport simulation code.

Many chemical transformation processes in the Earth's critical zone are catalyzed by micro-organisms. These reactions break down or build up complex molecules, couple elemental cycles, form interactions with plants, fungi and minerals, and change material mobility and phase, among many other things. Reactive transport models (RTMs) are used to integrate and test our understanding of coupled flow, transport, and biogeochemical reaction processes, and to quantitatively describe and predict chemical distributions in time and space (Steefel et al. 2005). Representations of microbial processes in RTMs, both implicit and explicit, have developed rapidly using a variety of approaches (Meile and Scheibe 2018). In this segment of the workshop, three tutorial-style presentations were given to illustrate various means for incorporating microbial reaction networks (such as those derived using KBase workflows) into RTMs. These demonstrations focused on the reactive transport code PFLOTRAN (https://www.pflotran.org/), which is widely used within the BER science community. PFLOTRAN is an open source and freely-accessible code for simulating subsurface multiphase flow and multicomponent biogeochemical transport. The object-oriented code is designed to run on computing platforms ranging from laptops to supercomputers. However, we note that a wide range of similar codes exist (Steefel et al. 2015) and can be used in similar manners.

3.3.1 Roelof Versteeg: Predictive assimilation framework and cloud-based PFLOTRAN modeling

Roelof Versteeg (Subsurface Insights, Inc.) has been funded by the DOE Small Business Innovation Research (SBIR) program to develop a web-based interface to the PFLOTRAN code as part of a predictive framework for hydrobiogeochemical processes. His background expertise is in geophysical methods for subsurface imaging, and the presentation began with five key perspectives learned from his geophysical application experiences:

1. There is never enough automation.
2. Geophysics alone is not sufficient; we need integration with models and other data.
3. Data-model integration is complex and difficult.
4. We need to couple surface/subsurface process understanding with observational data.
5. There remains a need for better sensing systems to fill critical data gaps.

The takeaway message from this introduction was that understanding ecosystem function requires both new science and enabling tools. These may include new –omics data tools (as discussed by other presenters) as well as new sensing technologies and scalable analysis tools (models). The Predictive Assimilation Framework (PAF) being developed by Subsurface Insights integrates data acquisition, data ingestion and management, analysis, processing, and result delivery as summarized in Figure 6.

One element of the PAF is computational analysis using RTMs such as PFLOTRAN. Subsurface Insights is developing a web-based interface to a cloud implementation of PFLOTRAN (Figure 7), and this interface was demonstrated at the workshop. Through this interface users can execute PFLOTRAN jobs using a standard PFLOTRAN executable, which can sit on any server. Through the interface a selected number of model parameters can be exposed so that users can rapidly change these parameters, run jobs, and see results. The models can also be controlled through an application programming interface (API). Attendees that registered early had been provided temporary user accounts ahead of time and could follow along on their own laptop. The demo problem was a simple 0D (batch) or 1D (column) reactive transport simulation (to minimize time required to actually run simulations across the internet); participants followed along as input parameters were changed, the model was executed, and results were plotted in real time. Of particular relevance to this workshop was one of the 0D models included in the demo, which represented a cybernetic model of microbial regulation (Li et al. 2017; Song et al. 2017) incorporated into PFLOTRAN using the reaction sandbox (see next subsection).

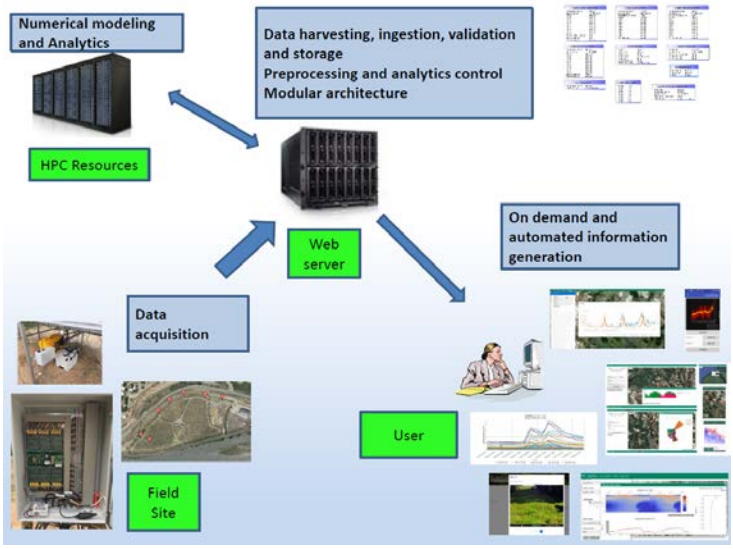


Figure 6. Schematic diagram of Predictive Assimilation Framework (Subsurface Insights)

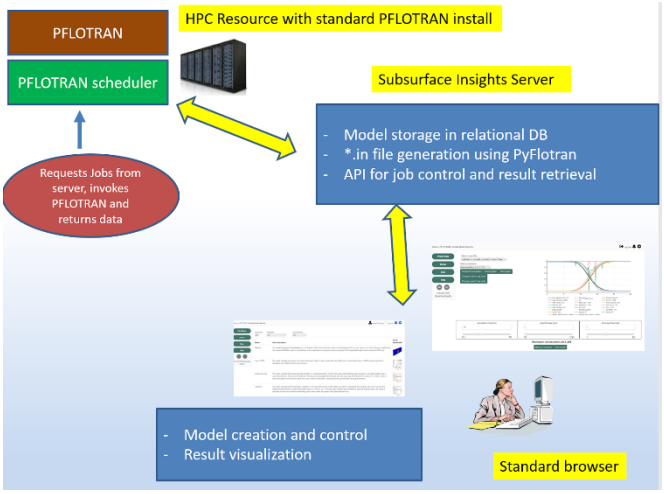


Figure 7. Schematic diagram of PFLOTRAN web interface

3.3.2 Xuehang Song: PFLOTRAN reaction sandbox – User experiences

A critical step in integrating –omics-based understanding into RTMs is implementing new and complex reaction models within existing RTM code frameworks. The PFLOTRAN code provides a "reaction sandbox" module within which custom reaction models can be imported and used with other PFLOTRAN capabilities. Xuehang Song (PNNL) has utilized the PFLOTRAN reaction sandbox as part of research being conducted under the PNNL SBR SFA (https://sbrsfa.pnnl.gov/) and presented an overview of the reaction sandbox in the context of his own experiences. "*Many PFLOTRAN users have asked for a means of implementing custom kinetic rate expressions for chemistry. The reaction sandbox fulfills this purpose by isolating PFLOTRAN's chemistry and providing a simplified reaction framework within which the researcher may quickly implement a kinetic reaction without completely learning/understanding PFLOTRAN's reaction process model. The reaction sandbox also serves as a tool for testing kinetic reactions prior to acceptance and integration within the code.*" (Hammond 2015). The reaction sandbox is a user-defined FORTRAN 2003/2008 module.

The presentation walked participants through the process of creating a new reaction in PFLOTRAN using the reaction sandbox. This process includes some or all of the following steps:

- Reaction class (define species and procedures)
- Procedure: create (allocate reaction objects)
- Procedure: read (read parameters from external file; note that parameters can also be hardwired if preferred)
- Procedure: setup (initialize reaction)
- Procedure: react (evaluates reaction—this is where the reaction network can be fully customized)
- Procedure: destroy (destroys allocable or pointer)
- Other user-defined procedures, such as subroutines to record cumulative consumption/production of reactants

An example reaction sandbox is provided as part of the PFLOTRAN source code as reaction_sandbox_example.F90. This can be used as a starting point to create a new reaction module by following the instructions in the comments contained in that file:

- Rename the module, reaction class, and module procedures.
- Add variables to the reaction class as needed.
- Populate module procedures with code that creates, reads, initializes, evaluates, and destroys the reaction class. Note that only the procedures that create and evaluate the reaction are required.
- Add the new reaction class to the reaction sandbox's linked list.

Once complete, PFLOTRAN must be recompiled and the new reaction system can then be tested.

This presentation concluded with a number of examples drawn from the PNNL SBR SFA that illustrate potential uses of the reaction sandbox to incorporate new kinetic models such as cybernetic models for regulatory control of microbial reactions (Song et al. 2017), temperature-dependent reaction rates, and a coupled thermal-hydro-biogeochemical model (Song et al. 2018). It was noted that new formulations that are well tested and useful to the community can eventually become part of the PFLOTRAN main branch, and that the cybernetic model is now available as $PFLOTRAN_DIR/src/pflotran/reaction_sandbox_pnnl_cyber.F90.

3.3.3 David Moulton: Alquimia – An application programming interface for geochemical codes

As has already been noted, there exist a multitude of codes to simulate reactive transport (Steefel et al. 2015). Similarly, there is an even larger body of codes that simulate hydrologic processes in a variety of environments and at different time and length scales. While tools like the reaction sandbox described above allow users to import new reaction models into PFLOTRAN, individual scientists may either (1) prefer to use another RTM code or (2) use a flow and transport code that does not simulate biogeochemical reactions.

David Moulton (LANL) presented an overview of Alquimia, an interoperable interface that provides access to the reaction models embodied in PFLOTRAN to other codes without requiring extensive modification of existing code bases. Alquimia was developed under the Interoperable Design of Extreme-scale Application Software (IDEAS) project, funded by DOE-BER. The IDEAS project (https://ideas-productivity.org) broadly addresses challenges related to scientific

software engineering for high quality and productivity and has generated a number of products related to software interoperability. Alquimia enables interoperability of PFLOTRAN (and CRUNCH, another RTM) with other codes that need reaction module components. Such interoperability is enabled by two primary elements: (1) use of an operator splitting approach that separates flow and transport calculations from reaction calculations within a single time step of simulation and (2) specification of an API that allows codes to communicate. The operator splitting allows each code to retain primary responsibility for one aspect of the overall simulation (reducing the need for code modifications) and the API specification enforces a signature for geochemical reactions that allows each participating code to communicate by writing only a single interface component. The geochemical engine is initialized through a preliminary geochemical speciation step that sets initial and boundary conditions. An example application was presented based on connecting the Amanzi RTM with PFLOTRAN through Alquimia. Amanzi (and its derivative Advanced Terrestrial Simulator) (Coon et al. 2016) simulates surface and subsurface flow in complex ecosystems, and uses reaction models in PFLOTRAN through the Alquimia interface.

4.0 Discussion and Next Steps

The workshop wrapped up with a moderated group discussion in which the presenters served as panelists and addressed questions raised by the organizers and participants. The discussion was wide-ranging and free-form and is therefore difficult to summarize in a succinct manner. Some of the key points of discussion are listed here in bullet form:

- How can we address plants within the resource framework discussed today (the next frontier)?
- What are the range of potential uses of an integrated resource framework? These include (1) generate new concepts and hypotheses; (2) enable prediction of system responses to perturbation by including biological regulation; (3) move biogeochemistry knowledge forward and integrate with other process domains; (4) translate process understanding to larger scales; and (5) understand how chemistry and biology co-evolve and respond to change.
- How can we communicate our integrated capabilities beyond the DOE science community (e.g., other agencies and research groups)? Key may be to demonstrate the utility of the approach first, then do workshops, webinars, and other forms of outreach. A challenge for computational elements may be limited access of others to high-performance computing systems.
- What roles can community-oriented approaches to science play? Open data and supporting metadata are critical (e.g., ESS-DIVE https://data.ess-dive.lbl.gov/). Community approaches to distributed generation of environmental datasets (e.g., WHONDRS (Stegen and Goldman 2018) https://whondrs.pnnl.gov) can provide highly impactful inputs to modeling workflows.
- Is this interactive, integrated approach sustainable? Data systems and software tend to come and go, how can we ensure longevity? One approach could be licensing to industrial partners for commercial deployment.
- What is the current bottleneck to integration of genome-scale metabolic models with reactive transport models? Complexity of natural microbial communities is daunting; methods are needed to reduce complexity of representations in appropriate ways. Also need to improve tools in KBase for assembly of complex environmental metagenomes.
- Can these approaches be applied in other domains beyond subsurface? Ocean science, perhaps atmospheric chemistry, also need similar approaches. Models need generalized interfaces to allow interoperability.
- Can KBase provide sets of experimental data to serve as benchmark problems for testing various workflows and apps? This is currently in development.

Many workshop participants expressed enthusiasm about the progress already made, and opportunities that lie ahead, for bringing the resources discussed together to solve challenging and important environmental problems. Progress will continue on specific scientific fronts (e.g., as SFA projects engage regularly with KBase), through community efforts such as WHONDRS and IDEAS, and through user engagement with EMSL, JGI, and ESS-DIVE resources. A vision imagined by the community for some time is now becoming reality, as communities of scientists work more collaboratively, enabled by DOE program investments in these resources (Figure 8) and increased openness to shared data and computational infrastructures.

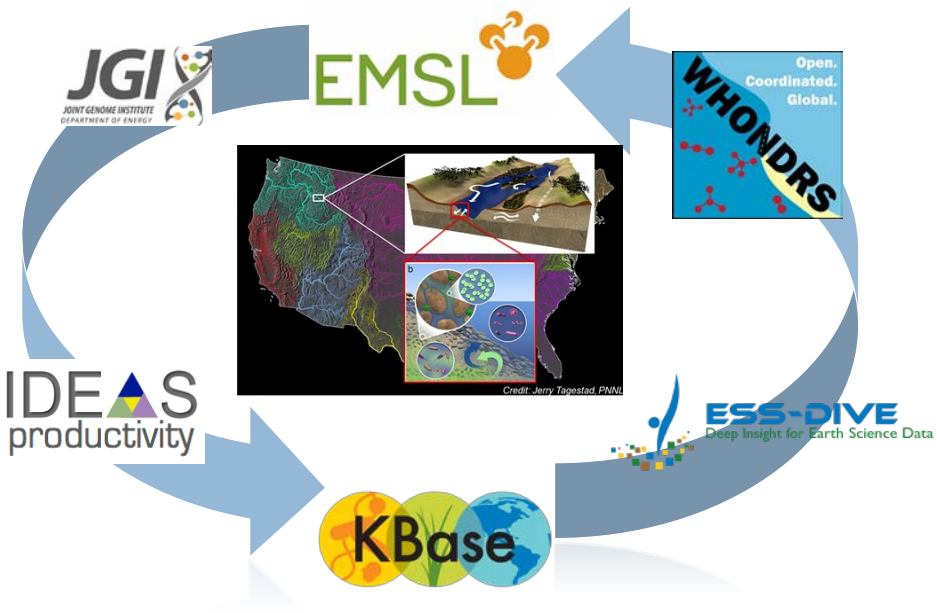


Figure 8. Schematic diagram of multiple interacting resources applied to solve scientific problems of national importance

5.0 References

Coon ET, JD Moulton and SL Painter. 2016. "Managing complexity in simulations of land surface and near-surface processes." *Environmental Modelling & Software* 78:134-149. DOI: 10.1016/j.envsoft.2015.12.017.

Graham EB, AE Goldman, AR Crump, AE Goldman, LM Bramer, E Arntzen, E Romero, CT Resch, DW Kennedy and JC Stegen. 2017. "Carbon Inputs From Riparian Vegetation Limit Oxidation of Physically Bound Organic Carbon Via Biochemical and Thermodynamic Processes." *Journal of Geophysical Research-Biogeosciences* 122(12):3188-3205. DOI: 10.1002/2017jg003967.

Graham EB, AR Crump, DW Kennedy, E Arntzen, S Fansler, SO Purvine, CD Nicora, W Nelson, MM Tfaily and JC Stege. 2018. "Multi 'omics Comparison Reveals Metabolome Biochemistry, Not Microbiome Composition or Gene Expression, Corresponds to Elevated Biogeochemical Function in the Hyporheic Zone." *Science of the Total Environment* 642:742-753. DOI: 10.1016/j.scitotenv.2018.05.256.

Hammond GE. 2015. "PFLOTRAN: Recent Developments Facilitating Massively-Parallel Reactive Biogeochemical Transport." In *American Geophysical Union Fall Meeting*, pp. Poster B43B-0547. December 14-18, San Francisco, CA.

References

Lehmann J and M Kleber. 2015. "The Contentious Nature of Soil Organic Matter." *Nature* 528(7580):60-68. DOI: 10.1038/nature16069.

Li MJ, YQ Gao, WJ Qian, L Shi, YY Liu, WC Nelson, CD Nicora, CT Resch, C Thompson, S Yan, JK Fredrickson, JM Zachara and CX Liu. 2017. "Targeted quantification of functional enzyme dynamics in environmental samples for microbially mediated biogeochemical processes." *Environmental Microbiology Reports* 9(5):512-521. DOI: 10.1111/1758-2229.12558.

McClure RS, CC Overall, EA Hill, HS Song, M Charania, HC Bernstein, JE McDermott and AS Beliaev. 2018. "Species-Specific Transcriptomic Network Inference of Interspecies Interactions." *ISME Journal* 12(8):2011-2023. DOI: 10.1038/s41396-018-0145-6.

Meile C and TD Scheibe. 2018. "Reactive transport modeling and biogeochemical cycling." In *Reactive Transport Modeling: Applications in Subsurface Energy and Environmental Problems*, pp. 485-510. eds: X Y., F Whitaker, T Xu and C Steefel. Ch. 10Wiley Press.

Scheibe TD, R Mahadevan, YL Fang, S Garg, PE Long and DR Lovley. 2009. "Coupling a Genome-Scale Metabolic Model with a Reactive Transport Model to Describe In Situ Uranium Bioremediation." *Microbial Biotechnology* 2(2):274-286. DOI: 10.1111/j.1751-7915.2009.00087.x.

Schmidt MWI, MS Torn, S Abiven, T Dittmar, G Guggenberger, IA Janssens, M Kleber, I Kogel-Knabner, J Lehmann, DAC Manning, P Nannipieri, DP Rasse, S Weiner and SE Trumbore. 2011. "Persistence of Soil Organic Matter as an Ecosystem Property." *Nature* 478(7367):49-56. DOI: 10.1038/nature10386.

Song HS, DG Thomas, JC Stegen, MJ Li, CX Liu, XH Song, XY Chen, JK Fredrickson, JM Zachara and TD Scheibe. 2017. "Regulation-Structured Dynamic Metabolic Model Provides a Potential Mechanism for Delayed Enzyme Response in Denitrification Process." *Frontiers in Microbiology* 8. DOI: ARTN 1866 10.3389/fmicb.2017.01866.

Song XH, XY Chen, J Stegen, G Hammond, HS Song, H Dai, E Graham and JM Zachara. 2018. "Drought Conditions Maximize the Impact of High-Frequency Flow Variations on Thermal Regimes and Biogeochemical Function in the Hyporheic Zone." *Water Resources Research* 54(10):7361-7382. DOI: 10.1029/2018wr022586.

Steefel CI, DJ DePaolo and PC Lichtner. 2005. "Reactive transport modeling: An essential tool and a new research approach for the Earth sciences." *Earth and Planetary Science Letters* 240(3-4):539-558. DOI: 10.1016/j.epsl.2005.09.017.

Steefel CI, CAJ Appelo, B Arora, D Jacques, T Kalbacher, O Kolditz, V Lagneau, PC Lichtner, KU Mayer, JCL Meeussen, S Molins, D Moulton, H Shao, J Simunek, N Spycher, SB Yabusaki and GT Yeh. 2015. "Reactive transport codes for subsurface environmental simulation." *Computational Geosciences* 19(3):445-478. DOI: 10.1007/s10596-014-9443-x.

Stegen JC and AE Goldman. 2018. "WHONDRS: a Community Resource for Studying Dynamic River Corridors." *Msystems* 3(5). DOI: ARTN e00151-18 10.1128/mSystems.00151-18.

Vishnivetskaya TA, CC Brandt, AS Madden, MM Drake, JE Kostka, DM Akob, K Kusel and AV Palumbo. 2010. "Microbial Community Changes in Response to Ethanol or Methanol Amendments for U(VI) Reduction." *Applied and Environmental Microbiology* 76(17):5728-5735. DOI: 10.1128/Aem.00308-10.

Wu XQ, LY Wu, Y Liu, P Zhang, QH Li, JZ Zhou, NJ Hess, TC Hazen, WL Yang and R Chakraborty. 2018. "Microbial Interactions With Dissolved Organic Matter Drive Carbon Dynamics and Community Succession." *Frontiers in Microbiology* 9. DOI: ARTN 1234
10.3389/fmicb.2018.01234.

Appendix A

Workshop Agenda

8:00–8:15 Tim Scheibe (PNNL/EMSL) – Welcome and introduction

8:15–8:45 Romy Chakraborty (LBNL) – Microbial interactions with dissolved organic matter

8:45–9:15 Hyun-Seob Song (PNNL) – A metabolic network-based approach to biogeochemical reaction modeling

9:15–9:45 Pamela Weisenhorn (ANL) – Microbiome heterogeneity across the redox dynamic zone

9:45–10:00 Break

10:00–Noon KBase workshop on microbiome modeling (Chris Henry, ANL)
- Assembly, binning, annotation of metagenomic data
- Reconstruction of isolate and community metabolic models
- Flux balance analysis and omics data integration
- Cheminformatics

Noon–1:00 Working Lunch – FTICR data visualization/analysis (LeeAnn McCue, PNNL)
- Visualization tools for FTICR data
- Tools to reconstruct pathways from FTICR data

1:00–3:00 PFLOTRAN workshop
- PFLOTRAN Predictive Assimilation Framework (Roelof Versteeg, Subsurface Insights)
- PFLOTRAN reaction sandbox tutorial (Xuehang Song, PNNL)
- IDEAS Alquimia interface (David Moulton, LANL)

3:00–3:15 Break

3:15–4:00 Working jam session, questions and answers, group discussion, feedback