Computational Tool Development for Integrative Systems Biology Data Analysis

Summary of projects awarded in 2020 under Funding Opportunity Announcement DE-FOA-0002217

Projects

- High-Throughput, Accurate Gene Annotation Through Artificial Intelligence and HPC-Enabled Structural Analysis
- Finding the Missing Pieces: Filling Gaps that Impede the Translation of Omics Data into Models
- Discovery of Signaling Small Molecules (e.g., Quorum-Sensing Molecules) from the Microbiome
- Overcoming Systems Biology Bottlenecks: A Pipeline for Metabolome Data Processing Analyses and Multi-Omics Integration
- Machine-Learning Approaches for Integrating Multi-Omics Data to Expand Microbiome Annotation
- Harnessing the Power of Big Omics Data: Novel Statistical Tools to Study the Role of Microbial Communities in Fundamental Biological Processes

Contact

BER Program Manager
Ramana Madupu
301.903.1398
Ramana.Madupu@science.doe.gov

Websites

BER Genomic Science program
genomicscience.energy.gov

DOE Office of Biological and Environmental Research
science.osti.gov/ber

DOE Office of Science
energy.gov/science

GSP Computational Biology
genomicscience.energy.gov/compbio

KBase
kbase.us

The Genomic Science program (GSP) of the Office of Biological and Environmental Research (BER), within the U.S. Department of Energy Office of Science, supports systems biology research on microbes, plants, plant-microbe interactions, and environmental microbial communities. Understanding and harnessing the metabolic and regulatory networks of plants and microbes will enable their design and re-engineering for improved energy resilience and sustainability, including advanced biofuels and bioproducts.

The widespread adoption of high-throughput, multiomic techniques has revolutionized biological research, providing a broader view and deeper understanding of cellular processes and the biological systems they drive. In pursuit of predictive modeling and genome-scale engineering of complex biological systems important for bioenergy, GSP-supported research generates vast amounts of complex omics data from a wide range of analytical technologies and experimental approaches. These data span many spatiotemporal scales, reflecting the organizational complexities of biological systems, and present significant computational challenges for identifying causal variants that influence phenotype. Accurate modeling of the underlying systems biology depends on surmounting those challenges.

To construct coherent knowledge of the systems underpinning and governing the diverse phenomics and functioning of environmental and host-associated microbial communities, detailed characterizations are essential for community genomic, transcriptomic, proteomic, and other systems processes—from a variety of samples and conditions. Such characterizations necessitate the ability to combine large, disparate datasets of heterogeneous types from multiple sources, integrated over time and space, and to represent emergent relationships in a coherent framework.

The breadth of plant and microbial community datasets and the complexities in the integration of different data layers present enormous challenges. Innovative approaches are needed to work effectively with, and glean useful insights from, complex, integrated molecular omics data. Computational simulation and rigorous hypothesis testing depend on the ability to incorporate multiple experimental and environmental conditions and associated metadata. Currently, the generation rate of multi-omic

datasets greatly exceeds the ability to analyze, integrate, and interpret them.

In fiscal year 2020, BER solicited applications proposing innovations in data integration approaches and new software frameworks for the management and analysis of large-scale, multimodal, and multiscale data. The program sought novel computational tools that will lead to scalable solutions for omics analysis, data mining, and knowledge extraction from complex datasets (experimental and calculated). Also sought were capabilities that are interoperable and effective for computationally intensive data processing and analyses for directing systems-level investigations. To aid the interpretation of multimodal data for environmental sciences, BER encouraged research focused on the enhancement of existing software or approaches already broadly used by the genomics community.

Applications were requested on research topics focused on the development of novel computational, bioinformatic, statistical, algorithmic, or analytical approaches, toolkits, and software. These capabilities include:

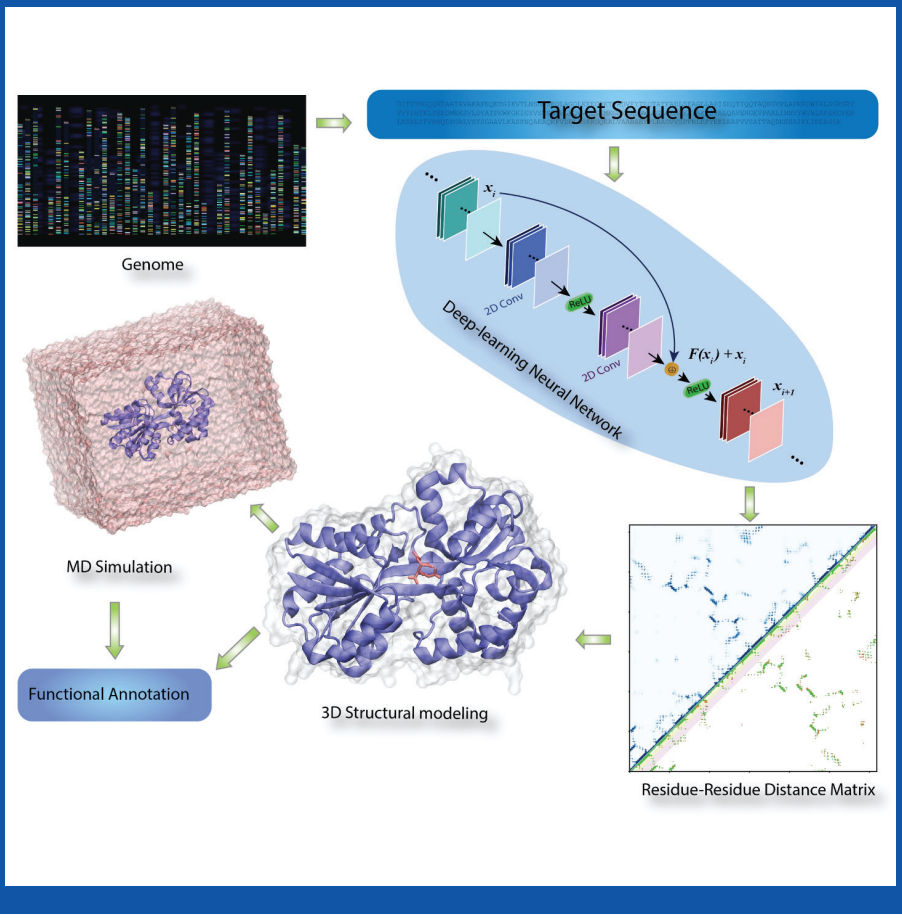
- Innovative computational strategies to enhance, scale, and optimize the management and processing throughput of large, complex, and heterogeneous systems biology data generated across scales for effective integration and interpretation.
- Integration of omics data with biochemical and biophysical measurements to provide insights into fundamental biological processes and to identify novel biological paradigms.
- Derivation of a systems-level understanding from orthogonal datasets of microbial cultures and communities via the development of integrated networks and computational models.
- Data integration approaches and new software frameworks for management and analysis of large-scale, multimodal, and multiscale data that enhance the transparency, effectiveness, and efficiency of data processing approaches.
- Data mining for the comparative analysis across large-scale datasets to infer microbial community composition and interactions or microbial community analysis to handle a wide range of functional genomics data types.

2020 Awards

High-Throughput, Accurate Gene Annotation Through Artificial Intelligence and HPC-Enabled Structural Analysis

- **Principal Investigator:** Jeffrey Skolnick (Georgia Institute of Technology)
- **Co-Investigators:** Jianlin Cheng (University of Missouri, Columbia); Ada Sedova and Jerry M. Parks (Oak Ridge National Laboratory)

The ability to predict the function of a protein-coding gene from its sequence is a grand challenge in biology. The goal of this project is to create a computational infrastructure to infer gene function from gene sequence using informatics, multiscale simulation based on high-performance computing (HPC), and machine-learning pipelines. Accurate gene annotation using computational methods will facilitate genomic science breakthroughs essential to understanding and harnessing life processes in bacteria, fungi, and plants. The incorporation of information from state-of-the-art structural modeling and simulation methods, together with evolutionary analysis and systems biology databases, will make this possible. This synergy, paired with HPC-enabled bioinformatics and machine learning, will provide vastly more powerful methodologies, tools, and results in gene annotation than previously existed. This project's success will advance one of BER's primary missions—translating nature's genetic code into predictive models of biological function—and will be facilitated by HPC resources provided by DOE leadership computing facilities.



Inferring Gene Function. Starting from a sequenced genome, each protein sequence is selected and used as input into a deep-learning neural network to predict the distance matrix between protein residues. Using this predicted distance matrix, a protein structural assembly algorithm is used to generate an initial model. For multiple domain proteins, molecular dynamics (MD) simulations are then employed to predict mutually exclusive sets of contacts to determine the conformational ensemble of the domain packing arrangements and refine the resulting protein structures. The final selected structural models will then be used to predict the protein's function. For enzymes, the Enzyme Commission number and placement of the protein in the appropriate metabolic network will be done. All tools will be made available as open-source programs and for use on supercomputers and academic clusters. [Courtesy Georgia Institute of Technology]

Finding the Missing Pieces: Filling Gaps that Impede the Translation of Omics Data into Models

- **Principal Investigator:** Christopher S. Miller (University of Colorado Denver)
- **Co-Investigators:** Kelly C. Wrighton (Colorado State University), Farnoush Banaei-Kashani (University of Colorado Denver), and Chris Henry (Argonne National Laboratory)

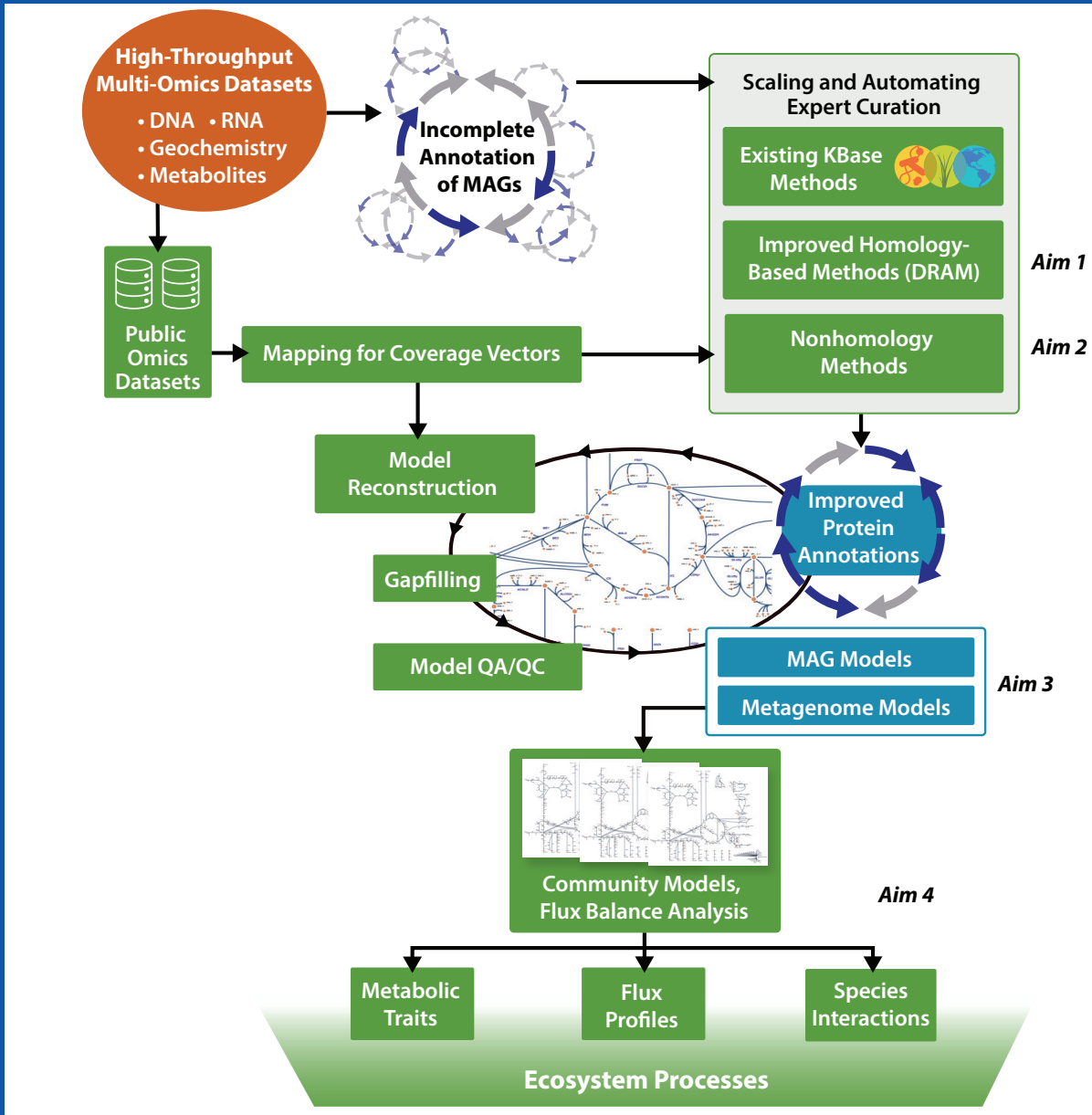
Advances in DNA sequencing and associated genome-enabled high-throughput technologies have made the assembly of microbial genomes and partial genomes recovered from the environment routine. In theory, computational inference of the protein products encoded by these genomes and the associated biochemical functions should allow for the accurate prediction and modeling of key microbial traits, organismal interactions, and ecosystem processes that drive biogeochemical cycles. In practice, however, a lack of scalable computational annotation tools means these outcomes are rarely achieved without expert manual curation, which scales extremely poorly. Scalable annotation should be informed directly by the time-consuming manual curation protocol of protein and metabolic annotations that researchers often enlist to understand and model microbial community metabolism.

The objective of this project is to innovate upon and scale up expert curation approaches to create a probabilistic annotation framework in the DOE Systems Biology Knowledgebase (KBase). This framework will enable easy-to-use and metabolism-centric protein annotation. Microbial proteins and metabolites function in the context of complex systems of inter- and intraorganismal interactions. To infer, understand, and model microbial and ecosystem traits and processes of relevance to biogeochemical cycling, protein and metabolite function need to be encoded, inferred, and studied in the context of this interconnected and dynamic network of interactions. Current approaches ignore these interactions and rely almost exclusively on limited sequence homology methods to infer protein function.

Finally, functional annotations are dynamic and can—and should—evolve when new evidence arises. Thus, annotations should be curatable, versioned, and probabilistic. Dissemination of new methods in KBase will make these annotations widely available to the scientific community. This research will ultimately enable prediction of microbial phenotypic traits and interorganism interactions in the context of community metabolic models for a wide variety of ecosystems important to biogeochemical cycling and bioenergy.

Key outcomes will be integrated within a new "Annotation Collective" in KBase for robust protein function and metabolism inference. This collective will include:

- **Improved protein annotations using expanded homology-based queries using DRAM.** DRAM annotation software will be integrated into KBase an d enhanced to include inference of a diverse set of curated exchange metabolisms (e.g., resource acquisition, quorum sensing, and antibiotic production) known to impact microbial community functionality.
- **New algorithms for nonhomology-based annotations.** Nonhomology methods can supplement homology-based methods with information about shared operon structure, coexpression of transcripts and proteins, and covariance in ecosystems. An additional deliverable from this research will be improved, fast-read mapping approaches, as they are key to scaling nonhomology methods across all public KBase data.
- **The ability to construct organism and community-level models and use them for annotation gap-filling.** Metabolic flux models have the capacity to evaluate annotations from a pathway context, either providing evidence for or against weak annotations if there is a "gap" in the models that could be filled by annotating candidate proteins.
- **Tools for utilizing robust annotations for predicting community-level models and interactions**. With robust protein functional annotations, an additional annotation layer of predicted production, reception, and expression for known exchange metabolisms can inform understanding of community-level interactions. This knowledge will be represented in new community-level metabolic models.



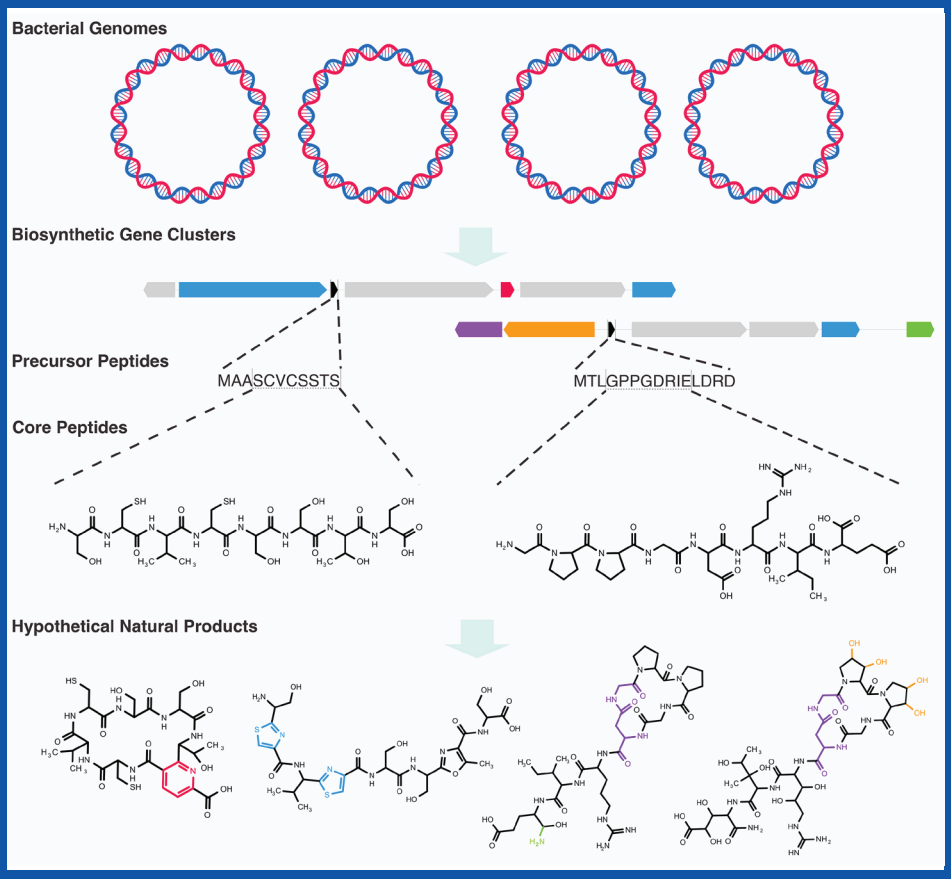
Improving Predictions of Microbial Community Function. This research scales and improves functional predictions for metagenome-assembled genome (MAG) protein products, which inform genome-centric and whole-metagenome metabolic models capable of predicting traits, flux, and interactions relevant to ecosystem processes. Multi-omic datasets (input) typically enable rapid generation of thousands of MAGs. Aim 1 of this project improves homology-based annotation of protein products predicted from these MAGs using the Distilled and Refined Annotation of Metabolism (DRAM) framework. Aim 2 improves nonhomology–based annotation methods. Metabolic models (Aim 3) are used to refine protein annotations in an iterative curation cycle. Refined models will enable community-scale flux balance analysis, predicting metabolic interactions that underpin ecosystem properties (Aim 4). Blue boxes highlight new curatable knowledge incorporated from Aims 1–3 into the DOE Systems Biology Knowl- edgebase (KBase). [Courtesy University of Colorado Denver]

Discovery of Signaling Small Molecules (e.g., Quorum-Sensing Molecules) from the Microbiome

- **Principal Investigator:** Hosein Mohimani (Carnegie Mellon University)
- **Co-Investigator:** Pieter Dorrestein (University of California, San Diego)

Microbial communities are regulated through the interactions between their microbial members. Most signal transduction pathways in the microbiome are known to be modulated through small-molecule products of microbial biosynthetic gene clusters (BGCs). Advances in 16S and shotgun metagenomics have revolutionized understanding of the microbial composition of various communities and their BGCs. Preliminary results from this and other research show that environmental metagenomes contain thousands of BGCs with uncharacterized small-molecule products that potentially play roles in signal transduction.

The aim of this project is to develop computational techniques for discovering these small molecules and characterizing their role in signal transduction. Recent analysis of tens of thousands of public isolated genomes and metagenomes has identified over 330,000 BGCs included in the Integrated Microbial Genome Atlas of Biosynthetic Gene Clusters (IMG-ABC). However, connecting these BGCs to their molecular products has not kept pace with the speed of microbial genome sequencing (fewer than 1% of the BGCs from IMG-ABC are connected to their molecular products). Discovering the chemical structure of these BGC products is the first step toward characterizing their activity. Moreover, many of these products might have novel chemistry or modifications, shedding light on the functionality of biosynthetic enzymes. The overarching goal of this project is to develop a high-throughput platform for determining the molecular products of BGCs in IMG-ABC using mass spectral data. The expected outcome is a catalog of microbial small molecules that play roles in signaling



Identifying Signaling Small Molecules. An Seq2NP algorithm automatically mines hundreds of thousands of microbial genomes and biosynthetic gene clusters from IMG/IMG-ABC repositories and constructs an atlas of hypothetical natural products. Missing post-translational and post-assembly modifications can be predicted by searching the molecules from the atlas against the spectra from the Global Natural Product Social molecular infrastructure using a Dereplicator+ *in silico* database search. [Courtesy Carnegie Mellon University]

in plant-associated microbial communities, along with their BGCs.

The research team recently developed computational techniques, including Dereplicator and Dereplicator+, for the identification of known small molecules and their variants from tandem mass spectra. Searching billions of mass spectra from publicly available datasets, such as Global Natural Product Social molecular infrastructure (GNPS), has revealed thousands of known small molecules and their novel variants. However, the majority of GNPS spectra remain unannotated. As a step forward, the research team hypothesizes that many of these unannotated spectra are the product of BGCs in microbial genomes and metagenomes. Building upon the team's previous work developing new computational tools for discovering natural products from mass spectral and genomic data, the immediate plan of this project is to develop new algorithms to elucidate the structure of novel signaling peptide natural products (PNPs) from microbial communities and to construct a catalog of signaling PNPs and their BGCs. This will be achieved by:

- Predicting modification of peptide natural products from their BGCs.
- Constructing an atlas of hypothetical small molecules by mining microbial genomes and IMG-ABC.
- Collecting liquid chromatography with tandem mass spectrometry (LC-MS-MS) data on plant microbial isolates and discovering PNPs with novel modifications in these spectra using the atlas.

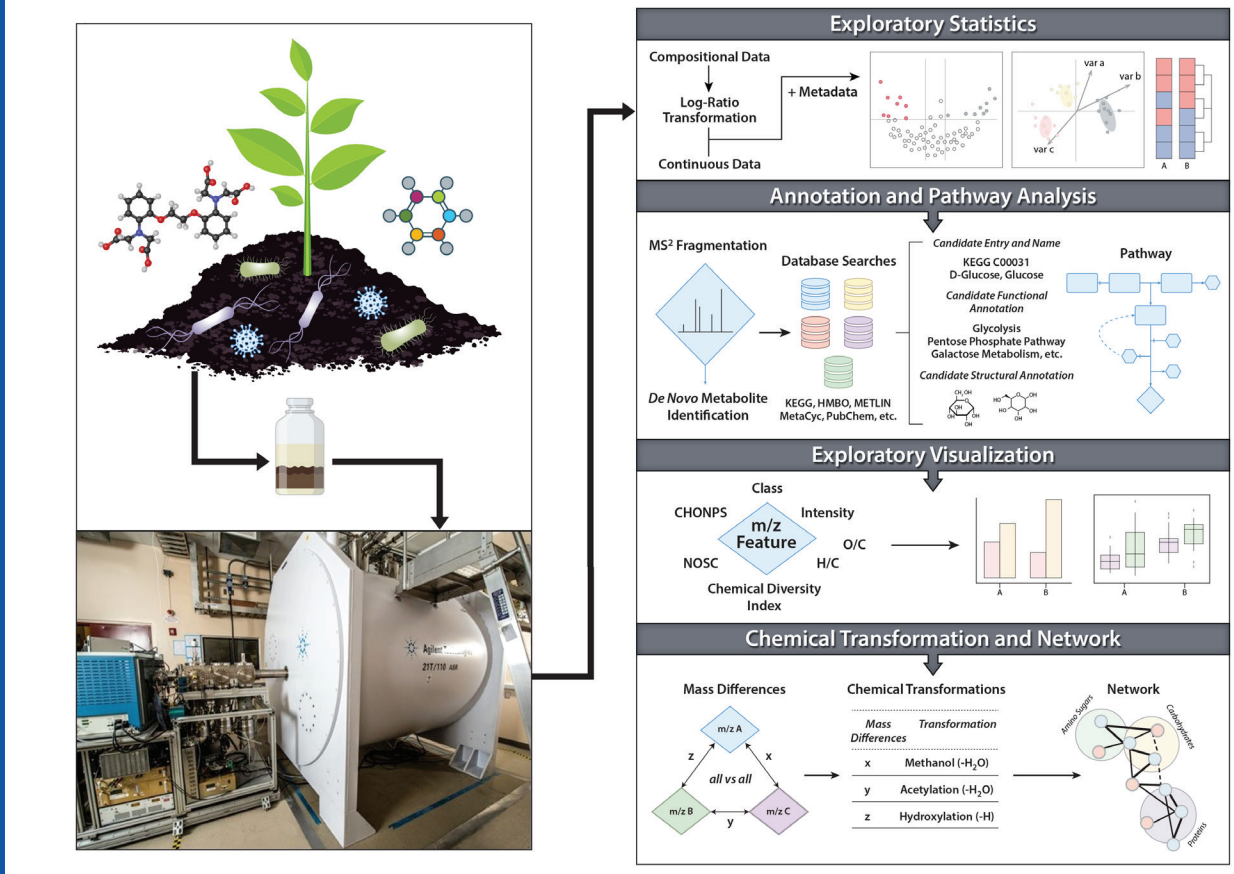
While the research team's computational methods are designed for general discovery of novel PNPs, special emphasis will be placed on the discovery of signaling PNPs from plant-associated microbes. All the data, software, and results developed during this project will be available through the GNPS infrastructure.

Overcoming Systems Biology Bottlenecks: A Pipeline for Metabolome Data Processing Analyses and Multi-Omics Integration

- **Principal Investigator:** Malak M. Tfaily (University of Arizona)
- **Co-Investigator:** Nancy Hess (Pacific Northwest National Laboratory)

During the past decade, advances in different omics technologies such as metagenomics, metatranscriptomics, metaproteomics, and metabolomics have revolutionized biological research by enabling high-throughput monitoring of biological processes at the molecular and organismal level and their responses to environmental perturbation. Metabolomics is a newer and fast-emerging technology in systems biology that aims to profile small compounds within a biological system that are often end products of complex biochemical cascades. Such compounds are the link from genome, transcriptome, and proteome to the phenotype. Thus, metabolomics provides a key tool in the discovery of the genetic basis of metabolic variation.

Despite advancements and increasing accessibility of multi-omics technologies, integration of multi-omics data in analysis pipelines remains a challenge, especially in the environmental field. In addition, there are still



Proposed Workflow for MetaboTandem Analysis Software Toolkit. A pipeline for coherent analysis of large-scale metabolomics datasets will extract metabolic features from tandem mass spectrometry experiments. After metabolite extraction, metabolite characterization will take place using high-resolution mass spectrometry. Data analysis will involve a five-step workflow: (1) raw data processing, (2) exploratory statistical analysis, (3) candidate annotation and pathway mapping, (4) exploratory visualization, and (5) predicted chemical transformations and network analysis. MetaboTandem will accept data from targeted and untargeted LC-MS/MS studies from both negative and positive ion modes by electrospray ionization. [Courtesy University of Arizona and U.S. Department of Energy Environmental Molecular Sciences Laboratory]

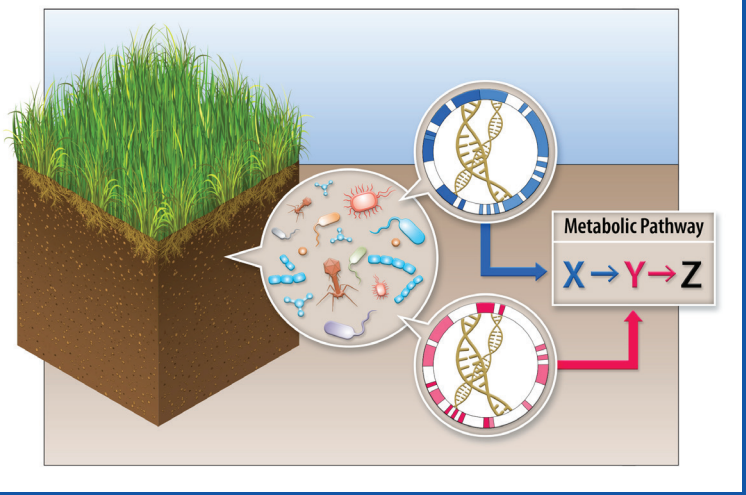
many associated bottlenecks to overcome in metabolomics before measurements will be considered robust. The overarching objective of this project is to optimize the analysis of complex and heterogeneous biological and environmental datasets by developing a user-friendly, open-source metabolomics data analysis pipeline that is integrable with other multi-omics datasets. These toolkits will be written in Python language and will incorporate well-established and community-specific software known as packages. Users can run the software as a stand-alone toolkit or through the DOE Systems Biology Knowledgebase (KBase). A website with a catalog of existing software products and best practices will be established. The website will link to DOE's Environmental Molecular Sciences Laboratory (EMSL) and Joint Genome Institute (JGI), where the experimental data will be housed. The project's large-scale multi-omics data integration approach is highly relevant to KBase's mission of achieving a predictive understanding of the role of compounds in diverse biological and environmental systems and will allow the scientific community to improve biological and metabolic genome-based predictive models by integrating "true" metabolic evidence. This research will further promote a new streamlined workflow-based approach for metabolomics and multi-omics data integration and interpretation that promotes transparent data analysis and reduces the technical expertise required to perform data import and processing.

Machine-Learning Approaches for Integrating Multi-Omics Data to Expand Microbiome Annotation

- **Principal Investigator:** Travis Wheeler (University of Montana)
- **Co-Investigator:** Jason McDermott (Pacific Northwest National Laboratory)

Research for this project is motivated by the need to understand soil communities that play a key role in the plant-soil dynamic, with impacts on food- and fuel-crop production. To understand the roles of these microbial communities, it is vital to maximally annotate their genomic and functional capacity, yet the majority of data from newly acquired microbiomes remains unannotated. This project will focus on the development of a novel method for incorporating nongenomic information into the process of annotating genomic sequence (Aim 1), and two complementary strategies building on recent advances in alignment-based and alignment-free labeling (Aims 2 and 3). In combination, these approaches are expected to substantially increase the completeness of labeling for difficult-to-annotate microbiome datasets. In addition to designing new methods, the research team will develop and release open-source software products that, where appropriate, will be integrated into existing frameworks such as the DOE Systems Biology Knowledgebase (KBase) for maximal benefit to the DOE community and the European Bioinformatic Institute's (EMBL-EBI) annotation systems for broader reach.

- **Aim 1. A Bayesian-inference framework for integrating multi-omics data.** The research team will develop a framework that establishes the utility of prior beliefs in sequence annotation, then develop those priors based on feedback among sequence annotation, metabolic pathways, protein-protein interaction networks, and sample-specific omic data. Efficacy will be validated on soil microbiome data, both real and simulated.
- **Aim 2. Accounting for sequencing error in alignment-based annotation.** Much of the unannotated content of genome sequences is the result of sequencing errors that induce false frameshifts. The research team has developed a frameshift-aware implementation of translated (protein-to-DNA) profile-hidden Markov model alignment that substantially improves database search sensitivity in the face of frameshift-inducing indels. The team will develop robust technology-specific error models and optimize for the speed necessary for metagenome-scale annotation. The software will be designed for natural integration with Aim 1 and validated in that context.
- **Aim 3. Deep-learning approaches for alignment-free annotation.** To address the large fraction of metagenomic datasets that are left unlabeled by sequence alignment, the team will develop a neural network framework to improve annotation at the levels of family and function. This strategy builds on approaches underlying modern advances in natural language processing and image segmentation, with emphasis on a novel peptide "word" embedding strategy. Through explicit handling of uncertainty, the framework will feed naturally into the Bayesian framework of Aim 1, both supporting and benefiting from its data integration mechanism.



Improving the Understanding of Microbial Communities in Soil. Communities of microbes in soil are key contributors to the plant-soil dynamic that supports production of food and fuel crops. This effort will improve understanding of soil microbial communities by using novel computational and deep-learning approaches that integrate data from multiple high-throughput sources, particularly linking the genomes of the microbial community with the activity of key metabolic pathways. [Courtesy University of Montana and Pacific Northwest National Laboratory]

Harnessing the Power of Big Omics Data: Novel Statistical Tools to Study the Role of Microbial Communities in Fundamental Biological Processes

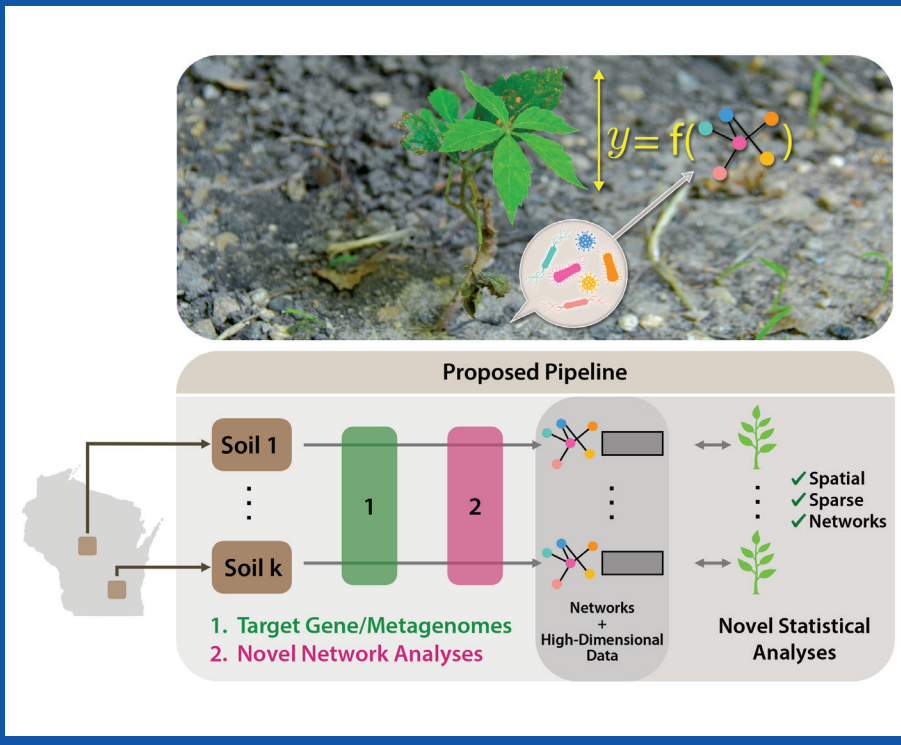
- **Principal Investigator:** Claudia Solís-Lemus (University of Wisconsin, Madison)

Microbial communities are among the main driving forces of biogeochemical processes in the biosphere. In particular, many critical soil processes such as mineral weathering and soil cycling of mineral-sorbed organic matter are governed by mineral-associated microbes. Understanding the composition of microbial communities and the environmental factors that play a role in shaping this composition is crucial to comprehending soil biological processes and predicting microbial responses to environmental changes.

To identify the driving factors in soil biological processes, researchers need robust statistical tools that can connect a set of predictors with a specific phenotype. However, innovations in statistical theory for biochemical and biophysical processes have not matched the increasing complexity of soil data. Existing statistical techniques have four main drawbacks: They (1) perform poorly on high-dimensional, highly sparse data, such as soil metagenomics; (2) ignore spatial correlation structure, which can be a key component in soil-related data; (3) do not provide valid p-values under high-dimensional settings, preventing detection of significant factors driving the phenotype of interest; and (4) tend to focus on abundance matrices to represent microbial compositions. Abundance data matrices are inherently flawed because they do not allow for easy propagation of statistical uncertainty in the data pipeline. For example, sequences are rarely a 100% match in the reference-based operational taxonomic unit (OTU) tables, which is especially troublesome for soil samples due to high microbial diversity and uneven distribution. Moreover, compositional data is restricted to sum to one, which affects how proportions behave in different experimental settings (i.e., changes in proportions in the microbial composition do not necessarily reflect *actual* biological changes in the interactions).

This project's objective is to pioneer for soil omics data the development of next-generation statistical theory (accompanied by open-source, publicly available software). The research team's novel statistical methods will overcome existing challenges in standard approaches in three ways:

- Inherently account for high-dimensional, highly interconnected data through the development of novel, mixed-effects, and sparse-learning models.



Next-Generation Statistical Methods for Soil Microbial Communities. Existing pipelines to study the role of microbial communities in biological processes tend to focus on abundance matrices in standard regression models. This project will produce sophisticated tools for high-dimensional data that will account for (1) correlations and interactions within microbial communities, (2) spatial covariance structure, and (3) adaptive model selection through post-selection inference. The project will also produce easy-to-use open-source software that will be incorporated into the DOE Systems Biology Knowledgebase (KBase). [Courtesy University of Wisconsin, Madison]

- Produce valid adaptive p-values through post-selection inference.
- Be implemented in open-source, publicly available software that will impact the broader scientific community.

By harnessing the power of big data through revolutionary new statistical theory in sparse learning and post-selection inference, the team's work will produce tools that can better understand the drivers in soil biological phenotypes to provide new insights into fundamental biological processes. The deliverables of this project are:

- Novel, mixed-effects, and sparse-learning models to predict biological phenotypes from high-dimensional omics data explicitly accounting for correlation due to spatial experimental design. As fragmented analyses implicitly assume independence in every stage of the model, the team hypothesizes that its unified approach will have more power to detect the phenotype-predictive effects, given its flexibility to account for multiple sources of dependency.
- The first post-selection inference framework to quantify adaptive p-values in factors driving biological phenotypes through a thorough study on likelihood penalties, incorporation of covariance structure, different levels of sparsity, and asymptotic theory for penalized estimates.
- An easy-to-use, open-source software that implements novel statistical theory derived in the other objectives. This user-friendly software, requiring no technical programming expertise, will be incorporated into the DOE Systems Biology Knowledgebase (KBase) with extensive documentation and step-by-step tutorials.