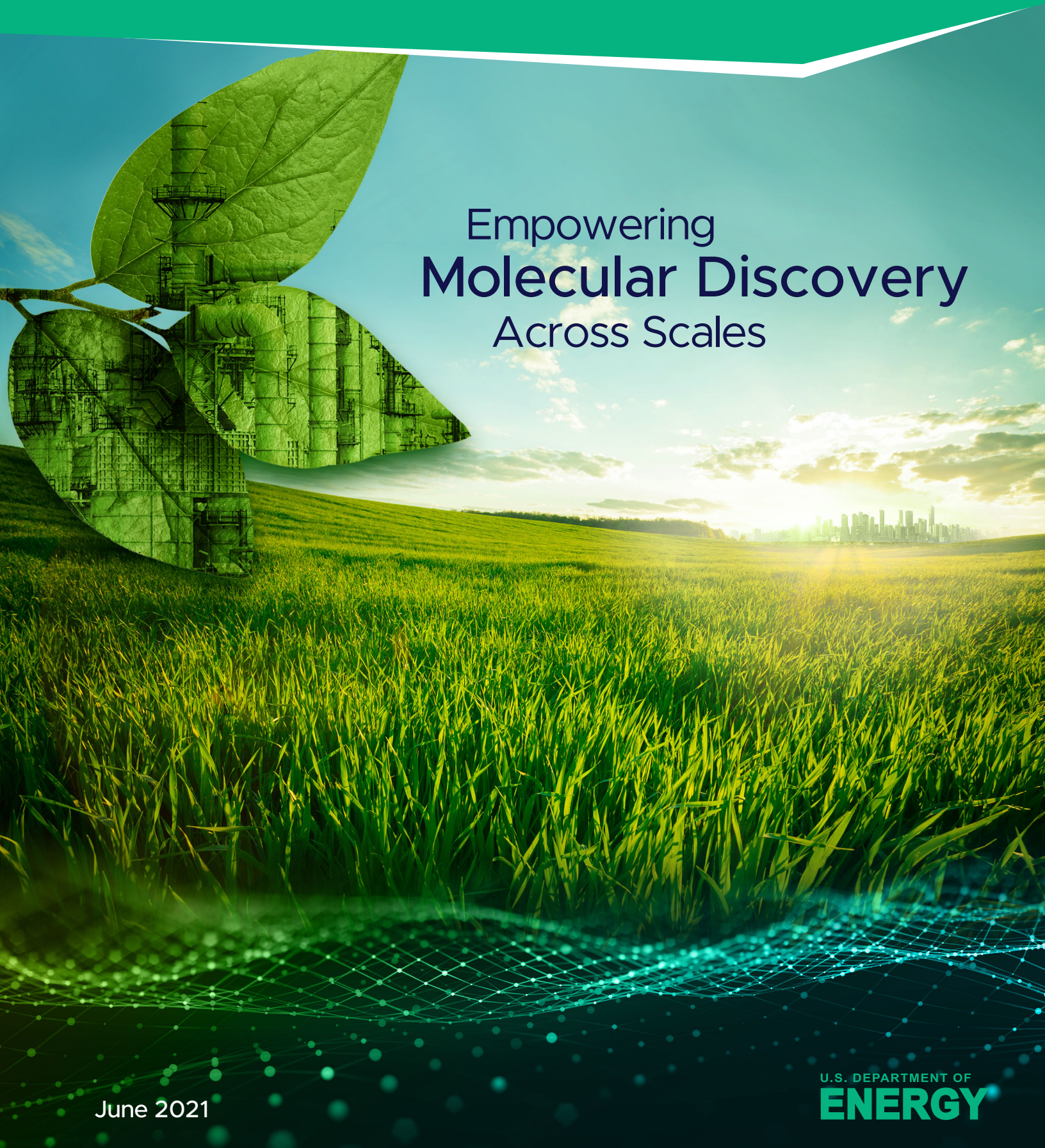
EMSL

Five-Year
Strategic Science Plan



Empowering Molecular Discovery Across Scales

June 2021

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
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

EMSL Five-Year Strategic Science Plan

Empowering Molecular Discovery Across Scales

June 2021

PNNL-SA-164144

Pacific Northwest National Laboratory
Richland, Washington 99354

ACRONYMS AND ABBREVIATIONS

AI	artificial intelligence
APS	Advanced Photon Source
APT	atom probe tomography
ARM	Atmospheric Radiation Measurement (User Facility)
BER	Biological and Environmental Research program
BERAC	BER Advisory Committee
BES	Basic Energy Sciences program
BRC	Bioenergy Research Center
BSSD	Biological Systems Science Division
CAM	Computing, Analytics, and Modeling
CAMERA	Center for Advanced Mathematics for Energy Research Applications
CBI	Chemical Biology Institute
CDAO	Chief Data and Analytics Officer
COMPASS	Coastal Observations, Mechanisms, and Predictions Across Systems and Scales
COO	Chief Operations Officer
CSMB	Center for Structural Molecular Biology
CSO	Chief Science Officer
CZO	Critical Zone Observatory
DOE	U.S. Department of Energy
EESSD	Earth and Environmental Systems Sciences Division
FAIR	findable, accessible, interoperable, and reusable
FICUS	Facilities Integrating Collaborations for User Science
FSB	Functional and Systems Biology
FTICR-MS	Fourier transform ion cyclotron resonance mass spectrometry
HPC	high-performance computing
HTP	high-throughput
IDEAS	Interoperable Design of Extreme-scale Application Software
IP	intellectual property
IRP	Integrated Research Platform
JGI	Joint Genome Institute
LBNL	Lawrence Berkeley National Laboratory
LDRD	laboratory-directed research and development
LEO	low Earth orbit
LTER	Long-Term Ecological Research
MDS	Modeling and Data Sciences
ML	machine learning
ModEx	Model–Experiment
MONet	Molecular Observation Network
NanoPOTS	Nanodroplet Processing in One pot for Trace Samples
NEON	National Ecological Observatory Network
NEXUS	Network for Execution of User Science

NIH	National Institutes of Health
NMDC	National Microbiome Data Collaborative
NMR	nuclear magnetic resonance
NSF	National Science Foundation
PNCC	Pacific Northwest Cryo-Electron Microscopy Center
SBIR	Small Business Innovation Research
SFA	Science Focus Area
SNS	Spallation Neutron Source
STAC	size- and time-resolved automated aerosol collector
STTR	Small Business Technology Transfer
TFS	Thermo Fisher Scientific
UEC	User Executive Committee
WHONDRS	Worldwide Hydrobiogeochemical Observation Network for Dynamic River Systems

CONTENTS

FIGURES

TABLES

LETTER FROM THE DIRECTOR

June 30, 2021

I am pleased to present the Environmental Molecular Sciences Laboratory (EMSL) 2021 Strategic Science Plan, describing EMSL's decadal objectives and the nearer-term research focuses and partnerships that will help achieve them. Our approach, having been enthusiastically endorsed by our advisory bodies (User Executive and Science & Technology Advisory Committees), is purposefully bold and ambitious. The objectives we describe are meant to provide a set of truly transformational capabilities for BER that will enable our user community to create deep functional understanding of complex biological and environmental systems across scales, from single proteins to ecosystems. Advancing the frontiers of scientific discovery is at the core of our strategy for engaging current and future EMSL users and partners and delivering our vision to empower research communities to discover molecular function across scales.

As a premier user facility for the U.S. Department of Energy's (DOE's) Biological and Environmental Research program (BER), EMSL provides unsurpassed access to premier molecular science capabilities to researchers discovering functions of biotic and abiotic processes for energy security and infrastructure resilience in support of DOE's research mission. Achieving that mission today requires a whole-institution shift from a historical focus on discrete instrument-based capabilities to truly multidisciplinary, integrated research—fundamental changes to our organizational, science, and technology focus. We have undertaken this transformation through the creation of seven Integrated Research Platforms (IRPs). Each platform represents an area of domain excellence for EMSL, including in-house scientific expertise, cutting-edge, next-generation capabilities, and facilities that foster multidisciplinary team research.

These IRPs and the leadership opportunities they comprise are the foundation of our strategy planning process. A series of workshops identified major trends and drivers within the areas of the user community, DOE and BER's scientific missions, and national and societal priorities; these trends and drivers motivate EMSL's strategic science and technology efforts within the IRPs over the next decade. Major recurring themes were the discovery of function of complex systems across scales from biomolecules to ecosystems, converting experimental and observational data into knowledge and models, the necessary drive toward open science, and the rapidly accelerating global pace of scientific advancements.

In response to the drivers identified in strategy workshops, this plan elaborates three audacious decadal objectives: the Digital Phenome (DigiPhen), the national Molecular Observations Network (MONet), and the Modeling and Data Sciences (MDS) capability for BER science. Because these objectives are too ambitious for EMSL to accomplish alone, we will purposefully expand our partnerships with the BER user community, companion Office of Science user facilities, other government agencies, and industry.

We envision a future where EMSL continues its legacy of transformative innovations for the BER research community, sustaining its unique role as the premier molecular sciences user facility for the DOE Office of Science, extending its history of contributions to U.S. bioeconomy and bioenergy leadership, and building the world's most complete and accurate process, ecosystem, and regional models based on experimentally derived data and information. Our success will be determined not only by our creativity in pursuit of scientific knowledge but by the planning, investments, and partnerships that enable our vision for BER user science.

Douglas Mans, EMSL Director

1.0 EMSL: A NATIONAL SCIENCE RESOURCE

The Biological and Environmental Research (BER) program within the U.S. Department of Energy Office of Science (DOE-SC) oversees the operation and stewardship of the Environmental Molecular Sciences Laboratory (EMSL) user facility, a national science resource supporting a broad national and international research community. EMSL delivers world-class facilities, advanced instrumentation, and scientific leadership that empower and enable this exceptional community of researchers to advance BER's mission to achieve a predictive understanding of complex biological, Earth, and environmental systems.

EMSL houses the multidisciplinary scientific expertise and advanced instrumentation required to continue its long history of innovation and pioneering developments in the biological and environmental molecular sciences for the user community. These critical advances accelerate scientific discovery to tackle our nation's energy and environmental challenges and address multiple BER goals and grand challenges, as well as goals identified by BER's Advisory Committee (BERAC) (Table A.1). EMSL occupies 234,000 ft^2 of laboratory and office space on the Pacific Northwest National Laboratory (PNNL) campus in Richland, Washington (Figure 1). The facility is supported by a staff of 160 scientists with expertise in the biological, chemical, environmental, computer, modeling, and data sciences. Over 150 advanced and often one-of-a-kind instruments are operated within the facility in addition to several highly specialized laboratory spaces. Included in these spaces is a state-of-the-art computing space for Tahoma, EMSL's 0.93 petaFLOPS hybrid architecture high-performance computer (HPC); the Quiet Wing, which features eight acoustically, vibrationally, and electromagnetically shielded bays housing advanced electron microscopes (e.g., a Krios cryo-EM system, a helium ion microscope, and a prototype dynamic transmission electron microscope [DTEM]); a plant sciences lab with controlled growth chambers, phytotrons, and various root and rhizosphere imaging capabilities; and a highly modified lab space to house our 21 tesla Fourier transform ion cyclotron resonance mass spectrometer (21T FTICR-MS), one of only two in the world. This Strategic Plan assures that EMSL's exceptional resources endure and evolve with the nation's science needs and are available to the user community.

Since its inception on October 1, 1997, EMSL has provided world-class leadership in the molecular sciences, driving predictive mechanistic understanding across the biological and environmental sciences for BER. EMSL researchers have supported more than 8,500 total publications, which have been cited 300,000 times (leading to an approximate cumulative *h*-index of 210). The more than 200 patent applications, approximately 15 actively licensed technologies, 10 active intellectual property (IP) licenses, 40 software copyrights, and 10 R&D 100 Awards granted to EMSL scientists illustrate EMSL's

Figure 1. The EMSL facility. EMSL provides user access to premier instrumentation and multidisciplinary science leadership in its 234,000-square-foot facility.

unique standing in the DOE-SC user facility community as a laboratory not only of scientific achievements, but also of cutting-edge technological advancements. For example, EMSL staff pioneered advanced capabilities in electron microscopy to see molecular interactions in situ and in operando, revealing critical observations in energy storage and biological transformations. Through our role as co-primary investigator for the Pacific Northwest Cryo-Electron Microscopy Center (PNCC), we have imaged over 80 protein structures—the basic machinery of life—with atomic precision, revealing mechanisms for the translation of the genetic code to functional metabolic pathways, the cycling of elements key to life on Earth (carbon, oxygen, nitrogen, and phosphorus), and microbial and plant resilience (PNCC 2021). EMSL has developed methods and instruments that accelerated the rapid and confident characterization of proteins and metabolites. EMSL's seminal contributions of high-resolution nano-separation and high-performance Fourier transform mass spectrometry led to the dominant paradigm in the field of proteomics, making EMSL a world leader in the analysis of biochemical pathways. Along the way, EMSL scientists developed, patented, and licensed the ion funnel technology found in virtually every mass spectrometer available today. With our expertise in multi-omics, EMSL published the first complete description of the proteins that comprise fungal cellulosomes, the multi-enzyme complex responsible for deconstructing biomass produced by anaerobic fungi (Haitjema et al. 2017). Our leadership in science-based cleanup solutions for DOE's Hanford, Savannah River, and other sites advanced understanding of the chemical interactions of toxic metals and radionuclides with mineral surfaces and microorganisms that control the rates by which these contaminants move through soils, sediments, and groundwater. This understanding of subsurface molecular transformations and transport allowed EMSL scientists and researchers to build pore flow models with enhanced accuracy for modeling the flow of molecules across nanometers to kilometers over femtoseconds to eons (Oostrom et al. 2016). In doing so, we have delivered technological advancements critical for securing future generations. Our work immediately benefits biological and environmental researchers and American citizens.

In the longer term, as we deliver on the promise of our vision, those benefits accrue to the nation and beyond. EMSL's breadth of scientific expertise spanning biologists, chemists, physicists, engineers, hydrobiogeochemists, atmospheric scientists, and production computing and data analytics scientists provides an unrivaled multidisciplinary approach to exploring and understanding fundamental molecular science frontiers. Our continuous focus on innovation and creativity has enabled the development of world-class and one-of-a-kind instruments for pursuing the most challenging molecular science questions. We have developed the world's most accurate mass spectrometer, the 21T FTICR, providing unsurpassed resolution and allowing researchers to observe mass differences of a single electron in molecular samples. More recently, we have also developed the Nanodroplet Processing in One pot for Trace Samples (nanoPOTS) system that provides unequaled recovery of biological molecules from ultra-small samples, enabling mass spectrometry on samples approaching single cells.

In accordance with EMSL's leading capabilities in instrument development, the world's foremost molecular computational chemistry software, NWChem, was created here. NWChem has been downloaded over 70,000 times since becoming open source in October 2010. Over the same period, NWChem reference papers (e.g., Valiev et al. 2010; Kendall et al. 2000) have been cited around 4,000 times. The code is installed on all major DOE computing facilities, many National Science Foundation computing centers, and academic computer clusters worldwide. This software has improved researchers' ability to model complex chemical and biochemical systems, including electron transfer from microbes to minerals, the stability of DNA, catalysis, hydrogen production and storage, material and surface properties, behavior of heavy elements in the environment, and biological processes.

Through the strategy described in the following sections, we endeavor to continue advancing these and to create wholly new capabilities in partnership with and for the BER research community.

1.1 Vision and Mission

Our nation's long-term energy and environmental security increasingly depend upon effective delivery of continuous innovation in multidisciplinary research to achieve a systems understanding to support DOE's, BER's, and the world's biological and environmental researchers. To create a secure bioeconomy and a predictive understanding of the Earth system, *EMSL's vision is for a research community empowered to study the role of molecular processes in controlling the function of biological and ecological systems across spatial and temporal scales and to enable a predictive understanding of the living Earth system* (Figure 2). EMSL contributes to this future state through its mission to provide access to premier multimodal molecular science instruments, data analytics, production computing, and multiscale modeling to enable researchers to study biotic and abiotic processes and understand their function in a systems context for energy and environmental security and infrastructure resilience (Figure 2). Engaging and empowering the user community is a critical element of EMSL's strategy to deliver on our mission and vision.

Figure 2. EMSL's vision and mission. Our vision is the future we seek to bring about; our mission describes EMSL's role in building that future. Both our vision and mission are in alignment with and in service of BER's vision and mission.

1.2 Strategic Planning

EMSL leadership produces and refreshes a Strategic Plan every five years to guide and support the review of the EMSL user facility, the user program, our leadership, and administration of this national scientific resource. This Strategic Plan also meets a key BER expectation for effective stewardship of EMSL.

The decadal scientific and operational objectives described in this 2021 EMSL Strategic Science Plan were developed over a 10-month period in 2020, starting with a series of in-person and virtual workshops that synthesized wide-ranging input from 50 senior science leaders, subject matter experts, members of EMSL User Executive and Science & Technology Committees (UEC and STAC, respectively), and representatives of the user community from across EMSL, PNNL, Lawrence Berkeley National Laboratory (LBNL), Oak Ridge National Laboratory (ORNL), industry, and academic institutions. These workshops were designed to inform our future directions through a broad survey of national and global trends in science, technology, and energy. That survey produced a prioritized set of emerging S&T areas that (1) are highly responsive to BER's mission, objectives, and grand challenges, as well as the future needs of the EMSL user community, and (2) build on historical or emerging areas of BER-focused scientific strength in EMSL. EMSL is uniquely positioned to lead efforts to meet these objectives in partnership with the user community and other DOE research organizations.

These strategic objectives and the research areas that support them were also shaped by input from numerous external research and advisory bodies, including the UEC (FY 2021 meeting), BER's two division directors, the EMSL program manager, BER program managers (July 2020), and the EMSL Science and Technology Advisory Committee (August 2020 and April 2021). EMSL also led multiple outreach efforts with the BER user community that explored interest in the scientific directions now presented in this 2021 EMSL Strategic Science Plan. These efforts, including the FY 2020 Multiscale Microbial Dynamics Modeling EMSL Summer School (July 2020) and FY 2021 EMSL User Integration Meeting on Visual Proteomics (October 2020), provided additional feedback and confirmed strong interest in the community for EMSL's strategic directions. Outreach and communication efforts will continue with the FY 2021 Multi-Omics Modeling of Biochemical Pathways EMSL Summer School (July 2021) and the FY 2022 EMSL User Integration Meeting on Biological and Environmental Sensors (October 2021).

This 2021 EMSL Strategic Science Plan describes our focus and planning for our three strategic science objectives (Sections 3, 4, and 5)—the Digital Phenome (DigiPhen), the Molecular Observations Network (MONet), and the Modeling and Data Science Center (MDS), respectively—and one strategic operations objective, Operations for Capacity and Pace (Section 6). This plan describes these objectives, highlights of ongoing and planned near-term activities and critical partnerships that support them, and how EMSL will leverage facilities and operations to drive progress toward these goals. The emphasis on partnership demonstrates EMSL's role in amplifying the value of BER investments in other BER and DOE resources and organizations. The plan also details how EMSL engages and empowers the user community (Section 7). Ultimately, execution of the EMSL Strategic Science Plan is driven and supported by leadership from EMSL's three science areas and supporting Integrated Research Platforms (IRPs) (see Section 2). The Strategic Science Plan is available in electronic format on EMSL's website.

2.0 TRANSFORMING EMSL'S LEADERSHIP FOR MULTIDISCIPLINARY USER SCIENCE

User science at EMSL is conducted within three foundational science areas (Figure 3): (1) Functional and Systems Biology (FSB), (2) Environmental Transformations and Interactions (ETI), and (3) Computing, Analytics, and Modeling (CAM). Establishing scientific leadership and strategic science objectives based on these three foundational science areas promotes close partnership with BER and continuous alignment of these areas with BER's Earth and Environmental Systems Sciences Division (EESSD) and Biological Systems Science Division (BSSD). The FSB and ETI science areas represent traditional areas of science focus and strength within BER and EMSL. EMSL users conducting research within these two science areas research the environmental impacts of energy production and resource use, detailing the fundamental need for resilient ecosystems and improved sources for sustainable energy and bioproducts production. BER has a longstanding commitment to understanding, modeling, and predicting the environmental impacts of energy production, as well as building a biological understanding that will serve as the foundation for sustainable bioenergy and bioproduct production through research aligned with EMSL's FSB and ETI science areas. Historically, much of this work has utilized computational and data analytics to support research, which continues to grow in importance and influence.

Over the past several decades, BER scientists have increasingly used a wide array of rapidly advancing computational and analytical methods to generate, manage, and analyze vast amounts of data to build simulation and predictive models of environmental and biological systems that drive iterative modeling and experimental design. BER has continued to highlight the significant need for data analytics, mid-range computing, software and code development and predictive modeling in BERAC reports, workshops and strategic plans from BSSD and EESSD (see Table A.1; BERAC Grand Challenges 6.1, 6.2, 6.4, and 8.5;

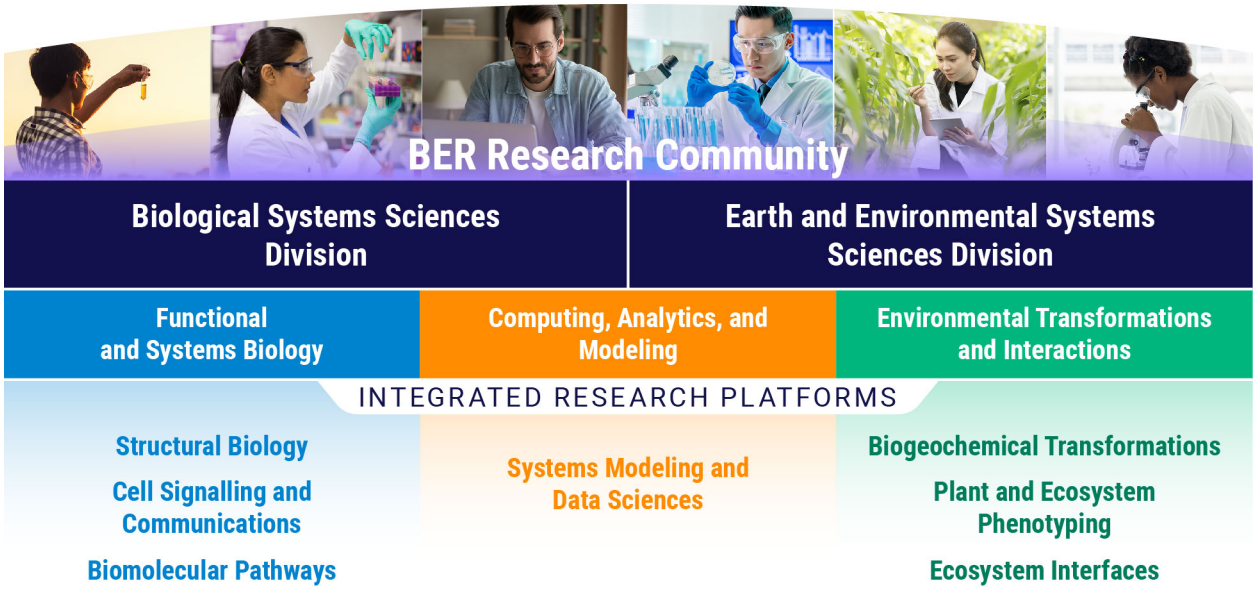


Figure 3. EMSL's science leadership is organized to empower the BER user community. The three science areas, Functional and Systems Biology, Computing, Analytics, and Modeling, and Environmental Transformations and Interactions, assure continuous alignment and close partnership between EMSL and BER's EESSD, and BSSD. EMSL's seven IRPs are aligned to EMSL's three foundational science areas, providing support for science area activities. They are centers of multidisciplinary scientific and technical domain excellence critical to execute delivery of EMSL's mission and ensure continued availability of premier S&T capabilities to users.

BERAC Facility Recommendations 4.4, 6.1, 6.3, 6.6, and 6.7; BSSD Goals 4 and 4-1; and EESSD Grand Challenge 5). Computational science has thus grown from a supporting activity where computers were used to analyze experimental data to a recognized scientific discipline where simulations are increasingly used to explain and predict scientific phenomena and generate scientific data that informs hypothesis generation and experimentation. Building on its unique position as a leading producer of multimodal environmental and biological data, analytical tools, and modeling, EMSL established the CAM science area as its newest foundational science area to accelerate integration of computing with EMSL's multidisciplinary science model, grow our leadership in these computing related fields and build the next generation of premier capabilities for BER and the user community. This addition was made in recognition of the escalating importance that computing and computational science have in advancing BER's predictive capabilities in the biological and environmental sciences.

To provide scientific direction and focus that builds science leadership, evolves multidisciplinary science, and advances EMSL's premier capabilities to meet the current and future needs of users, EMSL has established three strategic science objectives, one for each foundational science area: DigiPhen, for FSB; MONet, for ETI; and MDS, for CAM. Each strategic science objective establishes a bold scientific goal that directly supports BER mission and strategic directions. The three strategic science objectives were developed using input from a series of workshops conducted over a 10-month period in 2020 and other input from the user community and BER. They leverage and enhance EMSL's seven IRPs (Figure 3; see EMSL Leaders). Ultimately, these strategic science objectives are meant to deliver a set of transformational capabilities for BER users that facilitate deep

Working with EMSL

EMSL uses a flexible and fluid approach to engage other researchers and organizations in support of BER science missions. Early engagement may be an informal collaboration intended to advance a shared scientific goal, but later evolve into a more formal relationship or partnership supported by a contractual mechanism that delineates terms, responsibilities, and requirements. EMSL uses three programs, each with a different contractual mechanism and terms, to support formal research and development relationships, the User Program, Partner Program and Sponsored Research:

User Program

- Free to Users for non-proprietary work; science inquiry focused.
- Initiated by Users in response to proposal calls including FICUS
- Funded by EMSL User program
- Peer-reviewed for technical merit and BER relevance
- IP owned jointly; terms set by standard DOE user agreement
- Intent to publish; data: open access after standard embargo period

Partner Program

- Co-development of capabilities and technologies
- Jointly funded, utilizes EMSL intramural S&T funds
- BER relevant
- Reviewed for technical merit, strategic alignment, and User impact
- Terms and mechanism fit to IP and data protection needs

Sponsored Research

- Sponsor funded work
- Utilizes capacity hours
- Flexible terms fit to IP and data protection needs

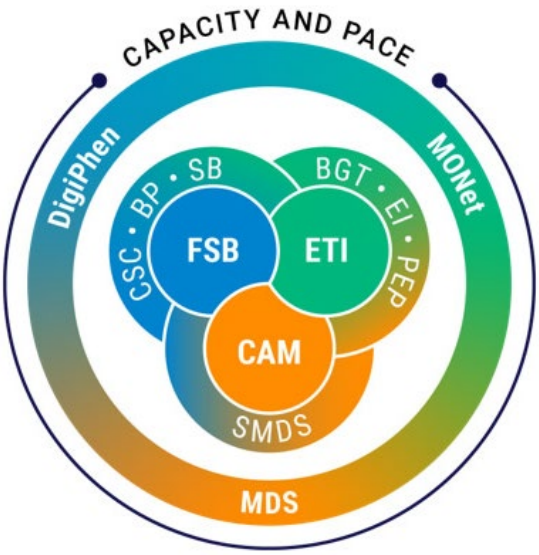


Figure 4. EMSL's Operations for Capacity and Pace objective provides direction for infrastructure development and operational activities that accelerate delivery of a suite of transformational capabilities for users through its three strategic science objectives DigiPhen, MONet, and MDS.

functional understanding of complex biological and environmental systems across scales, from single proteins to ecosystems.

A fourth objective, the Operations for Capacity and Pace objective, was established to give focus and direction on the alignment of operations to embrace, accelerate, and drive innovations that speed scientific discovery in EMSL's three strategic science objectives (Figure 4) through expanded capacity and pace. The Operations for Capacity and Pace objective provides direction for infrastructure development activities; the design, modification, and allocation of space; services to users and partners; and the life cycle of EMSL capabilities. The resulting alignment of resources and operations with our research activities amplifies the impact of each in the effort to deliver the outcomes of our three strategic science objectives. This operational objective was created to maximize the utilization and impact of the transformational capabilities EMSL is establishing through its strategic science objectives DigiPhen, MONet, and MDS for the user community. Easier and more effective user access, optimized processes and facilities, improved communications and partnership processes (see Working with EMSL), expanded space, computing and data storage supporting automation, and autonomous and remote operations are intended outcomes of the Operations for Capacity and Pace objective.

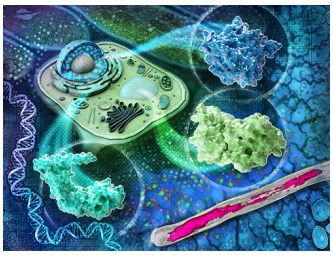
Fostering multidisciplinary user science and expanding user access to our integrative capabilities is a major goal of EMSL's strategy. EMSL's IRPs and scientists that lead the IRPs play a critical role in delivering this goal.

EMSL's IRPs (Figure 3) were constructed to serve as centers of multidisciplinary scientific and technical domain excellence in seven focus areas critical to support BER researchers and advance the needs of users through EMSL's foundational science areas and three strategic science objectives. They were purposefully selected to amplify EMSL core strengths and focus our science mission and strategic objectives on the most critical, unmet, and evolving areas of EMSL and BER science. The seven IRPs serve as the primary interface for EMSL users to connect with EMSL's science and technical capabilities, facilitate consultations on research design, and steward access to the unique scientific leadership available within the IRP multidisciplinary teams. The IRPs are EMSL's focal point for planning, leading, and executing our strategic science objectives with the user community. The IRP teams also operate across disciplines and engage with EMSL users to develop and evolve the leading science questions that drive continued evolution of EMSL capabilities. The move to IRPs was a critical strategic transformation of EMSL's scientific leadership.

Historically, EMSL had eight capability areas, each representing deep expertise in an instrument class or a specific analytical technique. The capability areas allowed users to interface with specific sets of instruments or capabilities. Over time, however, the value of the highly specialized but narrow technical focus of the capability areas diminished as the nature of science itself changed within and outside of EMSL. The rapid advances in science combined with the increasingly challenging research being pursued requires a level of multidisciplinary interactions that the instrument focus of the capability areas was not well positioned to deliver. The transition from capability areas to IRPs was a necessary organizational evolution in response to

the increasing complexity of BER science and the needs of EMSL users. That growing complexity required emphasizing close collaboration within the user-facing multidisciplinary teams as well as access to cutting-edge instruments rather than reliance on single areas of technical expertise. Toward that end, each IRP retains the unique technical and instrument expertise that evolved in EMSL but now intentionally bridges boundaries between multiple scientific disciplines. This platform approach brings together EMSL's scientific and technical staff from multiple disciplines, cutting-edge next-generation instruments, and data analytics and modeling that foster interdisciplinary team research. The IRPs are meant to serve as EMSL's nucleation points for innovations that breach traditional disciplinary boundaries to pioneer new areas of scientific research, propel leadership in our science areas, and empower users through access to these innovations.

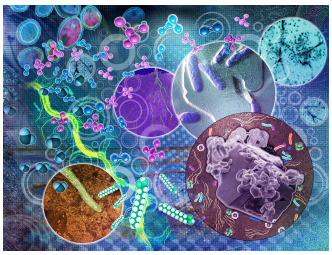
The seven IRPs are Structural Biology, Biomolecular Pathways, Cell Signaling and Communication, Biogeochemical Transformations, Ecosystem Interfaces, Plant and Ecosystem Phenotyping, and Systems Modeling and Data Sciences, detailed below.



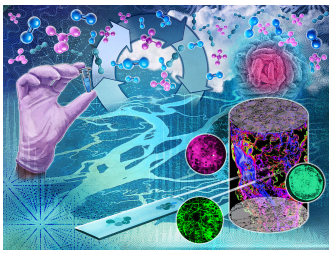
The **Structural Biology IRP** seeks structural, biochemical, and dynamic information about proteins, protein complexes, and other biomolecules at nanoscale spatial and temporal resolutions to infer function.

The **Biomolecular Pathways IRP** investigates the translation of genomic information into functional relationships among biomolecules within cells in response to changes in their internal or external environment.

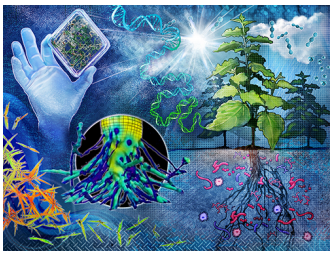
The **Cell Signaling and Communication IRP** reveals dynamic interactions and trafficking of molecular signals between cells, populations, and communities to understand complex interrelationships between organisms in response to their environment.



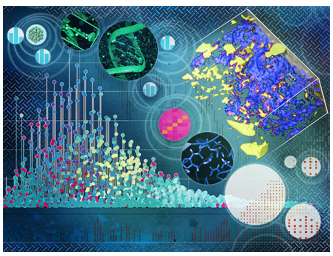
The **Biogeochemical Transformations IRP** investigates the biochemical, physical, and microbial interactions that affect chemical speciation, transport, and transformation of critical nutrients, contaminants, and compounds within the environment.



The **Ecosystem Interfaces IRP** investigates the chemical composition and transport phenomenon that result in fluxes and exchange of chemicals and nutrients, including biogenic and anthropogenic emissions, at the interfaces between ecosystem domains.



The **Plant and Ecosystem Phenotyping IRP** investigates interactions between genes and the environment at the molecular level to understand, predict, and control plant and ecosystem traits at the system scale.



The **Systems Modeling and Data Sciences IRP** advances the prediction and control of biological and environmental systems by developing approaches for advanced data analysis, data integration, multiscale modeling, and simulation of processes across scales.

Ultimately, the move to IRPs and the positioning of IRP leaders as the primary point of contact with users establishes a stronger partnership with users that expands awareness of opportunities to access and utilize the full suite of multidisciplinary capabilities available in EMSL. IRP leaders are integrated into EMSL's leadership structure to strengthen the connectivity between users and our strategic planning and investment efforts.

EMSL's matrixed science leadership model strengthens the connection to users, drives multidisciplinary science execution on a strategy

EMSL Leaders

We provide continuing access for users to EMSL's capabilities and technologies and execute our science mission and objectives through three classes of leadership roles.

1. **Our three science area leaders** are the primary liaisons to BER and the broader scientific user community, assuring programmatic alignment, coordination, and development of user programs, campaigns, and other user community research proposal calls. Science area leaders also steward our three strategic science objectives. Each science area is supported by one or more IRPs that represent more specific science domains within the broader science area.

2. **Our seven IRP leaders** are the principal points of contact with EMSL users, providing guidance and scientific partnership, coordinating access to EMSL capabilities, and spearheading capability development to meet emerging user needs.

3. The **CSO**, **CDAO**, **COO**, and **deputy of user services** enable science area and IRP leadership by optimizing investments, line management, facilities, and the user program in support of the larger EMSL strategy and user science.

that delivers the future scientific needs of BER, and advances a deeper partnership with BER and DOE. EMSL's chief science, data and analytics, and operations officers (CSO, CDAO, and COO, respectively) set the future direction for the science, technology, and operational capabilities of the user program and EMSL facilities in partnership with users and BER leadership. The deputy of user services is responsible for the setting and execution of the User Services strategy and is responsible for providing world-class customer service that enables access to EMSL's capabilities. The FSB, ETI, and CAM science area leads are EMSL's senior science liaisons to the EMSL program manager and broader BER scientific program management, assuring near-term focus and program alignment. The science area leads also serve critical roles as the primary liaisons to the broader scientific user community, building collaborations and seeding and nurturing new program and user campaign projects. The IRP leads, as experts in their scientific IRP domain, serve as the direct interface with EMSL users supporting access to multidisciplinary research tools and capabilities as well as partnering to identify unmet user needs for new capabilities. The IRP leads develop and execute roadmaps for key research areas within EMSL's DigiPhen, MONet, and MDS strategic science objectives. The FSB, ETI, and CAM science area leads work closely with the IRP leads aligned with their science area to harmonize strategic efforts with BER program managers in close coordination with the EMSL program manager for effective execution and shepherding of EMSL research calls and project support. In addition, each IRP lead liaises with the CSO, CDAO, and COO to represent and develop the respective IRP domain focus aligned to our strategic scientific and operational objectives.

3.0 FUNCTIONAL AND SYSTEMS BIOLOGY SCIENCE AREA

The Functional and Systems Biology (FSB) Science Area focuses on revealing the connections between protein structure and function, biochemical pathways, and complex phenotypic responses. Our rich approach to phenotyping incorporates interactions within cells, among cells in communities, and between cellular membrane surfaces and their environments, for microbes (archaea, bacteria, protists, viruses, algae, and fungi) and plants. FSB embraces multiscale, multimodal, and molecular experimental observations, reconstructing metabolic pathways, and modeling structure and function to improve strategies for designing plants and microbes for biofuels and biobased products, and ultimately to unravel the complexities of carbon, nutrient, and elemental cycles within cells and their immediate environment.

EMSL's FSB science area positions EMSL to lead the BER research community in addressing the compelling grand challenges in functional and systems biology by working directly with users to produce new data and knowledge necessary to translate genomes into functional knowledge and phenotypes. This science area is aligned with multiple Biological and Environmental Research Advisory Committee (BERAC) Grand Challenges and BER goals (Table A.1).

The establishment of the Structural Biology, Biomolecular Pathways, and Cell Signaling and Communication IRPs provides opportunities for users to pursue research in critical and emerging areas of importance for BER science in biosystems prediction and design through a flexible and modular approach that utilizes a single multidisciplinary IRP or a combination of IRPs. In this vein, the FSB-focused IRPs provide the ability for users to investigate targeted aspects of functional annotation and biomolecular phenotyping associated with individual proteins, metabolic pathways involving multiple interacting proteins, or perturbations to metabolic pathways and the resultant responses within and between cells leading to communal phenotyping. For EMSL users, the FSB science area IRPs present a faster, more comprehensive, and integrated approach to

discovery of wholly new functions in new and model organisms. Creating detailed functional knowledge of newly identified proteins and placing that function first in the context of metabolic pathways and then in the context of communities and community function under normal and perturbed conditions will advance phenotype characterization and prediction. The FSB science area's focus on deep functional annotation for phenotyping microbial and plant systems of critical importance to BER addresses a growing and urgent need to augment vast genomics and metagenomics datasets with biological or structural functional data to advance biosystems design and metabolic modeling efforts. The accelerating accumulation of sequence data and the corresponding need for functional data necessitates an ambitious and transformational capability for phenotyping that the FSB science area will steward in partnership with users over the next decade. Sections 3.1 and 3.2 provide background on the strategic science objective for FSB that emerged in response to EMSL's assessment of trends and drivers during our 2020 strategy workshops. The strategic science objective provides scientific direction and focus for the research efforts in the FSB science area.

3.1 Background for Strategic Science Objective 1

BER has invested in sequencing the genomes of plants and microbes to enable advances in sustainable biofuels and bioproducts, the design, modification, and optimization of plants and microbes, and in systems biology research. While genome sequencing reveals a "parts list" for a breadth of both known and uncharacterized biological processes, there are significant challenges to associating genes and their products with function and phenotype (Breaking the Bottleneck of Genomes, U.S. DOE 2019). This wealth of genomics data is heavily utilized by the BER user community to build metabolic models based on the historical approach to assign function of unassigned or new proteins by homology to known proteins. However, the genomics parts list is growing nearly exponentially without the requisite parallel increase in accurate assignment of gene and protein function needed to validate and improve metabolic and other models of biological function and phenotype.

Access to capabilities that enable experimentation incorporating both genetic and environmental variables necessary for accurate, deep, and high-throughput phenotyping by the user community is an urgent user community need. Moreover, those capabilities necessarily go beyond sequencing, and they require expansion to a broader array of multidisciplinary and multimodal molecular analytical approaches that align well with EMSL's strengths. This has resulted in an accelerating shift toward augmenting genome sequencing with a broad array of phenotyping methodologies—synthetic biology, structural biology, cellular imaging, and multi-omics—to improve the annotation and prediction of critical functions of key organisms and communities by the user community. Such multimodal and multidisciplinary approaches are a hallmark of EMSL. In addition, the rich tradition of pioneering new technologies for molecular measurements, innovating high-resolution and high-throughput multi-omics approaches and techniques, and advancing biomolecular imaging, biological modeling, computing, and analytics capabilities residing in EMSL make such endeavors in molecular phenotyping a logical, natural extension of EMSL's leading role in the environmental and biomolecular sciences.

Improving the functional understanding of plant and microbial biology is a national priority required to assure and strengthen U.S. leadership in the burgeoning bioeconomy and bioenergy arenas (White House Memo M-20-29). As the user community expands the number of microbial and plant systems genotyped, the ability to provide the corresponding phenotype is a prerequisite to building accurate metabolic models for harnessing plant and microbial systems for bioenergy applications and bioproduct use and production. Further, this ability is needed to build higher-fidelity land-based ecosystem models that reflect more accurate responses to perturbations from changing environments. Implicit in these requirements is a need to develop a fundamental understanding of genomic and regulatory principles for key biological functions to design, modify, and optimize plants, microbes, and biomass for beneficial purposes (BSSD Strategic Plan,

U.S. DOE 2021a). Acquiring this fundamental understanding calls for converting complex multimodal data from soil, water, plant, and microbial systems into integrated, modeled, and visualized simulations for accelerating the discovery of function. Each successive advancement in our fundamental understanding is driving science and scientific research to become more multidisciplinary, requiring more fluid teaming and greater open access to vast data streams (Scientific User Research Facilities and Biological and Environmental Research: Review and Recommendations, BERAC 2018). A direct corollary to the acceleration in integrated multidisciplinary science is an increased need for regular, strategic adoption and implementation of advanced technologies and approaches in computing and analytical/experimental instrumentation (Safeguarding the Bioeconomy, NASEM 2020).

To address the gap in accelerating the accurate assignment of function to predict the molecular basis of functional phenotypes, EMSL's first 10-year strategic science objective is to establish a comprehensive offering of advanced multimodal phenotyping capabilities for users. Strategic Science Objective 1 will facilitate the phenotyping of organisms important for BER missions, objectives, and goals at a scale and pace matching that of current and next-generation genotyping capabilities. EMSL's core expertise in structural biology, metabolic pathways, and cell communication position us well to drive the numerous advances required in creating these phenotyping workflows. EMSL also recognizes that leveraging complimentary advanced capabilities resident at other SC user facilities and federal agencies is a prerequisite for success in this bold but high-impact endeavor. For example, advanced genomics and gene library generation performed with our partner user facility, the Joint Genome Institute (JGI), and detailed atomic-level imaging available from the Basic Energy Sciences (BES) advanced light source and neutron source user facilities will complement the multi-omics, molecular and cellular imaging, and modeling capabilities at EMSL.

3.2 Strategic Science Objective 1: Create a Digital Phenome (DigiPhen) as a Platform for Predicting and Controlling Biology from Biomolecules to Organisms

EMSL will develop a digital phenome platform (DigiPhen) that (1) consists of experimental and analytical workflows supporting a digital representation of the functional basis of phenotype for whole microbial or plant systems, and (2) is composed of an expanding set of interchangeable, interconnecting modules that contain data and models representing the mechanistic determinants of phenotype. DigiPhen is a platform for data production and data–model integration across the fields of structural biology, biomolecular pathways, and cellular signaling and communication that connects the molecular basis of function to observable or desirable phenotypes. Molecular, structural, and functional data from EMSL and EMSL partner and collaborator analytical platforms form DigiPhen's data modules and inform connected mechanistic models of function developed in EMSL by users or by other researchers and organizations EMSL works with. When fully deployed, DigiPhen will provide the user community a massive and continually growing stream of well-curated multimodal phenotyping data and numeric or simulation models for accurately and expediently annotating biological function across species and taxa for BER-relevant microbial and plant systems.

Using DigiPhen, researchers and users will be able to interrogate experimental data to design, calibrate, parameterize, and validate metabolic and other models for the determinants of key phenotypes within well-studied model organisms to create more complete digital models of organism function. Ultimately, the data, models, and workflows can be combined to create complete digital representations of new pathways and even organisms that accurately model selected cellular and metabolic functions aligned to BER interests in bioenergy, bioproducts, and ecosystem modeling. Combined with the genomics data streams generated at our partner user facility—JGI—and systems modeling applications within the BER KnowledgeBase (KBase), a dramatic acceleration in the biodesign and systems modeling approaches within the Design-Build-Test-Learn paradigm will be achieved, expediting the development of beneficial plant and microbial systems. DigiPhen will accelerate the association of genes with phenotypes, help identify entirely new protein and metabolic

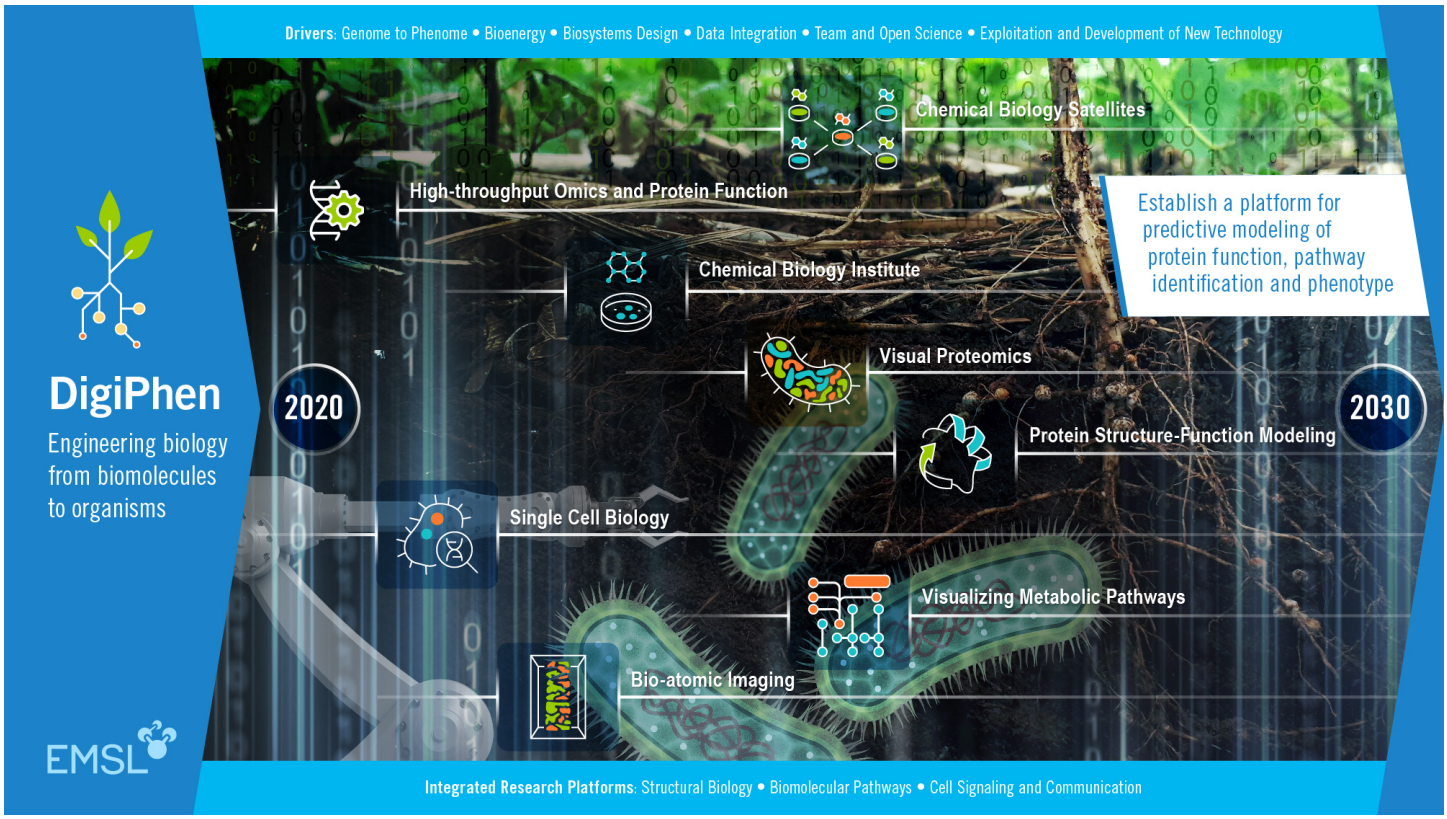


Figure 5. Overview, timeline, and research areas supporting Strategic Science Objective 1. This objective establishes EMSL's leadership in the development of a new generation of science and technology innovations that produce new knowledge about the molecular, cellular, and community foundation of phenotype required for the user community's effective translation of genomes into function and phenotype. Each research area is placed on the 2020–2030 timeline to show where we anticipate the most activity, although we expect work to begin before and to continue after.

functions, and enable model-driven design and engineering of plant and microorganism physical traits critical for environmental sustainability, scalable bioproducts, enhanced plant resilience, and feedstock productivity.

To meet the bold objective to establish DigiPhen and deliver the attendant benefits to users, EMSL in partnership with users and researchers in academia, industry, and other SC facilities, will focus efforts in eight research areas (Figure 5) over the next decade. During the first 2–5 years, our priority will be activities, programs, projects, and investments that primarily support three of these research areas: High-Throughput (HTP) Omics and Protein Function, Single Cell Biology, and Bio-Atomic Imaging. In each case, these research areas build the scientific foundation for DigiPhen while the user community makes use of the emerging science and technology for scientific inquiry that directly supports BER missions and goals.

3.2.1 HTP Omics and Protein Function Research Area

The rapid pace of genome sequencing continues to increase the catalog of predicted proteins without known functions. However, the ability to predict, control, and engineer biochemical pathways relies on understanding the function of proteins in cellular processes. EMSL will advance the multiple approaches now needed to extend beyond genome sequencing to fully characterize new and unknown proteins. Multi-omic mass spectrometry-based approaches, such as proteomics, phosphoproteomics, metabolomics, lipidomics, and glycomics, will be automated to increase throughput. The HTP Omics and Protein Function Research Area will also explore incorporation of chemical probes in automated workflows for protein function identification, laying the groundwork for establishing the Chemical Biology Institute research area and later

adopting probes and probe platforms for functional discovery. A pipeline of unknown proteins of interest can be identified with tools EMSL will develop, initially by using a preliminary amino acid sequence homology and then moving to a combination of cryo-electron microscopy (cryo-EM), mass spectrometry, and nuclear magnetic resonance (NMR) to determine the structures of novel proteins and their metabolites. This capability nears realization with the availability of a suite of modular HTP and automated platforms that allow rapid identification and functional and structural characterization of proteins and metabolites.

Supporting IRPs: Biomolecular Pathways, Structural Biology, Plant and Ecosystem Phenotyping, Systems Modeling and Data Science

Major External Engagements

- **Pacific Northwest Cryo-EM Center** (PNCC). *Ongoing:* PNCC is an NIH-funded cryo-EM structural biology center. Image processing for the center is performed at EMSL. This relationship provides a unique opportunity to leverage advances in image processing developed at PNCC for accurate atomic-level protein structure determination to infer protein or biomolecule function to benefit EMSL users and the broader structural biology capability in EMSL.
- **Thermo Fisher Scientific** (TFS). *Anticipated:* TFS is an analytical instrument manufacturer that develops both mass spectrometers and cryo-EMs. EMSL has started discussing and anticipates working with TFS to jointly advance technology development that would create the ability to soft-land proteins on EM grids to advance protein structure elucidation.

Recent and Near-Term Supporting Activities

- **Improve throughput and scope of screening for biological function and phenotypes.**
  Ongoing: EMSL Intramural S&T Research investments in (1) activity-based proteomic probes for unknown function and (2) an integrated multimodal approach to elucidate structure and function; PNNL's newly approved Predictive Phenomics Initiative and Lab Strategy (in development). The Predictive Phenomics Initiative and strategy are expected to produce investments that drive development of phenotyping capabilities at PNNL and within EMSL.
- **Increase throughput and resolution for protein structural elucidation.**
  Ongoing: EMSL Intramural S&T Research investments in (1) dynamic transmission electron microscopy and (2) a cell-free protein expression pipeline for structural biology; PNNL Laboratory Directed Research and Development [LDRD] project to develop mass spectrometry methods for soft-landing of proteins for cryo-EM imaging.
- **Expand metabolomics capacity and accelerate identification of unknown features.**
  Ongoing: EMSL Intramural S&T Research investment in an integrated multimodal approach to determine structure and function of lignin-forming protein complexes isolated from plants; a $10M PNNL LDRD investment to advance computationally enabled mass spectrometry (the m/q initiative) for PNNL, presenting opportunities for EMSL researchers to co-develop and deploy emerging capabilities to the BER user community through EMSL; and PNNL's lab-level Predictive Phenomics Initiative.

3.2.2 Single-Cell Biology Research Area

The ability to analyze single cells and single cell types will be developed to fully characterize and understand how variations in protein abundance, post-translational modification, and metabolite flux among populations of genetically identical cells give rise to the behavior of cell populations and microbial communities. This effort will dramatically improve mathematical models not only of cellular structure, but also biochemical

pathways and function of single cells by providing a complete spectrum of phenotypic variation possible for biochemical and regulatory processes within single cells. The single-cell-informed community models will enable an advanced AI/ML-based prediction of community response to environmental perturbations and function in engineered organisms. Synergies with advances in analytical chemistry platforms in the HTP Omics and Protein Function, Visual Proteomics, and Visualizing Metabolic Pathways Research Areas are expected. The availability of instruments, cell handling, and analytical workflows that produce single-cell and single-cell-type proteomics, transcriptomics, and metabolomics data for users is a key mark of success for this effort.

Supporting IRPs: Cell Signaling and Communication, Biomolecular Pathways, and Plant and Ecosystem Phenotyping

Major External Engagements

- **Scienion.** *Ongoing:* Scienion developed the world's first picoliter-volume printer capable of both automated single-cell isolation and reagent dispensing. EMSL researchers are working with leading experts in Scienion through EMSL's Partner Program (see Working with EMSL) to integrate their technology with EMSL's nanoPOTS system to deliver a transformative new technology for untargeted single-cell proteomics. This partnership will provide a foundational capability to generate a parts list of the single cell proteome essential for visual proteomics efforts.

Recent and Near-Term Supporting Activities

- **Improve single-cell-resolution mass spectrometry for proteomics and metabolomics.**
  Ongoing: Scienion-EMSL partner program project to develop a single-cell platform with nanoPOTS for combined proteomics and transcriptomics; EMSL Intramural S&T Research investments in (1) targeted spatial metabolomics, and (2) an automated approach to identify and analyze single cells in situ.
- **Expand single-cell proteomics and metabolomics into plants and fungi.**
  Ongoing: EMSL Intramural S&T Research investments to (1) further develop the fliFISH transcriptomics imaging technique (Cui et al. 2018) for higher throughput and (2) establish in-depth, single-cell proteomics for plant and fungal cells.

3.2.3 Bio-Atomic Imaging Research Area

The acquisition of atomic-level resolution and structural information of proteins, protein complexes, and enzyme active sites will be pursued to provide atomic-structure-based computational chemistry and simulation models of key biochemical functions and pathways. The realization of robust atomic-level resolution as part of structural biology analyses for proteins and protein complexes will directly inform a detailed understanding of enzyme active site chemistry. Comprehensive understanding of the active site can be used to assign function(s) for new and unknown proteins identified within the HTP Omics and Protein Function Research Area to amplify efforts in discovering and annotating function. Three-dimensional structural imaging of proteins and biomolecules will provide the space-filling information that will be used in the Visual Proteomics Research Area to generate accurate renderings and models of the cellular interior critical to understanding organismal phenotype. In addition, atomic-level resolution will provide detailed understanding of amino acid targets within the protein primary sequence for genetic engineering efforts to enhance and alter protein function in engineered systems within the Protein Structure–Function Modeling Research Area. Efforts within this area are synergistic and will leverage advancements and developments in the Image Processing Center Research (5.2.6) and Artificial Intelligence (AI)/Machine Learning (ML) for Automation Research (5.2.5) Areas as part of Strategic Science Objective 3 (Modeling and Data Sciences

Center). A progression of technical, instrument, and computational innovations that consistently advance atom probe tomography (APT) for use in imaging soft/biological materials, imaging mass spectrometry, and cryo-EM at atomic-level resolution for biomolecules is expected. The combined use of advancements in APT, mass spectrometry, and cryo-EM will reveal structure–function relationships and will represent maturation of this challenging capability development effort.

Supporting IRPs: Structural Biology, Systems Modeling and Data Science

Major External Engagements

- **National Institute of Standards and Technology (NIST).** *Ongoing:* NIST research increases U.S. competitiveness across a breadth of technologies. NIST is now developing and testing new lasers for APT. EMSL is engaging with NIST via the EMSL Intramural S&T program to test extreme UV laser systems for biological applications of APT.
- **CAMECA.** *Anticipated:* CAMECA is a leader in the development and manufacturing of APT instruments. EMSL completed an Intramural S&T project with CAMECA in 2020 and now anticipates continuing to work with CAMECA to develop and validate innovations in atom probe imaging of soft/organic matter for integration in the next generation of instruments from CAMECA by combining EMSL's expertise in bio-APT and biological applications with CAMECA's expertise in APT.
- **Advanced Photon Source (APS) and Stanford Synchrotron Radiation Light Source (SSRL).** *Anticipated:* These BES- and BER-funded user facilities provide access to specialized experimental capabilities at synchrotron light and neutron sources. EMSL will form partnerships through the Facilities Integrating Collaborations for User Science (FICUS) framework to facilitate research communities' access to capabilities that are complimentary to measurements made at EMSL and JGI. Of particular interest are techniques that provide highly resolved measurements of the biomolecular structure and dynamics of proteins.

Recent and Near-Term Supporting Activities

- **Advance APT technologies.**
  Ongoing: EMSL Intramural S&T Research and Partner Program activities with the National Institutes of Standards and Technology to develop and test extreme UV APT for biology applications.
- **Acquire next-generation APT instruments and analysis methods.**
  Ongoing: EMSL Intramural S&T Research investment in application of ML to APT data to trace nanoscale protein structure. *Anticipated:* Develop engagements with external researchers for beta-testing of next-generation APT at EMSL; Planned capital purchase of LEAP 5000-HR APT.
- **Optimize sample preparation methods for soft materials.**
  Ongoing: EMSL Intramural S&T Research investment to develop sample handling for bio-APT.
- **Advance partnerships with industry for next-generation instruments.**
  Anticipated: Pursuing industry engagements with key cryo-EM instrument vendors (e.g., ThermoFisher, JOEL, Hummingbird) to develop, prototype, and eventually produce next-generation cryo-EM instrumentation to rival the atomic resolution of X-ray crystallography.

As these initial three activities mature over the long term and begin enabling capabilities for the user program, our efforts will focus on the remaining research areas described in Sections 3.2.4–3.2.8 (Figure 5).

3.2.4 Chemical Biology Institute Research Area

As the HTP Omics and Protein Function, Bio-Atomic Imaging, and Single-Cell Biology Research Areas mature, there will be a need to more directly interrogate prioritized proteins and identify new protein functions in a manner readily automated to keep pace and capacity with the HTP omics and imaging workflows developed as part of DigiPhen. EMSL will establish a Chemical Biology Institute (CBI) that develops protocols and procedures for the design and synthesis of libraries of substrate-based chemical probes for an evolving broad spectrum of key biochemical functions relevant to BER science missions. The CBI's probes will be made available to users to characterize active site chemistry, visualize subcellular, cell, and community areas of activity, and isolate active proteins for subsequent identification and structural characterization using imaging via cryo-EM, tomography, mass spectrometry, and X-ray-based techniques. Importantly, the standardized protocols and procedures will serve as the basis for extending to and working with users to continually develop new functional probes and validated assays to increase the functional space that can be explored in annotating new and unknown proteins. The efforts in creating the CBI will be focused on research to develop standardized, reproducible, and scalable approaches to probe library creation and validation to enable a growing functional probe library to be made available to the user community. There is a direct linkage between the CBI efforts and output in enabling the creation of the CBI Satellite research area, which expands these efforts by involving the user community in the continuous evolution and growth of new functional assays.

Supporting IRPs: Biomolecular Pathways, Structural Biology, Systems Modeling and Data Sciences

Major External Engagements

- **BioLog.** *Anticipated:* Biolog is a world leader in cell-based phenotypic testing technologies and assays. Their Phenotype MicroArray technology, for example, enables researchers to evaluate nearly 2,000 phenotypes of a microbial cell in a single experiment. EMSL plans to engage BioLog to work toward an integrated workflow for identification of proteins of unknown function and their role in determining phenotype.
- **Charles River.** *Anticipated:* Charles River has developed AtomNet, which is a patented AI platform created by Atomwise for the prediction of small-molecule binding to protein targets. Working with Charles River would provide a complementary computational approach to identification of protein function by predicting molecular structures that bind a protein of known structure but unknown function. The approach would be leveraged to create chemical probes to isolate and identify proteins in the same functional class.

3.2.5 Visual Proteomics Research Area

The ability to perform controlled engineering of biological organisms relevant to the BER mission (plant, microbes, and fungi, among others) necessitates a fundamental understanding of how spatial and temporal changes in the proteome affect the linked metabolic pathways associated with the proteins. Spatially resolved information not available from conventional approaches that average data across populations of cells will advance our understanding of how the insertion of engineered pathways contributes to cellular phenotypes in host and model organisms. In this research area, protein structures and functions will be combined with visual spatial information to develop space-filling models of organelles, protein complexes, metabolons, and eventually whole cells with the regulatory structures that control important metabolic pathways. Visualization of proteins, local cell structures, and organelles will require detailed multi-omics, structural, and cellular (including intercellular) information generated from the HTP Omics and Protein Function, Single-Cell Biology, and Bio-Atomic Imaging Research Areas. For example, mapping post-

translational modifications in glycosylation and phosphorylation to assemble 4-D representation of the proteomes will provide a deeper understanding of regulation of metabolic pathways while contributing to the Visualizing Metabolic Pathways Research Area.

Supporting IRPs: Structural biology, Cell Signaling and Communication, Biomolecular Pathways, Plant and Ecosystem Phenotyping, and Systems Modeling and Data Sciences

Major External Engagements

- **Thermo Fisher Scientific** (TFS). *Ongoing:* TFS is a world leader in cryo-EM and advanced mass spectrometry for biological applications. EMSL and TFS have commenced discussions about a partnership (EMSL Partner Program) to combine powerful mass spectrometry chemical composition elucidation abilities with structural biology insights attainable by cryo-EM (single particle electron microscopy or diffraction and tomography) and multimodal imaging approaches to realize EMSL's visual proteomics goals. EMSL and TFS agreed that the initial focus of this partnership will be on furthering mass spectrometry-based soft-landing approaches to enable improved sample preparation for cryo-EM imaging. *Anticipated:* Future discussions with TFS will focus on better integration of the larger visual proteomics workflow and exploring EMSL's role as a beta-testing site for new cryo-TEM instrumentation.
- **Center for Structural Molecular Biology** (CSMB). *Ongoing:* CSMB supports the user access and science program of the Biological Small-Angle Neutron Scattering (Bio-SANS) instrument at the Spallation Neutron Source (SNS) at ORNL. Bio-SANS is dedicated to the analysis of the structure, function, and dynamics of complex biological systems. The CSMB also operates the Bio-Deuteration Laboratory for deuterium labeling of biological macromolecules. EMSL will pilot a FICUS partnership with CSMB through the FY 2022 FICUS call to provide access to complimentary tools that help researchers understand how macromolecular systems are formed and how they interact with other systems in living cells.

3.2.6 Visualizing Metabolic Pathways Research Area

As the DigiPhen platform tools produce new data streams for users, the assimilation and assembly of metabolome data from the HTP Omics and Protein Function, Single-Cell Biology, Visual Proteomics, and CBI Research Areas will become the next critical step to providing spatial information and visualization for understanding metabolic networks, the impacts of perturbations to those networks, and how they influence interactions with other cells and ultimately produce phenotypes. Current technology allows the user research community to visualize a limited set of proteins and transcripts within cells. This effort will develop and refine methods for simultaneous transcript and protein visualization (leveraging developments from the Visual Proteomics Research Area) and for tracking biological processes, from mRNA expression to protein translation and functional protein complexes. Over time, the number of biological processes being simultaneously tracked is anticipated to increase approximately 10× every 2–3 years as the development efforts continue, allowing BER users to study and interrogate complex and potentially multiple interacting metabolic pathways simultaneously.

Supporting IRPs: Cell Signaling and Communication, Biomolecular Pathways, and Plant and Ecosystem Phenotyping

Major External Engagements

- **University of California, Berkeley.** Ongoing: PNNL and EMSL are working with leading experts in the characterization and modification of photosynthetic algae for the synthesis of bioproducts to understand

the temporal and spatial molecular signatures underlying the synthetic pathways of molecules important for bioproducts. This engagement is part of the BER-funded project "Systems Analysis and Engineering of Biofuel Production in Chromochloris Zofingiensis, an Emerging Model Green Alga" lead by Krishna Niyogi (U.C. Berkeley).

- We anticipate additional engagements will emerge as this research area becomes our focus in the next several years.

3.2.7 CBI Satellite Researchers Research Area

The breadth of both model and "new" organisms (plant, microbial, fungal, archaea, etc.) being studied for bioproduct and sustainable energy applications continues to grow. Within each of these organisms are thousands and potentially tens of thousands of unknown and new proteins and metabolic process "targets" for chemical probe-based identification and characterization. The identification and characterization effort is beyond the capacity of singular users in the research community and beyond the current scope of EMSL. Empowering the user community to become more active developers and users of chemical probes will be required to develop and apply new functional assays at the scale necessary to support the entire breadth of organismal and phenotype research EMSL will develop through Strategic Science Objective 1 (DigiPhen). Toward that end, after successful creation of the CBI at EMSL, a CBI Satellite Researchers network will be created. Chemical probe libraries and experimental protocols for both development and application will be provided to collaborator labs, where biochemical, microscopic, and proteomic studies exponentially accelerate the discovery and localization of new protein functions. EMSL's Network for Execution of User Science (NEXUS) User Portal and Aurora data archive would be available to CBI satellite researchers to store the growing body of biological function data. CBI satellite researchers would grow capacity for functional analysis as EMSL enables the institutions to directly expand and create new libraries by providing standard protocols for chemical probe library synthesis. New probes developed by users and the broader research community will become part of a growing library of chemical probes managed by CBI, available to EMSL users, and incorporated into the DigiPhen platform as part of our Single-Cell Biology, HTP Omics and Protein Function, and Bio-Atomic Imaging research area analytical workflows. This would provide the capacity and pace for identification of function needed to match the pace of discovery in the genomics and functional omics fields, collectively accelerating assignment of phenotypes to genotypes for a much broader span of organisms than possible in any single facility. This effort would be synergistic with the Open-Source Data Analysis Software Suites and Metadata Capture and Findable, Accessible, Interoperable, and Reusable (FAIR) Data Management Research Areas within Strategic Science Objective 3 (see Sections 5.2.1 and 5.2.3).

Supporting IRPs: Biomolecular Pathways, Cell Signaling and Communications, Systems Modeling and Data Sciences

Major External Engagements

- We anticipate these user engagements will emerge during development of the CBI Institute and eventually mature to full partnerships (mix of EMSL User Program and Partner Program partnerships).

3.2.8 Protein Structure–Function Modeling Research Area

As the progress from DigiPhen advances beyond identifying protein function and begins to directly support engineering of desired properties into cells and communities, the ability to understand how the chemical and physical environment of active sites controls the mechanisms of protein function will be critical to optimizing and augmenting protein function. This research area effort will incorporate developments from the HTP Omics and Protein Function, Bio-Atomic Imaging, and CBI Research Areas with EMSL's NWChem and other computational tools used to simulate molecular dynamics to develop advanced, atomically precise simulations of fundamental chemistry and mechanisms for native and engineered enzyme active sites. This data–model integration will be crucial for facilitating large-scale, predictable, and controllable engineering of organism protein functions in support of BER objectives for bioproducts synthesis and ecosystem resilience. We therefore see efforts in the Open-Source Data Analysis Software Suites and the Data Integration Software Framework Research Areas (Sections 5.2.1 and 5.2.2) as highly synergistic and incorporated into activities and implementation for the Protein Structure–Function Modeling Research Area.

Supporting IRPs: Structural Biology, Systems Modeling and Data Sciences

Major External Engagements

- **NSF Center for Theoretical Biological Physics (CTBP) (Rice University, University of Houston, Northeastern University, and Baylor College of Medicine).** *Anticipated:* CTBP is the leading organization developing theories and software tools for advancing structural interpretation of molecular interactions, complexes, and machines by integrating -omics data into physical models from atomic, cellular, and system scales. EMSL will work with CTBP to investigate macromolecular dynamics of protein structures at mesoscopic time and length scales, filling a gap in the current capabilities of NWChem and KBase.
- **NIH Center for Macromolecular Modeling and Bioinformatics (University of Illinois at Urbana-Champaign)**. *Anticipated:* This center is the leading organization developing software for visualizing and simulating supramolecular systems in the living cell as well as the development of new algorithms and efficient computing tools for physical biology. The development and maintenance of widely distributed software tools, nanoscale molecular dynamics, and visual molecular dynamics are central to their work. EMSL will work with the Center for Macromolecular Modeling and Bioinformatics to bring advanced molecular modeling methods that bridge bio-atomic imaging and protein–structure–function modeling to the user community.
- **Sandia National Laboratories.** *Anticipated:* Sandia National Laboratories is the lead organization developing the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) molecular dynamics simulator supported by multiple DOE offices within SC. LAMMPS has potential for solid-state materials (metals, semiconductors), soft matter (biomolecules, polymers), and coarse-grained or mesoscopic systems. It can be used to model atoms or, more generically, as a parallel particle simulator at the atomic, meso, or continuum scales. EMSL will work with Sandia to achieve new developments in spatial modeling of protein complexes as potential targets for predicting their cellular functions.

4.0 ENVIRONMENTAL TRANSFORMATIONS AND INTERACTIONS SCIENCE AREA

The Environmental Transformations and Interactions (ETI) Science Area seeks to understand molecular transformation and transport across scales to predict ecosystem response. This science area focuses on the mechanistic and predictive understanding of environmental (physiochemical, hydrological, biogeochemical), microbial, plant, soil, and ecological processes in above- and belowground ecosystems, the atmosphere, and their interfaces. This understanding is obtained by investigating the cycling, transformation, and transport of critical biogeochemical elements, contaminants, atmospheric aerosols, specifically from biogenic and anthropogenic emissions to test, improve, and validate model predictions or identify sources of model uncertainty. Coupled experimental and modeling approaches will accelerate understanding of the mechanisms and dynamics of processes, their interdependencies, and feedbacks at molecular to ecosystem scales.

The ETI science area positions EMSL to lead the BER research community in addressing the coupled, exciting challenges of understanding molecular transformation and transport across scales, working directly with users to connect biotic and abiotic molecular processes to multiscale models forming a predictive view of emergent ecosystem properties. This science area is directly aligned with multiple BERAC Grand Challenges and BER goals (Table A.1). The establishment of the supporting Biogeochemical Transformations, Ecosystem Interfaces, and Plant and Ecosystem Phenotyping IRPs provide opportunities for users to pursue research in critical and emerging areas of importance for BER science processes controlling the flux and transformation of critical elements and molecules in ecosystems through a flexible and modular approach utilizing a single multidisciplinary IRP or a combination of IRPs. The ETI-focused IRPs allow users to pursue deep molecular analyses of organic matter across scales and systems.

EMSL's historical strength in the subsurface flow of nutrients and compounds within the Biogeochemical Transformations IRP facilitates a rich and detailed molecular understanding to enable extremely accurate predictive models for the sequestration, release, and transport of compounds within narrowly defined subsurface systems such as the vadose zone. Within the Ecosystem Interfaces IRP, molecular analyses can be expanded to encompass the interactions, transformations, and transport of compounds and biogenic emissions at ecosystem interfaces that drive land–atmosphere interactions. These interactions give rise to hot spot and hot moment occurrences that cause hyper-localized yet outsize proportions of chemical transport phenomena in these environments. Such interfacial interaction studies are critical to the user community's understanding of perturbations and effects across ecosystems. At the more extreme scale, the Plant and Ecosystem Phenotyping IRP provides users the ability to understand the effect of emergent ecosystem properties, such as plant metaphenomes, arising from the interactions of environmental stressors on transcriptional and translational activity of key biological functions encoded within plant genomes and metagenomes. A critical capability resident across the ETI IRPs is the unique approach of interrogating and exploring the rich, complex rhizosphere interactions that affect molecular transformation, transportation, and communication linking the subsurface, soil, microbial, plant, and ultimately land–atmosphere interactions responsible for the emergent properties observed at individual plant to ecosystem scales. Holistically, the IRPs present an opportunity for users to explore the breadth of the ETI scientific domain, from the molecular and observational to the continental and computational, building a comprehensive understanding of biotic and abiotic control of processes in soils, plants, microbial communities, and atmospheric elements. Sections 4.1 and 4.2 provide background on Strategic Science Objective 2, which emerged in response to EMSL's assessment of trends and drivers during our 2020 strategy workshops. Strategic Science Objective 2 provides scientific direction and focus for the research efforts in the ETI science area.

4.1 Background for Strategic Science Objective 2

There is a growing need for anchoring complex environmental and resource management decisions with robust predictive models of the future state of global and regional environments. To assure a secure and sustainable energy future even in the face of changing climatic conditions and related environmental impacts, BER and user community efforts to establish a predictive understanding of ecosystem responses to perturbations must accelerate. However, understanding and predicting climate-driven events, compounding disturbances, and potential future states of local and regional environments is now recognized to be highly dependent on advancing a next-generation understanding of coupled Earth system processes that span soil–water–atmosphere interfaces and their effects on the interconnected, dynamic, Earth–energy–human system (EESSD Strategic Plan, U.S. DOE 2018b). To address these challenges, it will be necessary for the BER research community and users to deepen and extend our understanding of natural and anthropogenic interactions and feedbacks alongside their associated uncertainties within atmospheric, terrestrial, watershed, and human systems.

Toward this end, BER has invested in a portfolio of observational and experimental research programs to unravel the complex processes and controls of the structure, function, feedbacks, and dynamics of ecosystems, spanning from the bedrock through the rhizosphere and vegetation to the land–atmosphere interface. The scope includes watersheds and coastal zones, terrestrial–aquatic interfaces, understudied ecosystems, and ecosystem interfaces that represent a significant knowledge gap in local and regional process models and predictive Earth system models that EMSL can help the user research community fill. These include augmenting poorly modeled Earth system and important regional-scale ecosystem phenomena with appropriately parameterized molecular and mechanistic understandings, such as redox reactions at terrestrial–aquatic ecosystem interfaces, hot spots and hot moments of biogeochemical activity, chronic or long-term perturbations or disturbances, and extreme events at larger spatial scales and multidecadal time scales. There is also a need to fully understand ecosystem dynamics through more

complete and deliberate data–model integration that incorporates molecular and mechanistic understanding. Greatly enhanced predictabilities of regional water cycles, coupled biogeochemical processes, land–atmosphere interactions, and interfaces with human systems and the built environment are urgently needed to reduce uncertainty in the response to short- and long-term perturbations. These coupled cycles are being increasingly linked to molecular-level understanding of hydrobiogeochemical processes that control the flux of materials in the environment and how these processes affect ecosystem function. As a result, the BER research community requires access to molecular and mechanistic measurement capabilities to move beyond correlative or idealized models of complex dynamic biological systems and toward developing multiscale dynamic models that represent biological systems with greater fidelity.

EMSL is exceptionally well positioned to meet the need to incorporate complex multimodal datasets from soil, water, plant, and microbial systems by virtue of its extensive experience in user science in these areas. In addition, EMSL's broad range of interconnected infrastructure capabilities and tools support integration and management of models, experiments, and observations across a hierarchy of spatial and temporal scales and complexity. Each successive advancement in our fundamental understanding drives science and scientific research to become more multi- and interdisciplinary, requiring more fluid teaming and greater open access to vast data streams. A direct corollary to the acceleration in integrated multidisciplinary science is an increasing need for regular, strategic adoption and implementation of advanced technologies and approaches that support science to include modeling and simulation of current and future states as well as analytical, experimental, and field-deployable observational capabilities (Scientific User Research Facilities and Biological and Environmental Research, BERAC 2018). These are needs EMSL is well suited to develop and deliver to the BER user community.

To address the needs identified in the ETI science area, EMSL's second 10-year Strategic Science Objective (Section 4.2) is to create a national network of remote sensing and observational capabilities coupled with standardized sampling, high-throughput molecular analysis capabilities, data analytics, and spatial and temporal modeling and simulation.

4.2 Strategic Science Objective 2: Establish MONet, a National Molecular Observations Network for Modeling from Elements to Ecosystems

EMSL will lead the effort to develop a national network of environmental sampling and sensing sites and fieldable sampling tools and methods that collectively produce molecular-level information on the composition and structure of soil, water, biogenic and more complex anthropogenic emissions, and resident microbial communities, as well as the site-specific metadata required to improve multiscale models of Earth systems. The network will work with stewards of selected natural and managed systems consisting of watershed, coastal, and continental observational and experimental networks as well as necessary atmospheric measurement facilities to collect samples and send data from deployed sensors directly to EMSL for the user community. Automated laboratory molecular analyses are supplemented with an extensive network of remote sensing applications at field sites. A national regional-scale-resolution model of key ecosystem processes and feedbacks between ecosystems will utilize the MONet data streams (biogeochemical, hydrologic, and microbial processes) to parameterize key variables for process models and for an eventual coupling to larger-scale regional, landscape, and DOE Earth system models. These more advanced and accurate models will improve prediction of ecosystem function and response, supporting the long-term goal of scientifically informed decision-making regarding energy and water security and sustainability for the United States.

To meet the bold objective to establish a national molecular observations network (MONet) and bring the benefits to multiple stakeholders and the research community, EMSL will focus efforts over the next decade

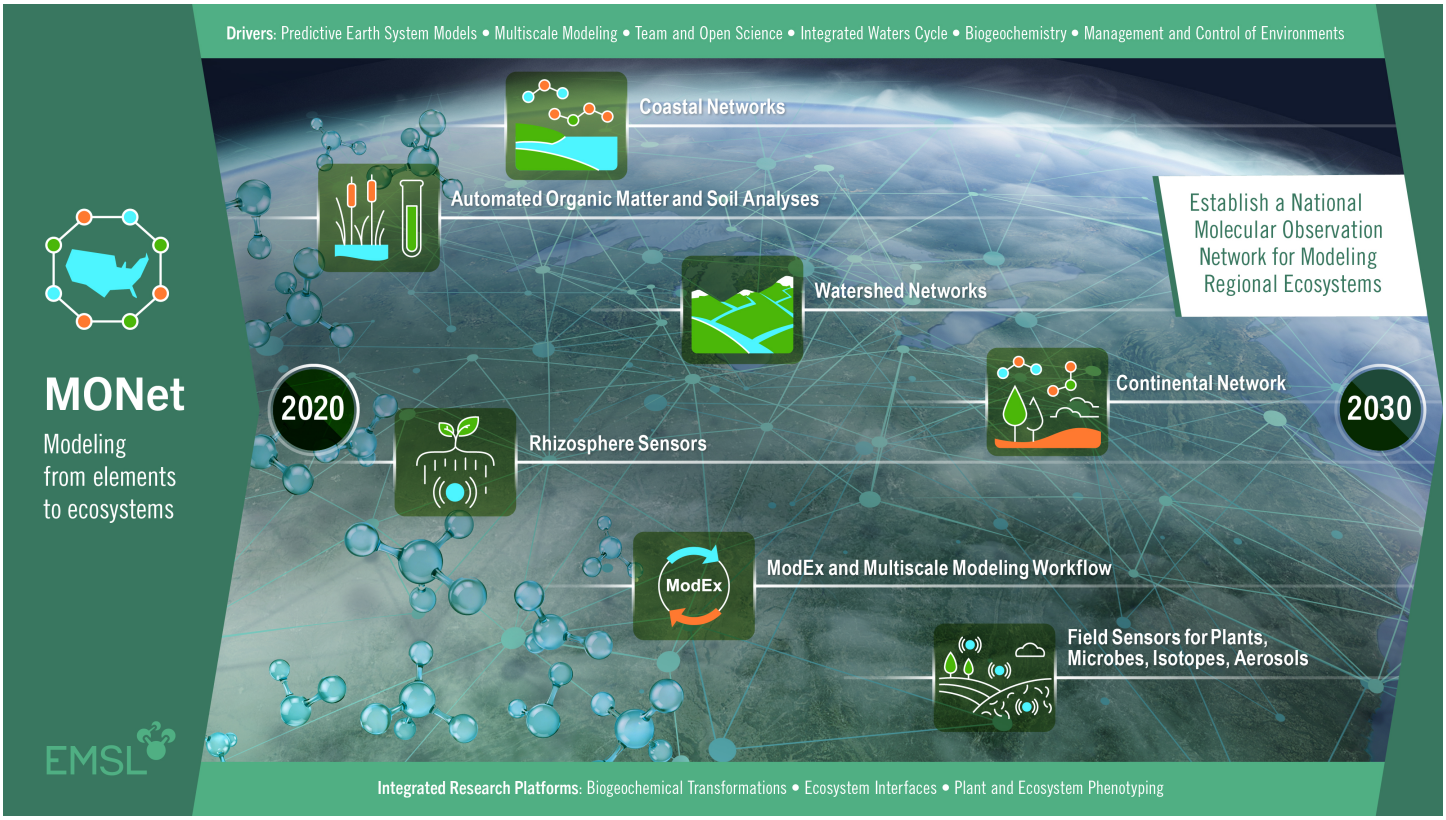


Figure 6. Overview, timeline, and research areas supporting Strategic Science Objective 2. Establishing a national molecular observations network will provide high-quality, interoperable, and integrated molecular-level information on key ecosystem processes at local and continental scales to enable predictive modeling of ecosystem function critical for energy and environmental security. Each research area is placed on the 2020–2030 timeline to show where we anticipate the most activity, though we expect work to begin before and to continue after, as required.

in seven research areas (Figure 6). During the initial 2–5 years, our efforts will be concentrated in four of those research areas, each with specific activities and supporting programs, projects, and investments: Automated Organic Matter and Soil Analysis; Rhizosphere Sensors; Model–Experiment Integration (ModEX) and Multiscale Modeling; and MONet Field Sites (Coastal, Watershed, Continental Networks). In each case, the research areas build the scientific foundation for MONet and the user Community makes use of the emerging science and technology for scientific inquiry that directly supports BER missions and goals.

4.2.1 Automated Organic Matter and Soil Analysis Research Area

The Automated Organic Matter and Soils Analysis Research Area will pioneer new field-ready sampling devices and automated "smart" variable workflows for physical and hydrogeochemical analysis of point-in-time soil and multimodal chemical characterization of organic matter samples from a set of diverse sampling sites across the United States managed by a broad group of partners and collaborators. These workflows will increase the pace and capacity of field sampling as well as laboratory analyses, improve sample and data quality, reduce costs, and deliver data to public-facing data repositories (e.g., through NEXUS) for MONet collaborators and the broader national research community. Increasing sample throughput from the large network of collaborator and partner sampling sites necessitates establishing workflows to collect, receive, process, and analyze core samples at a throughput several orders of magnitude greater than what current manual methods offer. The challenge extends beyond just throughput to include the need for development and inclusion of automated multimodal analyses (i.e., FTCIR-MS, X-ray computed tomography [XCT], and NMR, among others) in these future workflows. Critical activities include the creation of sample containers that stabilize soil organic matter and natural organic matter chemistries at the point and time of field sampling for shipment to EMSL and integration into the dissolved organic matter workflow. There will be many stages of success for this pioneering effort, but the availability of an extensible and modular platform with the annual capacity scale to meet the demands of MONet collaborators (initially, 1,000 core samples and scaling up with demand) is our target metric. Collaborations with EMSL peers to advance four research areas under Strategic Science Objective 3 (Modeling and Data Sciences Center)—Open-Source Data Analysis Software Suites, Data Integration Software Framework, Metadata Capture and FAIR Data Management, and ML and AI for Automation (see Sections 5.2.1, 5.2.2, 5.2.3, and 5.2.5)—reflect an important strategic alignment of capabilities and effort to deliver a soil analysis capability to the research community.

Supporting IRPs: Biogeochemical Transformations

Major External Engagements

- **PNNL's River Corridors SFA subtask Worldwide Hydrobiogeochemical Observation Network for Dynamic River Systems (WHONDRS) and COMPASS-Exchange.** *Ongoing:* These BER-supported research community efforts are pioneering widely distributed environmental sampling with easy-to-use sample kits and standardized metadata templates. EMSL will expand our work with the WHONDRS team for both access to samples and sampling networks, but also to advance new sampling methods, pioneer new field-deployable sample preparation and analysis technologies, and to expand from organic matter analyses to the more comprehensive soil analyses necessary to calibrate and parameterize crucial process, local, regional, landscape and Earth system models.
- **Advanced Photon Source (APS).** *Anticipated:* EMSL will form partnerships through the FICUS framework to facilitate the research communities' access to capabilities that are complimentary to measurements of organic matter made at EMSL. Of particular interest are techniques developed in Argonne's Subsurface Biogeochemical Research program that provide highly resolved measurements of chemical speciation, chemical imaging, and pore to core scale imaging resolution over large sample sizes.

Recent and Near-Term Supporting Activities

- **Improve extensibility of organic matter mass spectrometry measurements.** *Ongoing:* EMSL Intramural S&T Research investments in (1) chromatographic approaches for functional natural organic matter characterization and (2) understanding biases in natural organic matter FTICR-MS to more accurately model complex environmental systems.
- **Methods for tracing organic matter chemical fate.** *Ongoing:* EMSL Intramural S&T Research investment in position-specific stable isotope analysis for elucidating the provenance and fate of organic matter.
- **Design, purchase, and implement automated workflows for dissolved organic matter.** *Ongoing:* EMSL plans to develop working relationships with leaders in automation and laboratory information management systems to incorporate the recently acquired 7T FTICR-MS into an automated dissolved organic matter workflow; planned capital purchase of additional automation and analytical instruments for fully automated dissolved organic matter workflow.

4.2.2 Rhizosphere Sensors Research Area

Development and deployment of field-based sensors for continuous observations in the field is a critical complement to the point-in-time sampling and analyses that the Automated Organic Matter and Soil Analyses Research Area establishes. The Rhizosphere Sensors Research Area will focus on the critical interactions occurring between the soil microbiome and plant root structures, providing rich, real-time data streams that represent normal and perturbed states and regions associated with hot spot and hot moment occurrences of biogeochemical activity within the MONet network. The ability to non-destructively and continuously detect and quantify signaling, chemical exchange, and nutrient acquisition in the rhizosphere in situ would rapidly advance user understanding of these mechanisms and processes and inform fundamental models of C cycling, microbial dynamics and function, and inter-organismal interactions. This research area will work to design, test, validate, and deploy programmable molecular sensors of key microbial functions, root exudates, and nutrient acquisition enzymes. Sensor design and development efforts here would leverage synergies and developments from the HTP Omics and Protein Function, Single-Cell Biology, and CBI Research Areas within Strategic Science Objective 1 (Sections 3.2.1, 3.2.2, 3.2.4, and 3.2.7) to inform and guide next-generation sensor approaches and applications. Focus in this area would be directed at developing lab-based rhizosphere sensors to augment lab-based analyses as well as to guide development of long-term field-deployed sensors as part of the Field Sensors for Plants, Microbes, and Aerosols Research Area. Early success will be the development of sensors based on complex resistivity tomography to image and measure microbial-influenced chemical fluxes at the root–soil interface and sensors for the de novo design and assembly of microbial- or plant-based biosensors to detect the exchange of metabolites (amino acids, vitamins, and sugars) and signaling molecules (auxins and quorum-sensing molecules) across the root–microbe–soil interface in response to drought and nutrient limitation.

Supporting IRPs: Plant and Ecosystem Phenotyping, Biogeochemical Transformations, Ecosystem Interfaces

Major External Engagements

- **Rice University.** *Anticipated:* Lead researchers at Rice University have combined chemical engineering and synthetic biology disciplines to establish workflows for transforming native microbial species into reporters for specific biological functions. As part of the BER-funded Twin Ecosystems project, EMSL anticipates working with Rice University researchers to create microbial sensors for signaling molecules

and metabolites to detect metabolite exchange between plants and the associated rhizosphere microbial community.

- **Twin Ecosystems Project.** *Ongoing:* Through the BER-funded Twin Ecosystems project, EMSL would deploy a suite of environmental sensors in a managed field system at the Washington State University agriculture station in Prosser, Washington. The sensors will provide near continuous measurements of climatic and soil conditions and periodic measurement plant status using multispectral imaging and rhizotrons.
- **SBIR/STTR Programs.** *Anticipated:* When appropriately focused opportunities arise, EMSL will develop partnerships with small businesses through the DOE-SC SBIR/STTR program to co-develop field-deployable sensors of key biological activities. EMSL anticipates leveraging these partnerships and EMSL test beds/demonstration projects to create a prototype to commercial product pipeline for sensors.

Recent and Near-Term Supporting Activities

- **Platform to study reduced complexity microbial communities.**
  Anticipated: Trial Ecosystems for the Advancement of Microbiome Science (TEAMS) program at LBNL. EMSL anticipates engaging TEAMS to add a modified EcoFAB experimental platform for EMSL users to explore the principles of microbial community assembly and structure, understand the functions of genes, microbes, and metabolomes, and predict microbiome health and trajectory. Modified EcoFAB platforms are designed to couple to EMSL omics and imaging capabilities.
- **Discover targets for molecular sensors.**
  Ongoing: EMSL Intramural S&T Research investments in tracing rhizosphere carbon exchange processes and nutrient interactions; leveraging a PNNL LDRD initiative, Mathematics of Artificial Reasoning; capital purchase of an isotope ratio mass spectrometer and a nanoscale secondary ion mass spectrometry (NanoSIMS) O source.
- **Establish sensor mechanisms and platforms.**
  Ongoing: EMSL Intramural S&T Research investment into imaging and analysis of root system architecture with an automated noninvasive phenotyping system. *Anticipated:* As part of the Operations for Capacity and Pace Objective (see Section 6.2), EMSL will create a "maker space" to facilitate design, development, and testing of biological sensors.
- **Develop next-generation sensors of biological activity.**
  Ongoing: The Sensors Working Group and Biosensing Thrust in the PNNL Predictive Phenomics Strategy are focused on defining the molecular signatures of critical biological functions and responses to perturbations in situ to develop novel sensing approaches and technologies tuned to molecular signatures and functions of interest; BER-funded Secure BioDesign Science Focus Area (SFA); *Anticipated:* Engagements with academic and industry researchers; Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) opportunities.

4.2.3 Model–Experiment (ModEx) Integration and Multiscale Modeling Research Area

An explicit opportunity for the MONet objective is to facilitate multiscale modeling by incorporating experimentally validated data to greatly enhance the accuracy and predictive power of crucial process, local, regional, landscape, and Earth system models. This research area will develop methods to integrate experimental measurements (e.g., soils, rhizosphere, and biologic and anthropogenic emissions) into computational and modeling frameworks either directly for scale-appropriate models or through parameterizations. These will lead to model predictions that can be tested by subsequent rounds of

experimentation or by executing programmable remote sampling and measurements within MONet, producing greatly accelerated and efficient ModEx iteration cycles. Models will also pass process parameterizations across spatial and temporal scales. This effort is highly synergistic with the Visualizing Metabolic Pathways Research Area in Strategic Science Objective 1 (see Section 3.2.6), the Automated Organic Matter and Soils Analysis and Sensor Research Areas (4.2.1, 4.2.2, and 4.2.6), and the Data Integration Software Management and Metadata Capture and FAIR Data Management Research Areas in Strategic Science Objective 3 (see Section 5.2.2 and 5.2.3). Although we expect model evolution to be continuous, this research area will near completion with the demonstration and availability of 2–3 systems composed of coupled experimental and ecosystem models and with supporting innovations in multiscale modeling and model–data integration that advance real-time coupling of models and experimental systems.

Supporting IRPs: Biogeochemical Transformations, Ecosystem Interfaces, Systems Modeling and Data Sciences

Major External Engagements

- **JGI and KBase**. *Ongoing:* BER's Systems Biology Knowledgebase (KBase) hosts user community-generated narratives, or workflows, that integrate EMSL molecular mass spectrometry datatypes into genome-informed metabolic network models. EMSL and KBase coordinate development and web hosting of tutorials teaching materials that highlight these workflows. EMSL will continue to work synergistically with JGI to assure EMSL users can leverage the advanced analytical instruments and evolving technologies at both facilities that allow rapid integration and interpretation of genome sequence data and molecular data related to biological function.
- **BER's IDEAS-Watersheds (IDEAS-W) Project Research Community.** *Anticipated:* EMSL expects to engage with BER's domain science modelers from the IDEAS-W project to utilize the IDEAS-Watershed Alquimia interface to connect pore-scale models to the PFLOTRAN and CRUNCHFLOW geochemical transport models. Working with the IDEAS-W research community would advance development of the next generation of pore-scale reactive transport models and their integration into the KBase toolset.
- **ExaSheds.** *Anticipated:* EMSL expects to support BER domain science modelers in application of Exasheds-developed, AI-enabled, high-performance watershed models on EMSL computer systems and (through the CD-MII project and other avenues) to develop new multiscale modeling workflows that will facilitate use of EMSL-generated data in those models.

Recent and Near-Term Activities

- **Build datasets for ModEx and multiscale modeling.**
  Ongoing: EMSL Intramural S&T Research investment in the EMSL "1,000 Soils" project to conduct molecular and structural characterization on a large cross section of soils across the continental US.
- **Conceptual model frameworks across scales.**
  Ongoing: BER-funded CD-MII project "Model-Driven Datasets for Plant-Soil-Microbe Interactions using a RhizoCell Experimental System" (development of a conceptual multiscale modeling framework and event notification tools).
- **Exercise ModEx and multiscale modeling paradigm.**
  Ongoing: EMSL Intramural S&T Research investments in studying how mineralogy exerts control on organic molecule zonal structuring and reordering; Scientific engagement with the Montpellier Ecotron Facility within the Centre national de la recherche scientifique studying the effects of earthworms on ecosystem multifunctionality.

- **Combine biogeochemical data with numerical models of reaction networks.** *Ongoing:* EMSL Intramural S&T Research investment into simulating hydrobiogeochemical processes and states in the rhizosphere and adjacent soil locations.

4.2.4 MONet Field Sites Research Area

There are three research areas in Strategic Science Objective 2 that together contribute to a complete set of national sampling and observational sites for ecosystem modeling. Each of these three research areas focuses on a unique, mechanistically important geographical scale, from target regional areas (fresh and saltwater coastal networks) to entire watersheds (watershed networks) and finally to continental-scale networks (continental network). The sites are interdependent and connected by processes that cross scales and by multiscale Earth system models that utilize information at all three scales. Each of these three research areas share similar goals and approaches, including engagements with projects such as Coastal Observations, Mechanisms, and Predictions Across Systems and Scales (COMPASS)-Exchange, PNNL's River Corridors SFA subtask for WHONDRS, and new community science projects.

These external engagements will be leveraged to collect soil, organic matter, microbe, sediment, and related samples from distinct sites and to capture data from existing repositories to build a spatially distributed network of sampling at single points in time with rich metadata collection augmented by an array of advanced field-deployed sensors for continuous observations (MONet). These networks and engagements (users, partners, and collaborators) will focus on establishing unique sites for sampling as well as continuous sensor deployment, allowing the capture of chronic and episodic disturbances across diverse ecosystems. The samples support analyses to inform key biogeochemical, hydrological, ecological, and microbial processes for domain science modelers who plan to use EMSL's computational resources to model and simulate processes and systems across multiple spatial and temporal scales. There is a strong synergy between these networks and the ModEx and Multiscale Modeling Research Area (Section 4.2.3), which plays a key role in the translation of the data from these networks into regional-scale, fine-resolution process models. The success of this effort is measured by established collaborations and programs that deliver thousands (initially) to tens of thousands (eventually) of samples and field sensor datasets that produce high-quality data fit for modeling at all three scales. A comprehensive, integrated network of sites across the United States that is supported by a variety of funding sources is a longer-term view of success.

Supporting IRPs: Systems Modeling and Data Sciences, Biogeochemical Transformations, Ecosystem Interfaces

Major External Engagements

- **PNNL's River Corridors SFA subtask WHONDRS and COMPASS-Exchange.** *Ongoing:* These BER-supported research community efforts are pioneering widely distributed environmental sampling using easy-to-use sample kits and standardized metadata templates. Working with WHONDRS and COMPASS-Exchange will bring together EMSL's extensive analytical and soil analysis capabilities with these leading sampling networks to advance development of MONet.
- **Sampling Partnerships for Continental and Terrestrial Sampling Sites.** PNNL will extend sampling sites from watersheds and coastal regions to the terrestrial or continental scale by working with existing managed research sites to include those in the National Ecological Observatory Network (NEON), NSF-funded Critical Zone Observatory (CZO), USDA sites, AmeriFlux, and Next-Generation Ecosystem Experiments (NGEEs).

Recent and Near-Term Supporting Activities

- **Source samples for analyses.**
  Ongoing: PNNL's River Corridors SFA subtask WHONDRS; the EMSL 1000 Soils Intramural S&T Research investment; *Anticipated:* expanded engagements with BER-funded SFAs, AmeriFlux, and NGEEs; the NSF-funded CZO and NEON; and USDA-managed agricultural research sites.
- **Accelerate throughput of sample analyses.**
  Ongoing: Synergistic activities in the dissolved organic matter automation workflow associated with the Automation of Organic Matter and Soil Analysis Research Area.
- **Couple molecular measurements with remote spectral signatures.**
  Ongoing: FICUS Partnership with the NEON Biorepository; *Anticipated:* deployed assets program under which NEON staff collect samples and hyperspectral images or maintain instrumentation for an investigator at NEON sites.

Activities 3–5 Years Out

As the next generation of biological function sensors is developed, we will form new and extend existing relationships with other researchers and research organizations to place these sensors in distributed observational networks (e.g., the Long-Term Ecological Research [LTER] network and NEON) by leveraging site knowledge, ready-built infrastructure, and captured metadata to provide context for these measurements. We will also seek to utilize NASA remote imaging data. NASA's Earth-focused low Earth orbit satellites (e.g., SMAP) and space-station-based sensors (e.g., ECOSTRESS, GEDI) capture ecosystem function and status globally while being able to detect factors like soil moisture that are signs of environmental stress as well as measures of ecosystem function like photosynthetic efficiency and nutrient limitation. These data streams can be used to first validate next-generation sensors and then to extend inference of ecosystem status to ecosystems not directly sensed by the existing NEON, LTER, and CZO sites.

Over the longer term, our efforts will begin focusing on the following two additional research areas.

4.2.5 Automated Organic Matter and Soil Analysis/MONet Networks Research Area

As the MONet networks expand and the productivity of the organic matter and soil analyses workflows grows in parallel, we envision extending the measurement modalities deployed to analyze samples. This may involve creating low-cost sample kits that stabilize soil organic matter and natural organic matter chemistries at the point and time of field sampling for shipment to EMSL, including orthogonal measurements of soil matrix and soil properties such as pore structure, into the soil analysis workflows.

Supporting IRPs: Systems Modeling and Data Sciences, Biogeochemical Transformations, Ecosystem Interfaces, Biomolecular Pathways

Major External Engagements

- **PNNL's River Corridors SFA WHONDRS and COMPASS-Exchange.** *Ongoing:* These BER-supported research community efforts are pioneering widely distributed environmental sampling providing easy-to-use sample kits and standardized metadata templates. Working with the River Corridors SFA WHONDRS and COMPASS-Exchange will bring together EMSL's extensive analytical and soil analysis capabilities with these leading sampling networks to advance development of MONet.

- **Remote Imaging Services.** *Anticipated:* There are multiple international low Earth orbit (LEO) satellite and space station imaging platforms that provide ecosystem assessments though hyperspectral imaging. Working with NASA/JPL scientists, EMSL will connect sampling and imaging networks to correlate molecular measurements of ecosystem status with hyperspectral images collected at ecological network sites (NEON, LTER, CZO, etc.) and extend beyond these using LEO imaging platforms.

4.2.6 Field Sensors for Plants, Microbes, and Aerosols Research Area

Beyond efforts to observe key biological interactions within the rhizosphere in a laboratory setting, continuous observations in the field would enable a much more detailed understanding of the spatial and temporal heterogeneity of environmental processes as well as the dynamic response to perturbations. The ability to observe and measure biological and physiochemical activity across a range of systems is essential for initial building and then refining of multiscale models of environmental processes. The Field Sensors for Plants, Microbes, and Aerosols Research Area will focus on development and deployment of robust field-ready sensors using advances in microfluidics and novel sensors of biological function, including those leveraged from the CBI and HTP Omics and Protein Function Research Areas within Strategic Science Objective 1 (see Sections 3.2.1 and 3.2.4) and the Rhizosphere Sensors Research Area (Section 4.2.2). As the sensor network grows across systems and geography, a near-real-time observational network can be created that provides streaming data on both episodic and chronic disturbances. These data can be augmented with detailed multimodal analyses performed by the Automated Organic Matter and Soils Analysis Research Area and will leverage efforts within the Data Integration Software Framework and Metadata Capture and FAIR Data Management Research Areas (see Sections 5.2.2 and 5.2.3).

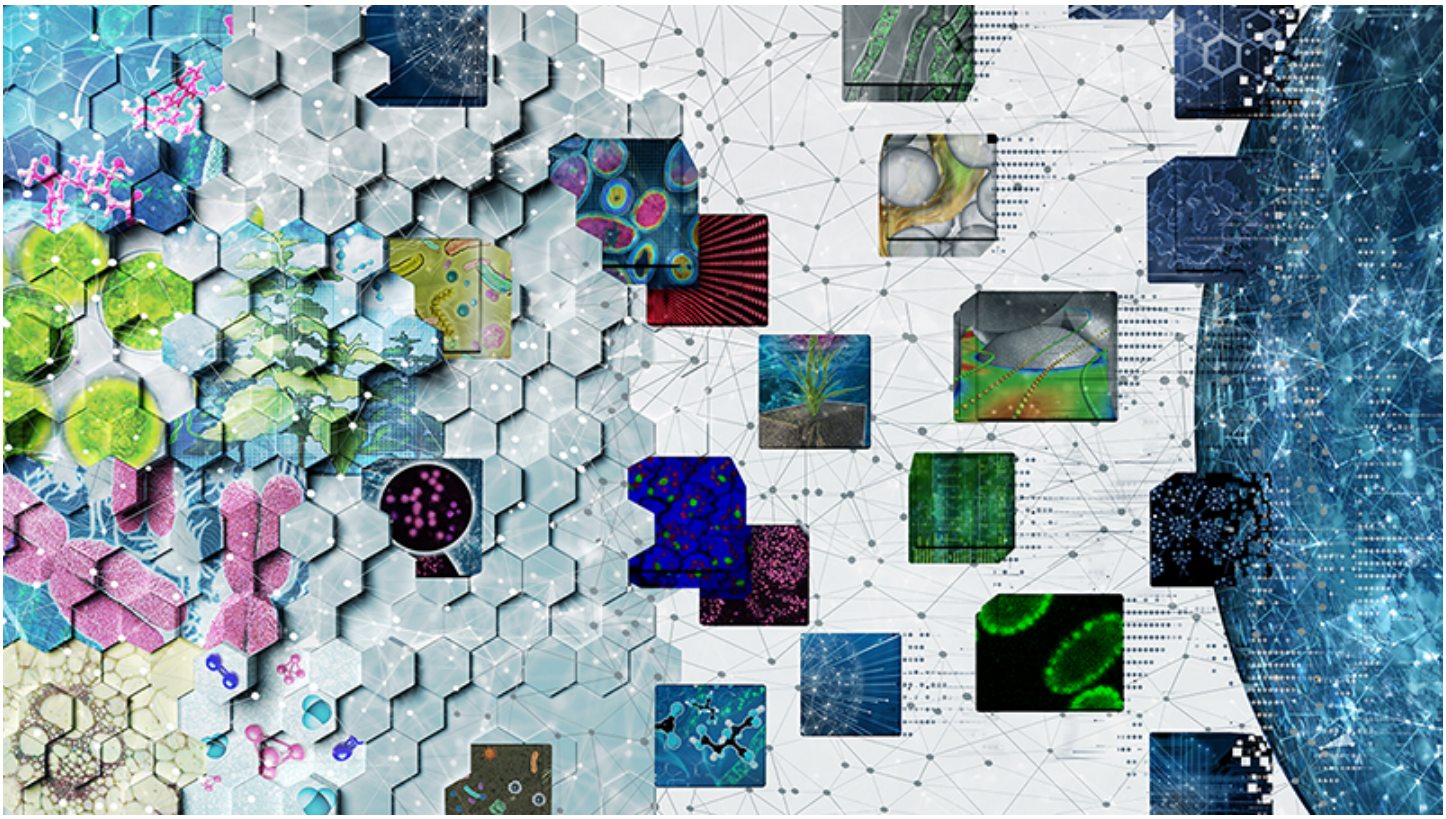
Supporting IRPs: Plant and Ecosystem Phenotyping, Ecosystem Interfaces, Cell Signaling and Communication, and Systems Modeling and Data Science

Major External Engagements

- **Ecological Research Networks.** *Anticipated:* There are numerous existing networks (e.g., AmeriFlux, NEON, LTER, and CZOs) that have the infrastructure to maintain and collect relevant ecological data over extended time periods. As a critical component of MONet, EMSL anticipates co-locating field sensors at these sites to enable collection of longitudinal molecular data that can be associated with co-located long-term ecological monitoring data.
- **ARM.** *Ongoing:* ARM provides the infrastructure for atmospheric field campaigns. EMSL will continue to partner with ARM for the deployment of advanced aerosol capture and sensor platforms such as the size- and time-resolved automated aerosol collector (STAC).

Recent and Near-Term Supporting Activities

- **Pioneer development of new sensing mechanisms, platforms, and sensors.** *Ongoing:* PNNL sensors workshops; continued developments in the Biosensing Thrust within the PNNL Predictive Phenomics Strategy focused on defining the molecular signatures of critical biological functions and responses to perturbations in situ to develop novel sensing approaches and technologies tuned to molecular signatures and functions of interest; EMSL Integration Meeting on Environmental Sensors.
- **Build and deploy suites of atmospheric aerosol sensors and samplers.**
  Ongoing: EMSL–ARM partner proposal and EMSL Intramural S&T Research investment in the development of the STAC for unmanned aerial systems; the EMSL–ARM FICUS program call for users to access EMSL's STAC for ARM field campaigns and to utilize EMSL analytical capabilities.



5.0 COMPUTING, ANALYTICS, AND MODELING SCIENCE AREA

The Computing, Analytics, and Modeling (CAM) Science Area brings advanced data analytics, visualization, and computational modeling and simulation to bear on increasingly complex multimodal experimental data to develop a predictive understanding of biological and environmental systems. This cohesive approach to integrating experimental and computational methods advances predictive approaches to biodesign for biofuel and bioproduct production and accelerates research to understand the molecular mechanisms underlying biological and hydrobiogeochemical processes controlling the cycling, flux, and movement of materials (e.g., carbon, nutrients, and contaminants) in the environment.

Because CAM is a new addition to EMSL science areas, the Systems Modeling and Data Sciences IRP currently provides the primary capabilities and science leadership supporting CAM. As the breadth and utilization of CAM science capabilities expand, additional CAM-focused IRPs will be formed as needed to meet the needs of EMSL users. EMSL's CAM Science Area positions us to lead the BER research community in addressing data analytics, modeling, and simulation challenges and user needs by working directly with users to optimally incorporate multimodal data streams into models of coupled physical and biological processes that span a broad range of temporal and spatial scales. The CAM science area focuses on generating knowledge from data, thereby directly supporting BERAC Grand Challenges (e.g., 2.3, 6.4 and 8.5) and several needs outlined in the BERAC User Facilities Report (Table A.1) as well as responding to the recent call (U.S. DOE 2021b) to address the daunting challenge of analysis of "vast quantities of disparate data" by developing wholly new approaches. The creation of the Systems Modeling and Data Sciences IRP places emphasis on advancing the computational tools and methods necessary to manage and process the increasingly complex, high-throughput, and multimodal data and images generated by EMSL users. This IRP also provides opportunities for users to incorporate these observational data into predictive

models to make inferences on the mechanisms and processes that govern system behavior under various environmental stresses. Together with the IRPs supporting the ETI and FSB science areas, the Systems Modeling and Data Sciences IRP creates a rare opportunity for users to access and utilize an exceptionally wide breadth of measurement, analytical, and modeling capabilities in a single institution. EMSL is thus exceptionally well positioned to lead the development of the next generation of computational approaches that will aid interpretation of observations, convert experimental data into scientific understanding, and fuel multiscale models to predict system behavior. Sections 5.1 and 5.2 provide background on Strategic Science Objective 3 that emerged in response to EMSL's assessment of trends and drivers during our 2020 strategy workshops. Strategic Science Objective 3 provides scientific direction and focus for the research efforts in the MDS and CAM science areas.

5.1 Background for Strategic Science Objective 3

BER has invested in computational infrastructure for data analysis and visualization (KBase 2021), modeling and simulation, and data management and archiving (NMDC 2021; ESS-DIVE 2021) while also investing in the biological and environmental research programs that leverage this infrastructure. However, there is a need to provide computation and data analytics capabilities in a more holistic and interconnected manner to facilitate seamless interaction among BER researchers and across BER facilities' capabilities. Science, especially BER science, is accelerating and inherently interdisciplinary, requiring more fluid teaming and greater open access to data streams.

Computational science and data analytics are now the dominant mechanisms for converting the exponentially growing amount of massive raw experimental and observational data into scientific understanding and knowledge. To create a secure bioeconomy and achieve a predictive understanding of the living Earth system, it is imperative that Earth system, ecosystem, local, and process models move beyond correlative or idealized models of complex dynamic biological systems. The development of multiscale, dynamic models that represent biological processes and their interactions with physical and chemical components of their surrounding environment with greater fidelity is crucial. The increased fidelity of these dynamic models must incorporate the fundamental understanding of the genomic and regulatory principles of key biological functions that plants and microbes perform in the presence of their surrounding environment to enable simulations of their current state and the ability to change parameters to project potential future state(s). To do so necessitates ingesting complex multimodal datasets from soil, water, plant, bioaerosol, and microbial systems to develop the next generation of methods, software, and visualization tools that will accelerate the iteration between modeling and experiments (the ModEx approach), thereby speeding the interpretation of cellular processes, community interactions with their environment, dynamical aspects of ecosystems and watersheds, and natural–human system interactions and assuring environmental security.

Software, and visualization tools are core capabilities that EMSL users and other BER researchers are increasingly dependent on but that rarely exist in the complete form that is needed. Moreover, the continuing expansion of high-throughput experimentation and analysis is creating a bottleneck in data analyses that limits discovery and understanding. Parallel advancements in multiscale and multi-process modeling, AI, and ML will be necessary to keep pace, improve throughput of data acquisition and analysis, enhance process-rich models of ecosystems to speed scientific discovery and hypothesis generation, and accelerate creation of new frontiers in technology (American Artificial Intelligence Initiative, OSTP 2020). Users will need these advancements and access to mid-range HPC computational resources tailored to these research needs if they are to take full advantage of the growing capabilities of EMSL and other BER user facilities to address BER priority research objectives through analysis, integration, and modeling of these growing data streams. Meeting these future needs will necessitate a broad range of interconnected

infrastructure capabilities and tools within and among DOE user facilities that support integration and management of models, experiments, and observations across a hierarchy of scales and complexity, with EMSL's collective capabilities forming a strong starting point. The inevitable drive to characterize and more completely model complex systems (e.g., organisms and Earth systems) demands a requirement across scientific domains to produce high-quality data through standardization, ontologies, metadata tagging, and adhering to FAIR principles. As a leader in the production, storage, management, integration, analysis, modeling, and accessing of the full span of data types and scales, EMSL is in a natural position to lead the evolution of new capabilities that meet the broad computational, analytical, and modeling needs of the user community.

To address this gap, EMSL's third 10-year strategic science objective (Strategic Science Objective 3) is centered on establishing core expertise and capacity in advanced multiscale modeling and data analyses as a broadly available capability to aid researchers and users across BER in amplifying the ModEx approach to advance BER's science missions. This strategic science objective is meant to form a community of experts to continually advance data integration through community standards, multiscale modeling approaches, and development of the most advanced applications of AI/ML for experiment execution and knowledge generation. As a result, EMSL users will be armed with the ability to dramatically expand the frontiers of biological and environmental science challenges to maintain U.S. leadership in energy and environmental security.

5.2 Strategic Science Objective 3: Build a BER-Focused Modeling and Data Sciences Capability to Visualize and Incorporate Biological and Environmental Data and Parameterizations into Simulations

EMSL will lead the effort to build an extensible MDS capability to support the BER research community through computational science advances that convert data into knowledge, as well as establish and provide mid-range HPC capabilities that are focused on delivering computational resources and expertise to support BER and EMSL mission science. MDS will foster development of domain software to process, integrate, and visualize high-throughput and multimodal data in such a way that promotes parameterization of crucial process, local, regional, landscape, and Earth system models. Modeling approaches will be enhanced by ML and AI at every step of the data life cycle, from data acquisition to analysis and integration into models. Combined with an infrastructure to capture metadata that conforms to community-accepted standards, EMSL's cohesive approach to data generation, management, and analysis will enhance the predictive modeling and simulation of biological and environmental processes.

The mid-range production HPC capability will host heterogeneous compute systems tailored to support diverse computational research in data mining, data processing, and multiscale modeling of biological and environmental processes. By providing a reliable and secure HPC computing environment together with scalable hierarchical data storage and archiving, as well as software codes optimized for running efficiently on these systems, this capability will provide seamless integration between measurements and model simulations and will form the backbone of a centralized computational resource for processing image data generated at EMSL and at collaborating institutions and user facilities.

To establish leadership in computational sciences and to build MDS, EMSL will focus efforts in six research areas over the next decade (Figure 7). During the initial 2–5 years, our activities will concentrate on three of these research areas: (1) Data Integration Software Framework, (2) Metadata Capture and FAIR Data Management, and (3) Open-Source Data Analysis Software Suites. Delivering the intended outcomes of each of these research areas involves significant development efforts to meet underlying infrastructure needs, such as supporting containerization technologies and streamlining access to high-performance, high-capacity data archives. The Systems Modeling and Data Sciences IRP provides science leadership for the

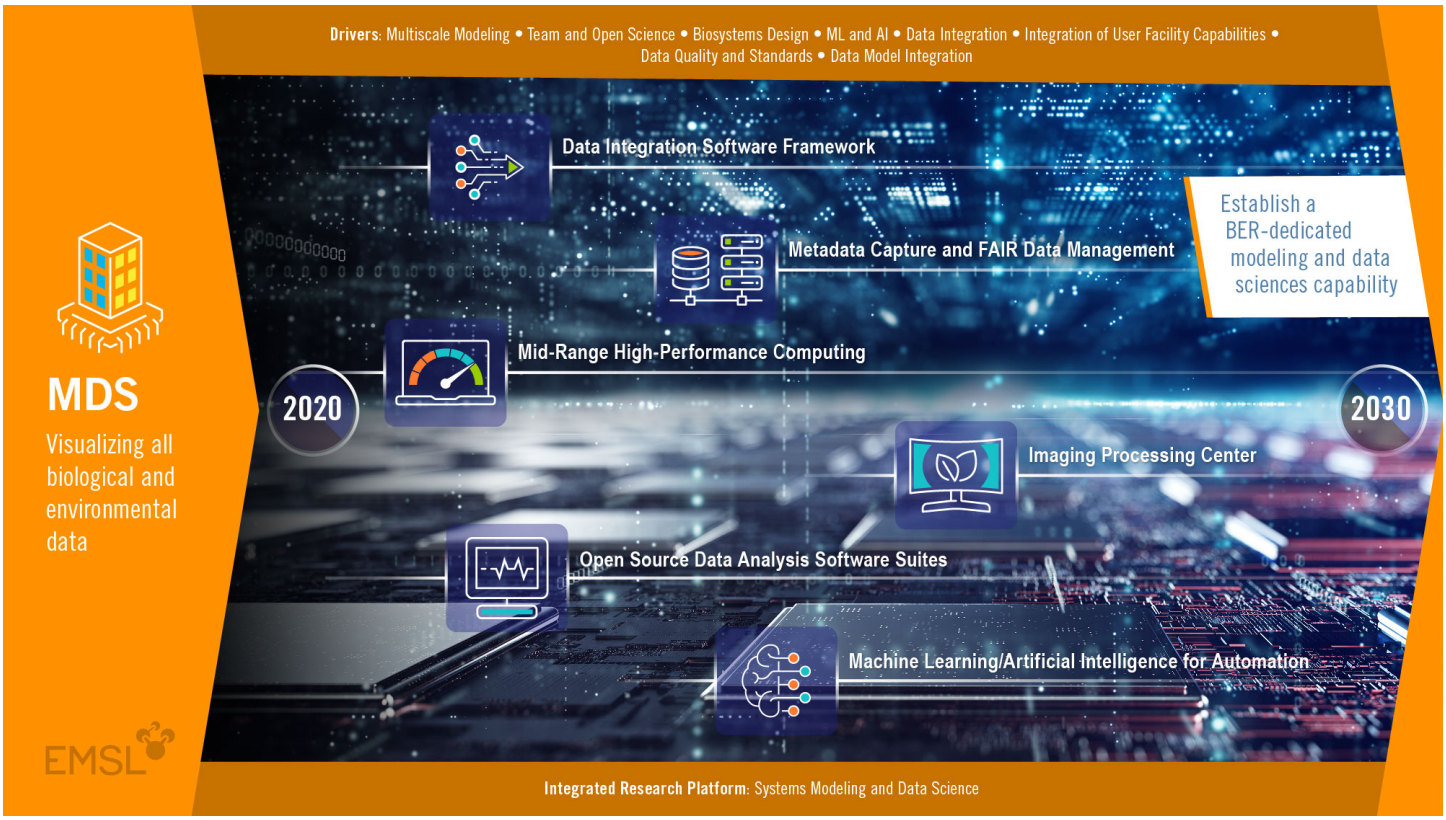


Figure 7. Overview, timeline, and research areas supporting Strategic Science Objective 3. The research areas within this objective build capacity, methodology, software, and new capabilities that together accelerate translation of data into knowledge by enabling integration visualization and utilization of all BER data streams. Each research area is placed on the 2020–2030 timeline to show where we anticipate the most activity, though we expect work to begin before and continue after, as required.

research elements of these activities. Other matrixed staff, under the direction of the Chief Data and Analytics Officer, support development of the data and computing infrastructure required to meet the outcomes in each of the research areas below. In each case, the research areas build the scientific foundation for MDS, and the user community makes use of the emerging science and technology for scientific inquiry that directly supports BER mission and goals.

5.2.1 Open-Source Data Analysis Software Suites Research Area

The continuing expansion of high-throughput experimentation and sample processing has created a constraint in data analyses that limits discovery and understanding. To move away from complex, manual, and time-consuming analysis processes, modular and interoperable software will be developed that take advantage of cutting-edge computational approaches (e.g., AI and ML) to automate standard processing, data reduction, and molecular identification, thereby accelerating discovery and improving reproducibility of research results. This activity will near maturity with the availability of effective, open-source software suites for mass spectrometry data of complex organic mixtures, metabolites, and proteins, as well as NMR-based metabolite data. These tools support the Automated Organic Matter and Soil Analysis (MONet Strategic Science Objective, Section 4.2.1), HTP Omics and Protein Function, Single-Cell Biology, and Visualizing Metabolic Pathways Research Areas (Strategic Science Objective 1; Sections 3.2.1, 3.2.2, and 3.2.6).

Supporting IRPs: All by virtue of their data producing roles, but principally Systems Modeling and Data Sciences.

Major External Engagements

- **National Microbiome Data Collaborative** (NMDC). *Ongoing:* The NMDC provides open, FAIR access (Section 5.2.3) to multiple high-throughput, complex data types, together with standardized data analysis workflows to support data interoperability. EMSL will continue partnering with NMDC to provide the community with open-source analysis workflows for processing a variety of omics data types.
- **Interoperable Design of Extreme-scale Application Software** (IDEAS). *Anticipated:* IDEAS is a family of projects (IDEAS-Classic, IDEAS-ECP, IDEAS-Watersheds) that support software productivity and sustainability for computational science and engineering applications targeting BER EESSD research areas. Working with IDEAS will accelerate EMSL's development of software supporting models of biogeochemical cycling in watersheds.

Recent and Near-Term Supporting Activities

- **Develop a software framework for mass spectrometry data.**
  Ongoing: EMSL Intramural S&T Research investment developing and maintaining the CoreMS framework (EMSL 2021).
- **Build mass spectrometry data analysis modules and workflows that leverage the CoreMS software framework.** *Ongoing:* EMSL Intramural S&T Research investments to develop (1) an analysis module designed for stable isotope-labeled peptide spectra, (2) a workflow for GC-MS metabolomics data processing, (3) a workflow for FTICR-MS natural organic matter data processing and annotation, and (4) a workflow for MS/MS metaproteomics data; the National Microbiome Data Collaborative (NMDC) partnership; a PNNL lab-level R&D investment through the *m/q* Initiative to address uncertainties in spectra matches and metabolite identification through a robust false discovery rate application.
- **Develop a standardized workflow for analysis of NMR data.**
  Ongoing: EMSL Intramural S&T Research investment in designing and testing an NMR data analysis workflow interface.

5.2.2 Data Integration Software Framework Research Area

Integrated analyses of complex, multimodal data are crucial to furthering our understanding of biological and environmental processes. The combination of separate data types (e.g., genomics, transcriptomics, proteomics, metabolomics, and images) requires knowledge of the wealth of information resident in biological databases, statistical programming skills, and an understanding of the underlying assumptions of statistical models. EMSL will develop a framework for discovery of the relationships between diverse omics data types that goes beyond traditional correlative methods, using ML, classification, or discriminatory models that incorporate diverse omics data types to identify the processes that are affected by experimental conditions. The framework will be modular in nature to allow the addition of new tools and algorithms for data integration and to support user-guided visualizations for data exploration. Keys to success in this activity area include openly available software and web services for data exploration, visualization, and integration for proteomics, metabolomics, and transcriptomics, as well as spatially resolved and time-series data over the long term. This effort is synergistic with the HTP Omics and Protein Function and Visualizing Metabolic Pathways Research Areas (Strategic Science Objective 1; Sections 3.2.1 and 3.2.6), as well as the Automated Organic Matter and Soils Analysis, Rhizosphere Sensors, and Field Sensors for

Plants, Microbes, and Aerosols Research Areas (Strategic Science Objective 2; Sections 4.2.1, 4.2.2, and 4.2.6).

Supporting IRPs: All by virtue of their data producing roles, but principally Systems Modeling and Data Sciences.

Major External Engagements

- **Systems Biology Knowledgebase** (KBase). *Ongoing:* KBase hosts computational tools and a reference database to build metabolic models. EMSL will continue to partner with KBase to provide access to open-source analysis tools, enhance the reference data with calculated thermodynamic properties of the metabolites and reactions, and provide users with the tools needed to integrate protein and metabolite measurements into metabolic pathway simulations that are informed by constraints on the energetics of biochemical reactions.

Recent and Near-Term Supporting Activities

- **Advance software tools and interfaces for multi-omics data exploration and visualization.** *Ongoing:* EMSL Intramural S&T Research investment in a web-based portal to provide a centralized platform for analysis and visualization of multiple omics data types.
- **Develop ML-based data integration approaches.** *Ongoing:* EMSL Intramural S&T Research investment in data integration software that identifies appropriate integration methods (e.g., canonical correlation analysis, reinforcement learning) and incorporates available knowledge of molecular structure and function.

5.2.3 Metadata Capture and FAIR Data Management Research Area

Biological and environmental research approaches are shifting from being empirical and observational toward data-driven exploration and model generation methods. This shift is being enabled by transformative data science strategies and sophisticated modeling methods that rely on data being FAIR. These FAIR principles thus guide EMSL's data management efforts and define data infrastructure needs. EMSL is implementing requirements for collection of sample metadata, using standards developed by the Genomic Standards Consortium and the Open Biological and Biomedical Ontology Foundry, producing schemas that track sample processing and analysis, and employing a data indexing strategy to support reliable search and retrieval through an openly accessible portal that both supports data sharing with the broader scientific community and encourages data citation through the use of persistent identifiers, known as Digital Object Identifiers (DOIs), for data. Deployment of a new system for the EMSL user project management portals and integration with data management systems to provide users with streamlined access to data, increasing adoption of FAIR principles for new projects, and development and deployment of sample and metadata tracking systems in EMSL are key metrics of success on the way to completion of this activity.

Supporting IRPs: All by virtue of their data producing roles.

Major External Engagements

- **Joint Genome Institute (JGI) and National Microbiome Data Collaborative (NMDC).** *Ongoing:* To assure broad access to high-throughput, complex data types generated from microbiome samples, JGI and NMDC support community-driven development of the standardized metadata formats that are essential for data to be findable and interoperable. EMSL will continue partnering with JGI and NMDC to

provide the community with access to EMSL data that are harmonized with complementary data distributed across existing resources.

- **Environmental Systems Science Data Infrastructure for a Virtual Ecosystem** (**ESS-DIVE**). *Ongoing:* ESS-DIVE is a BER-supported data repository located at LBNL that is designed to store and publicly distribute data from observational, experimental, and modeling research of terrestrial and subsurface ecosystems. EMSL will continue to partner with ESS-DIVE to provide the user community with access to relevant EMSL data and to support the adoption of globally resolvable sample identifiers and mechanisms for data attribution.
- **DOE Office of Scientific and Technical Information** (**OSTI**). *Ongoing:* OSTI has responsibility to collect, preserve, and disseminate scientific and technical information emanating from DOE-funded research and development. As part of that mission, OSTI assigns DOIs to datasets and registers the DOIs with DataCite to aid in data citation, discovery, retrieval, and reuse. EMSL will continue to work with OSTI to provide users with DOIs for EMSL-generated datasets.

Recent and Near-Term Supporting Activities

- **Upgrade and enhance the EMSL data management infrastructure.**
  Ongoing: EMSL operations, especially through NEXUS, which replaces the EMSL Usage System (EUS) for managing user projects and can be expanded to provide data access, data sharing, and metadata tracking.
- **Build partnerships that support distributed data systems and data citation.**
  Ongoing: EMSL operations working with NMDC, JGI, and ESS-DIVE to harmonize metadata collection for biological and environmental samples will enable making EMSL-generated data stored in EMSL's data repository findable and shareable through an open application programming interface (API); working with DOE-OSTI and DataCite to provide DOIs to datasets.

5.2.4 High-Performance Computing Center Research Area

Demand for mechanistic models, fine-scale land-based models, and analyses and visualizations of large datasets associated with multi-omics and imaging workflows has highlighted a gap in the availability of mid-range HPC as a key resource for data science and modeling efforts within the BER research community as well as across many areas of research of interest to the DOE Office of Science. With an existing mid-range computing capability and ample available production computing space, EMSL is ideally positioned to provide this computing resource for EMSL users and the BER scientific community. EMSL will expand its existing computational capability to support interaction across BER facilities. This center will host heterogeneous architecture HPC clusters interconnected with scalable data storage and archiving to facilitate ModEx and multiscale modeling. Data processing, web services, and visualization will be supported though a scalable container-orchestration system. Success will be measured by the efficient and maximal operation of Tahoma, demonstrable growth in EMSL HPC to support the user program, and planning, purchasing, and deploying a second mid-range HPC resource for the COMPASS Program. There are strong synergies with the Capacity and Pace Objective (Section 6) and the modeling and simulation efforts in Strategic Science Objectives 1 and 2 (Sections 3 and 4).

Supporting IRPs: Systems Modeling and Data Sciences

Major External Engagements

- **Coastal Observations, Mechanisms and Predictions Across Systems and Scales (COMPASS).** *Ongoing:* The COMPASS project requires substantial computing resources to support process-based models and seamless integration between measurements and model simulations of coastal processes. EMSL will continue to work with COMPASS to deploy and support the necessary HPC resources for COMPASS and complementary BER-supported research.
- **Exascale Computing Project (ECP).** *Anticipated:* The ECP supports research, development, and deployment of mission-critical applications, an integrated software stack, and exascale hardware technology advances. EMSL will explore ECP activities in the Application Development and Software Technology thrust areas that could enhance EMSL's HPC productivity and more formally engage ECP where warranted.

Recent and Near-Term Supporting Activities

- **Refresh the EMSL HPC capability.** *Ongoing:* EMSL operations, including the Tahoma mixed CPU/GPGPU computer purchase and deployment and expansion of Aurora data archive size.
- **Plan mid-range HPC resource for BER's coastal research program and related projects.** *Ongoing:* BER coastal program, leveraging EMSL computing expertise to expand the COMPASS project HPC resource to support BER research on coastal ecosystems more broadly.

We anticipate a gradual shift in these initial research areas after 4–6 years to the following additional research areas.

5.2.5 AI/ML for Automation Research Area

To fully realize EMSL's ambition of increasing sample processing throughput by several orders of magnitude through automation of experimental workflows, we will simultaneously address the challenge of automated data acquisition to assure data quality, integrity, and reproducibility. Advances in ML and AI will be leveraged, building on EMSL's vast library of archived data to create the training datasets needed to construct robust models. Such models, implemented at the time of data acquisition, will facilitate real-time quality control assessment, provide feedback to the instrument automation tools, and inform parameterization of downstream data analysis workflows. We expect an evolving set of successes as this activity develops. These include ML models of mass spectra attributes that provide automated quality control for organic matter characterization workflows that are integral to the Automated Organic Matter and Soil Analysis Research Area (Strategic Science Objective 2; Section 4.2.1) and AI/ML methods that allow mass spectrometry structural features to be accurately extracted from molecular imaging data within the Bio-Atomic Imaging and Visual Proteomics Research Area (Strategic Science Objective 1; Sections 3.2.3 and 3.2.5). Both goals are essential elements of our envisioned world-class Imaging Processing Center (Strategic Science Objective 3; Section 5.2.6).

Supporting IRPs: All by virtue of their data producing roles, but principally Systems Modeling and Data Sciences.

Major External Engagements

- **Center for Advanced Mathematics for Energy Research Applications (CAMERA).** *Anticipated:* This project is jointly funded by DOE's Office of Science, Advanced Scientific Computing Research and Basic Energy Sciences programs to develop and deliver fundamental new mathematical and computational methods and software required by complex experiments. EMSL will explore opportunities to partner with CAMERA investigators on approaches for Autonomous Discovery to guide experiments using data as they are collected and to optimize data acquisition.
- *Anticipated:* EMSL expects to expand from engagements with PNNL laboratory initiatives focused on AI and ML to working with external researchers and research organizations over the next several years as this effort matures.

Recent and Near-Term Supporting Activities

- **Improve data acquisition from instruments in EMSL's automation pipeline.**
  Ongoing: EMSL Intramural S&T Research investments in automating acquisition-time quality control and instrument parameter optimization of FTICR-MS.

5.2.6 Imaging Processing Center

To keep pace with the rapid advances in technologies that capture images at ever-increasing resolution, EMSL will establish an Imaging Processing Center to provide a centralized computational resource to manage, process, and analyze imaging data in near real time. The resources and expertise in this activity will support several activities in Strategic Science Objective 1 (the Bio-Atomic Imaging, Single-Cell Biology, and Visual Proteomics Research Areas).

Supporting IRPs: Structural Biology, Systems Modeling and Data Sciences

Major External Engagements

- **Pacific Northwest Cryo-EM Center (PNCC).** *Ongoing:* The NIH-funded PNCC generates cryo-EM data for a diverse user community. EMSL will continue to partner with PNCC to provide image processing and data storage resources.
- **EMDataResource**. *Anticipated:* This NIH-funded resource is a joint effort involving the Stanford/SLAC Cryo-EM Facility, the Research Collaboratory for Structural Bioinformatics, and the European Bioinformatics Institute that enables data archiving and retrieval of three-dimensional electron microscopy density maps, atomic models, and associated metadata. EMSL will pursue collaborations to further the development of software tools, data standards, and sharing of image data.
- **SLAC**, **SNS**. *Anticipated:* These BES-funded user facilities provide access to specialized experimental capabilities located at synchrotron light and neutron sources. EMSL will form partnerships through the FICUS program to provide computational resources and staff expertise for data processing and complementary molecular dynamics simulations to deliver mechanistic insights into structure and function; collaborations will focus on FICUS projects using the SSRL Imaging Group capabilities at the Stanford Linear Accelerator (SLAC) and the Biological Small-Angle Neutron Scattering (Bio-SANS) capability in the Center for Structural Molecular Biology at SNS.

Recent and Near-Term Supporting Activities

- **Connect collaborating imaging resources with high-performance data processing and storage.**
  Ongoing: Processing and distributing data generated by the NIH-funded PNCC at Oregon Health Sciences University; EMSL user program projects supporting data processing needs associated with FICUS program projects with SNS Bio-SANS (Small-Angle Neutron Scattering, ORNL) and SLAC (cryo-EM, XANES).
- **Resolve complex protein structure at the atomic to nano scales.**
  Ongoing: EMSL Intramural S&T Research investment using ML to extract molecular structure from atomic probe tomography data.

6.0 OPERATIONS FOR CAPACITY AND PACE

The Operations for Capacity and Pace Strategic Operational Objective aligns multiple operational activities to expand and accelerate our most significant and impactful workflows, analyses, and modeling activities as well as growing and optimizing our external engagements (users and other research partners; see Working with EMSL). Ultimately, these aligned activities will accelerate the pace of scientific discovery for the benefit of EMSL users and across the BER research community. This objective is strongly collaborative, involving partnerships with the three science areas and their associated IRPs. By coordinating investments, external research engagements, and facilities planning, the Operations for Capacity and Pace objective will amplify the value and impact of each, and of the supporting research activities in Strategic Science Objectives 1, 2, and 3.

6.1 Background for Strategic Operational Objective 1

A series of evolving national environmental and energy research priorities, combined with emerging obstacles to the collective production of the often immense and complex datasets required to meet these priorities, has created a need for a new level of productivity and efficiency in the management and operation of the EMSL user program to enable scientific discovery. In parallel, as echoed throughout this Strategic Science Plan, science is undergoing a dramatic acceleration in interdisciplinary research, requiring more fluid teaming and greater open access to capabilities available at EMSL and at other DOE-SC user facilities and community resources, as well as to their data streams. Broadening the integrative capabilities within and among DOE-SC user facilities is increasingly a prerequisite for this interdisciplinary approach to BER-relevant science.

Two national priorities that require greater capacity for team science and generation of essential mechanistic data streams are garnering and maintaining global leadership in the projected $4 trillion/year bioeconomy and creating the world's most advanced, most accurate biological, process, and Earth systems models (Safeguarding the Bioeconomy, NASEM 2020; White House Memo M-20-29 ; EESSD Strategic Plan, U.S. DOE 2018a; BSSD Strategic Plan, U.S. DOE 2021a). However, delivering scalable bioeconomy and bioenergy solutions and improved process, ecosystem, and Earth system models will also require a dramatic shift from a traditional emphasis on genome sequencing and genomic data to one that advances and combines a more complete array of phenotyping strategies such as synthetic biology, structural biology, and multi-omics approaches. That shift has started but cannot thrive without innovations that move phenotyping strategies into interconnected platforms that operate at the pace and scale achieved for genome sequencing and sequence analyses. A broad range of similarly interconnected infrastructure capabilities, instrumentation, and tools will also be necessary to drive integration and management of models, as well as experimental and observational data across a hierarchy of scales and complexities to meet expanding needs in Earth and ecosystem simulation and predictive modeling. The cascading growth in the complexity and size of these datasets is accelerating the need for AI and ML approaches to data analysis, modeling, and execution of biological and environmental research; this need has been formally recognized as another U.S. national priority (American Artificial Intelligence Initiative, OSTP 2020).

Collectively, these national science trends and drivers point to a broad capability gap in BER related to capacity and pace that is best addressed through a new, coordinated focus on automation and partnerships. The EMSL Operations for Capacity and Pace Strategic Operational Objective is centered on aligning operations to embrace, accelerate, and drive innovations that speed scientific discovery through expanded throughput and speed. This uniquely operational objective directly supports the following DOE goals and objectives (Department of Energy Strategic Plan, U.S. DOE 2014):

1. Objective 3: Deliver the scientific discoveries and major scientific tools that transform our understanding of nature and strengthen the connection between advances in fundamental science and technology innovation.
2. Objective 9: Manage our assets in a sustainable manner that supports the DOE mission.
3. Objective 10: Effectively manage projects, financial assistance agreements, contracts, and contractor performance.
4. Objective 11: Operate the DOE enterprise safely, securely, and efficiently.

To address the needs related to national priorities and the gaps in BER capabilities, as well as support the overarching DOE goals and objectives, EMSL's 10-year operations objective comprises three operational areas: (1) Operational Area 1 – automating processes to accelerate the pace and scale of scientific discovery, (2) Operational Area 2 – optimizing infrastructure, instrumentation, and operations, and (3) Operational Area 3 – building partnerships to accelerate interdisciplinary research and team science (Figure 8). Within these operational areas, we will focus on six activities: IRP-focused Infrastructure and Operations, Computing Infrastructure, Strategic Industry Partnerships, DOE Facilities Partnerships, Automation Partnerships and Projects, and Instrumentation Life Cycle Management.

Figure 8. Overview, timeline, and research areas supporting Operational Objective 1. Premier facilities, partnerships, and engagements and effective operations are the foundation on which EMSL continues to meet its vision and mission objectives. Our focus on automation, infrastructure optimization, and partnerships aligns operations in direct support of Strategic Science Objectives 1–3, creating synergies between operations and science that accelerate scientific discovery and progress toward our vision of a research community *empowered to study the role of molecular processes in controlling the function of biological and ecological systems across spatial and temporal scales and to enable a predictive understanding of the living Earth system.* Each research area is placed on the 2020–2030 timeline to show where we anticipate the most activity, though we expect work to begin before and continue after, as required to complete the work.

Overall, the Operations for Capacity and Pace Strategic Operational Objective drives development and implementation of the most advanced approaches to automating experimental and analytical workflows and provides facilities and operations that expand and accelerate building highly synergistic research and development partnerships. Through the automation and partnering operational areas, EMSL establishes a priority to continually optimize capabilities and operations that streamline and accelerate scientific discovery and advance technological innovation for the EMSL user community, DOE/BER, and the nation.

6.2 Operational Area 1: Automate Processes to Accelerate the Pace and Scale of Scientific Discovery

EMSL will align and focus operations, management structures, investments, and partnerships (industrial and academic engagements) to more deliberately support the automation of analytical workflows that are key metrics of success for Strategic Science Objectives 1–3. This new focus will not only drive delivery of the three scientific decadal science objectives and their near-term efforts but will also yield other equally important outcomes for DOE, users, and EMSL. Three main value drivers for pursuing automation as a core design principle in EMSL operations are: (1) acceleration of **smart** scientific discovery for BER and users; (2) democratization of scientific knowledge; and (3) increased access through enhanced resilience and remote operations. In this vein, pursuing automation is an emerging holistic approach that combines both traditional fixed (high-throughput) and variable autonomous operations, to support and more importantly amplify the complexity of science performable for BER researchers. Fixed automation provides greatly enhanced capacity and pace for highly subscribed but largely standardized analyses. Near-term efforts in this area are focused on supporting the HTP Omics and Protein Function Research Area as well as the Automated Organic Matter and Soils Analysis Research Area in Strategic Science Objectives 1 and 2, respectively.

As advances in automation, sample preparation, and instrument scheduling logistics are made, increasingly complex analyses will be incorporated into fixed automation platforms. Variable automation will allow EMSL users to customize a variety of experimental parameters and “à la carte” fixed analyses, allowing researchers to select a suite of the most appropriate HTP analytical workflows to generate data needed for answering the most significant scientific questions. The combination of fixed and variable automation drives the acceleration of smart scientific discovery—the rapid and directed generation of knowledge by performing the most impactful analyses in an informed manner without being restrained by scale. Automation can dramatically increase data quality, reproducibility, and interoperability, reducing variability in experimentation and data to a level not possible with conventional human-driven processes. Additionally, automation will increase opportunities for EMSL users to operate remotely by allowing them to collaborate with EMSL expertise and by providing access to capabilities without requiring an on-site presence. Smarter, faster, and more efficient operation of experimental and analytical workflows will free up time and intellectual and financial resources for researchers to pursue innovations, address more complex problems, and build leadership in the multidisciplinary scientific domains of our IRPs. The extremely large datasets generated from the automated workflows will heavily leverage all the activities associated with Strategic Science Objective 3 (see Section 5.2). An ancillary yet highly valuable benefit of standardized automation will be the ability to seamlessly connect experimental and analytical data generation sources to the data repositories within BER (ESS-DIVE and NMDC), providing well-curated data through spatial and temporal regimes as well as across variants, species, taxa, and other taxonomic levels at a scale not previously possible.

To enable the scale of automation and partnerships we envision, EMSL will focus efforts on the operational activity areas below over the next decade.

6.2.1 Execute Automation Projects and Partnerships

Automating organic matter and soil analyses, whole soil analyses, and HTP analyses for microbial phenotyping are critical activities for Strategic Science Objectives 1 and 2 that cannot be executed without strong support from the facilities, infrastructure, and contracting side of the EMSL user program. Freeing space, making space modifications, building new spaces, and refitting EMSL spaces for ideation, development, piloting, and finally deploying both the instrument and data aspects of automation will be required. To meet this need, EMSL will plan and execute facility modifications or additions to provide infrastructure for an evolving suite of increasingly complex automation capabilities in partnership with leaders in automation, EMSL users and the BER research community. To be successful, we expect to deliver on the required facilities and space modifications first for organic matter and soils analysis, as well as help plan and make modifications to existing facility space, or build a new facility for microbial phenotyping for Strategic Science Objective 1. These efforts amplify and synergize the automation efforts in the HTP Omics and Protein Function, Automated Organic Matter and Soil Analysis, and MONet Field Sites Research Areas (see Sections 3.2.1, 4.2.1, 4.2.4).

Partnering IRPs: All, but principally Systems Modeling and Data Sciences, Biochemical Pathways, and Biogeochemical Transformations.

Recent and Near-Term Supporting Activities

- **Microbial Molecular Phenotyping (M2P) Capability Planning.**
  Ongoing: DOE's multi-year project management process for scoping, designing, building, equipping, and opening an M2P Capability (CD-0 approval by DOE April 2021).
- **Creating infrastructure for organic matter analyses workflow.**
  Ongoing: Facilities planning for dissolved organic matter and soil organic matter automated workflows.
- **Establishing automation laboratory space and "maker" space to develop and pilot automated workflows.**
  Ongoing: Automated Organic Matter and Soils Analysis, Rhizosphere Sensors, and the Field Sensors for Plants, Microbes, and Aerosols Research Areas.
- **Developing partnerships with leaders in automation.**
  Anticipated: Contracts and partnerships with automation leaders in industry, academia, and other DOE or federal agencies.

Mid-Term and Future Supporting Activities (Supporting Projects in Development)

- EMSL anticipates making infrastructure upgrades and renovations to lab spaces to accommodate further automation of various multimodal analytical workflows, examples include rhizosphere imaging, plant root multi-omics, and biological and surface imaging platforms.

6.3 Operational Area 2: Optimize Infrastructure, Instrumentation, and Operations

Strategic Science Objectives 1–3 in the FSB, ETI, and CAM science areas require world-class facilities, operations, and instrumentation (EMSL 2020). Approximately 12,900 square feet of laboratory space in EMSL is becoming available because of unaligned capabilities exiting EMSL. This provides an opportunity to realign laboratory space in support of the science areas, strategic science objectives, and their associated IRPs. This realignment will bring two immediate benefits: (1) instrumentation can be co-located to improve workflow efficiency, thus contributing to greater output, and (2) multidisciplinary approaches to research are

nurtured by bringing together diverse expertise and domain knowledge into shared spaces. To assure EMSL's continued leadership, instruments will need to be managed strategically through their entire life cycle. Often, instruments are maintained far beyond their intended period of use, becoming unreliable and more costly to maintain. New instruments often have smaller footprints and execute their functions with greater speed, accuracy, and precision (Future Capabilties Investment Plan, EMSL 2019). A streamlined life cycle management process will also guarantee that automated workflows are maintained and improved alongside advances in robotics and instrumentation, assuring that the operational costs are minimized while maximizing productivity. Finally, operations need to be executed with clear lines of responsibility and authority such that scientific staff can focus on guiding the science of users and executing EMSL's strategic road map (EMSL Operations Manual, EMSL 2020).

To realize these goals, EMSL will focus on the three operational activity areas related to infrastructure, instrumentation, and operations described below:

6.3.1 Build and Support IRP Infrastructure and Operations

EMSL's IRPs were constructed as centers of scientific leadership, technical excellence, and foundational science critical for execution and delivery of EMSL's mission, strategic science objectives, and associated activities. Institutionally, EMSL is committed to align investments, operations, and line management approaches to provide premier resources as well as a rich, innovative, and multidisciplinary research environment to the EMSL user community. Toward this end, EMSL leadership will work closely with IRP leaders to plan and provide laboratory space, facilities, equipment, line management structures, and operations tools to support the success of the IRPs. This activity will be continuous, but metrics of success include reorganization of space that consolidates IRP activities in EMSL, providing state-of-the-art instrumentation, and line management support for hiring and strategic planning. We expect to see other measurable successes, such as users producing the highest-impact science, IRPs increasing domain expertise and leadership through increased planning and leadership in BER workshops, conferences, and symposia, and synergizing across IRPs measured by increasing numbers of proposals utilizing multiple IRPs.

Partnering IRPs: All

Recent and Near-Term Supporting Activities

- **Deployment of space for consolidation of IRP equipment.**
  Ongoing: EMSL facility investments to remodel vacated EMSL laboratory spaces.
- **Alignment of user operations and line organization with IRPs.**
  Ongoing: Executing the IRP-focused user program operations and line management model for IRP leadership.

6.3.2 Expand Computation, Data, and Analytics Capacity

Strategic Science Objectives 1 and 2 will collectively produce the need for an order of magnitude or more increase in data storage, analytics, on-the-fly analyses, modeling, and computation. The value of the accumulated data will also increase as its richness improves from multimodal analyses, FAIR compliance, and expanding breadth that includes molecular, cellular, soil, sediment, and aerosol samples, across ecosystem, regional, and continental scales. Furthermore, the advanced capabilities developed in Strategic Science Objective 3 will require expanded capacity for computing and data storage, as well as improved data accessibility through distributed data systems (Section 5.2.3). In response to this need, EMSL will prepare for and provide space, facilities modifications, devices, computing systems, networks, and partner and user

interfaces that provide the capacity, data, and project management access required for EMSL automation, user research and access, and the broad set of scientific engagements EMSL will establish. Success will follow the evolution of specific needs over time, infrastructure for building and maintaining mid-range computing, facilities support for increased data storage capacity, NEXUS support, and facilities and infrastructure for data sharing (5G, network bandwidth of 100+ Gbit/s).

Partnering IRPs: Systems Modeling and Data Science

Recent and Near-Term Supporting Activities

- **Infrastructure support for mid-range HPC computing resource.**
  Ongoing: Retire aging HPC systems (Cascade) to open space for new HPC and data systems. Manage space upgrades for new computing capabilities.
- **Expand data storage and movement capacity to support increased automation efforts.**
  Anticipated: Grow data archive capacity to 100 petabytes, upgrade ESnet connection when ESnet6 becomes available (400 Gbit/s), upgrade EMSL building internal bandwidth to at least 100 Gbit/s, and contribute to automation planning workshops for capacity upgrades.
- **Systems and interfaces for managing and tracking user projects.**
  Ongoing: Replace the aging EUS with the NEXUS user portal and staff portal to manage user projects, track reviews and approvals, and assign instrument allocations.

6.3.3 Manage Instrumentation Life Cycle

EMSL's mission acknowledges the institution's unique role of providing a continually improving suite of premier science instrumentation, data storage and analytics, and production HPC, which enables users to employ a ModEx approach to their research. Many of these capabilities are the products of technological innovations produced by EMSL staff and its other research partners. Maximizing the lifespan of instruments, optimizing life cycle ends, and managing the transition to the next generation of instruments is critical for maintaining research productivity and access for users. The instrument life cycle is managed by the COO in partnership with the IRPs, who are responsible for the instrument purchase and development planning in accordance with our strategy. Toward that end, EMSL employs operational processes to manage instruments from planning and purchase through maintenance and final divestment of retired instrumentation. Success here is measured by minimal instrument downtime, by regular purchase of high-impact state-of-the-art or unique instrumentation in accordance with our capital and expense equipment purchase plan that is represented by our 2018 Strategic Capital Investment Plan (EMSL 2019) and ongoing updates to that plan, and by effective divestment of aging instruments that provide space for improved instrumentation.

Partnering IRPs: All

Recent and Near-Term Supporting Activities

- **Alignment of equipment investments with Strategic Science Objectives 1–3.**
  Ongoing: In close partnership with the IRPs, EMSL will continue to align plans for capital and expense equipment purchases with the specific needs of the research areas that support the three strategic science objectives.
- **Evolution of premier instrumentation life cycle management.**
  Ongoing: EMSL is developing a system for managing the full instrument life cycle of major instruments

that will track instruments from purchase, through their lifespan, and finally to divestment. Additionally, the system will provide insight into instrument usage and maintenance costs to help guide instrument life cycle decision making.

6.4 Operational Area 3: Build Partnerships to Accelerate Interdisciplinary Research and Team Science

EMSL uses several mechanisms to engage other researchers and organizations in support of BER science missions (see Working with EMSL). The FICUS program encourages the scientific community to propose novel ways for user facilities to work together. FICUS began as a joint call for proposals in 2013 between EMSL and JGI. It focused on DOE missions in bioenergy and the environment and offered the ability to combine EMSL's unique imaging, multi-omics, and computational resources with cutting-edge genomics at JGI. To build on the success of the EMSL–JGI FICUS Program, EMSL has established additional collaborations with other DOE user facilities using the FICUS framework as a guide (Figure 9). The potential of these collaborations has been recognized by BER's Advisory Committee as well as the user community for the impact and value the partnerships bring to EMSL users and BER science. EMSL will continue ongoing efforts to identify areas of synergy beyond the BER user facilities to include the DOE-SC user facility complex more broadly. As a preliminary step to realizing this potential, EMSL scientists have attended joint workshops at other user facilities, and several scientists from other DOE user facilities (e.g., NSLS-II, SLAC, and SNS) participated in the 2018 EMSL Integration Meeting. A pilot FICUS program with the SNS leveraging the Bio-SANS capability for advanced imaging of complex biological processes was recently pursued in the FY 2022 FICUS call. A similar effort with the APS is also being planned. Most recently, EMSL has explored the expansion of FICUS to incorporate capabilities and facilities outside the DOE-SC complex. A very successful pilot in the FY 2021 FICUS call with NEON from the National Science Foundation (NSF) led to a letter of cooperation between the JGI, EMSL, and NEON facilities to incorporate continuous partnership within the FICUS network. Expanding partnerships with synergistic facilities will amplify and extend the impact of EMSL capabilities, driving a growing set of high-value opportunities to accelerate multidisciplinary research that links synergistic capabilities from laboratory to field sites.

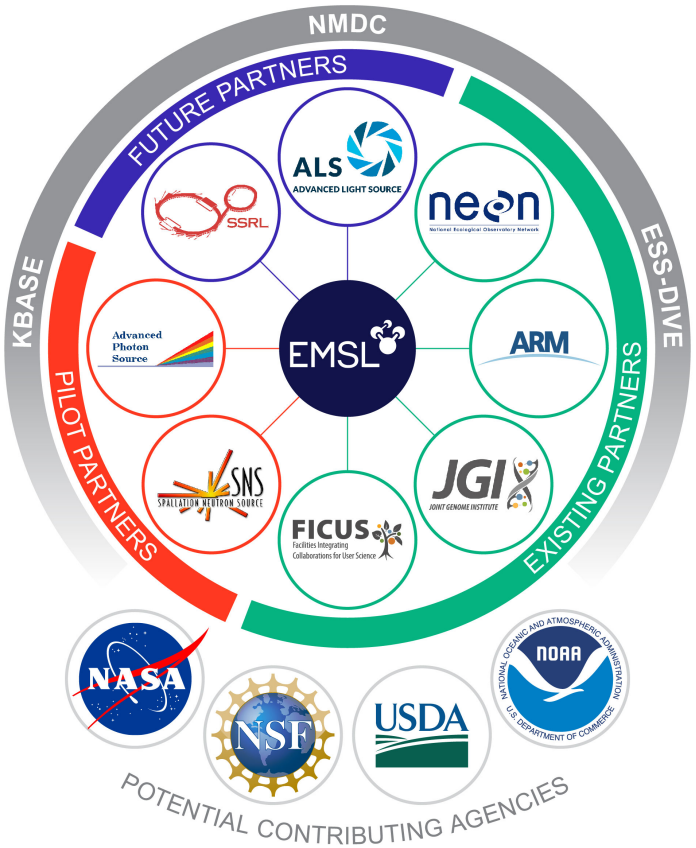


Figure 9. EMSL's growing landscape of partnering organization and agencies. EMSL has a broad, productive, and growing group of institutional partners that amplify and increase the impacts of EMSL research and capabilities.

Beyond growing partnerships with facilities capable of generating data, EMSL is concurrently pursuing partnerships with data repositories, including the NMDC and the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), to greatly increase access to and use of the massive well-curated datasets that EMSL and partners are producing with the user community.

In addition to these activities, which have largely been within the DOE or federal science complex, partnering with industry provides access to capabilities that will lead to the next generation of cutting-edge instrumentation. Further, industrial partnerships provide an outlet for EMSL-generated IP and lead to advanced instrumentation.

To achieve our desire for building partnerships to accelerate interdisciplinary research and team science, EMSL will focus on the following two operational activity areas in close collaboration with operations staff, the EMSL User Support Office, science area leads, and IRP leads, ultimately spearheading efforts to build, nurture, and contribute to our broad and growing list of scientific partnerships.

6.4.1 Establish Broader Partnerships with DOE Facilities

Partnerships, in their many forms, are the most important means for leveraging EMSL investments and resources to meet the objectives of this Strategic Plan, serving our mission in making high-impact capabilities available to users, and contributing to BER science missions and goals. Expanding partnerships through the DOE FICUS program is one of our highest priorities. EMSL is committed to growing partnerships across the DOE research ecosystem for the benefit of users and to accelerate progress toward EMSL's MONet, DigiPhen, and MDS Strategic Science Objectives. This activity area is by nature a joint venture for scientific leaders of Strategic Science Objectives 1–3 and our operations staff, particularly in the User Services Office. We see success in this effort as an addition of new partnerships and the expansion of existing partnerships each year.

Partnering IRPs: All

Major External Engagements: Existing and anticipated engagements are outlined in research area descriptions found under each of the three strategic science objectives.

Recent and Near-term Supporting Activities

- **Acceleration and development of aligned multidisciplinary partnerships across the DOE research community.**
  Ongoing: Execution of FICUS program calls. Active engagement with new DOE partners including SSRL, ARM, SNS/CSMB, and others to provide modeling and simulation support to the BER user community and expand participation in FICUS.
- **Create mechanism for access to DOE user facilities.**
  Ongoing: Development of mutually beneficial operating agreements—approvals process, review process, allocation of resources, and data policies and publications—with DOE user facilities to create a seamless user experience.

6.4.2 Develop Strategic Partnerships with Industry

EMSL has an established history of developing and co-developing leading technology and software in addition to driving new innovations in already commercialized instruments with industry leaders, often leveraging in-house expertise residing in EMSL's Instrument Development Laboratory. With the clear goals of advancing key analytical technologies such as bio-APT, cryo-EM, and mass spectrometry, strategic partnerships that leverage the expertise of both EMSL and technology innovators in industry will remain an important aspect of our long-term strategy to provide premier instrumentation. To meet this need, EMSL will invest in operational, contractual, and IP processes and partner engagements to establish partnerships with industry leaders to co-develop the next generation of cutting-edge instrumentation. The recent

commercial licensing of the NanoPOTS technology is an example of successful technology transfer from EMSL to industry; this technology development relied on capabilities in EMSL's Instrument Development Laboratory. This operational activity area supports and amplifies activities in Strategic Science Objectives 1 and 2. The most important metrics of success here are newly executed partnerships with key industry technology leaders and licensed IP; the latter is likely a long-term metric of success dependent on licensable or patentable innovations.

Supporting IRPs: All

Major External Engagements: Existing and anticipated engagements are outlined in research area descriptions found under each of the three strategic science objectives.

Recent and Near-Term Supporting Activities

- **Identify and pursue partnerships in alignment with strategy.**
  Ongoing: Planning activities to identify, assign, and manage space for collaboration with industry on the next generation of electron microscopes and assigning appropriate space for pilot testing and placement of automated organic matter and soil analyses. Contracts and IP: provide contract support and liaison with PNNL's Technology Development Office to assure timeline and efficient arrangements with partners. Provide contract and IP support for industry partnerships in automated organic matter and soil analysis.
- **Develop IP that attracts industry investment and partnerships.**
  Ongoing: Executing internal technology development projects and partner proposals for creation of IP in the areas of extreme UV-based bio-APT, mass spectrometry for soft-landing of proteins on EM grids, AI/ML for image analysis of APT data for soft biological materials, the nanoPOTS small sample and single-cell proteomics platform, definition of EMSL's metadata model structure, software to process mass spectrometry data from instrument data to molecular identifications, software for visualization and interpretation of organic matter data, and software for visualization of mass spectrometry proteomics data.

7.0 ENGAGING AND EMPOWERING THE USER COMMUNITY

As a national user facility, engaging and empowering our user community is the central focus of all aspects of this Strategic Science Plan. The elements of our plan, including investments, partnerships, strategic science objectives, operational objectives, and our models and mechanisms of user engagement, are all intended to expand on our long history of providing a transformational suite of science and technology capabilities to users pursuing DOE science missions and goals.

To foster an *engaged* user community, EMSL promotes awareness with effective communications by inviting user participation at conference sessions, user meetings, and workshops; building lasting partnerships through multiple programs including FICUS and our Partner Program; directly connecting users to our IRP leaders; providing platforms for data sharing and use; and seeking input from users and advisory boards (user executive and science and technology advisory committees). EMSL amplifies users' *productivity* by investing in automation to increase the capacity and pace of scientific discovery, improving user access to data and the user proposal processes through NEXUS, and continually advancing open-source tools for data analysis, modeling, and simulation to enhance experimental data interpretation. The UEC is charged with providing objective, timely advice and recommendations to the EMSL director and management team related to matters affecting EMSL users and evaluating our effectiveness in serving the user community. In addition, EMSL engages users in an advisory capacity through surveys and through interactions with our IRP leadership. EMSL maintains safety as a key aspect of its operations by reviewing planned experiments, assigning task-specific training, and assuring on-site guidance using PNNL's project and biosafety risk management and operations infrastructure.

7.1 Fostering User Community Engagement

An engaged EMSL user community is an essential element of our strategy to expand the number, diversity, and impact of our users across academia, industry, and other DOE research facilities. Equally important to

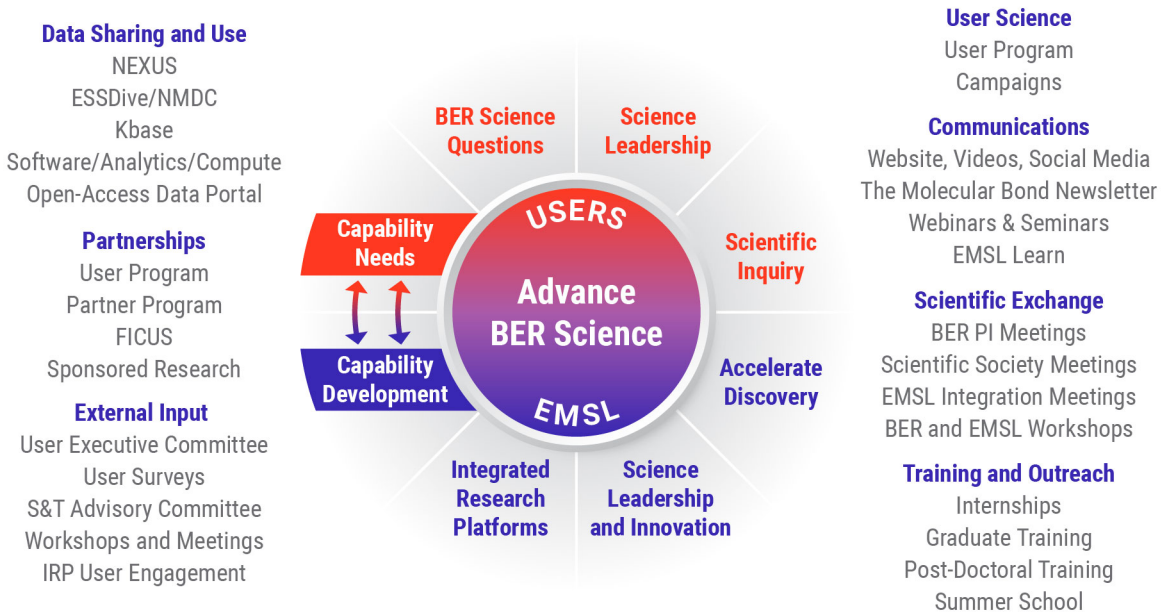


Figure 10. EMSL's strategy to engage and empower the user community. The multiple forms of engagement, communications, and partnership provide diverse forums for collaboration and recruitment of new users.

the EMSL Strategic Plan is the key role that an engaged user community plays in providing input that guides our strategic efforts to continually evolve our capabilities to meet the future needs of users through our Intramural Science and Technology R&D and capital investments. The myriad engagement mechanisms described below (Figure 10), as well as the UEC and Science and Technology Advisory Committee meetings will be used by EMSL to provide regular feedback that guides the evolution of strategic science directions described in the EMSL Strategic Science plan.

EMSL builds an engaged and enduring user community around important science questions relevant to BER missions through direct interaction with current and potential users, developing key partnerships, and employing far-reaching communication methods. Multiple elements of our strategy are directed toward user recruitment—social media communication platforms, scientific exchange at meetings, EMSL Program training, and outreach programs. These same elements play a critical role in effectively communicating the availability of new instruments, software, and experimental capabilities that emerge from EMSL's Intramural Science and Technology investments.

Multiple meetings and workshops will continue to be important elements of EMSL's engagement strategy. EMSL staff present EMSL science and capabilities at the BER PI meetings and scientific society meetings. Promotion of EMSL capabilities and expertise at BER PI meetings maximizes EMSL's exposure to BER-funded PIs, early career staff, and graduate students. Staff are encouraged to propose and chair sessions, organize townhalls, or participate in panel discussions at national meetings, particularly those that are BER-relevant (e.g., American Geophysical Union [AGU], International Society for Microbial Ecology [ISME], American Chemical Society, Biophysical Society, Society for Industrial Microbiology and Biotechnology [SIMB], and American Society for Plant Biology meetings).

The EMSL Integration Meeting is critical platform that EMSL uses to highlight user science, invite early career and prominent researchers in the community as plenary speakers, recruit new users to EMSL, and solicit input and ideas for research focuses aligned to EMSL's strategic science objectives. For example, the FY 2022 EMSL Integration Meeting focus on environmental sensors was deliberately chosen to both recruit new and prominent researchers in the sensor science field as well as inform our internal approaches to and priorities in developing and deploying fieldable sensors as part of our Rhizosphere Sensors Research Area and Field Sensors for Plants, Microbes, and Aerosols Research Area within MONet (Sections 4.2.2 and 4.2.6). These activities provide a venue for EMSL users to communicate the scientific impact of their research and the importance of EMSL as a national resource for state-of-the art capabilities and expertise.

EMSL also drives user engagement through its summer school activities, internships, and graduate student and postdoctoral training programs. Our strategy includes sponsoring workshops to engage the BER research community. EMSL invites established and early career scientists to these workshops to assure alignment of new capabilities with BER science mission and research directions and to increase awareness of our emerging capabilities. EMSL supports fellowships for visiting professors and postdoctoral researchers to develop key collaborative partnerships when sufficient funding is available.

EMSL will continue to develop one- to two-week summer school programs to train the user community on the fundamental biological and environmental theory and practice of experimentation in selected topical areas. For example, our 2020 summer school focused on the use of a multiscale microbial dynamics data integration pipeline being developed by the Office of Science, River Corridors Scientific Focus Area Program and other programs coupled to 1D reactive transport models. The 2021 summer school will provide training on using multi-omics data to model and engineer microbial metabolic pathways. In addition to being unique opportunities to build collaborations and the next generation of scientists, these workshops and summer schools attract potential users to EMSL by showcasing the areas of science, staff expertise, and

capabilities housed within our IRPs. The summer school teaching materials and videos of the lectures are available online for reuse by the research and academic community on EMSL's external website (EMSL Learn).

EMSL uses multiple mechanisms (see Working with EMSL) to actively pursue collaborations with the DOE Bioenergy Research Centers (BRCs), the NGEEs, and the National Laboratories Scientific Focus Area (SFA) projects to assure that these important programs are aware of EMSL's capabilities to advance their science objectives. The outreach includes planned visits and tours of EMSL to convey research capabilities and engage in scientific discussions to aid in high-quality proposal submissions by the DOE BER national laboratory funded research programs. EMSL recently issued a special invitation to the BRCs to submit exploratory proposals for the FY 2021 call cycle. During the review period, EMSL staff participated in or contributed to EESSD and BSSD SFAs and NGEEs.

Increasing the number of science partnerships with industry and expanding the opportunities to translate basic science knowledge to meet industry needs is an important aspect of EMSL's engagement strategy. EMSL extends awareness of EMSL's staff expertise and state-of-the-art instrumentation to new industrial users at science meetings, as well as direct peer-to-peer meetings initiated by staff or industry leaders. Past EMSL partnerships with industry have led to R&D 100 awards and patents and provide an opportunity to partner on SBIR or STTR funding.

EMSL employs a suite of modern, multi-platform electronic and social media communication mechanisms to raise awareness of EMSL within the scientific community, at BER and SC, and with key stakeholders. The EMSL external website is a primary communication mechanism for news and user proposal announcements and is continually updated with fresh news and content, such as the recently launched EMSL Learn section. The Molecular Bond newsletter, distributed quarterly to subscribers, features scientific perspectives from EMSL scientists, staff, and users, as well as emerging areas relevant to BER. Growing social media engagement further promotes and connects EMSL staff and user activities. Altogether, these electronic communication mechanisms comprise a key part of EMSL's strategy to build and maintain user awareness of new instruments, capabilities emerging from our internal investments, partnerships, and the user program.

Looking forward, EMSL will be evaluating the need for and opportunities to adopt new forms of communication and user engagement that reflect the expected evolution of our user community as we move into automation and autonomous activities, more remote users, analytics or compute and data users, and larger networks of users and partners as MONet, DigiPhen, and MDS come online.

7.2 Expanding User Community Productivity

EMSL's 2021 Strategic Science Plan takes a multifaceted approach to maximizing the productivity of our expanding user community. EMSL is implementing an IRP model (Section 2) for user access and engagement, launching a new user access system (NEXUS), providing open-source tools for data analysis, modeling, and simulation, expanding mid-range computing, and investing in automation of organic matter and molecular phenotyping workflows to expand capacity and pace of scientific discovery for users. These efforts are intended to support and grow the value of EMSL's mix of experimental, computational, and analytics capabilities, unique among DOE-SC facilities, for users.

EMSL implemented the IRP model to improve user awareness and access to its evolving suite of multidisciplinary capabilities with the goal of increasing the effectiveness, pace, and impact of user science. Our user engagement model was transformed to enable EMSL users to engage the seven IRP leaders and directly benefit from their expertise in the array of relevant scientific disciplines of value to their research, as

well as to provide greater connectivity and opportunities to leverage the broad capabilities available at EMSL.

EMSL will continue to ensure that users have the data, tools, and expert support necessary to be productive. Access to FAIR compliant data, metadata, and robust open-source tools for analysis and interpretation of those data are central to maintaining and growing a highly productive EMSL user community. We anticipate continued near-term gains in user productivity from EMSL's investments in improving access and analysis of open-access data streams available from EMSL and partner facilities.

EMSL's NEXUS is the centralized system for user project and data management, providing portals for user proposal submissions, project administration, and data retrieval. The NEXUS data repository portal supports data storage with community-established standards that assure interoperability with other BER data facilities, including JGI, NMDC, ESS-DIVE, and KBase and offers near-real-time access and open sharing of public data in accordance with EMSL's data management policy.

EMSL's mid-range, high-performance computing system, Tahoma, combined with new capabilities in cloud orchestration and edge computing, provide the diverse computing platforms needed to enhance the productivity of users through simulation and advanced data analytics and visualization to process and integrate our rapidly expanding array of multimodal biological and environmental data. As data production expands in EMSL through automation in the DigiPhen and MONet strategic science objectives, productivity will increasingly depend on the use of advanced computational tools to effectively translate molecular measurements into an understanding of biological and environmental processes across scales.

Making significant investments in automating experimentation and data collection is a key part of EMSL's strategy to dramatically expand the productivity of our user community. Initial priorities will be automation of organic matter and soil analysis and multi-omics measurements (Sections 4.2.1 and 3.2.1) and are expected to eventually evolve to include remote access for users of the many analytical and computational workflows offered by EMSL. Expanding both capacity and pace in these two areas is intended to amplify the impacts of similar expansion of the computing and analytics science area, bringing transformative increases in productivity achievable only though coordinated, integrated increases in both experimentation and data collection, storage, analytics, and access.

7.3 Operations for Remote, Satellite, and Data Researchers

Implementation of EMSL's strategic science objectives and operational objective will expand the community of remote, satellite, and data-focused researchers. Remote-access users may design and initiate experiments in EMSL's automated workflows as well as monitor data collection in real-time. Satellite user research groups might interact more regularly, with more coordination and cooperation with EMSL's experimental or computational capabilities and scientific leaders in the development of emerging EMSL capabilities supporting DigiPhen and MONet. The acceleration and expansive production of data emerging from MONet, DigiPhen, and MDS focused research is anticipated to dramatically increase the number of users conducting data analytics, modeling, and simulation-based research at EMSL.

To effectively anticipate and plan for the unique needs of this expanding community of remote, satellite, and data users, EMSL will engage the user community and BER leadership through the mechanisms described in Section 7.1 to identify necessary changes to our user program operations procedures and policies and build a roadmap for their development, implementation, and incorporation into our operations plan.

8.0 REFERENCES

BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision.* U.S. Department of Energy Office of Science, Biological and Environmental Research Advisory Committee. https://genomicscience.energy.gov/BERfiles/BERAC-2017-Grand-Challenges-Report.pdf.

BERAC. 2018. *Scientific User Research Facilities and Biological and Environmental Research: Review and Recommendations.* U.S. Department of Energy Office of Science, Biological and Environmental Research Advisory Committee. https://science.osti.gov/-/media/ber/pdf/community-resources/2018/BERAC_UserFacilities_Report.pdf.

Cui, Y., D. Hu, L. M. Markillie, W. B. Chrisler, M. J. Gaffrey, C. Ansong, L. Sussel, and G. Orr. 2018. "Fluctuation Localization Imaging-Based Fluorescence in Situ Hybridization (Flifish) for Accurate Detection and Counting of Rna Copies in Single Cells." *Nucleic Acids Research* 46 (2). https://doi.org/10.1093/nar/gkx874.

EMSL. 2019. *Future Capabilties Investment Plan.* Pacific Northwest National Laboratory. https://content-qa.emsl.pnl.gov/sites/default/files/2020-08/EMSL_Future_Capabilities_Investment_Plan.pdf.

EMSL. 2020. *EMSL Operations Manual.* Pacific Northwest National Laboratory. https://content-qa.emsl.pnl.gov/sites/default/files/2020-08/opsmanual.pdf.

EMSL. 2021. "Github: EMSL-Computing/CoreMS." https://github.com/EMSL-Computing/CoreMS.

ESS-DIVE. 2021. "ESS-DIVE: Deep Insight for Earth Science Data." U.S. Department of Energy Office of Science. Accessed March 21, 2021. https://ess-dive.lbl.gov/.

Haitjema, C. H., S. P. Gilmore, J. K. Henske, K. V. Solomon, R. de Groot, A. Kuo, S. J. Mondo, A. A. Salamov, K. LaButti, Z. Zhao, J. Chiniquy, K. Barry, H. M. Brewer, S. O. Purvine, A. T. Wright, M. Hainaut, B. Boxma, T. van Alen, J. H. P. Hackstein, B. Henrissat, S. E. Baker, I. V. Grigoriev, and M. A. O'Malley. 2017. "A Parts List for Fungal Cellulosomes Revealed by Comparative Genomics." *Nature Microbiology* 2 (8): 17087. https://doi.org/10.1038/nmicrobiol.2017.87.

KBase. 2021. "Welcome to KBase." U.S. Department of Energy Office of Science. Accessed March 21, 2021. https://www.kbase.us/.

Kendall, R. A., E. Aprà, D. E. Bernholdt, E. J. Bylaska, M. Dupuis, G. I. Fann, R. J. Harrison, J. Ju, J. A. Nichols, J. Nieplocha, T. P. Straatsma, T. L. Windus, and A. T. Wong. 2000. "High Performance Computational Chemistry: An Overview of NWChem a Distributed Parallel Application." *Computer Physics Communications* 128 (1): 260-283. https://doi.org/10.1016/S0010-4655(00)00065-5.

NASEM. 2020. *Safeguarding the Bioeconomy: Finding Strategies for Understanding, Evaluating, and Protecting the Bioeconomy While Sustaining Innovation and Growth.* The National Academies of Sciences, Engineering, and Medicine. https://www.nationalacademies.org/our-work/safeguarding-the-bioeconomy-finding-strategies-for-understanding-evaluating-and-protecting-the-bioeconomy-while-sustaining-innovation-and-growth.

NMDC. 2021. "National Microbiome Data Collective: An Open and Integrative Data Science System." Lawrence Berkeley National Laboratory. Accessed March 22, 2021. https://microbiomedata.org/.

Oostrom, M., Y. Mehmani, P. Romero-Gomez, Y. Tang, H. Liu, H. Yoon, Q. Kang, V. Joekar-Niasar, M. T. Balhoff, T. Dewers, G. D. Tartakovsky, E. A. Leist, N. J. Hess, W. A. Perkins, C. L. Rakowski, M. C. Richmond, J. A. Serkowski, C. J. Werth, A. J. Valocchi, T. W. Wietsma, and C. Zhang. 2016. "Pore-Scale and Continuum Simulations of Solute Transport Micromodel Benchmark Experiments." *Computational Geosciences* 20 (4): 857-879. https://doi.org/10.1007/s10596-014-9424-0.

OSTP. 2020. *American Artificial Intelligence Initiative: Year One Annual Report.* White House Office of Science and Technology Policy. https://trumpwhitehouse.archives.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf.

PNCC. 2021. "Publications | Pacific Northwest Cryo-EM Center." Accessed April 27, 2021. https://pncc.labworks.org/publications.

U.S. DOE. 2014. *U.S. Department of Energy Strategic Plan 2014–2018.* https://www.energy.gov/sites/prod/files/2014/04/f14/2014_dept_energy_strategic_plan.pdf.

U.S. DOE. 2018a. *Climate and Environmental Sciences Division Strategic Plan 2018–2023.* U.S. Department of Energy Office of Science DOE/SC-0192. https://science.osti.gov/-/media/ber/pdf/workshop-reports/2018_CESD_Strategic_Plan.pdf.

U.S. DOE. 2018b. *Climeate and Environmental Sciences Division Strategic Plan 2018–2023.* U.S. Department of Energy Office of Science DOE/SC-0192. https://science.osti.gov/-/media/ber/pdf/workshop-reports/2018_CESD_Strategic_Plan.pdf.

U.S. DOE. 2019. *Breaking the Bottleneck of Genomes: Understanding Gene Function across Taxa.* U.S. Department of Energy Office of Science DOE/SC-0199. https://genomicscience.energy.gov/genefunction/.

U.S. DOE. 2021a. *Biological Systems Science Division Strategic Plan.* U.S. Department of Energy Office of Science. https://science.osti.gov/-/media/ber/pdf/bssd/BSSD_Strategic_Plan_2021_HR.pdf.

U.S. DOE. 2021b. *Funding Opportunity Announcement: Data-Intensive Scientific Machine Learning and Analysis.* U.S. Department of Energy Office of Science, Advanced Scientific Computing Research Program. https://science.osti.gov/-/media/grants/pdf/foas/2021/SC_FOA_0002493.pdf.

Valiev, M., E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus, and W. A. De Jong. 2010. "NWChem: A Comprehensive and Scalable Open-Source Solution for Large Scale Molecular Simulations." *Computer Physics Communications* 181 (9): 1477-1489. https://doi.org/10.1016/j.cpc.2010.04.018.

White House Memo M-20-29. 2020. Fiscal Year (FY) 2022 Administration Research and Development Budget Priorities and Cross-Cutting Actions: Memorandum for the Heads of Executive Departmens and Agencies. Executive Office of the President. https://www.whitehouse.gov/wp-content/uploads/2020/08/M-20-29.pdf.

Appendix A – Science and Operations Drivers

Table A.1. Alignment of EMSL's strategic science areas and operational activities to BERAC Grand Challenges (BERAC 2017), BERAC User Facility Recommendations (BERAC 2018), EESSD's Grand Challenges (U.S. DOE 2018a), and BSSD's Goals (U.S. DOE 2021a).

FSB	ETI	CAM	C&P		BERAC Grand Challenges
	•			2.1	Understand the biological complexity of plant and microbial metabolism and interfaces across scales spanning molecules to ecosystems.
•				2.2	Develop technologies to identify DOE mission-relevant metabolic capabilities and engineering possibilities in bacteria, fungi, archaea, viruses, plants, and mixed communities.
		•		2.3	Optimize the use of large datasets that integrate omics surveys with biochemical and biophysical measurements to generate knowledge and identify biological paradigms.
•				2.4	Understand the links between genotype and phenotype in single but diverse organisms and in communities of organisms that interact in terrestrial ecosystems.
	•			3.2	Establish new observational technologies and use them to understand human and Earth system processes, such as land–atmosphere interactions, biogeochemical cycles, and subsurface soils, to estimate critical process parameters using novel analysis methods, such as machine learning and data science, and to quantify model errors.
	•			3.5	Characterize, understand, and model the complex, multiscale water cycle processes in the Earth system including the subsurface to understand and predict water availability and human system response to extremes.
	•			4.1	Characterize the biogeochemical exchanges driven by food web and plant-microbe interactions and evaluate their process-level impacts, sensitivity to disturbances, and shifting resource availability under changing environmental regimes.
	•			4.2	Define the sphere of influence and key elements of microbial communities in space and time relevant for predicting larger-scale ecosystem phenomena for Earth system understanding.
	•			4.3	Integrate molecular and process data to improve the ability to define ecologically significant traits of individual taxa and communities and use trait-based models to develop predictive links between community dynamics and ecosystem processes.
	•			4.4	Align and deepen connections among conceptual understanding, measurements and models related to the roles of microbes in determining the rate of transformation, uptake, and loss of chemical elements from ecosystems.
		•		6.1	Develop robust approaches for large-scale data collection, curation, annotation, and maintenance.
		•		6.2	Develop computing and software infrastructure to enable large-scale data storage and analysis.
		•		6.4	Engineer advanced computational modeling combined with data integration across temporal and spatial scales.
			•	7.1	Foster a spirit of collaboration to enable integrative capabilities among BER and SC user facilities, as well as other federal research facilities and infrastructure, thereby promoting a fully interdisciplinary approach to BER-relevant science.
			•	7.3	Develop innovative enabling technologies and construct and acquire state-of-the-art instruments that exploit the world-leading characteristics of each user facility. This will boost capabilities for basic research in biological systems and Earth and environmental science, thereby providing DOE and the nation with leading-edge capabilities for biological and environmental science.
	•			7.4	Develop multimodal imaging and remote sensing capabilities at user facilities for interrogating length scales ranging from atomic to mesoscale and time scales ranging from nanoseconds to days.
			•	7.5	Build upon existing investments and capabilities at the DOE-SC light and neutron science user facilities, continuing to align them with BER missions.

Notes—FSB = Functional and Systems Biology Science Area; ETI = Environmental Transformations and Interactions Science Area; CAM = Computing, Analysis, and Modeling Science Area; C&P = Operations for Capacity and Pace.

FSB	ETI	CAM	C&P		
					BERAC Grand Challenges (cont'd)
	•			7.6	Further develop the necessary infrastructure at user facilities to study organisms in their natural habitats.
•				7.7	Develop and adopt technologies to convert genome sequence data into functional understanding at appropriate BER user facilities.
•				8.1	Characterize the genotype and phenotype of individual cells, including genomics, transcriptomics, proteomics, and metabolomics to enable high-resolution predictive biology.
			•	8.2	Increase throughput and integration of genomics, transcriptomics, proteomics, and metabolomics to enable improved translation from the molecular to the cellular realm.
		•		8.5	Develop integrative and interpretive computational approaches that can handle large, disparate data types from multiple and heterogeneous sources using advanced and exascale computing.
					BERAC User Facility Recommendations
•				2.2	Develop structural libraries for metabolites and enzymes.
•				2.4	Develop methods for in situ measurements and single-cell measurements.
		•		2.6	Develop stoichiometric and kinetic models of metabolism that integrate omic data and allow the transition from observations of changes in gene expression to metabolic activity.
	•			2.14	Develop cellular sensors for monitoring metabolism and metabolic state in organisms and how they are influenced by their ecosystem.
•				2.25	Develop facilities to better characterize phenotypes resulting from altered gene function, including whole-organism and population growth and development, as well as high-resolution imaging and monitoring of metabolic changes.
•				2.27	Develop facilities so that researchers can perform imaging on a sample and then subject these samples to omics approaches.
	•			3.9	Develop a network of AmeriFlux omics-to-ecosystems supersites, where high–temporal resolution field and laboratory observations of omics, microhabitat-scale conditions, and fluctuating resources are generated automatically, and data are compared with ecosystem flux observations and models.
			•	3.11	Establish a joint facility activity among EMSL, JGI, and ARM, perhaps by extending existing Facilities Integrating Collaborations for User Science (FICUS) collaborations, to develop and implement a comprehensive observational strategy (field and laboratory) to measure and discern modes of ice nucleation under real atmospheric conditions.
	•			3.18	Establish a User Facility to enable manipulative experiments at field-relevant scales that are critical for advancing our understanding of the linkages between physical and biological systems and across scales of organization, from molecules to habitats to ecosystems.
			•	3.25	Further develop and implement a framework for joint calls, review, and decision making (perhaps via the FICUS program): (1) across multiple User Facilities to enable and incentivize cross-disciplinary research to address joint research priorities and Grand Challenges and (2) across User Facilities and appropriate science programs to ensure the availability and effective use of scientific resources. The primary focus for such a framework may be internal to BER, but it should also consider engagement from external agencies and facilities. Such joint calls could be supported through dedicated crosscutting budgets for integrative research.
		•		4.4	Enable process modeling and data-related computation by investing in midrange computing infrastructure and personnel time.
	•			4.5	Develop a robust computational framework that can connect and inform models at multiple scales and that facilitates iteration based on input from experimental and field data and modeling output.
	•			4.6	Develop field deployable, multimodal, remotely controlled sensors that ideally conduct nondestructive measurements to (1) characterize how microbial habitat–scale heterogeneity and dynamics influence biogeochemical processes and (2) validate relevance of lab experiments in field.
		•		6.1	Provide tools at facilities for labeling, metadata management, and data discovery both within one facility and across DOE and non-DOE facilities.
		•		6.3	Develop an infrastructure strategy that addresses data analysis and storage needs.

Notes—FSB = Functional and Systems Biology Science Area; ETI = Environmental Transformations and Interactions Science Area; CAM = Computing, Analysis, and Modeling Science Area; C&P = Operations for Capacity and Pace.

FSB	ETI	CAM	C&P		BERAC User Facility Recommendations (cont'd)
		•		6.6	Work with the research community and computational facilities to determine the hardware, software, and usage policies needed to support researchers' complex workflows.
		•		6.7	Address the needs of real-time streaming data and interactive computing as part of the recommended infrastructure strategy.
					EESSD Grand Challenges/BSSD Goals
	•			EESSD-1	Integrated Water Cycle: Advance understanding of the integrated water cycle by studying relevant processes involving the atmospheric, terrestrial, oceanic, and human system components and their interactions and feedbacks across local, regional, and global scales, thereby improving the predictability of the water cycle and reducing associated uncertainties in response to short- and long-term perturbations.
	•			EESSD-2	Biogeochemistry: Advance a robust, predictive understanding of coupled biogeochemical processes and cycles across spatial and temporal scales by investigating natural and anthropogenic interactions and feedbacks and their associated uncertainties within Earth and environmental systems.
	•			EESSD-5	Data–Model Integration: Develop a broad range of interconnected infrastructure capabilities and tools that support the integration and management of models, experiments, and observations across a hierarchy of scales and complexity to address CESD scientific grand challenges.
•				BSSD-1	Provide the basic science needed to convert renewable biomass to a range of fuels, chemicals, and other bioproducts in support of a burgeoning bioeconomy.
	•			BSSD-1-1	Gain a genome-level understanding of plant metabolism, physiology, and growth to develop new bioenergy feedstocks with traits tailored for bioenergy and bioproduct production.
	•			BSSD-1-2	Develop an understanding of microbial and fungal metabolism necessary to design new strains, communities, or enzymes capable of converting plant biomass components into fuels, chemicals, and bioproducts.
	•			BSSD-1-3	Understand the genomic properties of plants, microbes, and their interactions to enable the development of new approaches that improve the efficacy of bioenergy crop production on marginal lands with few or no agricultural inputs, while minimizing ecological impacts under changing environmental conditions.
•	•			BSSD-3	Develop a process-level understanding of microbiome function and be able to predict ecosystem impacts on the cycling of materials (carbon, nutrients, and contaminants) in the environment.
		•		BSSD-4	Support the development of computational and instrumental platforms to enable broader integration and analysis of large-scale complex data within BER's multidisciplinary research efforts.
		•		BSSD-4-1	Create open-access and integrated computational capabilities tailored to large-scale data science investigations for molecular, structural, genomic, and omics-enabled research on plants and microorganisms for a range of DOE mission goals.
•				BSSD-4-2	Improve or develop new multifunctional, multiscale imaging and measurement technologies that enable visualization of the spatiotemporal and functional relationships among biomolecules, cellular compartments, and higher-order organization of biological systems.
			•	BSSD-5	Build unique, best-in-class capabilities within Office of Science user facilities (including JGI, EMSL, and DOE's light and neutron sources) to enhance the multidisciplinary Bioenergy Research, Biosystems Design, and Environmental Microbiome Research supported by the Division.
			•	BSSD-5	Broaden the integrative capabilities within and among DOE user facilities to foster a more interdisciplinary approach to BER-relevant science and aid interpretation of plant, microbe, and microbial community biology.

Notes—FSB = Functional and Systems Biology Science Area; ETI = Environmental Transformations and Interactions Science Area; CAM = Computing, Analysis, and Modeling Science Area; C&P = Operations for Capacity and Pace.


EMSL