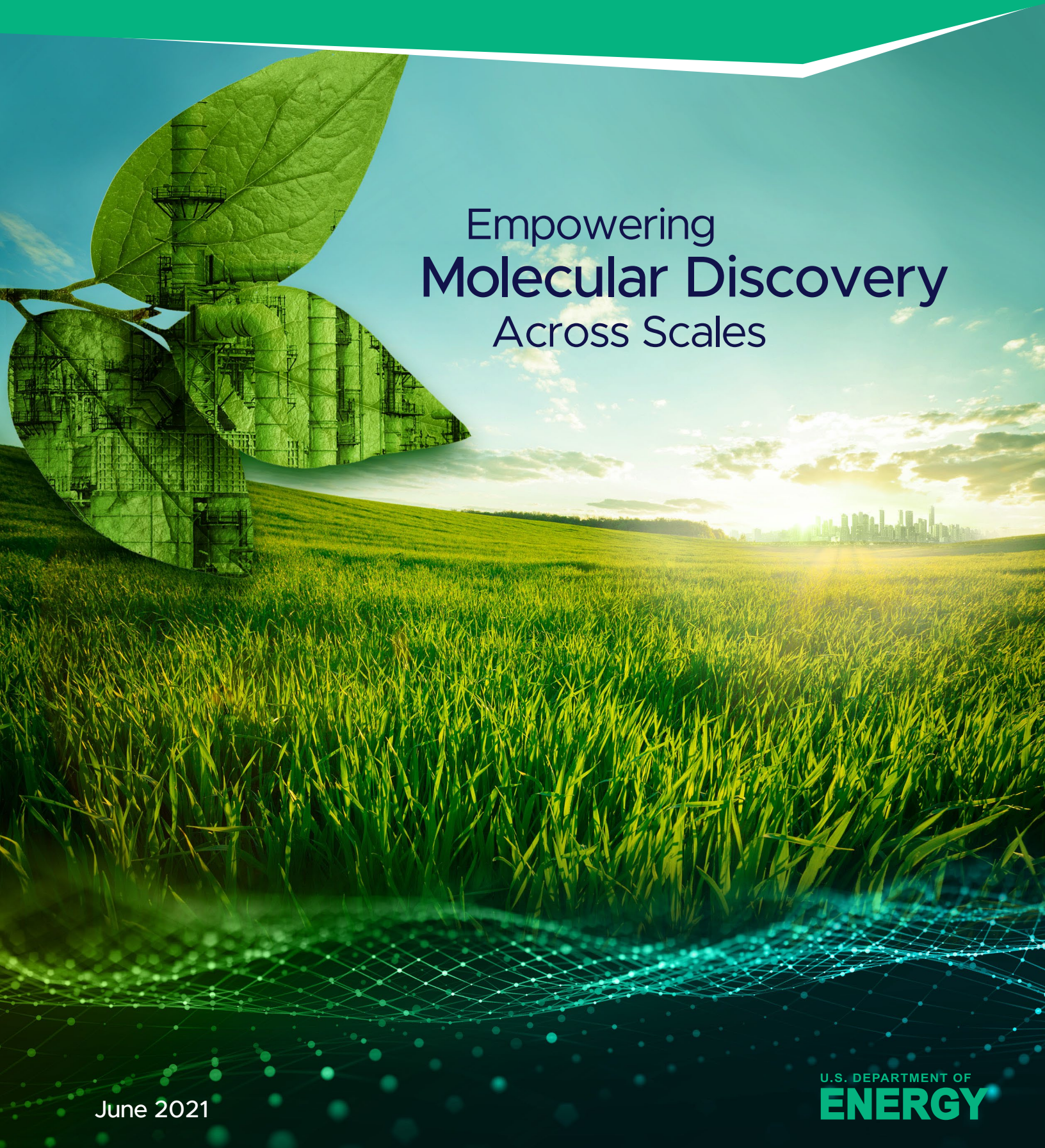
EMSL

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
EMSL	Environmental Molecular Sciences Laboratory
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

2022 ADDENDUMS

As part of its 2021 U.S. Department of Energy (DOE) triennial review, the Environmental Molecular Sciences Laboratory (EMSL) received three Major Actions related to the strategy presented in the 2021 Strategic Science Plan (this document). The outputs from the EMSL Plan of Actions and Milestones (POAMs) response to these actions are captured as three addendums to the strategic plan: one focused on additional detail in EMSL's Strategic Science Objective planning; one on defining a partnership framework that supports EMSL's future directions and objectives outcomes; and one on clarifying the integration between Integrated Research Platforms and EMSL's Strategic Science Objectives. These addendums are attached to the original EMSL Strategic Science Plan.

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;

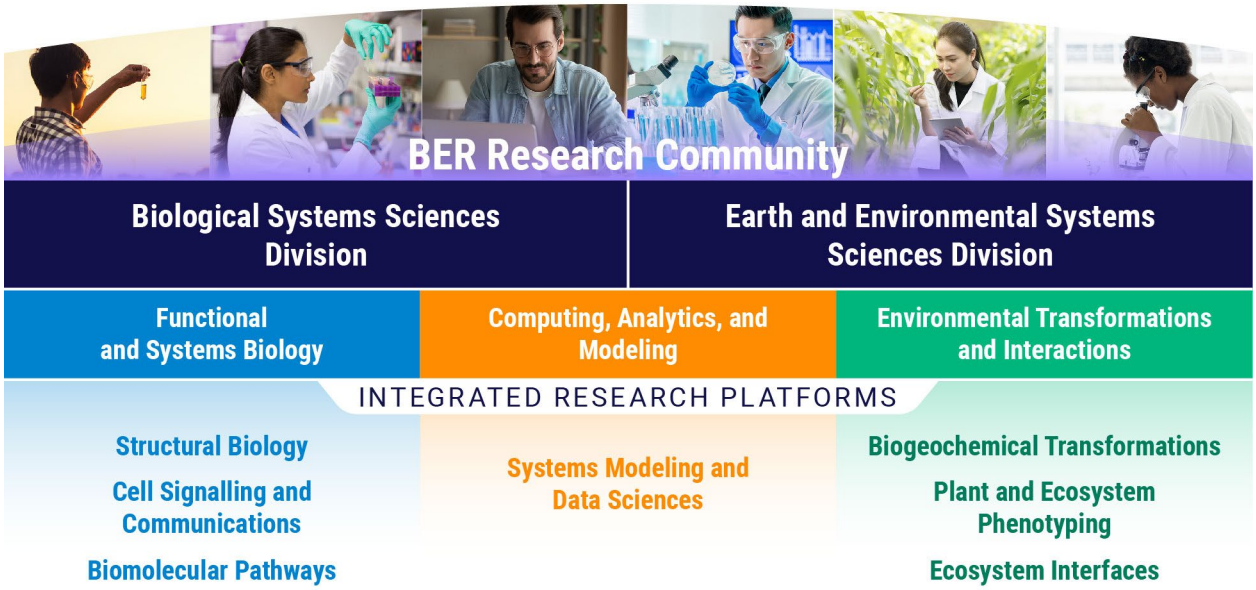


Figure 3. EMSL's science leadership is organized to empower the BER user community. The three science areas, Functional and Systems Biology, Computing, Analytics, and Modeling, and Environmental Transformations and Interactions, assure continuous alignment and close partnership between EMSL and BER's 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

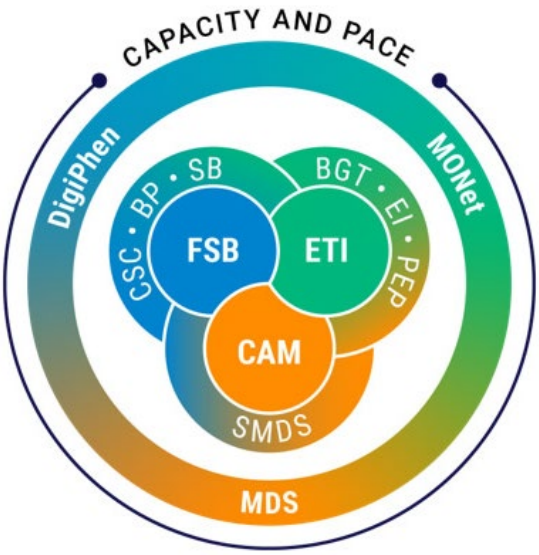


Figure 4. EMSL's Operations for Capacity and Pace objective provides direction for infrastructure development and operational activities that accelerate delivery of a suite of transformational capabilities for users through its three strategic science objectives DigiPhen, MONet, and MDS.

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.

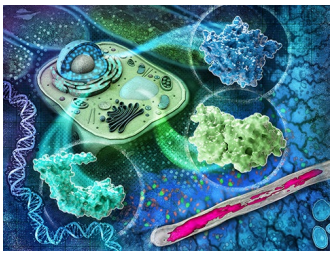
Fostering multidisciplinary user science and expanding user access to our integrative capabilities is a major goal of EMSL's strategy. EMSL's IRPs and scientists that lead the IRPs play a critical role in delivering this goal.

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.

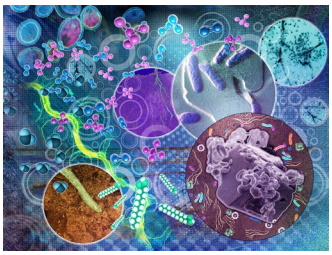
The seven IRPs are Structural Biology, Biomolecular Pathways, Cell Signaling and Communication, Biogeochemical Transformations, Ecosystem Interfaces, Plant and Ecosystem Phenotyping, and Systems Modeling and Data Sciences, detailed below.



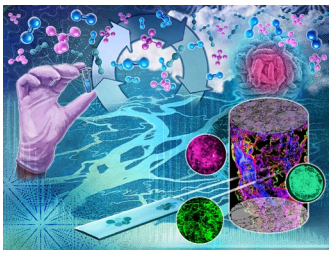
The **Structural Biology IRP** seeks structural, biochemical, and dynamic information about proteins, protein complexes, and other biomolecules at nanoscale spatial and temporal resolutions to infer function.

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.

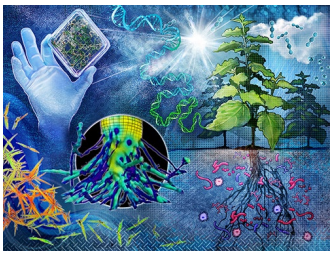
The **Cell Signaling and Communication IRP** reveals dynamic interactions and trafficking of molecular signals between cells, populations, and communities to understand complex interrelationships between organisms in response to their environment.



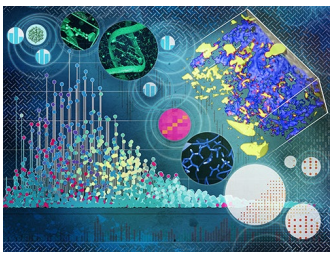
The **Biogeochemical Transformations IRP** investigates the biochemical, physical, and microbial interactions that affect chemical speciation, transport, and transformation of critical nutrients, contaminants, and compounds within the environment.



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

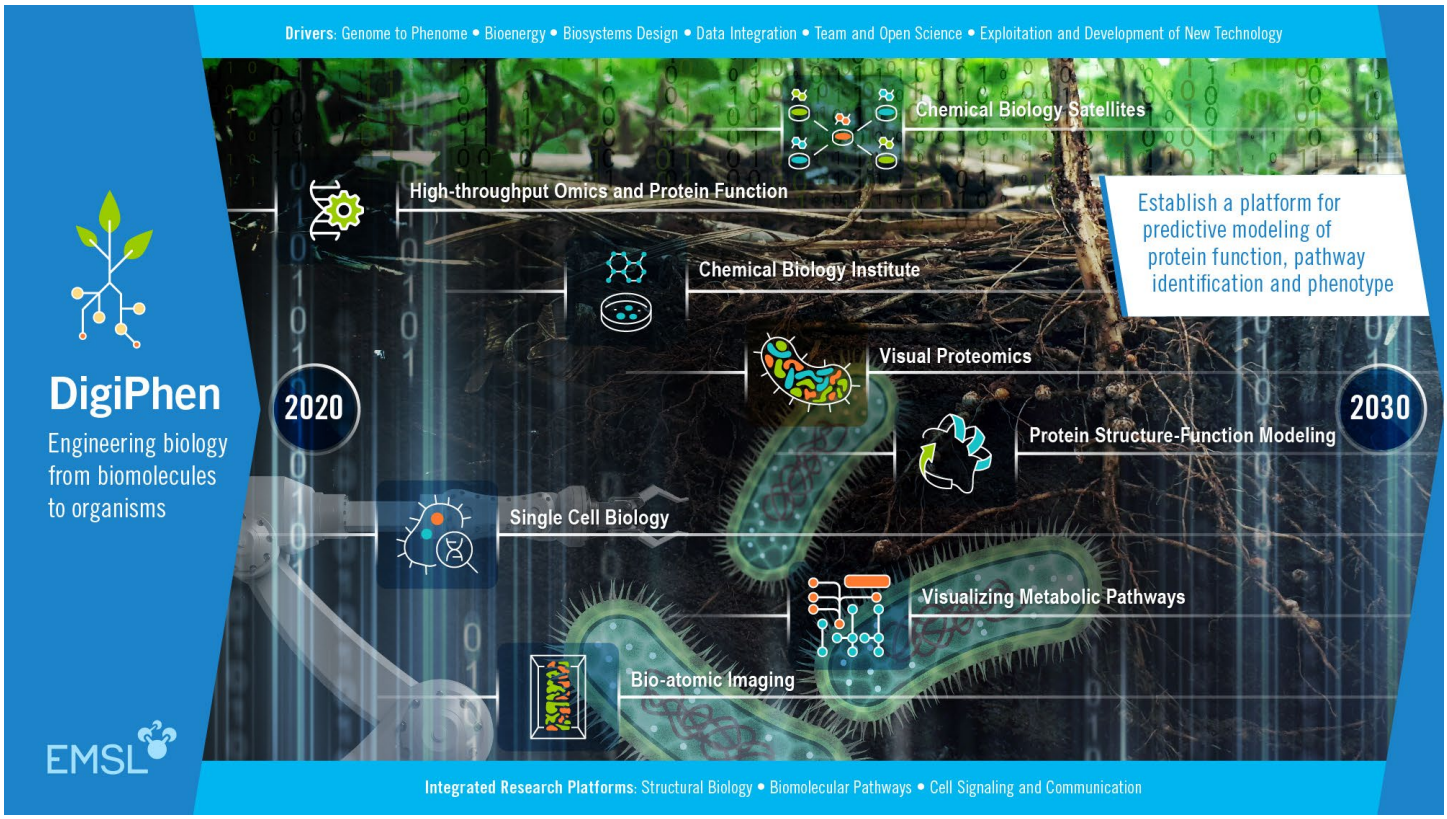


Figure 5. Overview, timeline, and research areas supporting Strategic Science Objective 1. This objective establishes EMSL's leadership in the development of a new generation of science and technology innovations that produce new knowledge about the molecular, cellular, and community foundation of phenotype 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

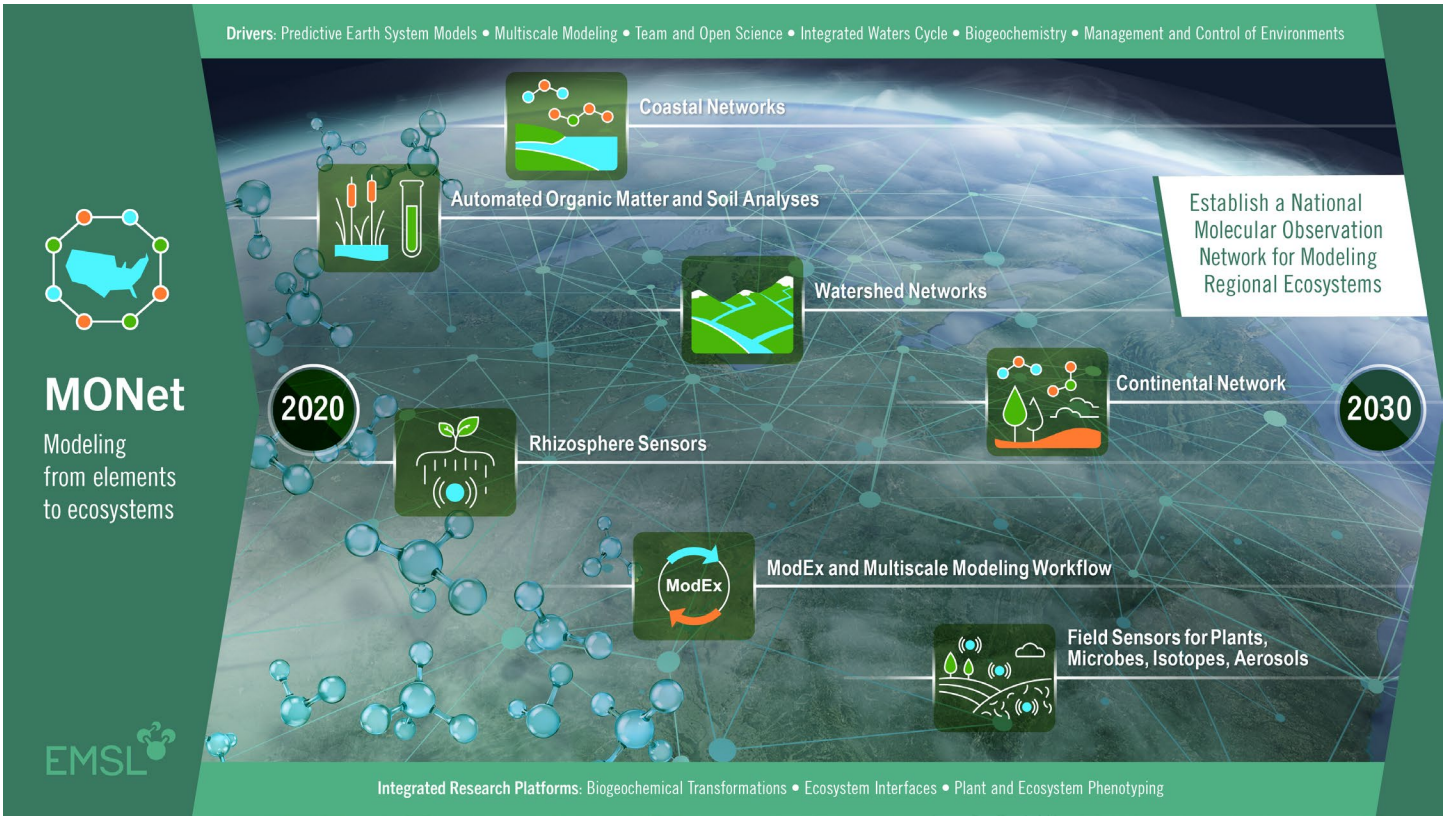


Figure 6. Overview, timeline, and research areas supporting Strategic Science Objective 2. Establishing a national molecular observations network will provide high-quality, interoperable, and integrated molecular-level information on key ecosystem processes at local and continental scales to enable predictive 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).

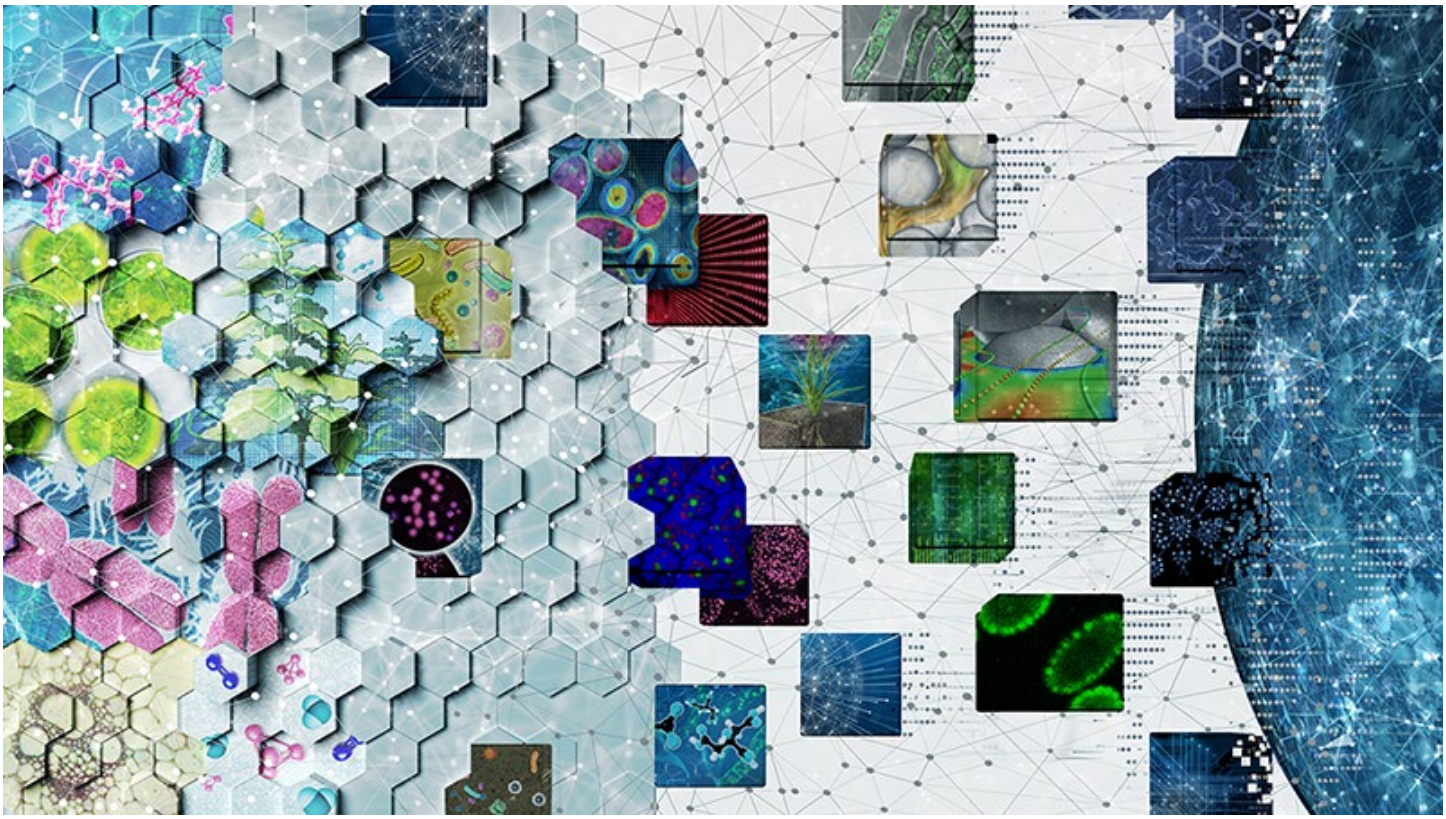
Supporting IRPs: Plant and Ecosystem Phenotyping, Ecosystem Interfaces, Cell Signaling and Communication, and Systems Modeling and Data Science

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

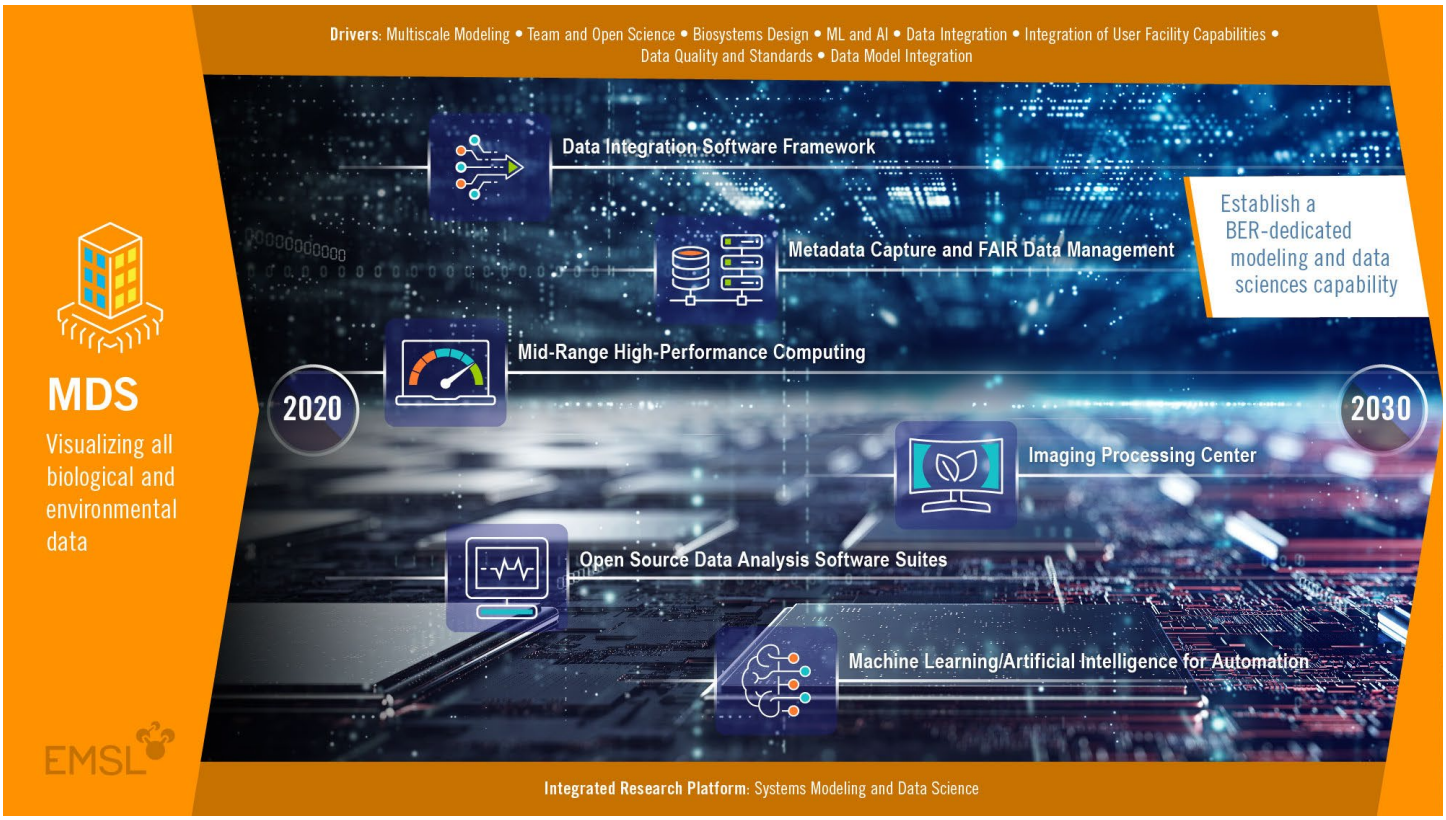


Figure 7. Overview, timeline, and research areas supporting Strategic Science Objective 3. The research areas within this objective build capacity, methodology, software, and new capabilities that together accelerate translation of data into knowledge by enabling integration visualization and utilization 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.

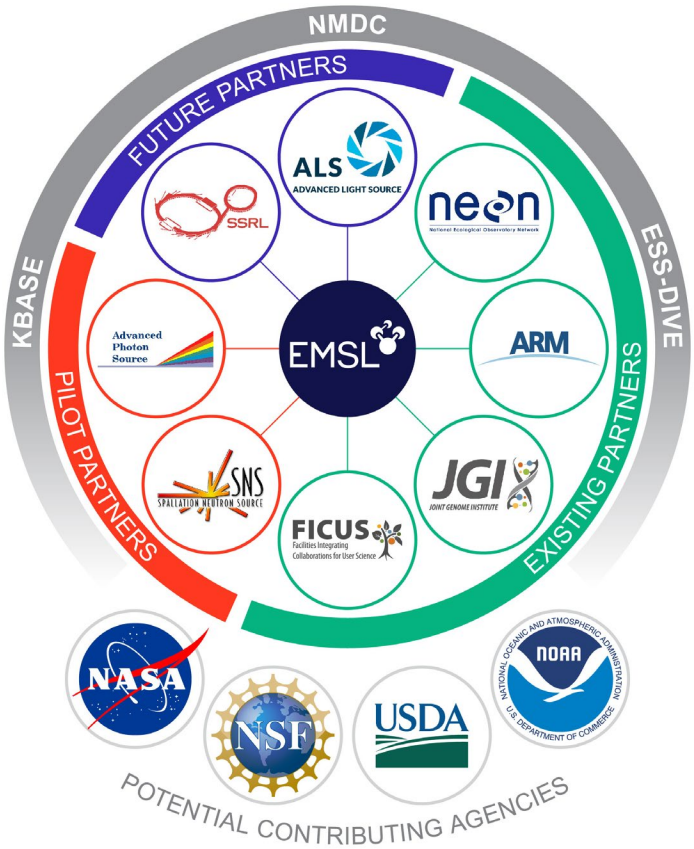


Figure 9. EMSL's growing landscape of partnering organization and agencies. EMSL has a broad, productive, and growing group of institutional partners that amplify and increase the impacts of EMSL research and capabilities.

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

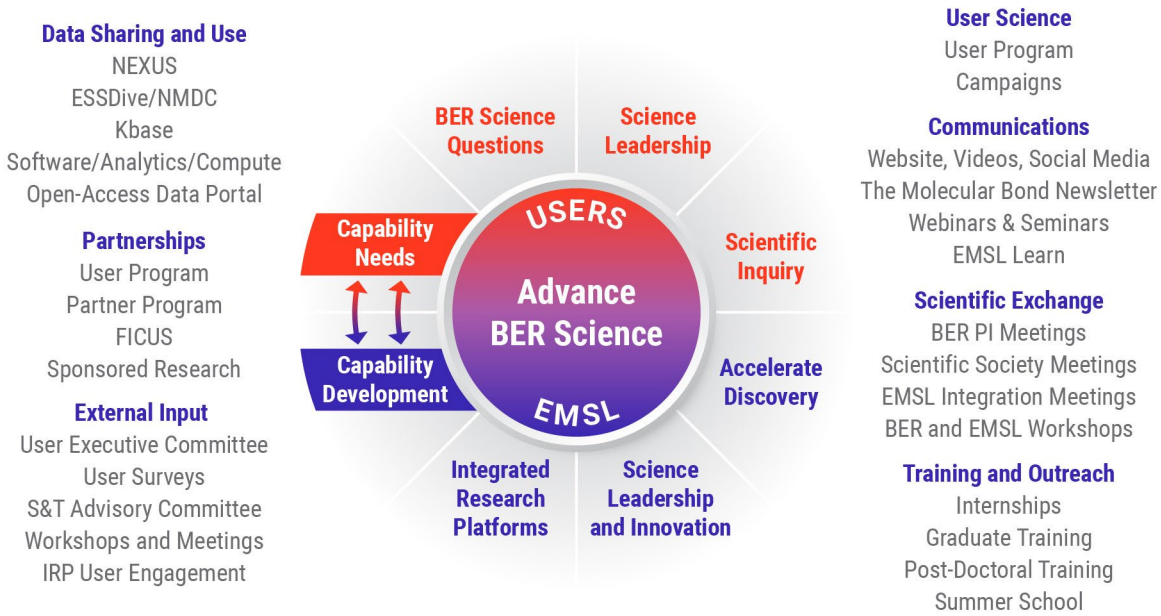


Figure 10. EMSL's strategy to engage and empower the user community. The multiple forms of engagement, communications, and partnership provide diverse forums for collaboration and recruitment of new users.

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

Addendum 1 Strategic Science Objective Planning

EMSL

2022 POAM S1

SSO Planning Addendum to the EMSL Strategic Plan

January 2023

Jay Bardhan, John Bargar, Scott Baker, Justin Teeguarden, Douglas Mans

PNNL-34044

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

2022 POAM S1

SSO Planning Addendum to the EMSL Strategic Plan

January 2023

Jay Bardhan, John Bargar, Scott Baker, Justin Teeguarden, Douglas Mans

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

As part of the 2021 Triennial Review, the Department of Energy Office of Science assigned the Environmental Molecular Sciences Laboratory (EMSL) a Major Action that would provide more detailed implementation plans for EMSL's three Strategic Science Objectives (SSOs): the Digital Phenome (DigiPhen), the national Molecular Observations Network (MONet), and the Modeling and Data Sciences (MDS) capability. *The SSOs direct and focus EMSL's efforts toward developing leading capabilities for the Biological and Environmental Research (BER) program user community and toward building science leadership. Ultimately, the planning and execution of the SSOs are directed at producing new scientific discoveries and capabilities for the BER research community.* The 2021 *EMSL Five-Year Strategic Science Plan* (hereafter, the Strategic Plan) presented high-level implementation plans for the SSOs in the form of a series of Research Areas, including planned and current activities and partners. Beyond the Strategic Plan, EMSL continues to develop and implement Research Area Roadmaps with the BER user community that collectively form the more detailed implementation plans for the SSOs. Each roadmap lays out across time the series of required science and technology developments, hires, facilities, instruments, and other resources, including partners.

In the Plan of Actions and Milestones (POAM) for Major Action S1, EMSL expands on the descriptions of our SSO implementation plans by developing an addendum to the Strategic Plan that provides additional details on our planning process, activities and leadership, Research Area roadmapping, and examples from each SSO of planning activities and outcomes. Major objectives and milestones are provided for each SSO. Examples demonstrating progress evolving the SSOs since completion of the Triennial Review Response, a timeline for developing additional Research Area Roadmaps, and comprehensive revisions to the Computing, Analytics, and Modeling Science Area, the MDS SSO, and the supporting Research Areas are also presented in this Strategic Plan Addendum.

Acronyms and Abbreviations

AI	artificial intelligence
BER	Biological and Environmental Research program
BERAC	Biological and Environmental Research Advisory Committee
CAM	Computing, Analytics, and Modeling
CSO	chief science officer
CD	Critical Decision
CP	Computing Platforms
DigiPhen	Digital Phenome
DS	Data Sciences
EESSD	Earth and Environmental System Sciences Division
EMSL	Environmental Molecular Sciences Laboratory
ETI	Environmental Transformations and Interactions
FICUS	Facilities Integrating Collaborations for User Science
FSB	Functional and Systems Biology
FTICR	Fourier-transform ion cyclotron resonance
FY	fiscal year
HPC	high-performance computing
HTP	high throughput
IRP	Integrated Research Platform
M2PC	Microbial Molecular Phenotyping Capability
MALDI-MS	matrix-assisted laser desorption/ionization mass spectrometry
MDS	Modeling and Data Sciences
MEDS	ModEx-Driven Simulations
ML	machine learning
ModEx	Model-Experiment Integration
MONet	Molecular Observation Network
MS	mass spectrometry
PI	principal investigator
PNNL	Pacific Northwest National Laboratory
POAM	Plan of Actions and Milestones
R&D	research and development
RF	Rhizosphere Function
S&T	science and technology
SSO	Strategic Science Objective
TAP	Terrestrial-Atmospheric Processes

Contents

Figures

Tables

1.0 Background

As a part of the 2021 Triennial Review, the Department of Energy, Office of Science assigned the Environmental Molecular Sciences Laboratory (EMSL) three Major Actions under the heading of Science. The first of these focused on detailed planning for EMSL's Strategic Science Objectives (SSOs):

> "*A POAM is needed that contains more detailed (implementation) plans for the SSOs in general, and especially for DigiPhen and MONet, the two primary SSOs, but also for MDS within the CAM SSO. Because many of the reviewers were science-focused, they had quite a few comments for EMSL management to consider, including the potential to learn from and also leverage on-going BER-funded activities, metrics for success, identification of science questions to drive the efforts, the likely use of existing EMSL capabilities at the potential expense of the user program, and a variety of other issues that needed clarification. Although the reviewers had few comments on MDS directly, they were supportive of the initial directions for this SSO. Nevertheless, the MDS SSO needs a considerable amount of planning, which should be included in this POAM. Reviewer comments in red text in section II of this attachment identify the variety of issues associated with the SSOs that need clarification as EMSL moves forward with SSO implementation*."

The 2021 *EMSL Five-Year Strategic Science Plan* (hereafter, the Strategic Plan) presented three SSOs: the Digital Phenome (DigiPhen), the national Molecular Observations Network (MONet), and the Modeling and Data Sciences capability (MDS). These SSOs provide focus for EMSL's three Science Areas—Functional and Systems Biology (FSB), Environmental Transformations and Interactions (ETI), and Computing, Analytics, and Modeling (CAM). The Strategic Plan presented high-level implementation plans for the SSOs in the form of a series of Research Areas, including planned and current activities and partners. Beyond the Strategic Plan, EMSL continues to develop and implement Research Area Roadmaps with the Biological and Environmental Research (BER) program user community that collectively form the more detailed implementation plans for the SSOs. Each roadmap lays out across time the series of required science and technology (S&T) developments, hires, facilities, instruments, and other resources, including partners.

In the Plan of Actions and Milestones (POAMs) for Major Action S1, EMSL expands on the descriptions of our SSO implementation plans by developing an addendum to the Strategic Plan that provides additional details on our planning process, activities and leadership, Research Area roadmapping, and examples from each SSO of planning activities and outcomes. Because the detailed plans for each SSO are closely linked to our partner planning and our Integrated Research Platform (IRP) planning POAMs, appropriate references and connections to these POAM responses will also be made.

2.0 Strategic Science Objective Planning

2.1 Overview

EMSL established three SSOs and one Strategic Operations Objective based on extensive input from 50 senior science leaders, subject matter experts, members of the EMSL User Executive and Science & Technology Committees, and representatives of the user community from across EMSL, Pacific Northwest National Laboratory (PNNL), Lawrence Berkeley National Laboratory, Oak Ridge National Laboratory, industry, and academic institutions, 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). The Strategic Plan describes the three SSOs (Sections 3, 4, and 5)—DigiPhen, MONet, and MDS, respectively—and our implementation plans. EMSL has one SSO aligned to each foundational Science Area, maintaining alignment with BER science foci—DigiPhen is aligned with FSB, MONet with ETI, and MDS with CAM.

The SSOs direct and focus EMSL's efforts toward developing leading capabilities for the BER user community and building science leadership. Ultimately, planning and execution of the SSOs are directed at producing new scientific discoveries and capabilities for the BER research community.

EMSL employs a plan, develop, and deploy model to produce SSO-aligned capabilities for the BER research community (Figure 1). Within this model, EMSL has established a high-level implementation plan, a planning process centered on S&T Roadmaps for specific Research Areas detailed activities that drive execution of those plans, and leadership assigned to the activities (Figure 1).

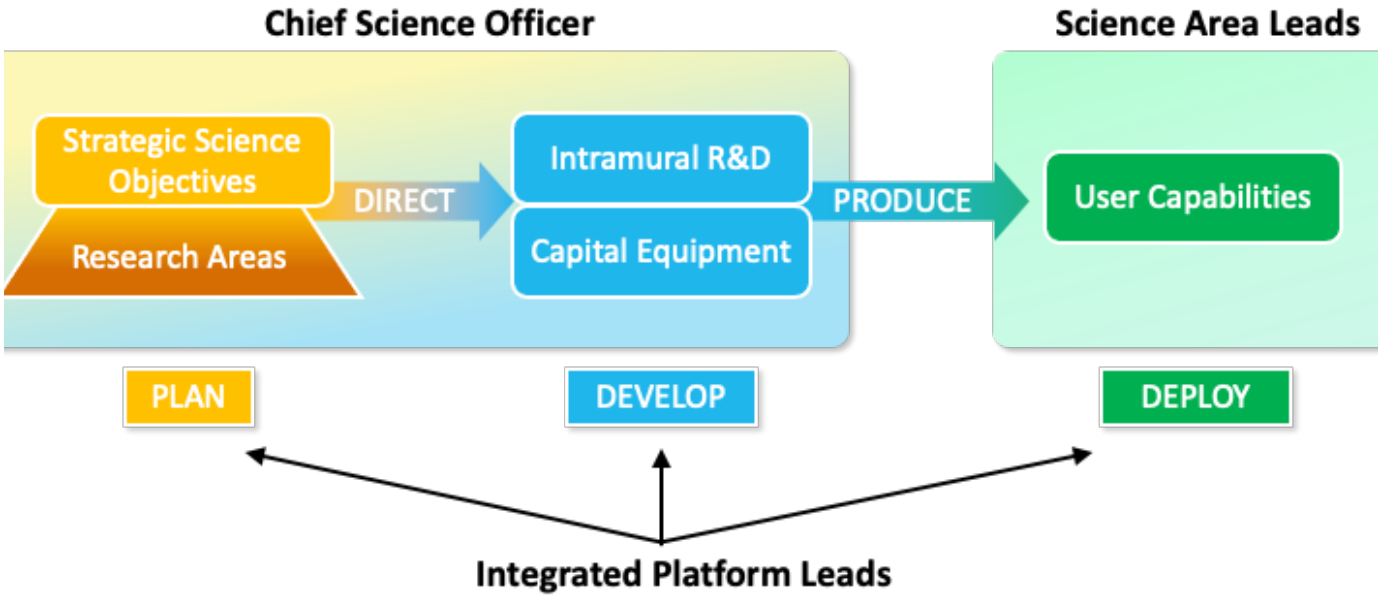


Figure 1. Summary of EMSL's plan, develop, and deploy model for creating new SSO-aligned capabilities for the BER research community. The detailed implementation plans are developed as part of the Research Area roadmapping and S&T research and development planning activities.

2.2 SSO Implementation Plans

Implementation plans for the three SSOs take the form of a series of layered S&T Roadmaps, with the level of detail increasing in each layer (see Box 1). The layered S&T roadmaps are informed by BER research community needs and the supporting implementation activities to sharpen BER alignment, increase user engagement, and assure development and deployment of new user capabilities.

Box 1: Elements of EMSL's S&T Roadmaps

Strategic Science Objectives: Provides scientific direction and focus within the broader Science Area that builds science leadership, evolves multidisciplinary science, and advances EMSL's premier capabilities to meet the current and future needs of users. A bold scientific objective provides scientific direction and focus for EMSL. EMSL has one SSO aligned to each foundational Science Area, maintaining alignment with BER science focus—DigiPhen (FSB aligned), MONet (ETI aligned), and MDS (CAM aligned).

Research Areas: A specific area of research and development needs that EMSL leads through S&T research projects, strategic partnerships, and user program calls to develop new capabilities. Each Research Area has its own goals and strategic roadmap that drive delivery of a fundamental element of an SSO. EMSL has 15 Research Areas, grouped by the SSO they support (see Figures 2, 3, and 4).

Research Area Roadmap: The detailed plan, including a timeline of specific S&T research and development (R&D) activities and supporting resources (facilities, instruments, and staff) for advancing a Research Area. Used to guide the Intramural S&T Research program, capital and expense equipment purchases, facilities planning, and hiring.

An initial collection of 21 Research Areas developed during the EMSL 2020 strategic S&T roadmapping activity (EMSL 2021 Strategic Science Plan, Section 3.2, 4.2 and 5.2), has now been revised to 14 and form the high-level implementation plan for each SSO. The group of Research Areas for each SSO is presented in the SSO graphics for DigiPhen, MONet, and MDS (Figures 2, 3, and 4). Each Research Area in turn represents a collection of S&T directions that, in the timeline proposed, together represent a roadmap for delivery of an SSO and the associated user capabilities. Research Area descriptions for each SSO were initially described in the Strategic Plan. Revised descriptions of some are found in Appendix A.

The more specific and more detailed implementation plans take the form of S&T roadmaps developed for each of the Research Areas (Box 1, Appendix A [completed roadmaps]). The process for developing these Research Area Roadmaps is presented in Section 2.4 of this addendum. These Research Area Roadmaps present a series of specific S&T R&D activities organized into thrusts and arrayed in time. Supporting resources, such as facilities, instruments, partnerships, and new staff, required to successfully complete the S&T activities and advance the Research Area in support of the SSO are also identified. EMSL's Research Area Roadmaps for the SSOs are in various forms of development. Six priority and near-term Research Area Roadmaps have been drafted—five of these have undergone external review in workshops led by EMSL, four have been revised accordingly (see Appendix A), and one (Model-Experiment Integration [ModEx]) is now undergoing additional revision as it moves to the MDS SSO.

The Research Area Roadmaps guide decision-making across EMSL. The S&T activities detailed in the thrusts of each roadmap are used to set priorities, create call topics, and guide review of projects for the EMSL

Intramural S&T Research program. This assures that the Intramural S&T Research program produces capabilities for users that align directly with the priorities established in the Strategic Plan, presented as SSOs. The Research Area Roadmaps are also used to guide allocation of resources for maintenance and improvement of existing capabilities, make new instrument purchasing decisions (capital and expense), plan for and make strategic hires, and build partnerships that provide critical complimentary technical expertise and capabilities. Maintenance, improvement, and purchase of capabilities is also supported by instrument life-cycle planning, stewarded by the chief operations officer as part of the Capacity and Pace Strategic Operational Objective. This coordination of EMSL's SSOs and Strategic Operational Objective leverages and amplifies the implementation activities for more effective stewardship, development, and delivery of capabilities for the user program.

EMSL's Intramural Research program is the final element of EMSL's implementation plans for the SSOs. The intramural program ultimately drives development of capabilities for the user program aligned with each of our SSOs. Thus, the detailed research plans in each intramural or partner program project provide the final and finest level of detail for our SSO implementation plans (Figure 1). These research plans detail the specific research activities that must be undertaken to deliver a new capability to users, and the timeline and nature of early engagement by users or others from the broader BER user community (see also Attachment 6, Section 2.0, Figure 1 for information on execution of the intramural program, how alignment is maintained with BER, and how users engage and influence the program).

2.3 SSO Planning Activities and Leadership

EMSL has established and implements multiple activities that review, revise, and/or execute our SSO implementation plans. Each activity sharpens BER alignment, increases user engagement, and improves the development of new capabilities for the user program. Many of the activities directly or indirectly involve a review of one or more levels of the implementation plans used to evolve the plans over the course of the year.

Research Area Roadmapping: This is the central element of EMSL's SSO planning process. The chief science officer (CSO) is accountable for these roadmapping efforts, which are led by the IRP leaders. Section 2.4 provides a detailed description of EMSL's roadmapping process.

Bi-Weekly S&T Leadership Meetings. The CSO leads bi-weekly S&T strategy meetings with the chief data officer, chief operations officer, Science Area leaders, and IRP leaders. These meetings are our primary venue for evaluating and evolving our implementation plans, planning Intramural S&T Research program call topics, selecting strategic hiring targets, prioritizing equipment purchases, managing the Intramural S&T Research portfolio, coordinating with operations and User Office activities, and identifying points of synergy and opportunities for leveraging and amplification across activities, including user projects and BER research (e.g., Science Focus Areas). Research Area Roadmaps and SSOs guide these discussions and decision-making. During these meetings, we specifically:

- Assess the relevance of Research Areas to the SSOs and to BER. Consider matriculation and revision of Research Areas (see Section 2.4 for examples).
- Assess the relevance of IRP capabilities and alignment with SSOs and the BER research community. Revise accordingly.
- Assess Research Area Roadmaps, plan for external review of Research Areas, and revise accordingly.

Intramural S&T Research Program Call Topic Planning: The Intramural S&T Research Program is the principal instrument for developing capabilities and is a key element of our SSO execution plans. The CSO leads annual

meetings with the Science Area leaders and IRP leaders to craft call topics for the EMSL Intramural S&T Research program that reflect the priorities specified in the Research Area Roadmaps.

Intramural S&T Research Project Review: Project review and management is an important element of EMSL's SSO implementation and execution plan. More than 100 quarterly project and other milestone review meetings are held each year with Intramural S&T Research project principal investigators (PIs) to evaluate progress in implementation, reassess feasibility, assure continued alignment with our Research Area Roadmaps, and find the earliest opportunities for user or BER researcher engagement/collaboration. The CSO leads these meetings. Progress is tracked and the portfolio and connections between projects are managed in SharpCloud.

User Call Topic Planning: EMSL deploys SSO-related capabilities through EMSL's user program. Science Area leaders, IRP leaders, and the User Services manager meet as necessary to craft call topics that connect users to new capabilities produced by executing the Research Area Roadmaps, such as those developed in the EMSL Intramural S&T Research program. The CSO provides the User Services manager with an evolving list and timeline of new user capabilities emerging from the Intramural Science and Technology Development Program for inclusion in call topics.

2.4 Research Area Roadmapping

An initial set of Research Areas supporting EMSL's three SSOs were established in the Strategic Plan. These Research Areas are initiated at a logical time in the developmental cycle for each SSO—some were initiated in 2021, and some will be initiated in each of the following 2 years. Table 1 presents the list of existing Research Areas, their level of development, and planned completion dates. Significant updates to the original 21 Research Areas were made as part of our continued SSO planning process.

Development and revision of Research Area Roadmaps is stewarded by IRP leaders. IRP leaders convene an internal workshop of subject matter experts from PNNL, approved by the CSO. Representation by other IRP leaders is included when the Research Area Roadmaps include S&T activities that involve or integrate multiple capabilities, such as instrument development, data processing, and high-performance computing. The resulting draft Research Area Roadmap is reviewed by the CSO and the S&T Leadership team and is subsequently revised as needed.

These draft Research Area Roadmaps undergo an external review by a panel of 5–8 domain experts drawn principally from the BER research community. The CSO convenes the review panels, which are then developed and led by the appropriate IRP lead. Feedback from the review panel is used to further revise the roadmaps. External reviews are planned for every few years to assess progress toward goals, assess the continued relevance of the proposed efforts, and evolve the thrusts and activities as necessary.

Six internal Research Area Roadmap workshops and five external Research Area Roadmap workshops were held in the first two quarters of fiscal year (FY) 2022:

- Rhizosphere sensors
- Single-cell biology
- Bio-atomic imaging
- Model-Experiment Integration (ModEx)
- High-throughput omics and protein function
- Visual proteomics (internal only).

EMSL's internal review of existing Research Areas resulted in several revisions to improve alignment with our SSOs and BER research priorities. For the MONet SSO (Figure 2, Table 1), the Coastal Networks, Watershed Networks, and Continental Networks Research Areas were combined into the single Field Sensing and Sampling Site Network Research Area. The Field Sensors for Plants, Microbes, Isotopes, and Aerosols Research Area was refocused to be more complimentary to the companion Field Sensing and Sampling Site Network Research Area. This new Research Area is called Field Sensing, Sampling, and Sensors to better reflect the need for S&T advancements in these three areas. The Field Sensors for Plants, Microbes, Isotopes, and Aerosols was further focused and aligned with the new Terrestrial-Atmospheric Processes IRP to produce the Atmospheric Aerosols and Particles Research Area.

In addition, as part of the planning process for DigiPhen and MDS, significant revisions were made to the original supporting Research Areas (Figure 3, Figure 4, Table 1). Additional details are provided in Section 3.0 of this addendum.

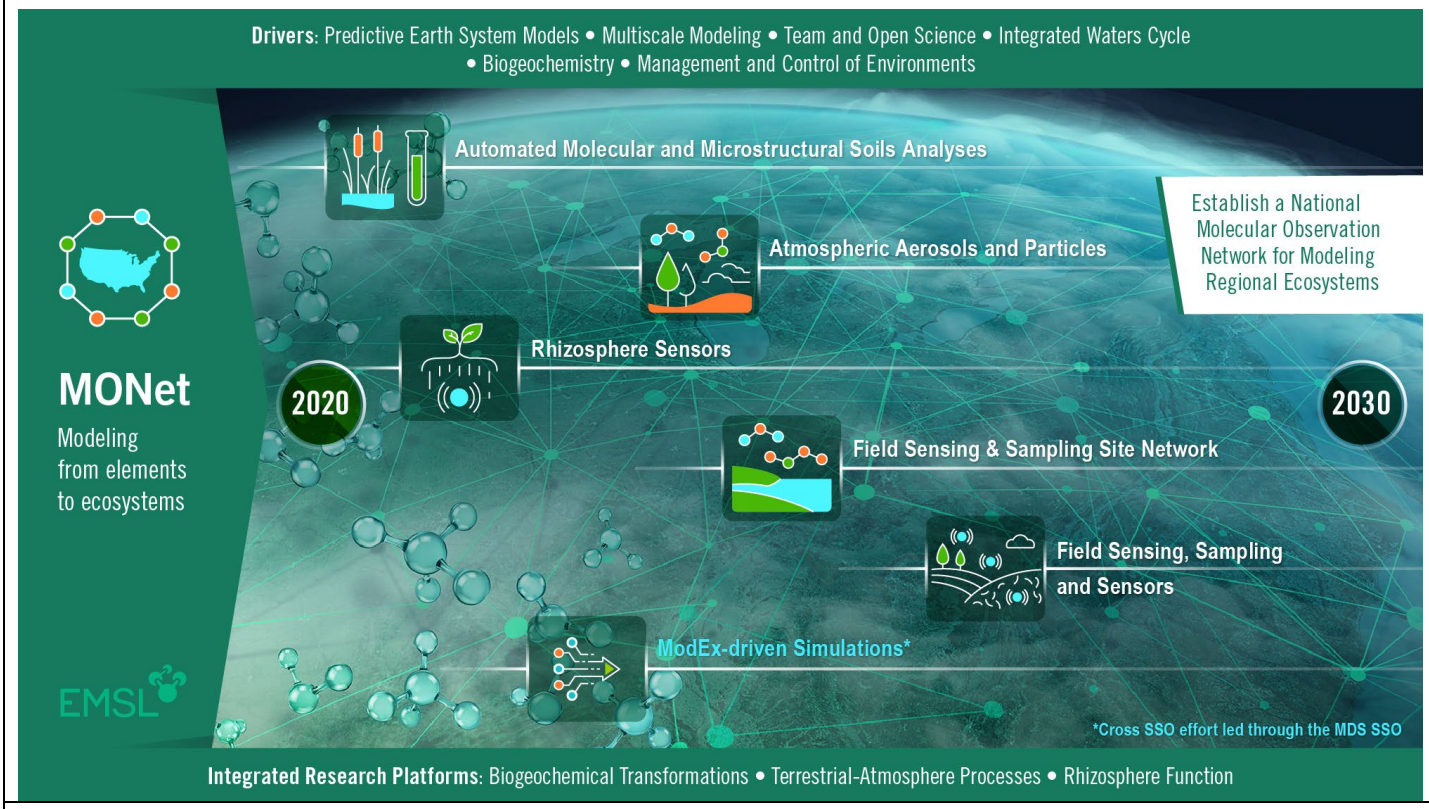


Figure 2. Overview, timeline, and Research Areas supporting MONet. Establishing a national Molecular Observations Network will provide high-quality, interoperable, and integrated molecular-level information on key ecosystem processes at local and continental scales to enable predictive modeling of ecosystem 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, although we expect work to begin before and to continue after, as required.

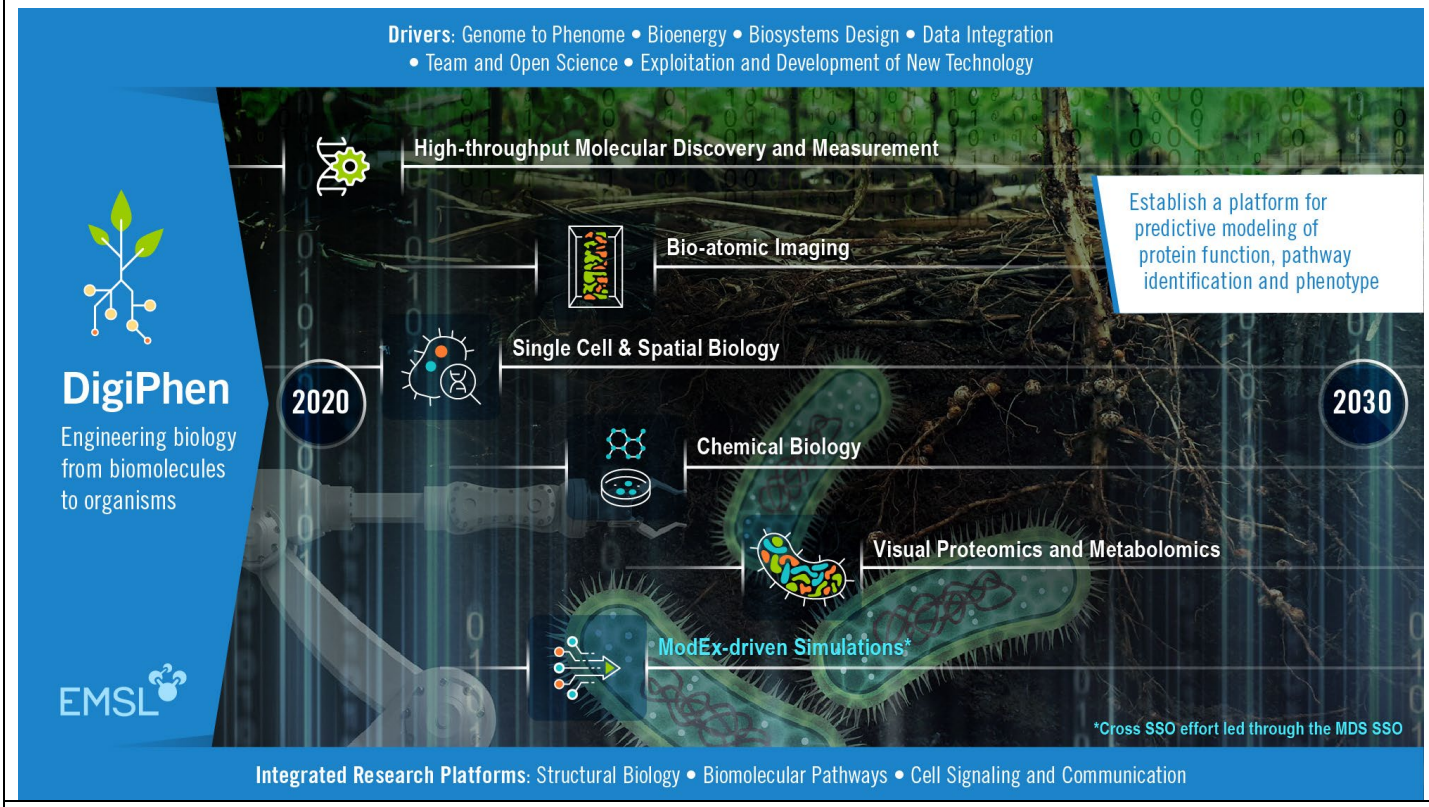


Figure 3. Overview, timeline, and Research Areas supporting DigiPhen. This objective establishes EMSL's leadership in the development of a new generation of S&T innovations that produce new knowledge about the molecular, cellular, and community foundation of phenotypes required for the user community's 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.

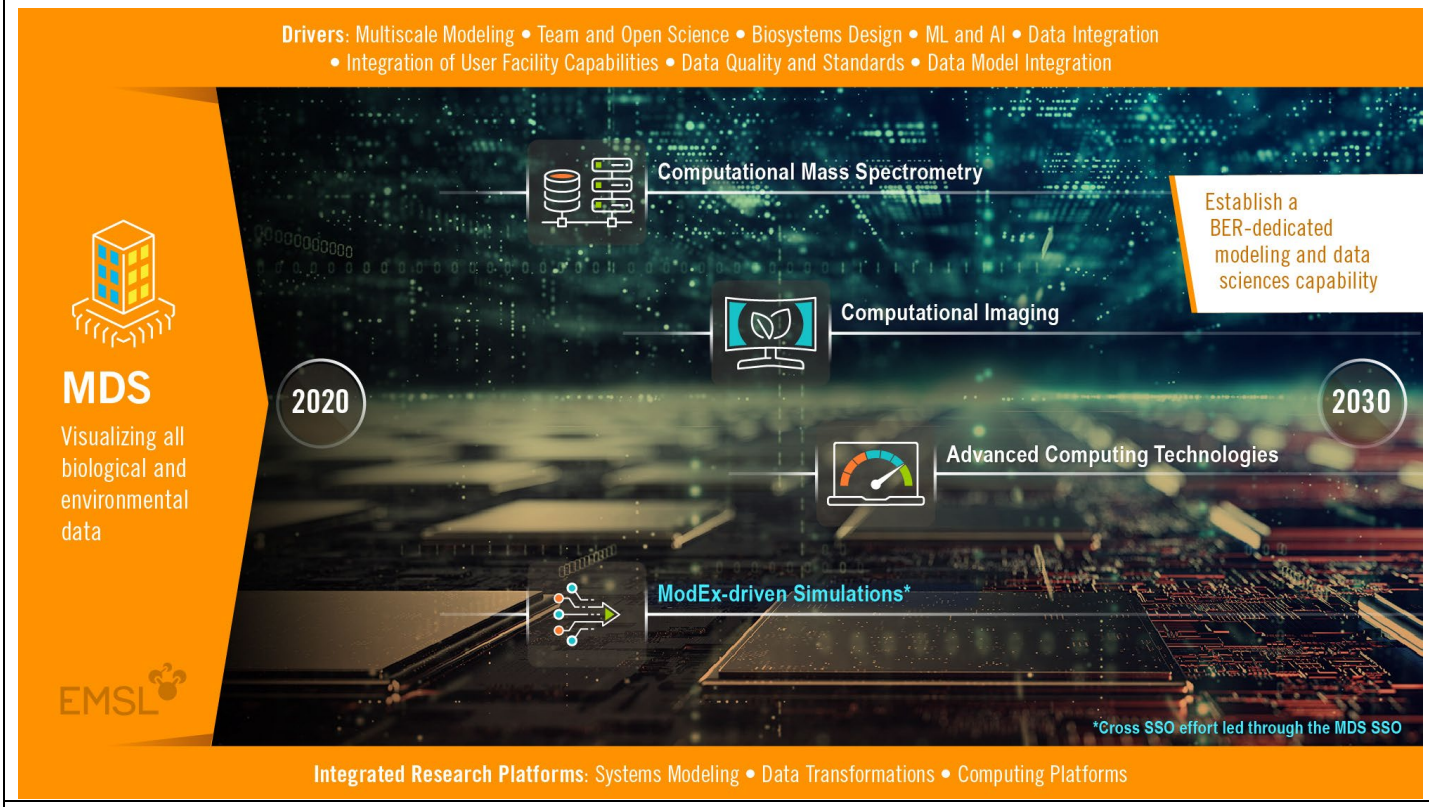


Figure 4. Overview, timeline, and Research Areas supporting MDS. The Research Areas within this objective build capacity, methodology, software, and new capabilities that together accelerate translation of data into knowledge by enabling integration visualization and utilization of all BER data streams. Each Research Area is placed on the 2020–2030 timeline to show where we anticipate the most activity, although we expect work to begin before and continue after, as required.

An important change to the Research Areas was the decision to elevate the ModEx Research Area to specifically reflect a joint project space across all three SSOs (see Figures 2, 3, and 4) that is led and administered through the MDS SSO. While most Research Areas have some activities that cross SSOs, ModEx requires sustained partnership across the SSOs and some of their supporting Research Areas to succeed.

Table 1. Research Area Roadmaps and Stage of Development

SSO	*Research Area*	Status	*Complete*
DigiPhen	High-throughput (HTP) Molecular Discovery and Measurement	Final	FY 2023 Q1
DigiPhen	Chemical Biology	Concept	FY 2023 Q4
DigiPhen	Visual Proteomics & Metabolomics	Drafted	FY 2023 Q4
DigiPhen	Single-Cell & Spatial Biology	Final	FY 2023 Q1
DigiPhen	Bio-Atomic Imaging	Final	FY 2023 Q1
MONet	Automated Soil Molecular and Microstructural Soils Analysis	Concept	FY 2023 Q4
MONet	Field Sensing & Sampling Site Network	Concept	FY 2023 Q4
MONet	Rhizosphere Sensors	Final	FY 2023 Q1
MONet	Field Sensing, Sampling, & Sensors	Concept	FY 2024 Q4
MONet	Atmospheric Aerosols and Particles	Concept	FY 2023 Q4
MDS	Computational Mass Spectrometry	Concept	FY 2023 Q4
MDS	Computational Imaging	Concept	FY 2023 Q4
MDS	ModEx-Driven Simulations	Drafted	FY 2023 Q3
MDS	Advanced Computing Technologies	Concept	FY 2023 Q4

3.0 FY 2022–23 Example SSO Evolution and Implementation Activities

3.1 MONet

Four major activities comprised most of the revision and planning for ETI and are as follows: (1) establishing FY 2023 milestones for MONet; (2) planning the launch of MONet in the user program; (3) review and revision of the IRPs supporting MONet; and (4) Research Area roadmapping. Each activity is detailed below.

Milestones: In October of 2022, EMSL established concrete FY 2023 milestones for the MONet SSO (Table 1). These milestones were submitted to BER leadership in our response to Minor Activity 1 – Updated Scientific Metrics and Milestones (Attachment 9).

Table 2. 2023 MONet Strategic Science Objective Milestones

Objective	Milestone	Status	Notes
MONet	M1: Publish MONet user solicitation	In progress	Publish the first MONet user sample solicitation
	M2: Complete new Roadmaps and workshops	Initiating	Complete roadmaps and workshops for Field Sensing and Sampling Sites Network Research Area
	M3: Complete MONet data architecture framework	In progress	Complete framework for the MONet Data Architecture
	M4: Initiate MONet data gateway build	In design	Contract the design/build for the MONet Data Gateway. Initiate build
	M5: Initiate automated organic matter analysis platform build	Planning	Contract the design/build for automated soil organic matter analysis platform. Initiate build
	M6: Publish EMSL Facilities Integrating Collaborations for User Science (FICUS) Call	In progress	Launch mineral–microbe–root biogeochemistry EMSL–Advanced Photon Source FICUS call
	M7: Launch new ETI IRPs	Complete	Launch Rhizosphere Function and Terrestrial-Atmospheric Processes IRPs

MONet User Program Launch: Planning for the launch of the MONet user program was informed by a meeting with the MONet Scientific Advisory Committee on August 30, 2022. The committee affirmed the value of the 23 data types and recommended that an additional data type (mineral-associated organic matter) be added. The committee also affirmed the overarching value of the envisioned MONet database to the modeling community and encouraged the MONet project to continue interacting robustly with the broader U.S. modeling community. Since that time, EMSL's ETI Science Area lead has:

- Developed and initiated implementation of the MONet soil characterization user program, including conducting risk analysis and mitigation (Table 2, Milestones M1, M3, M4), communications and outreach functions, sample handling and analysis functions, continuation of the data model/database builds, and user science engagement planning.
- Developed five activity teams tied to specific technical or operational aspects of the MONet capability.

- Coordinated emerging technical needs with the CAM/DigiPhen leadership and the Intramural S&T Research program. Two key Intramural S&T Research program activities are (1) developing automated capabilities within the S&T program targeted for implementation in MONet and more broadly in the user program (Milestone M5) and (2) generating data and database structures.
- Held an EMSL Learn Webinar (September 28, 2022), as well as American Geophysical Union presentation and poster sessions (December 12–16, 2022).
- Released the first call for MONet soil sample contributions (Milestone M1) on December 9, 2022, and announced plans to begin accepting soil characterization user requests in February 2023.

Align and focus the ETI IRPs with MONet SSO Priorities and with the Earth and Environmental System Sciences Division (EESSD) Strategic Plan: EMSL reviewed, refocused, and revised two of the three ETI IRPs to improve alignment of EMSL's capabilities and associated planning with the BER EESSD strategic plan and to improve the leverage and impact of the MONet SSO. Specifically, EMSL:

- Refocused the broad Plant and Ecosystem Phenotyping (PEP) IRP on the rhizosphere. The **Rhizosphere Function (RF) IRP** (https://www.emsl.pnnl.gov/science/expertise/rhizosphere-function/6) has formally replaced the PEP IRP.
- Launched the RF IRP through an EMSL Learn webinar on November 23, 2022, community engagement, and Large-Scale Research and FICUS science proposal call topics that feature RF capabilities and science themes (Milestone M7, in progress).
- Refined and refocused the former Ecosystem Interfaces IRP. The new IRP is the **Terrestrial-Atmospheric Processes (TAP) IRP**. The TAP IRP addresses critical gaps in our understanding of biogeochemical processes that occur when terrestrially sourced particles and gases mix in the near-surface atmosphere and how the resulting aerosols and chemical species affect cloud formation and Earth's radiation budget (https://www.emsl.pnnl.gov/science/expertise/terrestrial-atmosphere-processes/3). This IRP launch included hosting breakout sessions at the FY 2022 Environmental Systems Science PI meeting, the FY 2023 ARM/ASR PI meeting, and planning/roadmapping workshops (Milestones M2, M7).
- Initiated rollout of the new TAP IRP through EMSL Learn webinars, community engagement by the IRP lead, and Large-Scale Research and FICUS science proposal call topics that feature TAP capabilities and science themes (in progress) (Milestone M7).
- Began updates to MONet-supporting Research Area Roadmaps to leverage RF IRP capabilities and investments (in progress) (Milestone M2).

Research Area Roadmapping: Review of the MONet-supporting Research Areas after the launch of the two new IRPs revealed opportunities to revise and refocus several Research Areas. This is described in Section 2.4 (see also Table 1).

3.2 DigiPhen

Three major activities comprise most of the revision and planning activities for DigiPhen: (1) establishing FY 2023 milestones for DigiPhen, (2) planning for the Microbial Molecular Phenotyping Capability (M2PC) at EMSL, and (3) Research Area review and roadmapping, all of which are detailed below.

Milestones: In October of 2022, EMSL established concrete FY 2023 milestones for the DigiPhen SSO (Table 3). These milestones were submitted to BER leadership in our response to Major Action M1

Table 3. 2023 DigiPhen Strategic Science Objective Milestones

Objective	Milestone	Status	Notes
DigiPhen	Complete M2PC CD1	In Progress	Timely submission of materials and BER approval of CD1 milestone.
	Complete new Roadmaps and workshops	In Progress	Complete roadmaps and/or workshops for the Visual Proteomics and Chemical Biology Research Areas.
	Data integration workshop	Planning	Complete a joint DigiPhen/MDS external workshop on data integration methods/needs for emerging EMSL workflows.
	Launch single-cell proteomics/transcriptomics platform	Planning	Complete codevelopment (Scienion) of an integrated single-cell transcriptomics and proteomics platform and launch it in the EMSL user program.
	Launch hybrid Obitrap-21T Fourier-transform ion cyclotron resonance (FTICR) platform	Planning	Complete and launch first-of-its-kind hybrid Orbitrap-21T FTICR for advanced top-down proteomics.
	Launch matrix-assisted laser desorption/ionization mass spectrometry (MALDI)-MS for metabolites	Planning	Complete and launch a new capability for in situ derivatization and spatially resolved MALDI-MS imaging for metabolites.

M2PC Planning: EMSL is currently in the Critical Decision (CD) process for M2PC, which is planned as an expansion to the current EMSL facility. M2PC will directly support the DigiPhen SSO by providing automated systems for generating microbial genetic diversity, microbe cultivation, and high-throughput assays and measurement methods (metabolomics, proteomics, high-content imaging). The M2PC Science Driver Workshop was held in Q4 of FY 2021, and the associated report was completed and posted to the EMSL website in Q1 of FY 2022. The report detailed the breadth and types of microbes, culture conditions, analytical methods, and scientific outcomes that were desired for the future M2PC. The CD-1 process for M2PC is scheduled for completion in FY 2023.

Research Area Roadmapping: As mentioned previously, EMSL completed workshops for drafting detailed roadmaps for several Research Areas supporting the DigiPhen SSO. The Visual Proteomics, Single-Cell Biology, Bio-Atomic Imaging, and High-Throughput (HTP) Omics and Protein Function workshops were completed in FY 2022, and revisions were subsequently made to the Research Area Roadmaps (Appendix A). In addition to drafted Research Area Roadmaps, one conclusion from these workshops was the need for a stronger emphasis on integration of multi-model data. A joint DigiPhen/MDS workshop targeting data integration is planned for FY 2023. Other changes were also made to the DigiPhen Research Area Roadmaps. These include expanding the Single-Cell Biology Research Area title to more directly include the spatial biology focus, merging Visual Proteomics and Visualizing Metabolic Pathways into one Research Area, merging Chemical Biology Institute and Chemical Biology Satellites into one Research Area that focuses on Chemical Biology and will maintain the institute and satellite foci, and changing the title of the HTP Omics and Protein Function Research Area to better reflect its focus on molecular discovery and measurement (Figure 3, Table 4).

3.3 Modeling and Data Sciences

Extensive planning activities for MDS were undertaken in FY 2022 and will continue into FY 2023. Many of these were initiated in response to the Triennial Review. Activities include establishing concrete objectives for the MDS SSO (Table 4), revisions to the CAM Science Area description (Appendix B), revisions to the MDS SSO (Appendix B), revisions to the underlying Research Areas (Appendix B), and accompanying revisions to the IRPs (Attachment 6). These are detailed below.

Milestones: In October of 2022, EMSL established concrete FY 2023 milestones for the MDS SSO (Table 4). These milestones were submitted to BER leadership in our response to POAM M1 (Attachment 9).

Table 4. 2023 MDS Strategic Science Objective Milestones

Objective	Milestone	Status	Notes
MDS	Revise CAM SSO and Research Areas		Revise the CAM-associated SSO and the supporting Research Areas
	Complete new roadmaps and workshops		Complete two new Research Area Roadmaps and external review workshops
	Complete CAM IRP descriptions		Complete descriptions for the data science and computing platform IRPs
	Launch nuclear magnetic resonance profiling software		Complete and launch software for semi-automated nuclear magnetic resonance metabolite profiling software
	Launch software for omics data analysis		Launch the Multiomics Analysis Platform software suite for omics data analysis, integration, and visualization applications on the Network for the Execution of User Science (NEXUS)
	Launch quality control FTICR workflow		Complete and launch an acquisition time quality control workflow for FTICR mass spectrometry (MS)-based complex mixture analysis

CAM Revisions and MDS Revisions:

Planning activities have primarily been multiple discussion cycles between different sets of stakeholders (CSO, chief data officer, Science Area leaders, IRP leaders, the Computing Operations team, experimental scientists, data and computational scientists across PNNL, and, as appropriate, the entire EMSL Leadership team). Planning revisions have focused on four key areas: (1) identifying strategic differences between CAM and other Science Areas, (2) understanding capability gaps in the other two Science Areas where CAM science can play enabling roles, (3) understanding the context of EMSL's computing strengths for internal stakeholders as well as the BER community, and (4) ideating and prioritizing opportunities to leverage more of PNNL's expertise to advance BER science through the EMSL user program and EMSL's S&T investments.

Discussions about the strategic differences between CAM and the other two Science Areas have led to formulating a fast-follow strategy to enable sustained integration and delivery of new capabilities amid continuing exponential progress in computing and modeling research. More details on this strategy are available in Appendix B. The purpose of the MDS SSO has been reimagined with the goal of expanding both community demand for the capabilities it develops and stewards and its influence across the community and

scientific impact. The MDS SSO, associated Research Areas, and supporting IRPs are designed so that leading scientists can seek out CAM for user proposals and partnerships regardless of whether their research *applies* new analytics and modeling techniques to BER science or *develops* new analytics and modeling techniques.

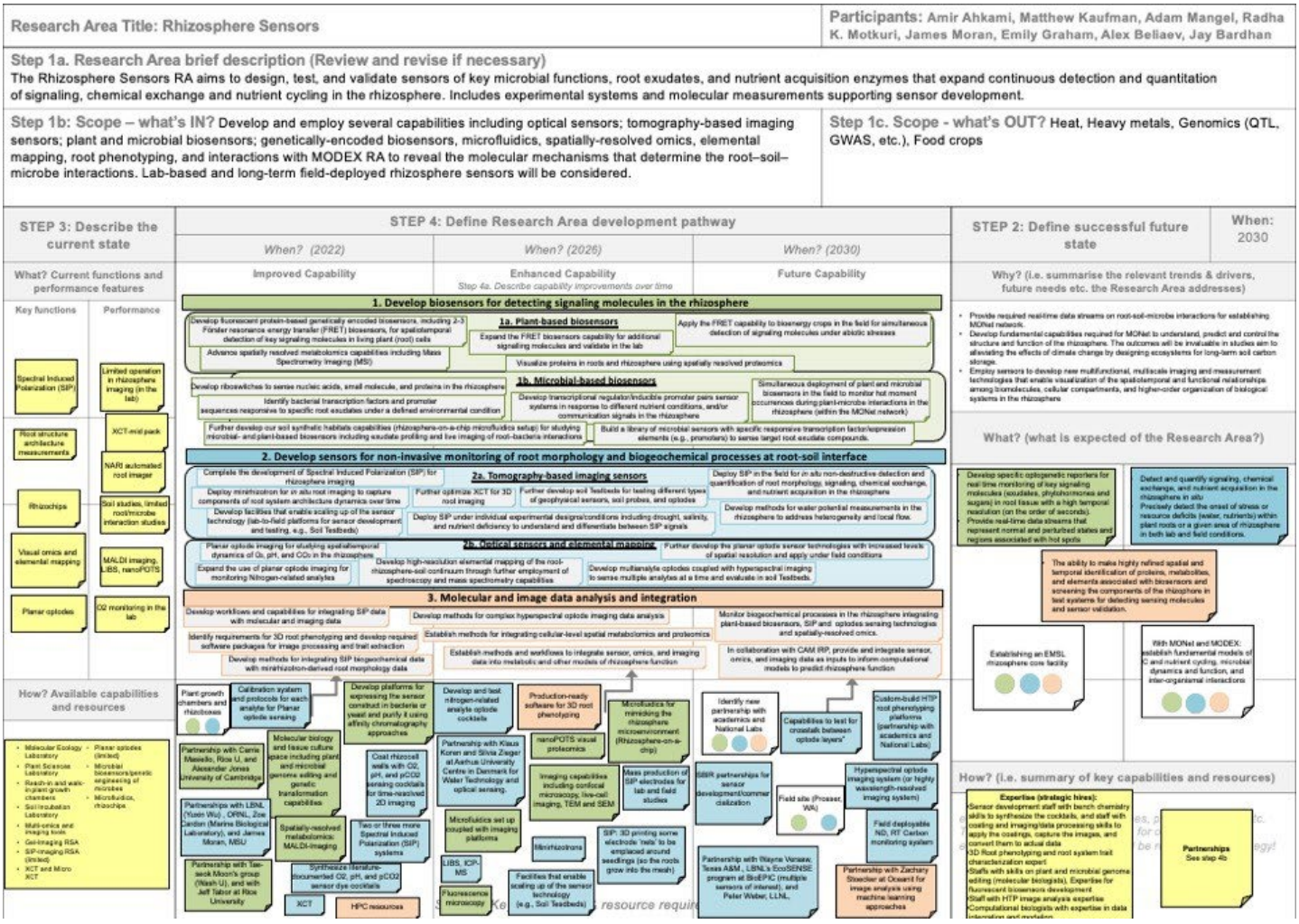
The FY 2022 discussions have yielded significant revisions to the plans for the CAM IRPs. During the 2021 Triennial Review, CAM had a single IRP, Systems Modeling and Data Sciences. EMSL now plans for CAM to have three IRPs. These will be (1) Systems Modeling, (2) Data Transformations, and (3) Computing Platforms. More detailed descriptions of the IRPs are found in the SSO-IRP Integration Addendum to the 2021 EMSL Strategic Science Plan. The Systems Modeling and Data Sciences IRP is being separated, and FY 2023 investments will grow the Data Transformations IRP activities so that it can be launched formally in FY 2024, delivering data analytics services throughout the project life cycle, from proposal recruitment through data analysis and long-term data management. The Computing Platforms IRP is intended to receive comparable investments in FY 2024, with a formal creation in FY 2025, and will enable BER science on new, potentially transformative computing hardware and software systems, such as the growing diversity of chips designed specifically for artificial intelligence (AI)/machine learning (ML) applications.

Engaging the EMSL user program as a full-fledged Science Area has yielded important learnings for planning discussions and activities. For example, in FY 2022, EMSL convened its first-ever proposal review panel for the CAM Science Area for the Large-Scale Research mechanism, following significant staff efforts to increase the number of user proposals submitted (12 in total, compared to 4 in FY 2021). Together, these activities revealed (1) the importance of seeing the entire process at a system level and the subtle complexities that FSB and ETI get correct through extensive experience, (2) opportunities to increase integration between CAM and the other IRPs, with different kinds of CAM expertise suitable for engagement at different points in the user project life cycle, and (3) strengths of the fast-follow strategy for accelerating BER mission science.

FY 2022 discussions have also led to re-conceptualization of the MDS SSO Research Areas. The four new Research Areas are (1) Computational Mass Spectrometry, (2) Computational Imaging, (3) ModEx-Driven Simulations, and (4) Advanced Computing Technologies. Together, these Research Areas capture key opportunities for synergy between EMSL's capabilities and delivery of leading developments to enable BER science. More details are available in the CAM MDS Research Area revision document (Appendix B of this addendum). The new Research Areas capture the key elements and scientific objectives of the CAM Research Areas presented during the 2021 Triennial Review, and the change serves to improve alignment between the Research Areas and the research communities that CAM can leverage. FY 2023/24 activities will include workshops to create detailed roadmaps for these Research Areas.

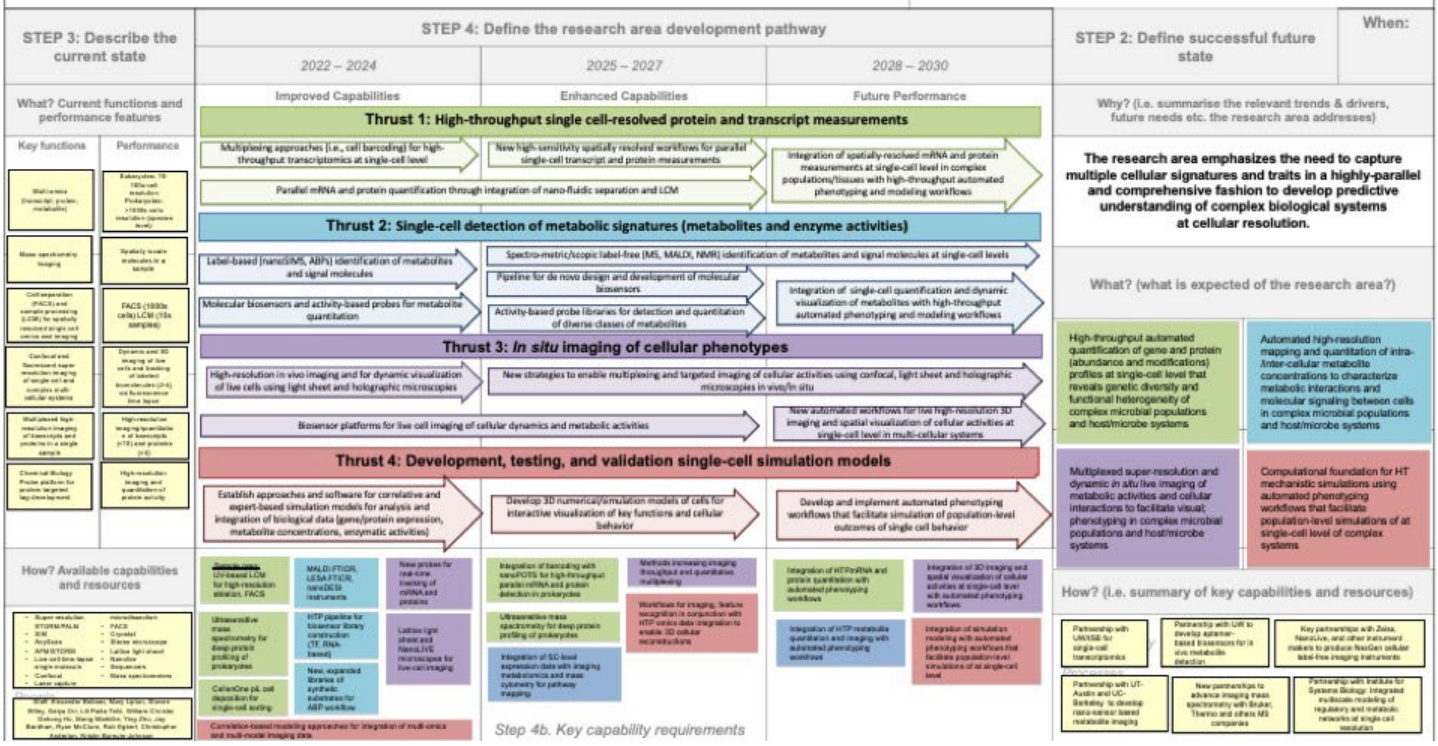
Appendix A – Completed Research Area Roadmaps

A.1 Rhizosphere Sensors



A.2 Single-Cell Biology

Research Area Title: Single-Cell and Spatial Biology	Participants: Alex Beliaev, Ryan McClure, Kristin Burnum-Johnson, Ljiljana Paša-Tolić, Mary Lipton, Jay Bardhan
Step 1a. Research Area Brief Description: This area focuses on quantitative assessment and dynamic measurements of specific biological outputs (i.e., mRNA, protein and metabolite abundance, metabolite flux, and post-translational modifications) in single cells and/or subpopulations to understand how individual phenotypes give rise to the behavior of multi-cellular and/or multi-species communities and tissues.	
Step 1b: Scope – what's IN? Bacterial, archaeal, viral, algal, fungal, plant, and mammalian (non-clinical) systems of natural and synthetic origin that do not exceed BSL2 safety classification; single-cell, species-resolved characterization of biological functions and behaviour driven by inter-cellular signalling, metabolite exchange and trafficking; population-wide measurements of transcript, protein, and metabolite profiles carried out at a single-cell resolution; dynamic imaging of biological molecules, features, and functions aimed at characterization of complex multi-cellular systems and populations at a single-cell level	**Step 1c. Scope - what's OUT?** *(what's not included?)* Clinical (human-related) bacterial, viral and mammalian samples and ≥BSL3 classified samples; population-averaged bulk measurements; HT translation of genomic information describing relationships among biomolecules; structural biology and molecular structure characterization



A.3 Bio-atomic Imaging

Research Area: Bio-Atomic Imaging

Contributors: Scott Lea, James Evans, Irina Novikova, Danny Perea, Mowei Zhou, Lili Paša-Tolić, John Cort, Margaret Cheung. **External workshop participants:** Mohammed Al-Quraishi, Crysten Blaby, Ann Chiaramoni, Joe Loo, Carrie Partch, Jose Rodriguez, Soichi Wakatsuki

Research area description: Acquire atomic-level resolution structural information of proteins, protein complexes, and enzyme active sites and use it to develop 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 interaction interfaces between proteins, nucleic acids, and small molecules, and reveal the chemistry and mechanisms linking these proteins to their cellular role or function. Comprehensive understanding of these interfaces 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. Atomic-level resolution will provide a detailed understanding of amino acid targets within the protein for genetic engineering efforts to enhance and alter protein function in engineered systems. Efforts within this research area will leverage advancements and developments in the Image Processing research area as part of the Modeling and Data Sciences Center Strategic Science Objective and benefit from development pathways in the Data Integration Software Framework and Open-Source Data Analysis Software Suites research areas to fulfill goals in automated structure determination and enable visual proteomics. A progression of computational, technological, and instrumental, innovations along with multimodal integration of different approaches to reveal structure–function relationships will represent the maturation of this capability development effort.

Scope – what's IN? cryo APT, cryoEM, MS, solution NMR, image processing/modeling, de novo structure prediction

Scope - what's OUT? Live cell imaging, tissue imaging, anything not subcellular, CARS, SPR, ToFSIMS

Current state

What? Current functions and performance features

Key functions	Performance
Sample screening (EM and MS)	Manual (every sample)
Data processing (EM)	Near real-time data processing
• CryoEM resolution • MicroED resolution • Cryo-tomography resolution	• Atomic and near atomic (1.8Å onsite) • Atomic resolution (0.8Å onsite) • Sub-nm to nm resolution
• APT resolution for biological material • APT elemental/ molecular identity	• Sub-nm to atomic • Small molecule
• NMR structure elucidation • NMR dynamics	Chemical shift assignment
• Computational structure prediction	de novo AI approaches

How? Available capabilities and resources

IRPs: Structural Biology

Technology: cryoFIB, Krios G3i, Leica GP and Vitrobot II, Cell-free protein expression robot, Native MS, Tava, Bokan, Hood, Baker, LEAP 4000, Rosetta, AlphaFold.

People: Evans, Novikova, Perea, Parvate, Moser, Zhou, Paša-Tolić, Cort, Buchko, Powell, Cheung, O'Callahan, Lea

Partners: SSGCID, UCLA DOE Institute, SLAC, NSLS, Cameca, NIST, NIH P41 Center, SSRL, UW, Center for Integrated Structural Biology, Thermo, Gatan

Research area development pathway

2022 – 2023 — Improved Capability

Thrust 1: Computationally guided structural prediction and inference from limited in experimental information

- Develop deep learning algorithms that extract structure and identity of biomolecules from sparse APT data
- Deploy ML/DL methods for NMR-based *de novo* structure prediction

Thrust 2: Multimodal correlative data integration for structural elucidation

- Increase ability to solve structures for small molecules through microED/crystallography
- Integration of NMR and XRD analyses for chemical shift perturbation studies

Thrust 3: High resolution, high sensitivity measurements

- Develop APT methods that increase fragmentation of biomolecules to improve biomolecular identification
- Establish/optimize bio-APT methods for direct mapping of macromolecular structure in native hydrated states

Thrust 4: Automation

- Automate data collection and analysis for APT through script-based instrument control and computationally enabled 3D point cloud data mining
- Develop open source or purchase commercial software for automated native MS data processing

- Partner with CAM IRPs (i.e., Computing Platforms, Data Sciences and Modeling Sciences)
- Computational approaches and requirements include AI methods for *de novo* structure prediction, protein complex spatial arrangement, and functional state annotation
- SSRL partnership for complementary XRD analyses
- NIST partnership to test and evaluate EUV laser enabled BioAPT
- Partnership with Cameca/Steam to develop advanced instrumentation for bio-APT
- NIH P41 partnership to deploy automated native MS for HTP batch analyses
- Acquire automated grid vitrification systems
- Strategic hire in sample prep for membrane protein production and characterization of user samples

2024 – 2027 — Enhanced Capability

Thrust 1: Computationally guided structural prediction and inference from limited in experimental information

- Utilization of low-resolution cryoEM data by ML models to provide constraints for high-resolution structure determination
- Develop ML/DL methods for determining protein structure and interfaces from sparse NMR data

Thrust 2: Multimodal correlative data integration for structural elucidation

- Integration of NMR and native MS data to feed into biophysical models for structural dynamics
- Integration of APT data with cryoEM to provide *in-situ* composition and molecular structure correlation
- Integration of APT data with native MS providing structural constraints for computational modeling

Thrust 3: High resolution, high sensitivity measurements

- Develop capabilities to obtain structure and resolve structural heterogeneity in complex samples and native environments
- Create native MS activation methods to obtain amino acid resolution structural information (interfaces, flexible region)

Thrust 4: Automation

- Automate APT data collection and analysis through script-based instrument control and computational 3D point cloud data mining
- Establish methods for rapid biochemical production and sample optimization for MS, NMR and cryo-EM
- Automate cryo-EM sample deposition, screening and grid optimization and data collection and analysis workflows for HTP

- Increased capacity for near real-time data processing and storage requires an HPC with significantly greater GPU's and expansion of data storage capacity for HTP applications
- Biophysicist hire with fluency in science-informed ML and HPC
- Strategic hire with expertise in molecular simulations, structural characterization and data integration
- Partnership with Thermo for site targeted BioAPT cryo specimen preparation around cryo-compatible GIS gas chemistries
- Partnership with FasmaTech and/or UCLA DOE Institute for development of new integrated MS dissociation methods
- Next gen APT instrument with enhanced detection sensitivity, increased FOV, and EUV laser with lab space to house instrument
- MS instruments with custom ionization, soft-landing and separation methods with lab space to house instruments
- Cameca instruments partnership for automated APT data collection and analysis
- Taheri/Gatan partnership for automated/autonomous cryoEM
- Strategic hire in laboratory automation

2028 – 2030 — Future Capability

Thrust 1: Computationally guided structural prediction and inference from limited in experimental information

- Advanced computation for simulating protein conformation and assembly dynamics to elucidate mechanism of action
- Develop ML/DL workflows to infer spatial arrangement of protein complexes and annotate protein functional states
- Develop hybrid data driven ML/physics-based approaches for low-data regimes

Thrust 2: Multimodal correlative data integration for structural elucidation

- Integrate analysis of different proteomic, structural and computional data types to define heterogeneous and dynamic structures
- Integration of cryoET with native MS to enable visual proteomics

Thrust 3: High resolution, high sensitivity measurements

- Establish an NMR capability and approaches to obtain structure and dynamics of biomolecules up to ~50kDa

Thrust 4: Automation

- Create and optimize near-real time image processing for structural analysis including sub-tomogram averaging methods
- Develop workflows for routine semi-automated HTP protein/biomolecular NMR in liquids

- Join consortia interested in systematically tackling AI/ML based structural determination of biomolecules
- Partnership with Thermo/Bruker/Waters to co-develop customized native MS instrumentation
- Next generation cryoEM with direct electron detectors, Selectris-type energy filter, and laser phase plate
- Acquisition of a dedicated ultra-high field NMR instrument with lab space to house instrument
- Acquire automated protein production (including isotope labeling), purification and fractionation systems with high flexibility in experimental design

Successful future state — When: 2030

Why? (i.e. summarise the relevant trends and drivers, future needs etc. the research area addresses)

The rapid pace of genome sequencing continues to increase the catalog of predicted proteins without known functions. As the gap between genomics knowledge and phenomics understanding is growing exponentially, the ability to close that gap through increased capacity and pace is essential. Higher throughput and robust protein function characterization combined with an understanding of complex metabolic and regulatory pathways are needed.

Atomic resolution of proteins and protein complexes and detailed understanding of enzyme active site chemistry will allow for detailed modeling and simulation of biochemical pathways that involve multiple proteins and/or enzyme active sites. Atomic-level resolution will provide detailed understanding of amino acids targets for genetic engineering efforts to enhance/alter protein function in engineered systems.

What? (what is expected of the research area?)

Computational approaches and requirements include AI/ML methods and physics-based models and their integration with experimental approaches for structure prediction, protein dynamics, and functional state annotation. Increased capability for near real-time data processing and storage.

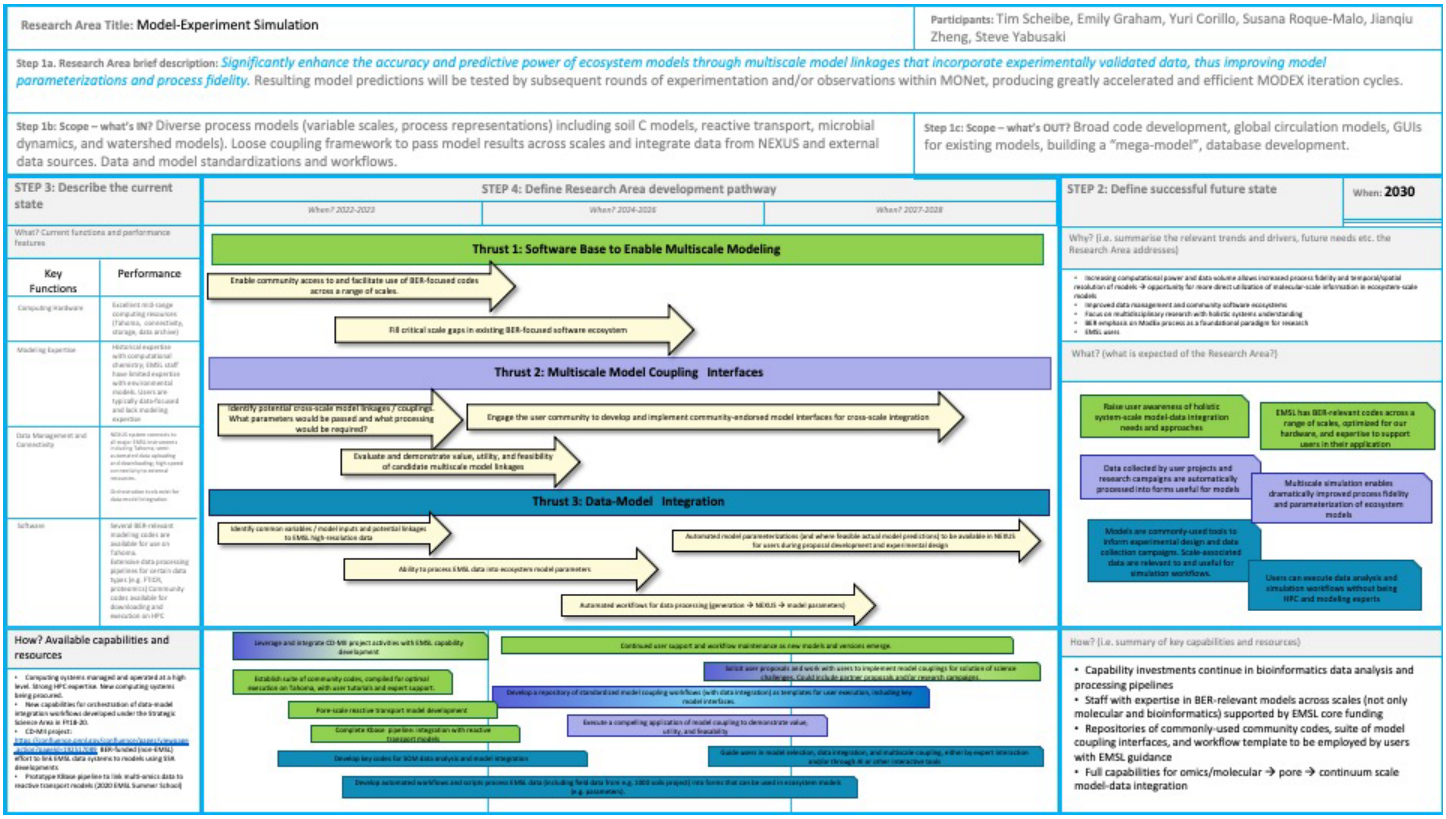
- Characterization of protein-protein, protein-nucleic acid, and protein-ligand interactions to aid in functional characterization for genetic engineering efforts
- The ability to resolve structural heterogeneity in complex systems (crude samples vs. a few protein component system currently)
- Automation needed in sample preparation, data collection, and data analysis workflows. Instrument based automation will result in 5-10x increase in throughput in sample preparation and data collection workflows. Data analysis will increase >20x through the development of automated computational workflows.
- Direct mapping of macromolecular structure of biomolecules in their native hydrated environment
- Advanced instrumentation for higher resolution, increased detection sensitivity, and the ability to extract more/new information from the same sample

How? (i.e. summary of key capabilities and resources)

- For APT, need EUV source, higher instrument detection sensitivity, increased field of view and advanced deep learning methods to extract biomolecular identification and orientation.
- MS requires automation for batch analyses (partnership), instruments for advanced activation, soft-landing approaches for sub-population enrichment, and methods for integrated analysis of different data types.
- Path forward in cryoEM includes automated sample prep and screening methods, additional (screening) instruments for increased throughput, and more sensitive electron detectors for faster analyses.
- NMR requires dedicated higher-field instruments, partnership for complementary XRD analyses, AI based *de novo* structure prediction, and additional expression platforms.

A.4 ModEx/ModEx-Simulations

This initial version of the ModEx Research Area Roadmap is under additional revision by the MDS SSO team.



A.5 High-Throughput Molecular Discovery and Measurement

Research Area Title: HTP Molecular Discovery and Measurement

Participants: Mary Lipton, Kristin Burnum-Johnson, Paul Piehowski, Ljiljana Paša-Tolić, Chris Brandvold, John Cort, Irina Novikova, Will Kew

Step 1a. Research area brief description (Review and revise if necessary). The HTP Omics and Protein Function Research Area seeks to increase the capacity and pace of mass spectrometry (MS) based molecular discovery, identification, isolation and structural elucidation of new protein targets.

Step 1b: Scope – what's IN? We combine advanced mass spectrometers, nuclear magnetic resonance (NMR) spectrometers, and cryo-EM/NMR elucidation of protein structure and function to conduct high throughput proteomics, metabolomics, lipidomics, and protein function elucidation for biological discovery investigations.

Step 1c. Scope - what's OUT? (*what's not included?*) Genomics, transcriptomics, epigenetics (DNA methylation)

STEP 3: Describe the current state

What? Current functions and performance features

Key functions | Performance

Integrated sample preparation for all analysis types | Sample preparation is based on manual methods

Proteomics (conventional) | Conventional proteomics use standard separations and instrumentation

Metabolomics | Metabolite identification tied to known and limited libraries of compounds

Lipidomics | Limited identification based on libraries of lipids from bacteria, fungi, algae

Proteomics (proteoforms and protein complexes) | Low throughput characterization based on top-down proteomics

Protein function | Activity based probes to produce/work with little enzyme kinetic information

Data interpretation and integration | One off analysis —no pipeline and different questions

How? Available capabilities and resources

LC-MS Orbitrap (11)
21T/ 15T/ 12T FTICR suite
Cellenone
GC-MS/GC-QE (metabolomics) (5)
Ion mobility IMS-TOF (8)
LC triple quadrupole (4)
Waters Synapt G-2
Ultra high mass range Orbitrap MS
NanoPOTS
Solution-state NMR, 5 total
Solid-state NMR, 4 total
CryoEM Krios G3i, cryo-FIB, Leica GP
Protein expression: cell-free expression pipeline
Heterologous protein overexpression: Pichia, E. coli (insect and mammalian cell culture—not implemented at EMSL yet)

STEP 4: Define research area development pathway

When? 2022 – 2024	When? 2025 – 2027	When? 2028 – 2030
Improved Capability	Enhanced Capability	Future Capability

Thrust 1: Advances in Protein Discovery and Identification

- Improve depth of proteomic coverage
- Advance metabolic flux modeling and activity measurements
- Develop a multiplexed spatial proteomics workflow.
- Implement protein stable isotope labeling to flux studies
- Next Generation Proteomics.

Thrust 2: Next Generation Metabolomics

- Deploy high throughput, reproducible library-based LC-MS metabolomics platforms
- HTP and Automated Lipid Analysis Workflows
- Library independent metabolomics, lipidomics, and glycomics
- Design, build and implement faster NMR-based experiments (both liquid and solid)
- Link Global Metabolomics Data with Imaging Approaches
- Connect stable isotope labeling to metabolomics techniques

Thrust 3: Data Analytics and Integration for Discovery

- Automated Data Capture and RT Data Processing
- Unified computational platforms that integrate prediction and analysis across technologies
- Automated and data independent omics analysis platforms
- Computationally Predict Protein and Metabolite Targets
- Create and deploy data analysis methods for stable isotopically labeled proteins and substrates from MS and NMR experiments

Thrust 4: Protein Functional and Structural Characterization

- Automated Characterization of Proteins of Unknown Function
- Develop and Automate Identification of Post-translational Modifications to Proteins
- Develop and Automate Top Down Proteomics Workflows
- HTP Structural Analysis of Protein-Metabolite Complexes

When? 2022 – 2024	When? 2025 – 2027	When? 2028 – 2030
Next generation MS for proteomics and metabolomics TIMS-ToF; Eclipse etc.cc	Next generation NMR for metabolomics with liquid handlers	Advanced and automated data independent acquisition methods for proteomics and metabolomics	Advanced and automated glycomics methods: skilled NMR/MS glycomics expert	Develop a multiplexed spatial proteomics workflow.	Automated solid phase synthesis with microscale chromatography
Dedicated staff for the development and implementation metabolomic separations	Dedicated staff for the automation pipeline development	Dedicated staff for integrated proteomics and metabolomics workflows aligned with spatial measurements	Platform for multiplexed enzyme assays for determination of kinetic parameters	Integrated fluorescence and MS capabilities from 384 well plates	
Next Generation NanoPOTS workflow for proteomics and metabolomics	Staff hire of a data scientist supporting development of data integration methods.	Staff hire with expertise in microbial lipidomics for environmental systems	Integrated stable isotope labeling data analysis infrastructure	Staff hire with expertise in new proteomics technologies (Nano-Pore)	Advanced and automated glycomics methods
	Automated screening for protein ligand interactions	Instrument automation wing	Develop a database free peptide identification workflow	Develop a multiplexed spatial metabolomics workflow.	Computationally driven multi omics data integration platforms
Automation Partnership for HTP workflows	Partnership for library free metabolomics	Computational partnerships with KBase, NMDC and PNNL computing	Partnership for next generation glycomics and lipidomics	Partnership with next generation MS for more rapid fractionation	

STEP 2: Define successful future state — **When: 2030**

Why? (i.e. summarize the relevant trends and drivers, future needs etc. the research area addresses)

The research area emphasizes the need to use advanced automated analytics with improved sensitivity and throughput to translate genotype to phenotype and define protein and molecular function.

What? (what is expected of the research area?)

Comprehensive spatial, temporal, quantitative, accurate high throughput measurements identification of proteins involved in the metabolic and signaling pathways expressed in microbes, plants fungi, algae and microbial communities, including conditions specific to the changes in protein abundance.

Comprehensive, spatial, temporal, quantitative, accurate high throughput measurements of metabolites, lipids, glycans and sugars from microbes, plants, algae, fungi and microbial communities without the use of standard analytical chemistry libraries.

Automated platforms and algorithms for the analysis and integration of multiple omics data types for biological discovery, modeling, and simulation.

Integration of analytical, computational and experimental methods for structural and functional characterization of proteins variants/isoforms from a variety of organisms.

How? (i.e. summary of key capabilities and resources)

e.g. Technolgy resources, processes, people, partners, etc. THIS IS CRITICAL! Forms inputs for our S&T, staff and equipment investments. They will be rolled into our strategy!

Appendix B – CAM and MDS SSO and Research Area Revision Document

Segments taken and amended from the EMSL Strategic Plan.

B.1 Computing, Analytics, and Modeling Science Area

Just as BER's Energy Exascale Earth System Model (E3SM) leaders exert enormous influence on the evolution of leadership-class high-performance computing (HPC) through the scientific importance of their use of HPC, an aspirational goal for the CAM Science Area is to enable BER science to influence and even drive the evolution of computing, analytics, and modeling platforms associated with small- to mid-range HPC. To achieve this goal, CAM's activities take two complementary aims and two complementary approaches. The first aim is to increase the impact of scientific computing on BER's central challenges; the second is to increase the impact that BER science has on scientific computing. The first approach is to apply existing computational methods to BER problems, and the second is to enable new approaches to BER problems. The first approach focuses on applied science questions, while the second focuses on the fundamental science of the CAM Science Area. Simply put, scientific computing platforms are complex technological and techno-social systems for large-scale mathematics. Due to this complexity, the scientific value of innovations can be hypothesized but must be validated by rigorous tests. For example, some innovations increase value through calculation speed, and others make it easier for non-specialists to make effective use of HPC. The sheer complexity of modern HPC systems, combined with the exponential pace of progress and the diversity of BER problems, necessitates hypothesis-driven testing of new methods as a core activity to achieve CAM's aims of advancing BER's technological edge.

CAM's emphasis on testing new methods within the user program is unlike the primary activities of ETI and FSB. This distinct characteristic arises from four key differences between CAM and the other two science areas. First, CAM adds substantial value by using leading software and data in addition to leading physical instruments. In contrast to physical instruments, which are expensive, occupy finite floor space, and are produced by a small number of vendors, software and data are often free or low cost, are not subject to zero-sum limitations on space, and are produced by a very large number of labs and companies. Second, CAM has an additional set of force multipliers—that is, mechanisms by which its activities advance BER science. Because software and data are not limited, CAM research that establishes the value of new software or data can spur adoption by the larger community. Third, in ETI and FSB, the user project teams perform most of the sample preparation (fieldwork, wet lab, etc.); however, following an equivalent paradigm for CAM is not valuable because a user project team that can perform its own sample preparation can usually perform the simulation or modeling studies themselves. Fourth, FSB and ETI steward a large set of world-leading experimental capabilities, which together allow studies supporting a wide range of BER mission science. In contrast, CAM's world-leading capabilities represent an important but relatively modest fraction of possible contributions (e.g., the NWChem package). Thus, to broaden CAM's coverage of BER mission science, it is important for CAM to seek and pilot applications of emerging computational methods.

CAM's activities and focus are designed to provide complementary value to the HPC resources already available to many researchers. During EMSL's early years, it was difficult to obtain access to HPC resources (supercomputers) and expert assistance. As a result, EMSL's supercomputing and scientific computing expertise were well-tailored to the needs of the BER community. With the increasing ubiquity of HPC resources (both through Advanced Scientific Computing Research Leadership Computing Facilities and institutional computing at many universities) and with the growing complexity of computing tasks, CAM has evolved to focus on the next challenge—empowering scientists to leverage more powerful and sophisticated computing approaches. CAM's three IRPs reflect the three primary paths toward this goal. The Systems Modeling IRP follows simulation-based approaches, building on EMSL's traditional strengths in quantum chemistry modeling (NWChem software package) and PNNL's strengths in hydrology and transport modeling (petascale reactive multiphase flow and multicomponent transport [PFLOTRAN]). The Data Transformations IRP focuses on the value that computing and

modeling can add throughout the data life cycle, from the statistical design of experiments to metadata, data collection, and on to processing, analyzing, and integrating the large, complex datasets EMSL produces. Finally, the Computing Platforms IRP focuses on ensuring that the BER research community stays at the cutting edge of hardware and software platforms. Each of these IRPs is tasked with three primary functions: (1) embed modeling capabilities into multiscale paradigms, (2) catalyze the transfer of new simulation capabilities into the BER mission space, and (3) advance AI/ML-based approaches in model building and simulation. The entire set of nine complement and reinforce each other, covering the broad computing landscape while retaining a tight focus on both the needs and priorities of BER science and the synergies possible through close integration with the other IRPs and Science Areas.

CAM's focus within scientific computing centers on advanced data analytics, visualization, and computational modeling and simulation. These approaches are leveraged to maximize the impact of increasingly complex multimodal experimental data toward developing a predictive understanding of biological and environmental systems. EMSL's unique ability to offer integrated experimental, computational, and modeling 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. By focusing on transforming data into knowledge and expanding BER's tools for future transformations, CAM directly supports the Biological and Environmental Research Advisory Committee (BERAC) Grand Challenges (e.g., 2.3, 6.4, and 8.5) (BERAC 2017) and multiple needs outlined in the BERAC User Facilities Report (Table A.1) (BERAC 2018). CAM also responds to the recent Department of Energy (DOE) call (U.S. DOE 2021) to address the daunting challenge of analyzing "vast quantities of disparate data" by developing wholly new approaches.

EMSL is thus exceptionally well positioned to lead the development of the next generation of computational approaches that will aid the interpretation of observations, convert experimental data into scientific understanding, and fuel multiscale models to predict system behavior to aid BER science. Sections 5.1 and 5.2 in the Strategic Plan provide background on the MDS SSO that emerged in response to EMSL's assessment of trends and drivers during our 2020 strategy workshops. The MDS SSO provides scientific direction and focus for the research efforts in the CAM Science Area.

B.1.1 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. These investments are critical because the need for advanced computing and modeling to better integrate with the complex data of BER science is noted throughout the BERAC Grand Challenges report (BERAC 2017) and the BER User Facilities report. In addition, the scientific challenges driving these needs are detailed extensively in the Biological Systems Science Division and EESSD Strategic Plans. In brief, it is critical to develop higher-fidelity, multiscale, and dynamic models of biological and environmental processes. These 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. They must also be able to project potential future state(s) via the changing of parameters. Meeting these challenges requires 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.

For CAM to maximize its impact on BER mission science over many years, the broader dynamics of the global computing community are essential to understand and frame the MDS SSO. Two dynamics are particularly

relevant—massive ongoing investments in AI/ML capabilities, and the increasingly diverse technologies for HPC. We detail each of these briefly below.

First, global investments in AI/ML have exceeded $20 billion per year since 2018, peaking in 2021 at $50 billion (Cbinsights.com). The new protein-folding AI software AlphaFold was developed by DeepMind, which Google purchased in 2014 for approximately $500 million, just as deep learning entered the computing mainstream. Despite DeepMind's numerous impressive advances after the acquisition, the company lost hundreds of millions of dollars per year until 2021—highlighting that even world-leading organizations face challenges in turning cutting-edge capabilities into valuable assets. Looking across all sectors, AI investments tend to focus on comparable challenges to EMSL's (data size, quality, and variety, with emphasis on images and time-series data), making transfer of capabilities a compelling approach if the valuable ones can be identified expeditiously (a fast-follow approach).

Second, computational capabilities continue to grow rapidly in both power and complexity. The past decade saw two contrasting dominant paradigms: (1) Exascale computing, led by DOE investments, and (2) cloud computing, led largely by industry. In 2021, cloud computing expenditures exceeded $400 billion (Fortune Business Insights 2022), far surpassing known expenditures on Exascale-type capabilities. The growth and scale of cloud computing have driven the proliferation of powerful new tools (e.g., Kubernetes) that support the development, deployment, and maintenance of complex computing workflows. These tools are emerging as valuable paradigms for reliable, reproducible computing and are enabling compliance with the findability, accessibility, interoperability, and reusability principles. Furthermore, advances in semiconductor technology suggest that the next decade is likely to see an explosion of diverse new approaches to HPC, such as chips designed specifically for AI/ML. These new computing platforms are well-suited to be integral components in automated workflows that integrate modeling and experiment, making it important that the BER community have exposure to the potential of these platforms as they emerge.

Unfortunately, almost all the communities in this external landscape are incentivized toward advances that serve other ends, ranging from human health diagnostics, to drug discovery, to other "big data" problems such as business data and internet traffic. The strategic challenge for CAM is how to identify the most valuable new capabilities from this complex landscape and integrate them into complex BER science studies. On one hand, resource constraints preclude EMSL CAM from re-implementing existing work with the necessary adaptations. On the other hand, due to the diversity of external actors and their interests, a small number of strategic partnerships will be inadequate to cover the breadth of EMSL science.

We therefore choose to adopt what is known in competitive strategy as an *ecosystem-shaping strategy*; the term "ecosystem" here refers to that broader landscape of science and technology in computing and modeling. The ecosystem-shaping approach focuses on realizing goals by influencing the dynamics and incentives among the other entities in the ecosystem. This perspective transforms size and scope disparities into sources of shared value. CAM's strategy is to help EMSL's unique scope and societally important science challenges make EMSL a "first port of call" for researchers and developers building new capabilities by creating new opportunities for them to showcase their work's value to their own stakeholders.

Accordingly, the revised CAM Research Areas (Computational Mass Spectrometry, Computational Imaging, ModEx-driven Simulations, and Advanced Computing Technologies) are designed to serve multiple purposes. First, and most clearly related to the ecosystem-shaping concept, these areas represent clear opportunities for external communities to appreciate their capacity to contribute to BER's societally important missions. Second, the four Research Areas build on existing strengths within EMSL, including momentum of current activities and research staff recognized as leaders within those external communities. Third, the areas prioritize high-value opportunities for EMSL to add unique capabilities to BER science. Fourth, they provide deep connections between the IRPs, and especially represent unifying goals for the three CAM IRPs to work together. Each of the

CAM IRP functions can contribute to all four—that is, no one IRP has a controlling interest in any of the four Research Areas.

To maximize EMSL's science impact through computing, the MDS SSO) centers on accelerating ModEx for BER science missions by establishing core expertise and capacity to drive emerging synergies between modeling and simulation, data science and AI/ML, and computing technologies. In accordance with the ecosystem-shaping strategy, this SSO is meant to nucleate a community of experts who are highly collaborative and agile, thereby attracting the broader computational community to bring its research advances to EMSL as a proving ground, where the scientific value of their work will be established through application to BER mission science. As a result, EMSL users will be empowered to dramatically expand the frontiers of biological and environmental science challenges to maintain U.S. leadership in energy and environmental security.

B.1.2 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 focused on delivering computational resources and expertise to support BER and EMSL mission science. MDS will foster the 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 computing 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, CAM will focus efforts in four Research Areas over the next decade (Figure 5). During the initial 2–5 years, our activities will concentrate on three of these Research Areas: (1) ModEx-Driven Simulations, (2) Computational Mass Spectrometry, and (3) Computational Imaging. Key opportunities for the fourth, Advanced Computing Technologies, will emerge in the 4–5-year time frame. Investments made before that will be limited and highly targeted (e.g., technologies that are available elsewhere at PNNL). Realizing the potential of these Research Areas will involve significant coordination between the CAM Science Area leader, the CAM IRP leaders, the chief data officer, and the leadership of the teams under the chief data officer. Taken together, the Research Areas build the scientific and technological foundations for MDS, as well as a diverse, highly collaborative user community for which CAM plays a catalytic role in delivering and growing capabilities for BER mission science.

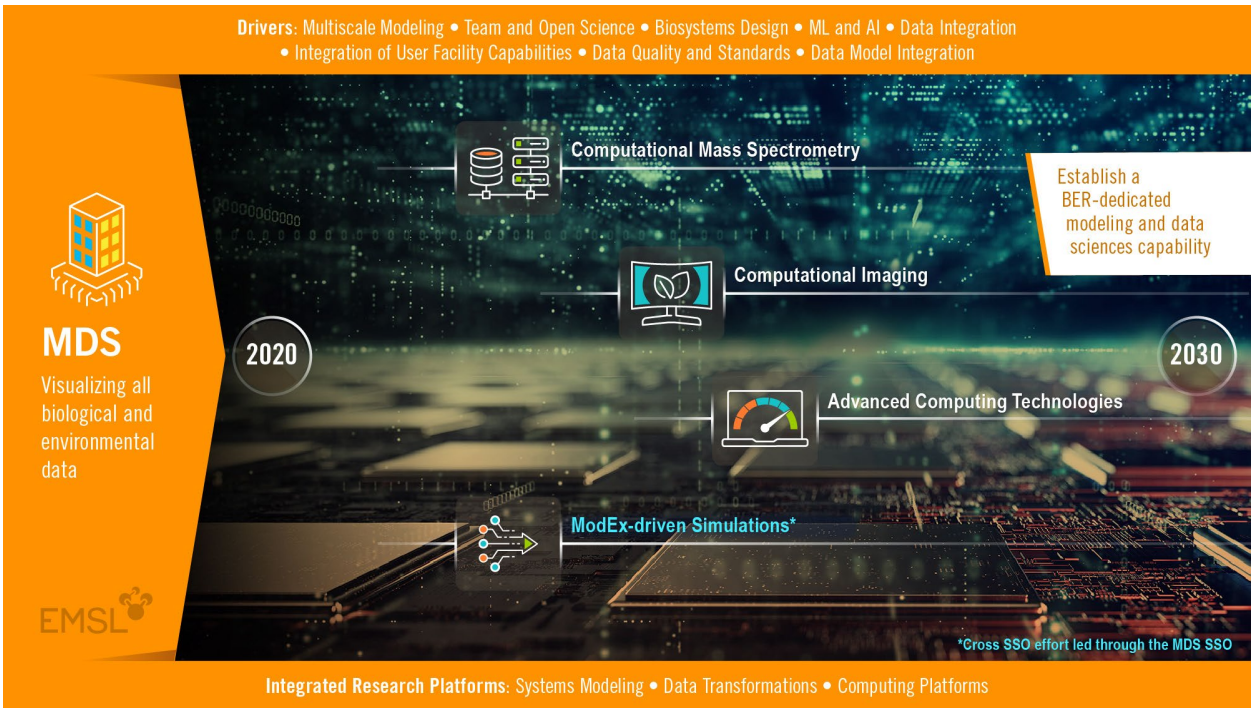


Figure 5. Overview, timeline, and Research Areas supporting SSO 3. The Research Areas within this objective build capacity, methodology, software, and new capabilities that together accelerate translation of data into knowledge by enabling integration, visualization, and utilization of all BER data streams. 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.

B.1.2.1 Computational Mass Spectrometry Research Area

EMSL offers leading instrumentation and expertise for multiple mass spectrometry (MS) based studies, including organic matter analysis, soil metal analyses, and plant, fungi, and microbial community omics. MS-based studies represent a high volume and substantial fraction of EMSL's user program projects. Due to the complexities of MS data and their myriad applications to EMSL science, the processing, analysis, and integration of MS data represent a major area for MDS to contribute original and high-impact research and development that benefits the BER research community and EMSL users. Near-term goals include building out the components and architecture of the CoreMS (MS data processing) ecosystem and integrating new computational and data-science methods to enhance data processing, such as AI/ML. Medium-term goals include efficient capabilities for analysis and integration of MS imaging data. To leverage the overlap between analysis methods and integration with data and metadata systems, the Data Transformations IRP will serve as a primary steward of the Computational Mass Spectrometry Research Area.

B.1.2.2 Computational Imaging Research Area

The collection and analysis of molecular (cryogenic electron microscopy/electron tomography) and microstructural (X-ray computed tomography [XCT]) data presents unique challenges to computing and data analysis. Advancing the imaging capabilities at EMSL requires confronting a complicated landscape of software tools emerging from a sizable computing community. The Computational Imaging Research Area prioritizes and adheres to the "fast-follow" strategy in developing advanced computational tools for image processing, analysis, and integration. Given the predominance of data science and AI/ML for image analysis, the Data Sciences (DS) IRP stewards this Research Area but collaborates closely with the Computing Platforms (CP)

IRP for novel HPC capabilities, as well as key imaging IRPs such as Biogeochemical Transformations (BGT) and Structural Biology.

B.1.2.3 ModEx-Driven Simulations Research Area

EMSL and PNNL have historical strengths in modeling and simulation at multiple scales, ranging from the subatomic (the quantum mechanical NWChem software) to subsurface flow (PFLOTRAN) and beyond (E3SM). The ModEx-Driven Simulations (MEDS) Research Area will advance EMSL's simulation expertise and capacity in ways that align with BER mission science through the ModEx paradigm. MEDS has two main focuses: the first applies cutting-edge simulation methods to transform EMSL data into discoveries that are not possible through data analysis alone, and the second uses these simulation approaches as a tool for experiment design, complementing the insights of domain experts. Examples include (1) integrating top-down proteomics with molecular modeling at molecular scales, (2) improving metabolic modeling using multi-omics data at cellular scales, and (3) understanding soils by building simulation models from imaging data (e.g., XCT) at the pore scale. The Systems Modeling IRP holds primary responsibility for this Research Area, adhering to the "fast-follow" strategy and working with both the DS and CP IRPs to develop capabilities that uniquely leverage the breadth of EMSL experimental data. Investments centered on particular classes of simulation approaches will involve close collaboration with the other IRPs as necessary.

B.1.2.4 Advanced Computing Technologies Research Area

The Advanced Computing Technologies (ACT) Research Area will continue EMSL's longstanding leadership in providing cutting-edge HPC capabilities to the BER research community into the post-Moore's-law era. The coming decade will be highly dynamic for new computing platforms. ranging from specialized AI/ML computing hardware to edge computing that leverages high-performance wireless networking (e.g., 5G). Given the expected proliferation of computing architectures, the ACT Research Area focuses on testing and leveraging systems that do not yet operate even at the scale of mid-range HPC, which distinguishes it from EMSL's mid-range HPC Research Area as well as DOE Leadership Computing Facilities. This Research Area will also focus on advancing novel software frameworks to accelerate ModEx, such as managing and automating computational workflows. The CP IRP will serve as the primary steward for this Research Area, and the DS IRP will likely play critical enabling roles thanks to growth in AI/ML applications.

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://science.osti.gov/-/media/ber/berac/pdf/Reports/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.

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/.

Fortune Business Insights. 2022. "Cloud Computing Market Size, Share & COVID-19 Impact Analysis." Fortune Business Insights. https://www.fortunebusinessinsights.com/cloud-computing-market-102697.

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

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

U.S. DOE 2021. “Environmental System Science.” U.S. Department of Energy Office of Science. Accessed March 21, 2021. https://ess.science.energy.gov/.


EMSL

Addendum 2 Partnership Planning

EMSL

2022 POAM S2

Partnership Planning Addendum to the EMSL Strategic Science Plan

January 2023

Jay Bardhan, John Bargar, Scott Baker, Douglas Mans, Justin Teeguarden

PNNL-34044

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

2022 POAM S2

Partnership Planning Addendum to the EMSL Strategic Science Plan

January 2023

Jay Bardhan, John Bargar, Scott Baker, Douglas Mans, Justin Teeguarden

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

Leveraging the expertise of a broad set of external organizations and individuals in partnerships that enhance and accelerate the Environmental Molecular Sciences Laboratory's (EMSL's) development and delivery of premier capabilities for the Biological and Environmental Research program community is an essential component of our long-term strategy. EMSL partners bring unique technical expertise, cutting-edge innovations, instrumentation, facilities, capabilities, and other resources that complement those available within EMSL for new capability development. They strengthen EMSL capabilities and accelerate new capability development.

Partnerships represent a form of engagement where each party contributes their own resources toward a common research or technology development goal that benefits both. Resources such as materials, software, code, or experimental samples may transfer between partners, but funding does not.

EMSL's approach to selecting partners is driven by our overall science strategy reflected in our Digital Phenome, Molecular Observation Network, and Modeling and Data Sciences Strategic Science Objectives and their supporting Research Area Roadmaps. This science strategy is presented in the 2021 *EMSL Five-Year Strategic Science Plan* (hereafter, the Strategic Plan), with additional specifics detailed in our Strategic Science Objective Planning Addendum to the Strategic Plan.

For Major Action S2, EMSL was asked to clarify its partnership strategy and supporting partnership mechanisms. In response, EMSL developed this addendum to the Strategic Plan that describes our strategy for partnerships, the types of partnerships we support, and the mechanisms used to execute the partnerships.

Acronyms and Abbreviations

BER	Biological and Environmental Research program
DigiPhen	Digital Phenome
EMSL	Environmental Molecular Sciences Laboratory
FICUS	Facilities Integrating Collaborations for User Science
JGI	Joint Genome Institute
MONet	Molecular Observation Network
S&T	science and technology
SFA	Scientific Focus Area
SSO	Strategic Science Objective

Contents

Tables

1.0 Introduction and Background

Leveraging the expertise of a broad set of external organizations and individuals in partnerships that enhance and accelerate the Environmental Molecular Sciences Laboratory's (EMSL's) development and delivery of premier capabilities for the Biological and Environmental Research (BER) program community is an essential component of our long-term strategy. EMSL partners bring unique technical expertise, cutting-edge innovations, instrumentation, facilities, capabilities, and other resources that complement those available within EMSL for new capability development. They strengthen EMSL capabilities and accelerate new capability development.

Partnerships represent a form of engagement where each party contributes their own resources toward a common research or technology development goal that benefits both. Resources such as materials, software, code, or experimental samples may transfer between partners, but funding does not.

The 2021 *EMSL Five-Year Strategic Science Plan* (hereafter, the Strategic Plan) presents industry, academic, and other partners as key engagements supporting the Research Areas composing our three Strategic Science Objectives (SSOs) and delineates a group of current partners in other Department of Energy, Office of Science (DOE-SC) facilities and U.S. organizations. The Strategic Plan also plans for strategic expansion of our partnerships (EMSL 2021 Strategic Science Plan, Section 6.4). These partnerships are detailed in the Major External Engagements section of each of the Research Area descriptions in the Strategic Plan. Each of these partners were identified with the goal of accelerating interdisciplinary and team science and providing the synergistic scientific and technical capabilities EMSL needs to achieve our SSOs and meet the future needs of the research community.

In Plan of Actions and Milestones (POAM) for Major Action S2, EMSL was asked to clarify its partnership strategy and supporting partnership mechanisms. In response, this addendum to the Strategic Plan describes our strategy for partnerships, the types of partnerships we support, and the mechanisms used to execute those partnerships.

2.0 Partnership Strategy

EMSL does not have a separate strategy for partnerships. Instead, EMSL's approach to selecting partners is driven by our overall science strategy reflected in our Digital Phenome (DigiPhen), Molecular Observation Network (MONet), and Modeling and Data Sciences (MDS) SSOs and in their supporting Research Area Roadmaps. This science strategy is presented in the Strategic Plan, with additional specifics detailed in our Strategic Science Objective Planning Addendum.

Needs for partnerships are identified in the Research Area Roadmaps, which outline the science and technology development necessary to establish one or more components of an SSO. Integrated Research Platform leaders and their teams develop Research Area Roadmaps in collaboration with the BER program research community to assure alignment with our SSOs and the specific needs of the EMSL user community.

Partnerships are evaluated based on the following criteria:

- Alignment with a BER strategic objective and one or more EMSL SSOs
- Scale and breadth of impacts
- Opportunity to amplify and leverage other investments
- Feasibility
- Timing.

Our approach is to prioritize and select partners that have the greatest and broadest impact across EMSL SSOs—those that offer to bring greater value to other EMSL investments and activities and to balance those impacts against feasibility (risk-reward). Science and technology (S&T) roadmaps provide guidance regarding the appropriate time for partnerships. Box 1 provides one example of timed, high-impact partnerships supporting the MONet SSO and its 2023 launch. In addition, EMSL and the Joint Genome Institute (JGI) recently agreed to partner in support of the MONet science objectives. EMSL will provide soil samples from its field sampling program to JGI for sequencing and metagenomics data provided without cost to EMSL. This is an example of a timely complementary resource partnership.

BOX 1: PARTNERSHIPS ENABLE MONET SSO

EMSL's partnership strategies that are key to achieving MONet goals include leveraging existing ecological research networks and NASA's low-Earth orbit (LEO) satellites. Networks such as AmeriFlux, the National Ecological Observatory Network, the Long-Term Ecological Research Network, and the Critical Zone Network have the infrastructure to maintain and collect relevant ecological data over extended time periods. As a critical component of MONet, EMSL anticipates co-locating above- and belowground field sensors of molecular processes at these sites to enable collection of longitudinal molecular data that can be associated with co-located long-term ecological monitoring data measured by research network eddy covariance towers. Correlation with hyperspectral imaging from LEOs can extend evaluation of ecosystem status to even larger spatial scales.

3.0 Partnership Types and Resources

EMSL collaborates with three types of partners (Table 1). Research and development partners work with EMSL to advance S&T developments aligned with the EMSL mission, the SSOs, and the long-term needs of the BER research community. These partners may be from industry, academia, or government. EMSL's activity is supported through a competitive process and funded by facilities research funds administered through our Scientific Partner Program. EMSL also partners to provide or leverage existing complementary capabilities (expertise, facilities, and technologies) of other institutions to conduct unique integrated research not possible for a single institution. The Facilities Integrating Collaborations for User Science (FICUS) program is a very successful and growing example of this kind of partnership. Other BER-funded research projects such as Scientific Focus Areas (SFAs) and early career projects often develop complementary resources, including field or experimental samples or tools and approaches. Those programs and investigators present our third type of partners, complementary resource partners (Table 1). These partnerships are funded as part of our facilities research (EMSL Intramural S&T Research Program) through a competitive process that encourages early engagement and involvement with users.

EMSL expects to continue funding strategic partnerships using current funding allocations for facilities research and the user program, which currently precludes large-scale engagements with SFAs like the Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) and Coastal Observations, Mechanisms, and Predictions Across Systems and Scales (COMPASS). Future engagements with these potentially critical partners and other SFA partners would be contingent on additional funding. Similarly, expansion of a complementary capability (e.g., FICUS) or research and development partnerships may depend on additional funding.

Examples of existing and potential future partners can be found in Sections 3, 4, and 5 of the Strategic Plan under the SSO Research Area descriptions.

Table 1. Partnership Types and Resources Required

Partner Type	Example	Funding Source*	New Funds	Mechanism
Research and Development	Industry	Facility Research	None	Scientific Partner Program
Research and Development	Academia	Facility Research	None	Scientific Partner Program
Research and Development	Government	Facility Research	None	Scientific Partner Program
Complementary Capability	FICUS	User Program	None	User Program
Complementary Capability	National Labs	Facility Research	None	Intramural S&T Research Program
Complementary Resource	BER SFA	Facility Research	None	Intramural S&T Research Program
Complementary Resource	EMSL Users	Facility Research	None	Intramural S&T Research Program

*EMSL funds support EMSL activities only. Funds are not provided to partner organizations. Facility Research is the DOE funding element of the EMSL operational budget that supports EMSL's Intramural S&T Research program.

4.0 Partner Engagement Mechanisms

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 may later evolve into a more formal relationship or partnership supported by a contractual mechanism that delineates terms, responsibilities, and requirements.

EMSL uses three engagement mechanisms to support partnerships. The agreement types, accountabilities, and responsibilities are presented in Table 2.

EMSL's Scientific Partner Program (Strategic Plan, page 6; Operations Manual Section 7) supports co-development of BER-relevant capabilities and technologies between research and development partners (Table 1) and the EMSL user program. Scientific partner proposals may be submitted at any time by individuals or groups to partner with EMSL staff to enhance an existing capability or develop and build unique new capabilities that enhance EMSL's user program. Capability development efforts are encouraged that utilize collaborative multidisciplinary teams, pooled or leveraged resources, unique operating environments, or other resources that may be beyond those available to individual researchers or teams. Scientific partner proposals are intended to leverage the combined resources, expertise, and capabilities of the partner institution to maximize impact for EMSL, the partner, and future EMSL users. In return for co-development, EMSL scientific partners may have priority access to the new capability for a negotiated and specified period. The award and timing of EMSL scientific partner projects are contingent upon EMSL strategic needs and the availability of EMSL resources.

Projects in the Scientific Partner Program are jointly funded—each institution contributes their own funds, and funds do not transfer between institutions. The terms and mechanisms are fit to IP and data protection needs. Projects are reviewed for technical merit, strategic alignment, and potential user program impact as described in the EMSL Operations Manual (Section 7). Facilities research funds support the EMSL Partner Program.

The FICUS program is the primary mechanism for supporting our Complementary Capability Partners (Table 1). FICUS was created in 2013 to accelerate ambitious user research projects. The program, spearheaded by EMSL and JGI, provides researchers with access to the world-class resources of multiple user facilities via a single user proposal. The intent of the program is to allow EMSL users to more easily leverage and integrate the capabilities of multiple DOE-SC user facilities to conduct otherwise difficult multidisciplinary research. FICUS projects are supported by funds from EMSL's user program and funds from the partnering user programs. Projects are selected for funding through a competitive peer-review process (EMSL Operations Manual, Section 5; Strategic Plan, Section 6.4).

Complementary capability partner engagements can also be supported by EMSL's Intramural S&T Research program where the project produces a capability that requires technologies that exist across more than one institution. For example, EMSL's Intramural S&T Research program has produced synthetic soil habits (RhizoChip) designed for imaging technologies available at several of BER's light sources. Those light sources were partnered with to support validation of the capability and set the stage for user access through the FICUS mechanism.

EMSL's Intramural S&T Research program (EMSL Operations Manual, Section 9) provides a competitive funding opportunity for EMSL investigators to create new capabilities for the BER research community. Funded EMSL investigators are expected to engage partners in the broad BER research community to leverage complementary resources that accelerate the research program, increase its impacts, and

demonstrate the utility of the research for potential future users of the new capability. Complementary resources provided by the BER research community may include unique biological or field samples, technical expertise, and/or time invested in beta testing software.

The program prioritizes and selects proposals through a scientific and technical peer-review process to provide both instrument time and operations funding to achieve the specific aims of intramural projects. Intramural program funding is not provided to collaborator labs or institutions external to Pacific Northwest National Laboratory.

Table 2. Partnership Mechanisms and Accountabilities

Partner Mechanism	Agreement	Accountable	Responsible
Scientific Partner Program	User Office	Chief Science Officer	IRP Leader
FICUS	User Office	Science Area Leader	IRP Leader
Facilities Research	User Office	Chief Science Officer	IRP Leader


EMSL

Addendum 3 Strategic Science Objective-Integrated Research Platform Integration

EMSL

2022 POAM S3

SSO-IRP Integration Addendum to the EMSL Strategic Science Plan

January 2023

Jay Bardhan, John Bargar, Scott Baker, Douglas Mans, Justin Teeguarden

PNNL-34044

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

2022 POAM S3

SSO-IRP Integration Addendum to the EMSL Strategic Science Plan

January 2023

Jay Bardhan, John Bargar, Scott Baker, Douglas Mans, Justin Teeguarden

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

The Integrated Research Platform (IRP) leaders, Science Area leaders, and chief science officer are responsible for the Strategic Science Objectives (SSOs), IRP capabilities, and science and technology (S&T) investments to maintain and advance premier capabilities for the Biological and Environmental Research program user community. In addition to biweekly meetings of these staff, the Environmental Molecular Sciences Laboratory (EMSL) manages multiple activities designed to continuously assess progress toward creation of unique, high-impact capabilities for users and to assess and evolve our IRPs. These activities include IRP leadership engagement with the user community, User Executive Committee and Science and Technology Advisory Committee meetings, S&T Leadership Team management of the Intramural S&T Research program and workshops to develop Research Area Roadmaps within each SSO. The 2021 *EMSL Five-Year Strategic Science Plan* provides a high-level view of the general relationships among these activities, assuring integration and alignment of the IRPs, SSOs, and other key elements. EMSL will expand on this description and provide additional details that inform our process for assuring integration of SSOs and IRPs and for assessing and evolving IRPs in an addendum to the Strategic Plan. Because the detailed plans for IRPs are intimately linked to our Strategic Plan partner planning and SSO planning addendums, appropriate references and connection to these addendums will be made.

Acronyms and Abbreviations

AI	artificial intelligence
BER	Biological and Environmental Research program
BGT	Biogeochemical Transformations
BSSD	Biological Systems Science Division
CAM	Computing, Analytics, and Modeling
CDO	chief data officer
COO	chief operations officer
CP	Computing Platforms
CSO	chief science officer
DigiPhen	Digital Phenome
DOE	Department of Energy
DT	Data Transformations
EESSD	Earth and Environmental Systems Sciences Division
EMSL	Environmental Molecular Sciences Laboratory
FY	fiscal year
HPC	high-performance computing
IRP	Integrated Research Platform
MDS	Modeling and Data Sciences
ML	Machine learning
ModEx	Model-Experiment Integration
MONet	Molecular Observation Network
POAM	Plan of Actions and Milestones
RF	Rhizosphere Function
S&T	Science & Technology
SAL	Science Area Lead
SM	Systems Modeling
SSO	Strategic Science Objective
TAP	Terrestrial–Atmospheric Processes

Contents

Figures

Tables

1.0 Background and Introduction

In Major Action S3, the Department of Energy (DOE) requested that the Environmental Molecular Sciences Laboratory (EMSL) provide additional details and clarification surrounding the role Integrated Research Platforms (IRPs) and their leadership play in assuring sufficient capabilities for users in the three Strategic Science Objectives (SSOs) into the future and how the IRPs will continue evolving over time to meet the science and technology (S&T) needs of the Biological and Environmental Research (BER) program user community.

This addendum to the 2021 *EMSL Five-Year Strategic Science Plan* (hereafter, the Strategic Plan) provides details on EMSL's approach and activities for building and deploying user capabilities aligned with our SSOs. This addendum also discusses how EMSL leadership, including the chief science officer (CSO), Science Area leaders, IRP leaders, and the User Office, interact to create those capabilities and how capability development and deployment are funded in alignment with user needs. Additional details regarding how the IRPs evolve over time with specific recent examples are also provided.

2.0 EMSL's Model for User Capability Development and Deployment

EMSL employs a plan, develop, and deploy model to produce capabilities for the BER research community (Figure 1). These activities are directed by our SSOs and coordinated across our S&T leadership to assure that (1) activities are integrated and (2) the capabilities in the IRPs evolve over time and are resourced in a manner that aligns with the SSOs and evolving user community needs. The CSO is accountable for capability planning and development aligned to the SSOs. Science Area leaders are accountable for user capability deployment.

IRP leaders have responsibilities across all three activities, serving as key integrators across SSOs and across the plan, develop, and deploy model.

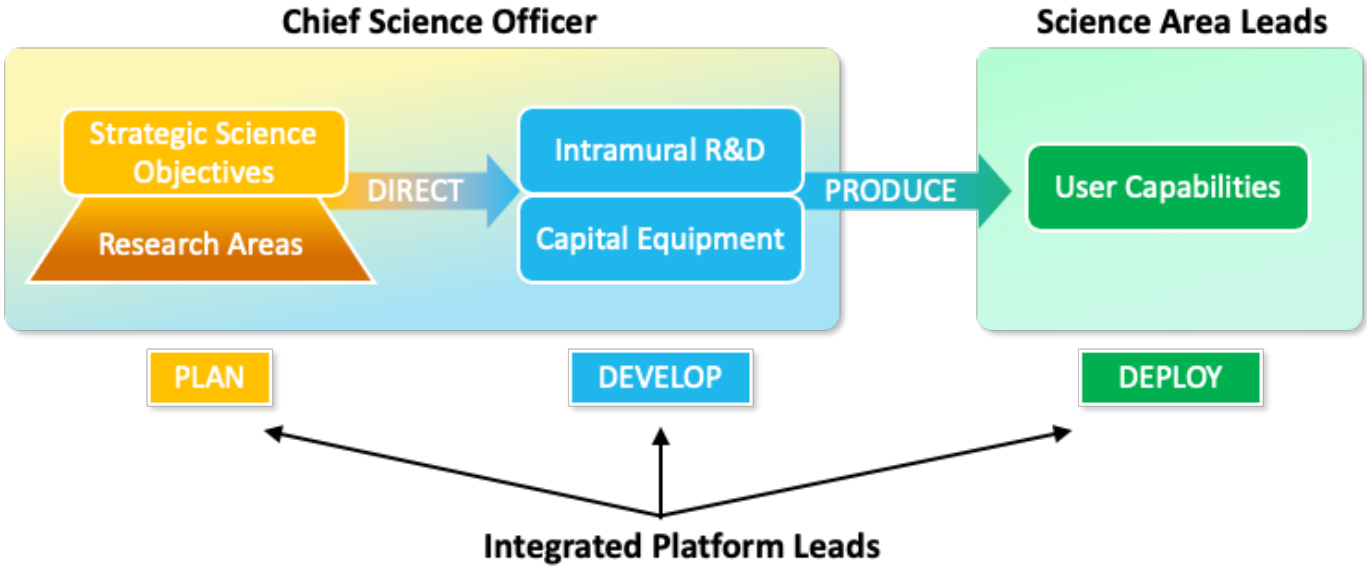


Figure 1. Summary of EMSL's plan, develop, and deploy approach to capability development that highlights the critical integration role that IRPs and IRP leaders play.

EMSL's IRPs play critical roles in each of the three main activities that produce user capabilities (bottom of Figure 1). Each IRP represents an area of domain excellence for EMSL, including in-house scientific expertise, cutting-edge, next-generation instrumentation, and specialized facilities that foster multidisciplinary team research. The IRPs were selected to amplify EMSL's core strengths and focus our science mission and SSOs on the most critical, unmet, and evolving areas of EMSL and BER science. IRP leaders serve as the primary interface for EMSL users to connect with EMSL's S&T capabilities, facilitate consultations on research design, and steward access to the scientific leadership available within the IRP multidisciplinary teams (Strategic Plan, Section 2 and Figure 3). Additionally, IRP leaders are EMSL's focal point for planning, leading, and executing our SSOs as well as capital and space planning with the user community. The EMSL 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 IRPs evolve over time to meet the changing science needs of the user community and changes in demand for IRP capabilities. EMSL's process for assessing and revising IRPs with current examples is provided in Section 3.0 of this document.

2.1 Capability Planning: SSOs and Research Area Roadmaps

The Strategic Plan describes our three SSOs — the Digital Phenome (DigiPhen), the Molecular Observation Network (MONet), and the Modeling and Data Sciences (MDS) Capability—and the S&T roadmapping and community engagement that produced them.

The purpose of EMSL's SSOs is to provide scientific direction and focus for EMSL and the BER research community toward delivery of a set of transformational experimental and computational capabilities for BER users and the nation. The scientific focus of the objectives directs and accelerates the evolution of our multidisciplinary science and advances EMSL's premier capabilities to meet the current and future needs of BER users and to produce high-impact scientific outcomes for the nation. Details on the implementation plans for the SSOs can be found in Strategic Plan SSO Planning addendum.

A group of Research Areas comprise each of the three SSOs. Each Research Area has its own goals and strategic roadmap that drive delivery of a fundamental element of an SSO. These roadmaps present the concise timeline of scientific developments needed to support an area of research and identify the resources—partnerships, instruments, computational tools, staff, and scientific developments—required to achieve that future state (Strategic Plan, Sections 3.2, 4.2, and 5.2; and SSO Planning Addendum, Appendix A). The Research Area Roadmaps are built from an assessment of the current state of our capabilities and resources and a clear definition of the desired future state. They are the more granular implementation plans that EMSL uses to execute each SSO.

The Research Area Roadmaps direct our priorities for capital instrument investments and Intramural S&T Research investments, both of which are used to develop new instruments, tools, software, and capabilities for users (Figure 1).

The EMSL CSO is accountable for establishing and executing the SSOs and the Research Area Roadmaps. Science Area leaders are responsible for developing the SSOs, and IRP leaders are the architects and stewards of the supporting Research Area Roadmaps. In practice, development and implementation is integrated across these members of EMSL's S&T Leadership Team.

2.2 Capability Development: S&T Research and Development, Capital Equipment

The two main driving areas for developing capabilities are EMSL's Intramural S&T Research and Development program and capital equipment program.

The Intramural S&T Research program creates new capabilities for the BER research community by advancing the goals laid out in EMSL's S&T roadmap presented in the Strategic Plan. Projects funded by the intramural program develop new scientific Research Areas and capabilities for the broader scientific community that are aligned with EMSL's S&T roadmap that directly affect EMSL's user program and drive progress in achieving BER objectives.

The program prioritizes and selects proposals through an S&T peer-review process to provide both instrument time and operations funding to achieve the specific aims of intramural projects. Intramural program funding is not available to support collaborators external to Pacific Northwest National Laboratory. Instrument hours

allocated for this program will be based on the instrument's total operating hours minus time scheduled for maintenance, upgrades, or repair. More detailed descriptions of the funding cycles and process can be found in EMSL's Operations Manual, Section 19. Figure 2 presents the overall approach and activities that assure the alignment and impact of projects.

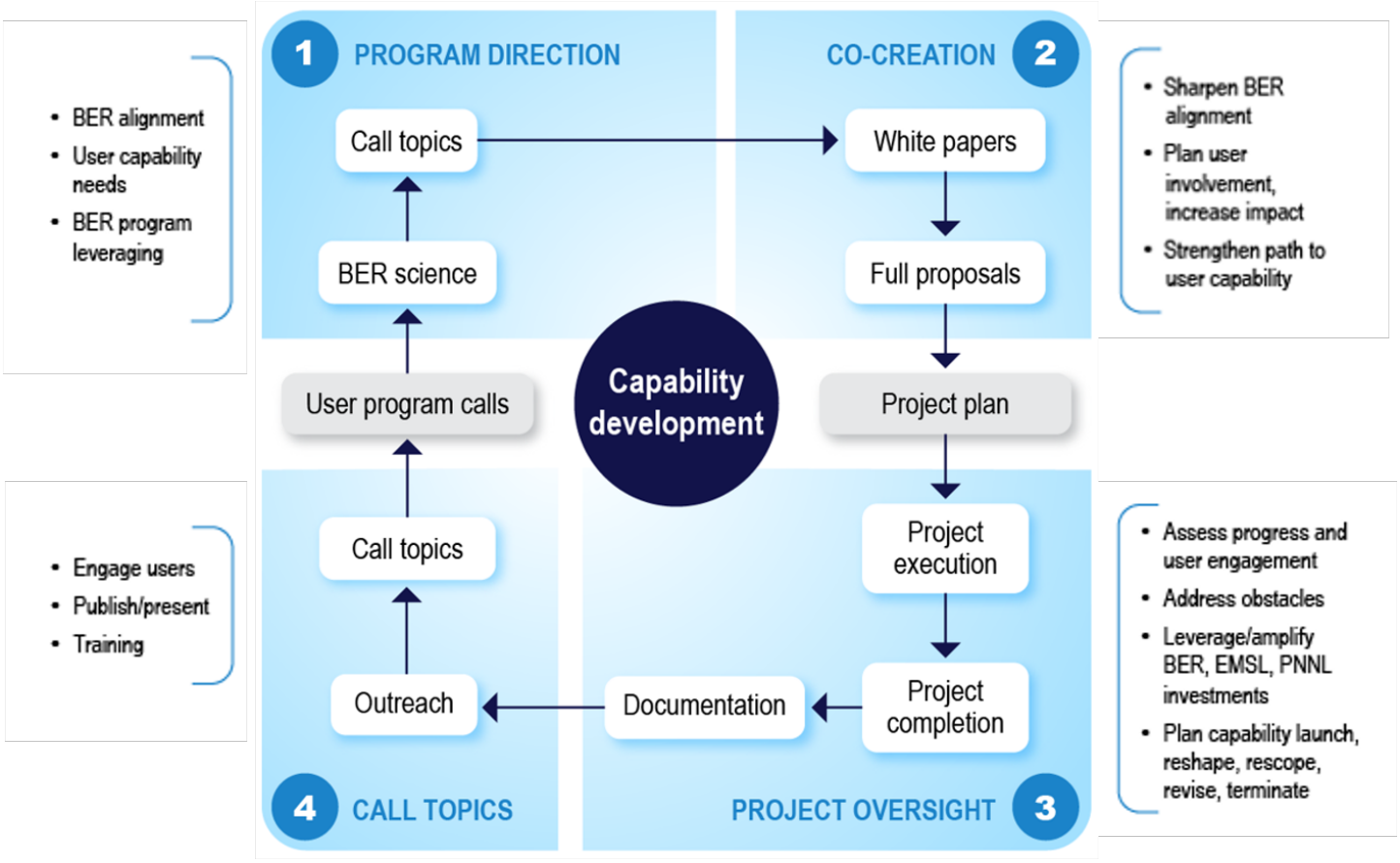


Figure 2. Summary of EMSL's Intramural S&T Research and Development program funding cycle. At all stages of the cycle, EMSL's S&T leadership, including the CSO, Science Area leaders (SALs), and IRP leaders, strives to sharpen alignment of these investments with BER user community needs.

The CSO is accountable for the EMSL Intramural S&T Research program. IRP leaders play several critical roles in the program. As the architects and stewards of the Research Area Roadmaps that provide specific targets for internal research and development, they drive project selection and execution through the following activities:

- Writing call topics
- Assuring alignment with user community needs
- Driving the review and co-creation processes
- Influencing final funding decisions
- Participating in quarterly project reviews.

As the primary interface with the user community accessing EMSL capabilities, IRP leaders play a critical role in recruiting users to participate in the EMSL Intramural S&T Research program as complimentary capability partners IRP leaders are supported in these activities by the SAL to assure coordination and resource balance across our SSOs and Science Areas.

EMSL also creates new premier capabilities for the user community through purchasing, modifying, and incorporating new instruments into unique workflows. The purchase of capital equipment, like investments made in the EMSL Intramural S&T Research program, is directed by our SSOs and the supporting Research Area Roadmaps, maintaining clear focus on the long-term needs of the BER research community. Decisions are guided by a set of requirements and priorities.

Requirements for capital equipment purchase:

- Direct benefit to the BER research community
- Align with EMSL SSOs.

Priorities for capital equipment purchase:

- Increases the capacity of one or more high-impact workflows to meet user capability demands
- Replaces and upgrades an aging workhorse instrument/component
- Satisfies a new capability required for roadmap/SSO success
- Significantly improves existing in-demand user capability
- Amplifies, compliments, or leverages other investments and activities (e.g., EMSL Intramural S&T Research rogram).

The CSO is accountable for capital equipment planning and purchase. The process is detailed in EMSL's Operations Manual, Section 17. Equipment requests will be prioritized by the CSO in consultation with the EMSL director, SALs, chief operations officer (COO), chief data officer (CDO), and IRP leaders before making final decisions on items to approve. In fiscal year (FY) 2023, EMSL revised the capital equipment planning process to include wider solicitation from EMSL staff interacting with users within the IRPs, as well as deeper involvement of IRP leaders in the review, prioritization, timing, and selection of capital equipment purchases. These activities include:

- Lunch and Learn seminar launch soliciting requests for capital equipment
- CSO, COO, SAL, and IRP leaders will perform initial screening and down selection (see guidance/criteria above)
- Selected quad chart development by proposers
- CSO, COO, SAL, and IRP leader selection
- EMSL Leadership Team review, and final decisions by CSO.

2.3 Capability Deployment: User Program and Integrated Research Platforms

SALs, working closely with IRP leaders, are accountable for deploying capabilities in the user program (Figure 1). SALs and IRP leaders work closely with EMSL's communications and technical staff to engage the user community in developing new capabilities, advertising emerging capabilities to BER and the research community, incorporating emerging capabilities into EMSL user program calls, and maintaining and improving

capabilities over time (Table 1). IRP leaders also coordinate capability use across IRPs and allocate resources for the user program.

Table 1. Capability deployment activities, accountabilities, and responsibilities.

Action Title	Accountable Staff	Responsible Staff
User engagement in capability development	CSO	IRP Leaders
New capability communications	SAL	Communications staff
New capability integration with user program calls	SAL	IRP Leaders
Apply capability in user program, resource capability	IRP Leaders	Staff
Steward, maintain, improve capability	IRP Leaders	Staff
Tracking utilization	COO	USO Manager

2.4 Resourcing Capability Development and Deployment

Resources for capability development come from funds allocated to facilities research capital equipment. Currently, the EMSL budget targets an allocation of $6 million per year for facilities research that supports capability development in the EMSL Intramural S&T Research program. This is the minimum sustained level of funding necessary to continue advancing innovations that produce premier molecular science, as well as computational and analytics capabilities, for users. Similarly, the EMSL budget targets $6 million per year in capital equipment investments. EMSL's capital equipment upgrade/replacement list, which includes urgently needed replacement of end-of-life-cycle, high-demand instrument and other near-term investments (including EMSL data archive replacement) required to make meaningful progress toward EMSL's SSOs, cannot be met with projected capital funding allocations within anticipated FY budgets. Looking forward, meeting the long-term needs of the user community by continuously improving EMSL capabilities will require a minimum sustained capital equipment budget of $6 million per year. Collectively, these budgets are required to assure that sufficient resources are available to meet the needs of the user community into the future.

3.0 Integrated Research Platform Evolution

IRPs are intended to evolve over time to meet the changing needs of the BER research community in EMSL's scientific areas described in our three SSOs and the supporting Research Areas. As the Research Areas and their S&T roadmaps progress and evolve over time, the IRPs that steward the growing collection of capabilities also evolve to better represent the scientific focus of those capabilities and the needs of EMSL users.

To effectively evolve in response to emerging needs, EMSL will continue using a wide array of regular user and BER engagement methods to identify those needs and update our Research Areas and Research Area Roadmaps accordingly. Maintaining awareness of the needs and directions of the user community and of BER is the principal means by which EMSL plans to meet evolving user needs. We use multiple mechanisms of communication and engagement to maintain a steady stream of information (Strategic Plan, Section 7). These include our Science and Technology Advisory Committee meetings, scientific exchange at BER PI meetings, scientific society meetings, EMSL Integration Meetings, BER- and EMSL-hosted workshops, and EMSL-led Research Area roadmapping workshops (five held in FY 2022) used to obtain BER research community feedback on our scientific directions. They also include regular input from the User Executive Committee and user surveys. We also use myriad modern communication mechanisms to signal our directions and attract users, including the EMSL website, videos, social media, The Molecular Bond newsletter, and EMSL Learn.

Aligning IRPs with user community needs, resourcing IRP capabilities with respect to user demand, and integrating IRPs are all discussion topics for the bi-weekly EMSL S&T leadership meetings. Thus, the S&T leadership, including the CSO, CDO, SALs, and IRP leaders, regularly assesses the IRPs and plans for and executes strategic changes.

For example, in response to the DOE Triennial Review, the Plant and Ecosystem Phenotyping IRP was refocused into the Rhizosphere Function (RF) IRP (https://www.emsl.pnnl.gov/science/expertise/rhizosphere-function/6). The RF IRP investigates the molecular mechanisms of root–soil–microbe interactions and the effects of root-controlled processes, including rhizodeposition on belowground C flux, biogeochemical nutrient cycling, plant resilience, and microbial community structure and function.

The RF IRP specifically addresses the impact of root system architecture and root exudates on highly interlinked rhizosphere components (microbial communities, organic matter, and soil mineralogy) in response to environmental perturbations. Research in the Biogeochemical Transformations (BGT) IRP compliments the RF IRP by focusing more fundamentally on processes common to all these systems, e.g., soil organic matter decomposition or mineral weathering (Figure 3), and on subsurface processes that occur outside the rhizosphere.

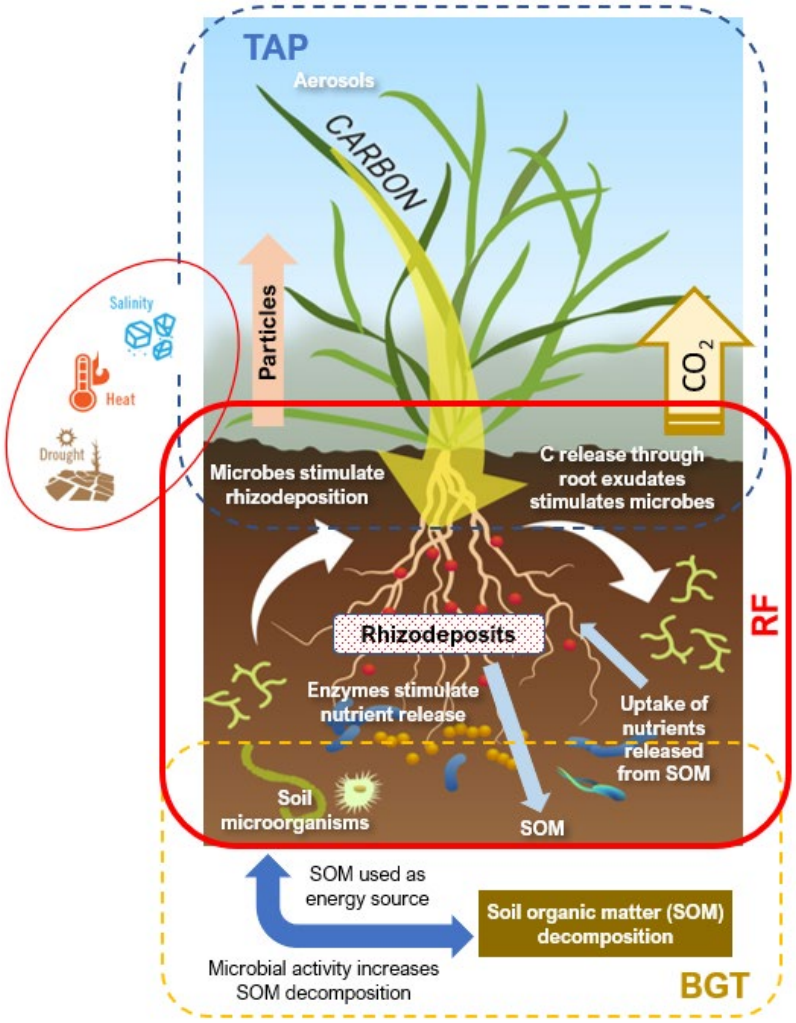


Figure 3. The science and scope of the RF IRP and its relationship with the BGT and Terrestrial-Atmospheric Processes (TAP) IRPs.

EMSL also recognized the opportunity to refine the focus for the Ecosystem Interfaces IRP, improve its connection to the other Environmental Transformations and Interactions Science Area IRPs, and meet the growing need for research that connects land and atmospheric processes. The Terrestrial-Atmospheric Processes (TAP) IRP was launched in FY 2023 and is led by EMSL's Swarup China (https://www.emsl.pnnl.gov/science/expertise/terrestrial-atmospheric-processes/3). The terrestrial ecosystem is a major source of aerosols and chemical species for the atmosphere. These include biological aerosol particles such as pollen and soil microorganisms or cell fragments, soil minerals ejected during rainfall or entrained during wind erosion, and volatile organic compounds released by plants. These particles, ranging in size from nano- to micrometers, remain aloft and undergo extensive chemical transformations as they react with other atmospheric constituents. Notably, these processes include warm and cold cloud droplet formation. Aerosols thus play important roles in regulating Earth's climate and in hydrological and biogeochemical cycles. The molecular processes by which aerosols transform and age strongly influences these behaviors. EMSL's TAP IRP is investigating these molecular transformations, the physical processes that control them, and the coupling of terrestrial and atmospheric processes. Knowledge from these studies allows us to understand the mechanisms by which these processes control air quality, Earth's radiation budget, cloud formation, and nutrient deposition.

Research in the TAP IRP examines interactions between volatiles and particles emitted by soils and plants and subsequent atmospheric processes, starting within the rhizosphere and extending up to the top of the troposphere. We will investigate how aerosols participate in warm and cold cloud formation by acting as cloud condensation nuclei or ice-nucleating particles and how these affect Earth's radiative budget. We will also study deposition of aerosols on terrestrial ecosystems. These activities will generate critical experimental and observational data to support Earth systems and climate models.

The focus on emissions from plants and soils is highly complementary to the RF IRP focus on belowground root–soil interactions and the BGT IRP focus on soil organic matter, microbiomes, and minerals. The TAP IRP mechanistically connects the BGT and RF IRP research domains to atmospheric processes and illuminates how belowground, surface, and atmospheric components interact as a complex system-of-systems. TAP research supports strategic MONet Research Areas—including Field Sensing, Sampling and Sensors and Atmospheric Aerosols and Particles—and MDS Research Area Model-Experiment Integration (ModEx) driven simulations.

Within the Computing, Analytics, and Modeling (CAM) Science Area, the single Systems Modeling and Data Sciences IRP has been divided, refined, and expanded into the Systems Modeling (SM), Data Transformations, and Computing Platforms IRPs (Figure 4). These three IRPs develop and deliver three highly complementary sets of capabilities for users that connect highly accessible modeling tools and frameworks with growing data streams supported by mid-range high-performance computing.

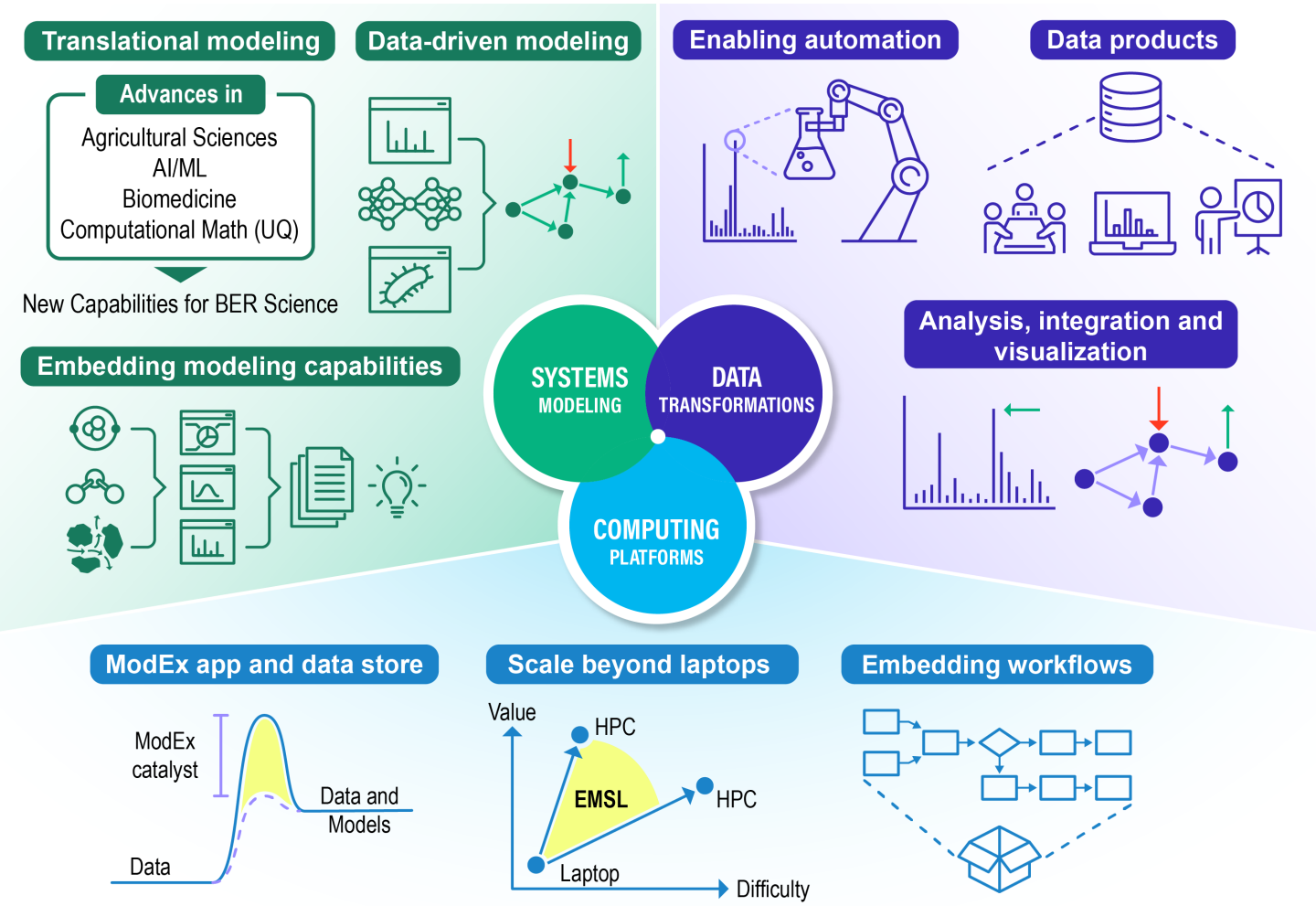


Figure 4. Overview of CAM IRPs—Systems Modeling, Data Transformations, and Computing Platforms.

Systems Modeling IRP: The SM IRP focuses on delivering state-of-the-art modeling approaches to accelerate prediction and control of complex systems in BER priority areas, focusing on leveraging the latest technologies in high-performance computing (HPC) and artificial intelligence and machine learning (AI/ML) to integrate EMSL experiments with modeling (ModEx). The IRP is named "Systems Modeling" to be inclusive of the diverse research questions and data types that advance BER science while maintaining focus on problems in which the inputs and outputs are controlling variables or priority outcomes for biological and environmental processes.

As shown in Figure 4, the SM IRP's three primary functions are to (1) embed modeling capabilities into multiscale paradigms, (2) catalyze the transfer of new simulation capabilities into the BER mission space, and (3) advance AI/ML-based approaches in model building and simulation. With these focuses, the SM IRP supports the DigiPhen SSO through modeling dynamic molecular assemblies (Structural Biology [SB] IRP) and metabolic modeling (Biomolecular Pathways [BP] and Cell Signaling and Communications [CSC] IRPs) and the MONet SSO through developing reactive transport models and data-driven multiscale models (BGT IRP and RF IRP). In total, the SM IRP's activities advance the multiscale modeling needs emphasized in the Earth and Environmental Systems Sciences Division (EESSD) and Biological Systems Science Division (BSSD) Strategic

Plans, and in multiple Grand Challenges identified by the Biological and Environmental Research Advisory Committee (BERAC), particularly "Develop integrative and interpretive computational approaches that can handle large, disparate data types from multiple and heterogeneous sources using advanced and exascale computing" (Grand Challenge 8.5).

Key capabilities for the SM IRP include EMSL's Tahoma supercomputer and its experienced team of HPC engineers and computational scientists, as well as EMSL's Network for the Execution of User Science (NEXUS) data management system.

Data Transformations (DT) IRP: The DT IRP specifically focuses on enhancing scientific discovery through data transformations, which could be embedded in interactive systems (for instance, exploratory graphical interfaces) or implemented in semi-automated pipelines. As highlighted in Figure 4, DT IRP work includes three key functions: (1) expert data analysis, integration, and visualization of EMSL experimental data, (2) enabling automation of EMSL instruments and workflows, and (3) curating and stewarding data products, that is, data sets released to the community to accelerate research in directions important to BER. These functions will be served in two forms: first, leveraging current capabilities (e.g., the Multiomics Analysis Portal [MAP]), and second, leveraging staff domain expertise to advance the use of novel techniques for BER problems.

The DT IRP will support the DigiPhen and MONet SSOs through multiple means, including direct application of developed capabilities, the transfer of new data transformation methods (e.g., AI/ML) from the external research community into EMSL practice, and leading two of CAM's redefined Research Areas, Computational Mass Spectrometry and Computational Imaging. The DT IRP's activities in study design, data processing, and data analysis will complement the work of numerous other IRPs, particularly those that leverage mass-spectrometry and imaging data. The DT IRP will work closely with the Computing Platforms (CP) IRP and the SM IRP where synergies exist with novel platforms and modeling approaches.

The DT IRP is designed to support the numerous needs and challenges for data analysis named in the BSSD and EESSD Strategic Plans and the Grand Challenges identified by BERAC. For example, the IRP directly advances BSSD's strategic goal to "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-4-1), and EESSD's strategic goal to "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" (EESSD-5).

Key capabilities for the DT IRP include the NEXUS data management system, the CoreMS framework for EMSL mass spectrometry data, the MAP for integrating and visualizing omics data, and the Open OnDemand platform allowing web-based access to the Tahoma supercomputer.

Computing Platforms IRP: The CP IRP focuses on enabling BER science to better leverage state-of-the-art computing capabilities ranging from emerging hardware (e.g., for AI/ML) to new software frameworks and platforms. As shown in Figure 4, the CP IRP has three key functions: (1) enabling codes and workflows to run efficiently on HPC resources where they did not previously, (2) embedding new workflow capabilities, such as accelerating ModEx, and (3) nucleating a hub to automate reproducibility and enhance provenance and portability, lowering barriers for ModEx involving EMSL data. Thus defined, the IRP's activities exclude both leadership-scale computing and research that is algorithmic or mathematical in nature. It is also important to distinguish the CP IRP's intent and scope from those managed by EMSL's CDO—in particular, direct, user-driven scientific impact in a specific science domain represents the scope of the CP IRP, whereas work driven by EMSL's operational activities are the scope of the CDO.

Key synergies for the CP IRP to advance DigiPhen lie with the SB IRP (for instance, AI predictions of protein structures and HPC for cryogenic electron microscopy data analysis) and the BP and CSC IRPs (platforms for metabolic modeling and AI-based data integration and image analysis). Synergies for advancing MONet lie primarily with the BGT IRP (automating ModEx integration of soil imaging with biological data). Through these activities and leadership of the Advanced Computing Technologies Research Area Roadmap, the CP IRP will directly advance EESSD's strategic goal to "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" (EESSD-5) and several of BSSD's strategic goals, especially to "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" (BSSD 5-1).

Key capabilities for the CP IRP include the Tahoma supercomputer, EMSL's Visualization Lab, and Pacific Northwest National Laboratory's Advanced Scientific Computing Research funded Center for Advanced Technology Evaluation, which acquires next-generation computing technologies for assessment and can be leveraged to improve the speed and quality of analyzing large datasets.


EMSL