DOE Office of Science

DOE/SC-2022

Envisioning Frontiers in AI and Computing for Biological Research

Workshop Report

U.S. DEPARTMENT of ENERGY | Office of Science

Advanced Scientific Computing Research Program

Biological and Environmental Research Program

Envisioning Frontiers in AI and Computing for Biological Research Workshop

February 4–6, 2025
Rockville, MD

Convened by

U.S. Department of Energy Office of Science

Advanced Scientific Computing Research Program

Biological and Environmental Research Program

Organizing Committee

Co-Chairs

Daniela Ushizima
Lawrence Berkeley National Laboratory

Christopher Henry
Argonne National Laboratory

Program Committee

Prasanna Balaprakash
Oak Ridge National Laboratory

Ayan Biswas
Los Alamos National Laboratory

Adrienne Hoarfrost
University of Georgia

Kirsten Hofmockel
Pacific Northwest National Laboratory

Neeraj Kumar
Pacific Northwest National Laboratory

Arvind Ramanathan
Argonne National Laboratory

Trent Northen
Lawrence Berkeley National Laboratory

Strategic Committee

Margaret Lentz
DOE Advanced Scientific Computing Research Program

Ramana Madupu
DOE Biological and Environmental Research Program

Todd Munson
Argonne National Laboratory

About ASCR

The Advanced Scientific Computing Research (ASCR) program within DOE's Office of Science advances science and U.S. competitiveness through investments in computational science, applied mathematics, computer science, networking, and software research. ASCR also develops and operates user facilities for high-performance and leadership computing and high-performance networking.

About BER

The Biological and Environmental Research (BER) program within DOE's Office of Science supports transformative science and scientific user facilities to harness the genomic potential found in nature, achieve a predictive understanding of complex systems, and provide fundamental research leading to solutions for U.S. energy and national security challenges.

Suggested Citation

U.S. DOE. 2026. *Envisioning Frontiers in AI and Computing for Biological Research Workshop Report*, DOE/SC-2022. U.S. Department of Energy Office of Science. https://doi.org/10.2172/2566158.

Cover graphic courtesy Argonne National Laboratory.

DOE/SC-2022

Envisioning Frontiers in AI and Computing for Biological Research

Workshop Report

January 2026

U.S. DEPARTMENT of ENERGY | Office of Science

Advanced Scientific Computing Research and
Biological and Environmental Research Programs

Contents

Executive Summary

Artificial intelligence (AI), machine learning (ML), and high-performance computing (HPC) are poised to transform biological research, spurring innovation in biotechnology and biosystems design. This transformation will bring an explosion of new capabilities to control the expression of genomic information in living organisms and harness that information to invent new biobased technologies (Jinek et al. 2012; NASEM 2025).

A recent report by the National Security Commission on Emerging Biotechnology states the widespread impacts of biotechnology "are not just matters of scientific achievement; they are questions of national security, economic power, and global influence" (NSCEB 2025). The U.S. Department of Energy (DOE) national laboratories are the world's greatest scientific infrastructure and are uniquely positioned to provide the resources and domain expertise needed to usher in this new era of AI-driven biotechnology, including creating and analyzing massive, open, and AI-ready datasets.

To better understand the opportunities and basic research needs at the interface of AI and biology, the Advanced Scientific Computing Research (ASCR) and Biological and Environmental Research (BER) programs in the DOE Office of Science (SC) organized a workshop on Envisioning Frontiers in AI and Computing for Biological Research (see Appendix A: Workshop Agenda, p. 49, and Appendix B: Workshop Attendees, p. 51). This workshop explored research intersections between BER and ASCR that will harness the power of AI and exascale computing to advance biotechnology.

Applying the full potential of artificial intelligence, machine learning, and computational sciences to biological research will drive transformative discoveries and enable unprecedented capabilities.

These new technologies will unleash and empower a new U.S. bioeconomy by (1) advancing predictive understanding and manipulation of biological systems, (2) enabling researchers to organize and simulate biological processes across vast scales, and (3) facilitating the discovery and design of new behaviors, mechanisms, and biological processes relevant to DOE missions. The workshop culminated in four priority research directions (see Fig. ES.1, p. iv) to guide future research and development within SC programs: Multimodal Data Assembly, Multiscale Biosystems Simulation, AI-Enabled Drivers for Experimental Systems, and Novel Algorithms for Genomics.

Integrating computation, experimentation, and next-generation automated technologies is expected to lead to the discovery and design of new biological behaviors and mechanisms. The workshop identified ways advanced computational methods can impact this mission by exploring novel algorithms, multiscale and multimodal modeling, data fusion, foundation models,

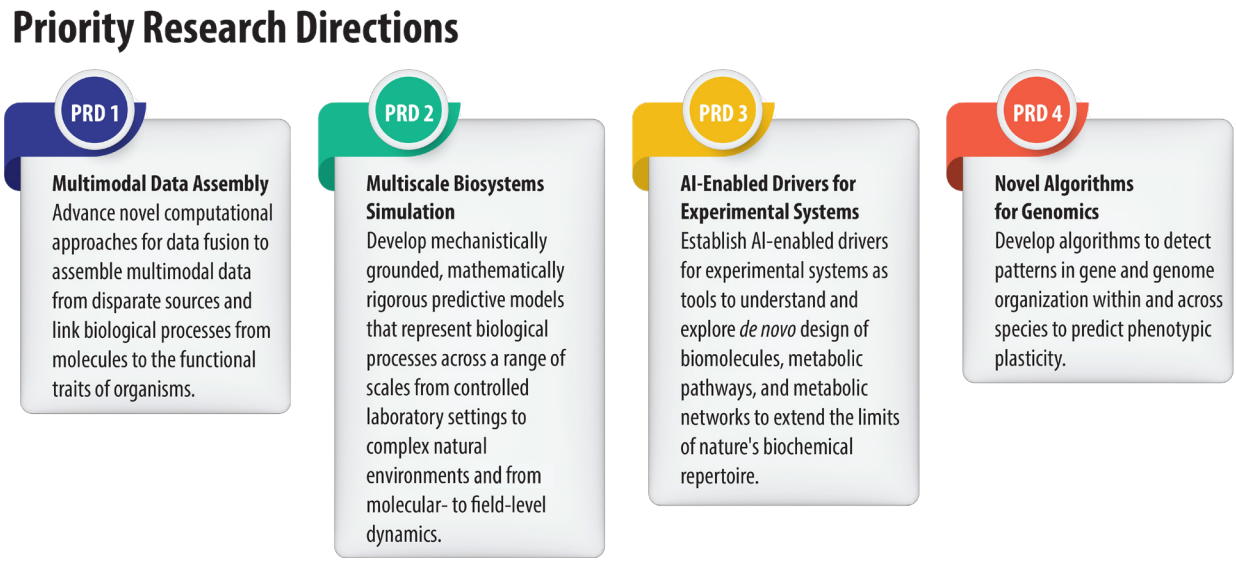


Fig. ES.1. Priority Research Directions. Workshop participants identified four priority research directions at the intersection of biology and AI to advance DOE missions and the U.S. bioeconomy.

digital twins, verification and validation, experiment design, and automated laboratories.

Participants assessed the current state, trends, and AI challenges at the interface of biology and computational science to identify new high-impact research opportunities with significant potential economic and environmental benefits.

Progress can be made by capitalizing on the profound computational capabilities spearheaded through ASCR investments in exascale architectures, HPC platforms, mathematics and computer science, and the wealth of BER-supported efforts to collect and analyze complex biological data at a scale unmatched by any other government or academic entity.

Chapter 1

Background

Within DOE, ASCR has led the development of new artificial intelligence (AI), applied math, and computer science capabilities, and has pioneered exascale computing architectures and the application of these groundbreaking machines across a host of scientific applications. The Argonne Leadership Computing Facility (ALCF), Oak Ridge Leadership Computing Facility (OLCF), and the National Energy Research Scientific Computing Center (NERSC) have been at the forefront of using high-performance computing (HPC) to tackle some of the most challenging scientific problems facing DOE's energy mission.

BER is a leader in large-scale biological data generation and cutting-edge research designed to understand the mechanisms and processes underlying complex biological phenomena, with biotechnology innovation as a primary goal. The program supports crosscutting synthesis across biological fields as well as user facilities and scientific resources that extract, organize, and classify biological data. These facilities and capabilities include the DOE Joint Genome Institute, Environmental Molecular Sciences Laboratory, DOE Systems Biology Knowledgebase (KBase), National Microbiome Data Collaborative, and structural biology and imaging resources at DOE light and neutron facilities across the country.

AI offers a unique and powerful opportunity to merge these two worlds. Advanced computer architectures operating at unprecedented scales and speeds, coupled with carefully designed new mathematical algorithms, can assemble and analyze vast biological data to extract meaning, reveal new insights, and autonomously guide experiments to both efficiently target knowledge gaps and home in on potentially groundbreaking processes and mechanisms (see Fig. 1.1, p. 2). Embedding deep biological knowledge into these algorithms will ensure computation provides scientifically relevant and meaningful results that power new, more accurate predictions and improve biosystems design. By integrating computation, experimentation, and next-generation technologies, researchers aim to simulate and manipulate biological systems across scales.

The Envisioning Frontiers in AI and Computing for Biological Research workshop hosted by ASCR and BER assessed current trends and challenges at the intersection of biology and AI (see sidebar, Supplemental

Supplemental Materials

Prior to the workshop, attendees were invited to submit position papers discussing key challenges and opportunities in formulating, implementing, and applying AI/ML frameworks for biological systems relevant to BER's mission space. This community input shaped the workshop agenda, panelist discussions, and workshop report.

The position papers and a report overview are available online:

- Position papers: DOI:10.2172/2512398
- Overview brochure: DOI:10.2172/2566160

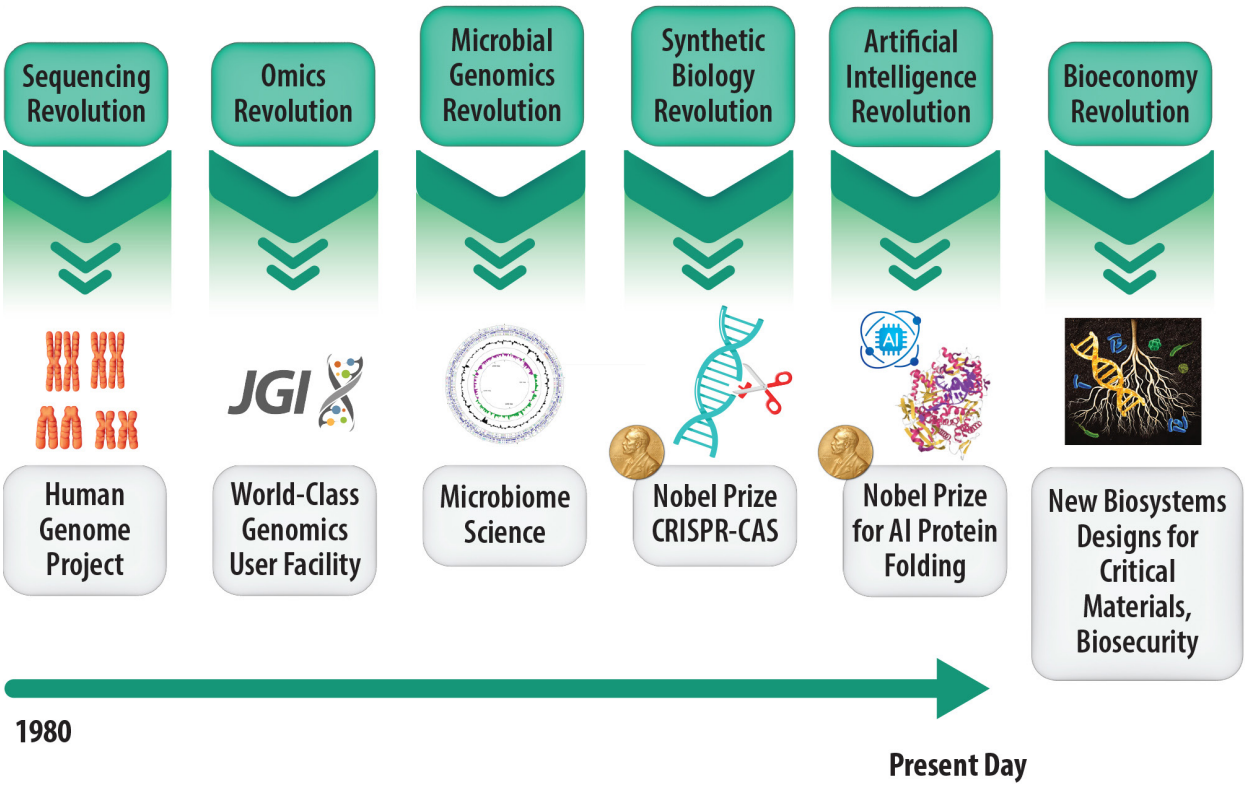


Fig. 1.1. Building a Legacy of Biological Research. In the last 50 years, researchers across the BER portfolio have made groundbreaking discoveries with everyday applications, facilitating the scientific revolutions that have powered U.S. leadership in biology.

Materials, p. 1). The workshop identified four transformative priority research directions (PRDs) that leverage ASCR's computational expertise and BER's leadership in biological systems research to advance DOE missions and the U.S. bioeconomy. Ch. 2 (see p. 11) describes the PRDs; Ch. 3 (see p. 25) discusses shared themes around data generation; Ch. 4 (see p. 29) discusses crosscutting focus areas for AI, including novel algorithms, multiscale multimodal modeling, data fusion, foundation models, digital twins, AI verification and validation, experiment design, and automated laboratories (see Fig 1.2, p. 3).

1.1 Overview of Computation and Mathematics Capabilities in DOE

For over 70 years, DOE and its predecessors have led the nation's development and application of advanced mathematics and computer science research to address the world's most formidable scientific challenges (ASCAC 2020). This work has pioneered advanced computational techniques in optimization and core mathematics, including differential equations, linear algebra, discrete mathematics and graph theory, as well as core computer science areas (i.e., massively parallel processing, scalable input/output, large-scale data analysis and visualization, and network protocols). Taken together, DOE research, technical advances, and leadership have produced groundbreaking results in various fields, including fluid and solid mechanics, materials sciences, computational chemistry, and biological modeling.

In 2023, the facilities subcommittee for the Advanced Scientific Computing Advisory Committee (ASCAC) was charged with assessing the necessity for new or upgraded facilities to ensure the Office of Science

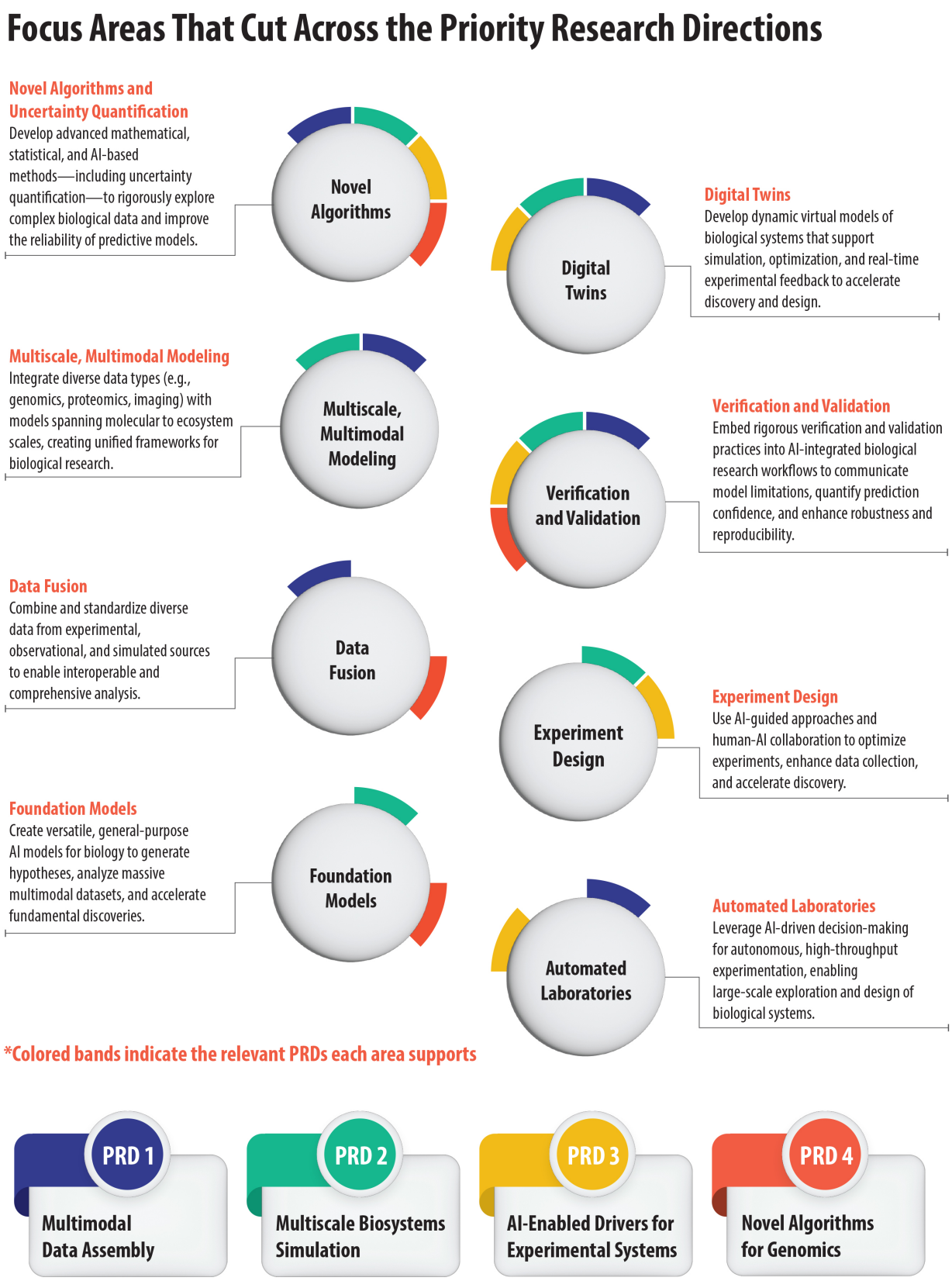


Fig. 1.2. Crosscutting Focus Areas. Workshop participants identified eight key focus areas spanning the priority research directions (PRD). These focus areas represent strategic opportunities for collaboration across ASCR and BER.

(SC) remains at the forefront of scientific discovery. The resulting report emphasized that ASCR systems are essential for maintaining this leadership, especially as science becomes increasingly interdisciplinary, integrated, and digital (ASCAC 2024). These facilities enable complex and diverse workloads running on petaflops and exascale supercomputers, many with hundreds of thousands of cores and hundreds of gigabytes of graphics processing unit (GPU) memory. As several DOE science programs produce large amounts of data, ASCR facilities are best utilized as a large, integrated ecosystem supporting SC programs alongside other ASCR efforts in software, algorithms, workforce, and science application components (ASCAC 2024).

Recent breakthroughs in DOE-based mathematics and computer science (U.S. DOE 2023a) have been instrumental in advancing AI across science and engineering. These innovations are ready to deliver reliable AI methodologies that will deepen fundamental understanding of biological processes. However, the complex, multiscale nature of biological dynamics is difficult to bridge. In addition, the underlying dynamic equations are unknown. Crosscutting, interdisciplinary AI research offers a unique opportunity to overcome these challenges by coupling experimental measurements directly with data from simulations and data-driven models.

1.2 Application Targets Within AI for Biology

ASCR is developing advanced exascale compute infrastructure, new algorithms, new data paradigms, and mathematical abstractions. These capabilities—particularly in AI—can address numerous challenges, knowledge gaps, and bottlenecks impacting BER's mission of advancing understanding of complex multiscale biological systems and their interactions. Discussions of these capabilities built on previous workshops held by ASCR and BER that explored how AI can advance biology research (see sidebar, Insights from Previous DOE Workshops on AI and Biology, p. 5). Critically, AI can improve the efficiency and efficacy of the laboratory and field experiments that feed AI analyses—creating a beneficial feedback loop to rapidly address key challenges.

Workshop participants identified five target areas in which AI could accelerate biological insights: (1) functional genomics; (2) metabolic engineering and synthetic biology; (3) microbiome analysis and engineering; (4) ecosystems analysis, prediction, and manipulation; and (5) data integration and knowledge representation.

Functional Genomics

Determining and manipulating protein function is a fundamental activity in biology, providing insights into the capabilities and design of protein molecules that act as sensors, motors, and biochemical catalysts. Accurate annotation requires solutions to many significant challenges the biological research community faces today (Liu et al. 2025), including (1) ensuring that functional annotations are accurately propagated across isofunctional protein families; (2) correcting numerous errors in functional assignments in existing databases; (3) discovering completely novel functions that have not yet been characterized by molecular biologists or biochemists; and (4) considering biological context during annotation, since functions vary with context. AI tools, particularly agents and foundation models, either trained or fine-tuned using exascale machines, could help to determine the combination of techniques and evidence that can be used to properly annotate proteins.

Understanding metabolic processes, particularly in plants and microbes, is also central to DOE missions in biology, including understanding biologically mediated chemical transformations and harnessing biology for the bioeconomy. However, detailed insights into gene and protein function are lacking, along with the ability to identify most metabolites and their intricate biological roles. Integrating computational approaches—such as deep learning–based spectral analysis, graph neural networks for metabolite interaction prediction, and probabilistic inference methods—can significantly enhance metabolite identification accuracy, deepen functional annotation, and enable more effective manipulation of metabolic processes.

Insights from Previous DOE Workshops on AI and Biology

Previous workshops held by ASCR and BER have explored how AI can advance biological research. This workshop builds upon insights from those efforts. In addition, a report led by the DOE national laboratories—*Advanced Research Directions on AI for Science, Energy, and Security* (U.S. DOE 2023a)—lays out a long-term vision for AI across DOE.

2019

Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence

U.S. DOE 2019

Organizer: ASCR

Priority Research Directions

- Integrate domain knowledge to improve accuracy and reduce data needs.
- Develop methods to interpret complex models and quantify model differences.
- Ensure methods are robust and reliable.
- Handle large-scale, noisy, and uncertain data effectively.
- Integrate AI into simulation codes to improve performance and robustness.
- Address challenges in simulation-based decisions, such as efficient exploration, data combination, and human–automation interaction.

2023

Artificial Intelligence and Machine Learning for Bioenergy Research: Opportunities and Challenges

U.S. DOE 2023b

Organizers: BER and the DOE Bioenergy Technologies Office

Opportunities

- Accelerate discovery with AI to analyze vast datasets to identify patterns and trends, leading to faster breakthroughs.
- Optimize bioprocesses through automated experimentation and AI-driven approaches to increase efficiency and yield.
- Engineer microorganisms with AI to design microbes with specific functions, such as producing biofuels or breaking down pollutants.

Challenges

- Address gaps in high-quality data, robust AI tools, and a skilled workforce by significantly investing in research and development and strengthening collaborations among academia, industry, and government agencies.

2024

Artificial Intelligence for the Methane Cycle

U.S. DOE 2024

Organizer: BER's Environmental System Science program

Opportunities

- Enhance the understanding and prediction of methane fluxes across various scales (from microbial populations to global systems) by improving data collection and integration and enhancing model design and accuracy.
- Bridge gaps between top-down and bottom-up methane flux estimates by developing comprehensive datasets and innovative modeling techniques and through infrastructure investment.

2024

A Unified Data Infrastructure for Biological and Environmental Research

BERAC 2024

Organizer: BER Advisory Committee

Goal: Review BER's existing data infrastructure and recommend a strategy for next-generation data management.

Recommendations

- Ensure infrastructure developers engage with the research community during the design and development process.
- Target high-impact science goals early to empower early adopters who can lead the charge on testing and leveraging the infrastructure.
- Use existing BER and ASCR resources as much as possible so new tools can focus on integration.
- Encourage use of new computer science methods (e.g., AI) through dedicated training, validation, and verification frameworks.

Metabolic Engineering and Synthetic Biology

Once functions for proteins and molecules are unveiled, it is possible to rationally re-engineer and modulate those functions to harness and optimize biological systems for bioenergy production and to develop solutions that address energy and crop-resilience challenges (Wu et al. 2025). This requires a deep mechanistic understanding of how a protein's sequence impacts its function. AI has already been revolutionary in this space by massively advancing the protein folding challenge with AlphaFold (Abramson et al. 2024), but much work remains. Emerging protein language model–based approaches also show promising progress in this area (Zvyagin et al. 2023). Furthermore, this metabolic engineering must be accomplished in a manner that supports continuous uploads of and updates to new datasets and ongoing improvements in analysis and understanding of existing datasets.

By combining imaging produced by DOE's world-leading high-energy light sources with protein structure data and mechanistic modeling approaches across scales, BER can unleash the potential to design not just individual proteins, but whole pathways, whole organisms, or even complex plant and fungal systems. This challenge requires a holistic understanding of these organisms at the mechanistic level, as well as knowledge of thermodynamic and kinetic parameters. If these research efforts are aligned with ASCR strategies, [e.g., accelerating hyperparameter optimization by using a fraction of a training dataset (Yu et al. 2024)], they could aid in (1) making whole-cell simulations more scalable, (2) predicting parameters, (3) correcting gaps in models, and (4) predicting modifications needed to achieve desired phenotypes.

Microbiome Analysis and Engineering

BER's mission also requires an understanding of how microbiome systems function, how they are connected to growth conditions, and how they respond to perturbations and manipulations, including individual organism behavior, interspecies interactions, and interactions with the environment (Knight et al. 2024). An additional challenge stems from the realization that many organisms in these systems cannot be isolated, and so even their genomic capabilities are uncertain. Knowledge of these systems can facilitate microbiome engineering, which will enable the optimization of individual microbes for particular steps in complex metabolic processes. If AI can operate in concert with microbiome modeling systems, digital twins, and cross-scale frameworks to predict interactions and behavior and fill in missing information in incomplete genomes, microbiome design will be enabled. AI can also greatly facilitate efforts to isolate currently unculturable microbes. These AI systems will need to operate across multiple scales and data modalities given the size and complexity of most microbiome systems of interest.

Ecosystems Analysis, Prediction, and Manipulation

Beyond the microbiome level, the BER mission involves the study of a wide range of natural systems. Of particular interest are soils, which represent one of the most complex biological systems and are critical to the bioeconomy, crop resilience, and understanding ecosystem response to a range of conditions (Knight et al. 2024). Despite over 100 years of study, significant knowledge gaps remain concerning the molecular-scale mechanisms that drive organismal and interkingdom interactions and how these processes scale to shape ecosystem-level dynamics and responses to extreme weather (Jansson and Hofmockel 2020), highlighting the tremendous complexity and multidisciplinary challenges that must be overcome to understand soil systems.

Rapidly advancing a causal and mechanistic understanding of soil systems will require the integration of diverse datasets, including multidomain omics analyses, abiotic controls, and plant communities, all within a 3D environment with physical limitations on flow and gas exchange. Tracking the associated spatial and temporal dynamics requires concerted analysis by laboratory, field, and computational scientists. Multiscale, multimodal data integration can benefit from computational models capable of analyzing physicochemical and biological processes through the fusion of diverse analytical modalities, such as multiomics and sensor

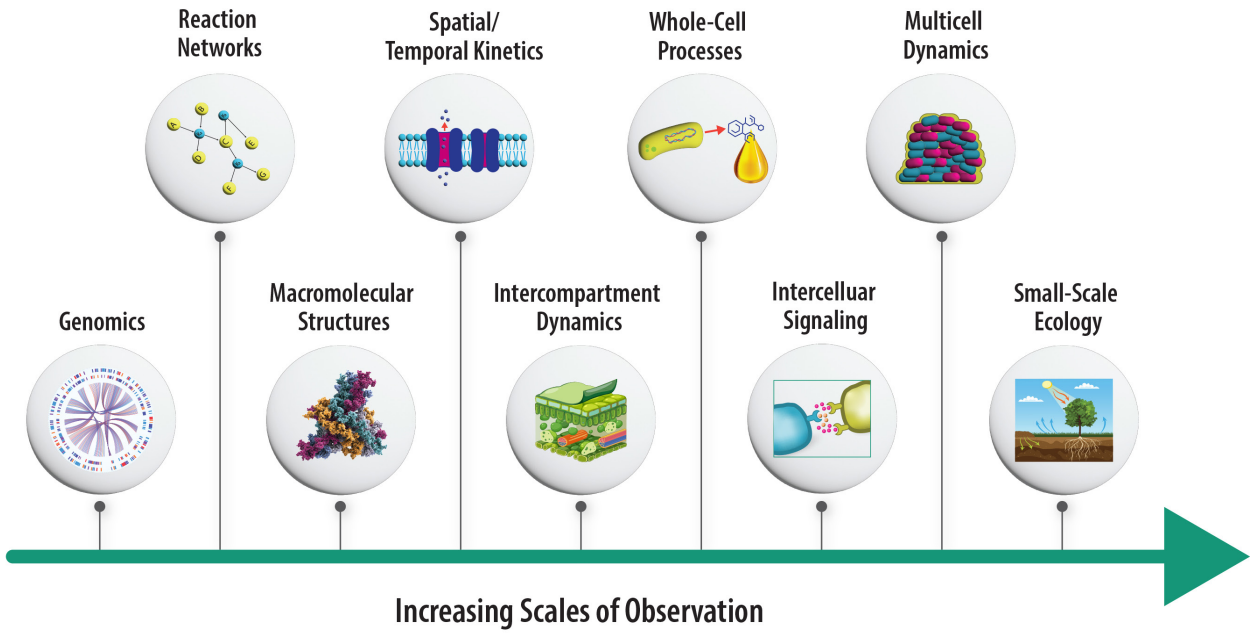


Fig. 1.3. Integrative Systems Biology. Understanding biological systems across scales of observation and complexity, from molecules to ecosystems, is a central challenge within BER that must be solved to discover the factors influencing function and processes within biological and environmental systems.

data. Leveraging ASCR's advanced computational tools can systematically address extensive laboratory and field datasets, enabling rigorous hypothesis generation (Mukhtar et al. 2022). These tools, such as accurate multiscale direct mathematical solvers for complex heterogeneous systems, are relevant across wide biological scales (see Fig. 1.3, this page) and can be coupled with domain-informed AI and parallelized deep networks optimized for exascale architectures. Using exascale machines, it is possible to train deep learning approaches from scratch, a method that could be used to scrutinize existing scientific understanding, quantify uncertainty, and identify knowledge gaps to improve models and their predictions.

Data Integration and Knowledge Representation

Data integration and knowledge representation form the foundation upon which all other biological understanding and discovery rest. The capacity to capture and synthesize data at scale, and to associate that data as evidence for consistently and accurately represented biological knowledge, is paramount to advancing biology as a science.

DOE is mobilizing its national laboratories to partner with industry, leveraging its unique role as a data generator (One Big Beautiful Bill Act 2025)—particularly through its user facilities—to curate and preprocess high-quality, AI-ready scientific data, which will then be made accessible to the research community along with specialized AI models via the American Science Cloud, a dedicated platform for scientific research, data sharing, and computational analysis.

Multiomics data from plants and microbes relevant to DOE now exist for billions of genes, vast numbers of biochemical molecules, millions of genomes, and hundreds of thousands of samples with inconsistent metadata, IDs, and analytical protocols (Anderson et al. 2025). Additionally, numerous competing ontologies represent knowledge of protein functions, metabolites, environments, cell types, and biological phenomena, with incomplete mapping to associated molecular representations (e.g., metabolites, reactions,

and macromolecules). AI can aid in reconciling and mapping ontologies to one another and in proposing relevant molecular representations. Advanced mathematics will be instrumental in multimodal registration of differing imaging modalities and in identifying missing or redundant representations. Exascale-aware libraries can also help address needs for latency hiding, improved vectorization, threading, and strong scaling in tasks involving comparison between long molecular representations (ECP 2025).

While outside the scope of this workshop, data management is essential to biology. AI can benefit efforts to integrate and reconcile sample metadata; map, query, and interpret data; and prioritize data acquisition. ASCR's report on *Management and Storage of Scientific Data* describes the benefit of data management to DOE research (U.S. DOE 2022a).

1.3 AI-Enabled Success Stories in Biology

AI-driven tools support the modeling of complex biological systems such as virtual cells, allowing scientists to simulate and study cellular processes in unprecedented detail. By combining computational power with biological insights, researchers can achieve more efficient and effective outcomes, driving innovation and addressing pressing global issues in health, energy, and the environment.

Protein Folding

One of the greatest recent successes of AI in biology was the development of AlphaFold and RoseTTAFold (Baek et al. 2021; Jumper et al. 2021), which led to a Nobel Prize in 2024 (see sidebar, DOE Powers Discovery, p. 9).

Although protein folding is one of the most complex challenges in biology, it became a target application and early success because of numerous advantages. First, all experimental protein structure data were aggregated, mapped, annotated, curated, and neatly organized in a single public repository, the Protein Data Bank (PDB; Burley 2025). The data were never cross-contaminated with computational predictions but were stably stored in a single public location for decades. The PDB facilitated the identification of promising, varied targets for novel structure determination, which significantly contributed to protein structure discovery. Second, Critical Assessment of Structure Prediction (CASP) contests led to the creation of objective benchmarks against which folding prediction tools could be tested, and these contests also inspired rich competitions from tool builders (Kryshtafovych et al. 2023). CASP events also led to the standardized development and deployment of folding tools, paving the way for the development of hybrid combinations of approaches (e.g., machine learning and molecular dynamics). These approaches provided insights into generalized features of the protein structure problem that were instrumental in creating successful AI solutions (e.g., conserved folds and conserved links between sequence, structure, and function). Third, protein folding benefited from the massive amount of highly interrelated protein sequence data that provided an evolutionary context to the folding problem. Lastly, protein folding and simulation efforts have always been at the forefront of the application of ASCR's HPC platforms to biology. Protein simulation efforts are one of the driving problems motivating the use of exascale platforms for biology.

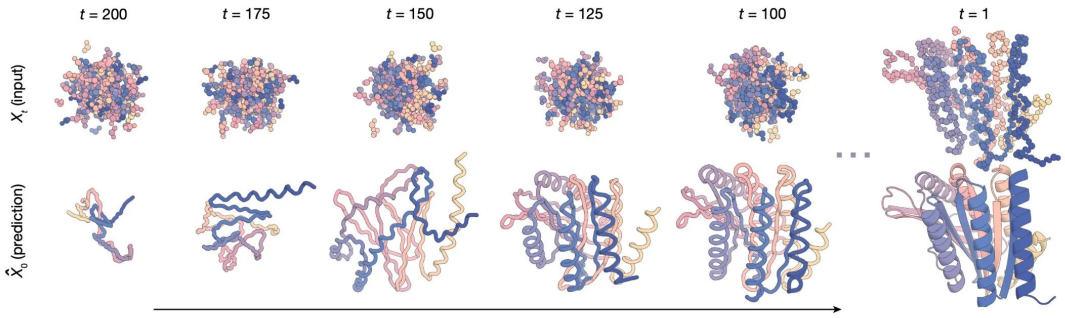
For definitions of "threading" and other discipline-specific terms found throughout the report, see Appendix C: Glossary, p. 53.

Molecular-Scale Techniques

MetaHipMer is another example of successful collaboration between computer scientists, biologists, and HPC. This computational tool efficiently assembles DNA sequences from complex microbial communities using a GPU-accelerated implementation to address challenges in metagenome assembly, including irregular memory access patterns and the need for dynamic data structures, by leveraging GPU optimization techniques and memory management strategies (Awan et al. 2021). MetaHipMer2 shows significant overall performance improvement, as tested on DOE's

DOE Powers Discovery: An AI Success Story

University of Washington biochemist and computational biologist Dr. David Baker, along with two colleagues, was awarded the 2024 Nobel Prize in Chemistry for his pioneering work in computational protein design using diffusion models, a breakthrough event that has accelerated the entire field of protein engineering and design of novel biomolecules not found in nature. The Nobel-winning research was performed using DOE's high-performance computing resources at the National Energy Research Scientific Computing Center. This landmark achievement exemplifies how AI-driven approaches, enabled by DOE supercomputing, have advanced computational protein design and protein structure prediction, leading to innovative applications in biotechnology, biomanufacturing, energy, agriculture, and medicine.



Diffusion models can generate new protein backbones, geometries, and sequences that were not included in the datasets used for training such models. [Reprinted under a Creative Commons Attribution 4.0 International License (CC BY 4.0) from Watson, J. L., et al. 2023. "*De novo* Design of Protein Structure and Function with RFdiffusion," *Nature* **620**, 1089–100. DOI:10.1038/s41586-023-06415-8.]

Summit supercomputer. This work highlights significant progress toward adapting metagenomic workflows to GPU-dominated exascale computing systems.

AI has also had a profound impact on instrumentation in biology (e.g., in mass spectrometry). Mass spectrometry metabolomics provides direct biochemical measures of biological processes and is often used to provide a functional complement to DNA sequencing. AI methods are already proving powerful in extracting additional information from these high-dimensional datasets to improve the currently small fraction of metabolites that can be identified in a metabolomics experiment. One recent example of this is BUDDY, a software tool that is able to accurately determine molecular formulas for metabolites through bottom-up interrogation of mass spectrometry data (Xing et al. 2023).

Other DOE Efforts

The collaboration between DOE and the National Cancer Institute (NCI) to advance precision oncology and scientific computing is another success story for AI-enabled biology. This effort resulted in the development of new scalable deep learning algorithms operating on DOE exascale platforms, which catalyzed computational drug discovery for cancer therapeutics (e.g., Lawrence Livermore National Laboratory–BridgeBio partnership). Collaborators also released

datasets and new AI models that are now heavily used by the cancer research community.

DOE's contributions to the BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies) have accelerated the pace of innovation in neuroscience, data integration, and cross-disciplinary collaboration. The BRAIN Initiative, which is a public–private partnership, supports the development of novel neurotechnologies and tools for monitoring brain function, including dynamic imaging.

During the COVID-19 pandemic, DOE established the National Virtual Biotechnology Laboratory (NVBL), a consortium leveraging AI and computational models. NVBL significantly contributed to the pandemic response by developing exascale-aware tools such as epidemiological models, simulations, and new testing protocols (Clyde et al. 2021; U.S. DOE 2022b). AI has also been instrumental in accelerating drug discovery processes, which enable rapid discovery and optimization of new materials. These collaborations demonstrate the power of AI and cross-disciplinary efforts in driving scientific breakthroughs.

Chapter 2

Priority Research Directions

The Envisioning Frontiers in AI and Computing for Biological Research workshop identified four priority research directions (PRDs) representing critical scientific areas for applying and advancing AI in biological research (see Table 2.1, this page). These PRDs, while presented as distinct thrusts, are inherently interconnected, and all share two significant themes: the need for (1) for data generation (see Ch. 3: Data Generation for AI, p. 25) and (2) crosscutting AI algorithms, methods, and models (see Ch. 4: Crosscutting Approaches, p. 29).

Workshop participants discussed the reasoning behind each PRD, detailed the impact it can have on biological research, defined the key science questions it could answer, and identified target activities (i.e., ideal strategies for the execution of the PRD).

Table 2.1. Crosscutting Tasks and Challenges Associated with Proposed Priority Research Directions

Priority Research Direction	Example Biological Task	Corresponding Computer Science/Math Challenge
PRD 1 Multimodal Data Assembly	Integrate imaging, omics, and text metadata to discover determinants of function and novel pathways	Exascale-scalable manifold alignment and optimal-transport fusion with uncertainty quantification
PRD 2 Multiscale Biosystems Simulation	Predict plant–soil–microbe interactions and phenotypes over time and space	Multigrid partial differential equation solvers, surrogate molecular dynamics and ordinary differential equation models, and adaptive mesh refinement
PRD 3 AI-Enabled Drivers for Experimental Systems	Design enzymes, pathways, and microbiomes *de novo* to manipulate expressed phenotypes	Novel, scalable optimization strategies and algorithms and reinforcement learning controllers
PRD 4 Novel Algorithms for Genomics	Detect regulatory motifs and community network modules	Beyond attention-based transformers; graph neural networks, transfer learning, and uncertainty quantification

2.1 Multimodal Data Assembly

> **PRD 1:** Advance novel computational approaches for data fusion to assemble multimodal data from disparate sources and link biological processes from molecules to the functional traits of organisms.

Rationale (Challenges and Opportunities)

Biological data are typically sparse, noisy, uncertain, and often lack standardization. Further, knowledge of the molecular mechanisms and even fundamental principles governing the behavior of most systems is incomplete and often fragmented. Fundamental limitations in measurement strategies make interrogating molecular entities within even model systems and diverse conditions challenging. To address and overcome these challenges, exascale computing for reasoning models, data fusion, optimal experimental design strategies, and verification approaches are needed, along with new integrative experimental, computational, and theoretical strategies informed by advances in AI foundation models. These strategies will require broad advances in automated laboratories and digital twins that enable reasoning models with experimental feedback to accelerate the generation of multiscale biological datasets.

Key Questions

- What computational approaches can be developed to fuse complex biological data (e.g., imaging, omics, abiotic conditions, and natural language text) to enable the discovery of new biological behaviors, mechanisms, and design principles, while simultaneously addressing data interoperability, noise, standardization, and uncertainty quantification?
- How can these approaches best leverage emerging data integration infrastructures like the BER Data Lakehouse and the ASCR American Science Cloud, and how can this infrastructure best serve data fusion needs?

Impact

Multimodal data assembly approaches will improve the capacity to integrate and synthesize extensive collections of existing biological data and to use existing data to rapidly contextualize new data as it is generated. This will improve both experimental design and the quality and value of data generated, enhancing queries across resources and the effectiveness of all ongoing biological research programs. These capabilities can support a wide range of DOE-relevant challenges in biotechnology innovation, bioenergy, biomaterials, and phytomining, including:

- Integrating omics and geochemical data to understand microbial community responses to abiotic stressors across diverse soil and plant systems
- Modeling drought-induced shifts in root exudate chemistry
- Understanding rhizosphere microbiome dynamics
- Developing advanced biodesign concepts to engineer molecules, microbes, plants, and microbial communities to extract and recover critical minerals and materials (CMM) with enhanced selectivity from natural and complex environments

Target Activities

Improve Experimental Strategies and Data Standardization. Many areas of DOE interest lack datasets of sufficient size to support deep AI analysis, particularly for critical conditional data on cellular and molecular dynamics, physiology, fitness, and activity in diverse relevant conditions. Implementing strategic experimental design and replicable data acquisition roadmaps is essential for producing high-quality, standardized datasets that comply with FAIR principles (Findable, Accessible, Interoperable, and Reusable; Wilkinson et al. 2016, 2019). These datasets are needed to create robust multimodal biological models (see Ch. 3: Data Generation for AI, p. 25). Importantly, this strategy should leverage DOE's existing strengths in data generation, such as extensive genomic sequence libraries and macromolecular structure datasets (Berman et al. 2000; Arkin et al. 2018) while addressing critical gaps in contextual information

(e.g., conditional data on cellular physiology and dynamics under diverse conditions, along with more complete metadata).

AI approaches can partially compensate for sparse or lower-quality data, but their greater promise lies in guiding data acquisition itself. For example, within an active learning framework (Lookman et al. 2019), AI models could identify high-value knowledge gaps and suggest targeted new experiments or measurements (e.g., adding specific controls, internal standards, or undersampled conditions) that would most improve predictive accuracy or reduce model uncertainty. By tightly integrating exascale computing for AI guidance with improved data standards, a feedback loop can be established in which better data leads to better models, and those models in turn inform more strategic experiments to advance progress toward a more predictive, cross-scale understanding of complex biological systems in line with DOE's mission.

Develop Scalable Methods To Manage Data Uncertainty and Sparsity. Advancing robust uncertainty quantification requires developing theoretically grounded and scalable methods specifically engineered to handle the sparsity and high dimensionality inherent in biological datasets. These methods should be capable of generating statistically rigorous confidence intervals that accurately reflect prediction reliability across complex molecular interactions. To ensure data integrity prior to downstream analysis, it is vital to establish automated data quality assessment architectures that implement sophisticated statistical approaches to detect, characterize, and remediate experimental artifacts and systematic biases inherent in multiomics datasets.

Innovations are needed in ensemble methodologies that integrate predictions from diverse modeling paradigms [e.g., physics-based simulations (Abramson et al. 2024), deep learning architectures (Ballard et al. 2024), multiplex network learning (Sullivan et al. 2024), and knowledge-driven approaches (Li et al. 2024)]. Ensemble methodologies should incorporate principled uncertainty propagation to enhance the robustness and reliability of gene function predictions. Furthermore, developing advanced transfer learning frameworks is essential to quantitatively characterize domain shifts when transposing models across phylogenetically diverse species or variable conditions, enabling precise evaluation of model generalizability boundaries and facilitating targeted refinement through domain adaptation techniques that preserve biological relevance. These methodological advances collectively establish a mathematical foundation for confidence-aware biological discovery systems that rigorously account for uncertainty throughout the analytical pipeline. If such approaches were implemented in computational platforms like the DOE Systems Biology Knowledgebase (KBase) or the National Microbiome Data Collaborative (NMDC), these services could provide researchers with a detailed understanding of how gene annotations, microbial traits, or insights from samples propagate among similar entities, or quantify confidence in these types of inferences.

Improve Capacity To Investigate Sources of Data Variability and Noise Amid Data Scarcity. One of the great challenges associated with integrating biological data from disparate sources is understanding the causes of variation across datasets. Replicates within a single laboratory are often extremely similar, while replicates across laboratories display greater variability; therefore, understanding and reducing interexperimental and interlaboratory variability is critical to data assembly (Novak et al. 2025). AI methods employing statistical anomaly detection, domain-adaptive learning, and Bayesian uncertainty quantification can mitigate this variability by systematically identifying which types of biological data are most susceptible to discrepancies caused by different laboratory protocols. Given the inherent scarcity, variability, and noise in biological data, robust computational methods (e.g., tensor-based imputation, sparse representation learning, graph-based denoising algorithms, and probabilistic generative modeling) are needed for training and inference from incomplete data.

Integrate Multimodal Data To Handle Incomplete Data. Leveraging existing multidisciplinary data, including epigenetic, multiomic, abiotic conditions, and phenotypic data, is required to develop models

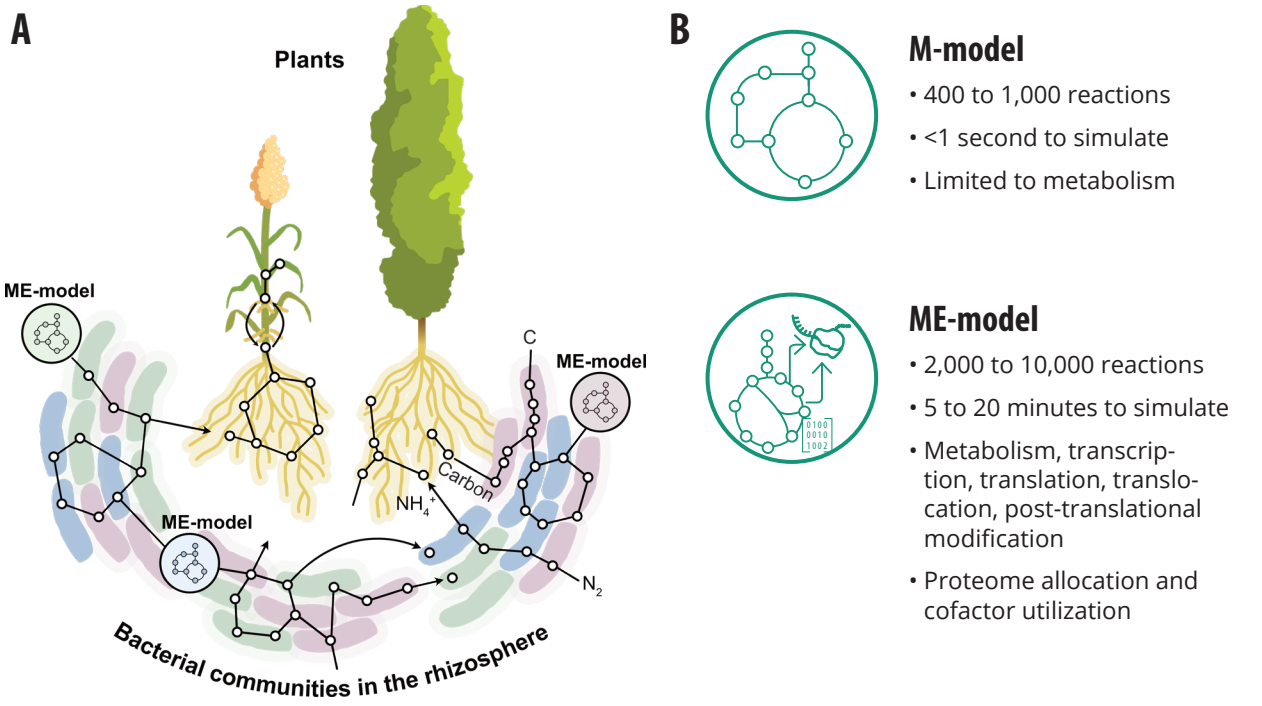


Fig. 2.1. Complexity of Plant–Microbe and Microbe–Microbe Interactions in Rhizosphere Systems. (A) An example of a mechanistic digital twin for rhizosphere systems is a community metabolic (M) or metabolic-enzyme (ME) model. These models can capture, represent, and mechanistically explain dynamics in molecules, enzymes, and species over time, enabling them to guide the manipulation of experimental and engineered biological systems to meet desired objectives. **(B)** Mechanistic models of metabolism only (M-models) have three advantages: they are smaller, have fewer parameters, and are less computationally intensive to run. However, these models lack the capacity to fully capture dynamics outside of metabolism (e.g., gene expression, regulation, and shifts in macromolecular processes). Expanding M-models to include these systems creates ME-models, which are far more capable of representing biology comprehensively. However, ME-models are also larger, more parameter-intensive, and more computationally costly to run. Proper model selection is crucial for optimizing digital twinning. [Courtesy University of California–San Diego]

that improve the understanding and prediction of functionality from molecules to systems. AI models have the potential to analyze large and complex multimodal datasets (Ushizima et al. 2021), like those for multiomics (Yetgin 2025); identify patterns; and make predictions about the changing behavior of dynamic hierarchical systems. Moreover, this approach opens new opportunities for model-informed experimental design, in which data used for multiscale analysis (Yoon et al. 2024) both guides and is guided by models, enabling rapid iterative learning. Computational tools and models capable of effectively integrating sparse, heterogeneous, and high-dimensional data are critical for making accurate predictions despite incomplete information.

Use Deep Learning Architectures for Multimodal Representation and Integration. Developing new computational approaches and deep learning architectures that can simultaneously handle heterogeneous data types (e.g., genomic, transcriptomic, metabolomic, proteomic, imaging, and metadata), particularly with sparse or incomplete multimodal data (Argelaguet et al. 2020), is essential for an integrative understanding of biological data. This task includes creating models that learn unified embeddings

(Gayoso et al. 2021), provide latent space interpretation to map into actionable decisions (Avsec et al. 2021), and develop techniques that are interoperable across modalities.

Integrate Multiomic and Environmental Data Using Physically or Biologically Informed Machine Learning Models. Advanced research is required to develop comprehensive digital twin platforms that combine multiomics datasets with environmental variables through physically or biologically informed machine learning (ML) or mechanistic models (Karniadakis et al. 2021), enabling predictive simulation of complex plant–microbe and microbe–microbe interactions across spatiotemporal scales (Corral-Acero et al. 2020; see Fig. 2.1, p. 14). For example, a digital twin of switchgrass roots under nutrient limitation could integrate real-time soil data, transcriptomic feedback, and data-informed microbiome interactions to optimize carbon allocation strategies and inform root trait engineering (Sasse et al. 2018).

Creating advanced self-supervised learning methodologies that effectively leverage vast repositories of unlabeled genomic data is necessary to establish fundamental representations (Ji et al. 2021) that substantially improve downstream prediction tasks while reducing dependence on limited labeled datasets. Curated datasets, benchmarks, and data pairings are needed to support contrastive learning and fuel these approaches (Frazer et al. 2021; Peng et al. 2025).

Significant innovation is required to implement meta-learning frameworks capable of rapidly developing and adapting foundation models to novel organisms with minimal labeled data through strategic knowledge transfer across phylogenetic boundaries (Theodoris et al. 2023). Furthermore, agent-based modeling approaches that establish quantitative bridges between molecular-level mechanisms and emergent community-level dynamics in microbial ecosystems will be essential for connecting genomic information to observable environmental phenomena through principled computational abstractions.

2.2 Multiscale Biosystems Simulation

> **PRD 2:** Develop mechanistically grounded, mathematically rigorous predictive models that represent biological processes across a range of scales, from controlled laboratory settings to complex natural environments and from molecular- to field-level dynamics.

Rationale (Challenges and Opportunities)

A core challenge in building mathematically consistent models that capture complex biological processes and systems is the inherently high dimensionality of those systems, including sparse, incomplete, and noisy experimental observations. Connecting genome-based molecular models with ecosystem-scale simulations remains computationally challenging due to the vast spatial and temporal scales involved, from nanometers to meters and seconds to years. AI-driven approaches such as multiresolution modeling, hierarchical neural networks, and causal inference algorithms could link these scales effectively (see Fig. 2.2, p. 16). Achieving this integration requires overcoming computational bottlenecks associated with high-dimensional simulations and memory-intensive calculations, demanding significant advancements in algorithms designed for high-performance computing (HPC) systems. Algorithmic innovations (e.g., adaptive mesh refinement and causal learning) coupled to scale-spanning experimental approaches could be integrated into novel predictive models of biological processes to enable more scalable computational simulations optimized for exascale platforms.

Key Questions

- What new mathematical and computational approaches are required to bridge genome-based molecular models with ecosystem-scale experimental studies, ensuring consistency across biological scales while using sparse multimodal data?
- How can AI-driven multiscale modeling be integrated with laboratory and field ecosystems through digital twins to enhance the accuracy,

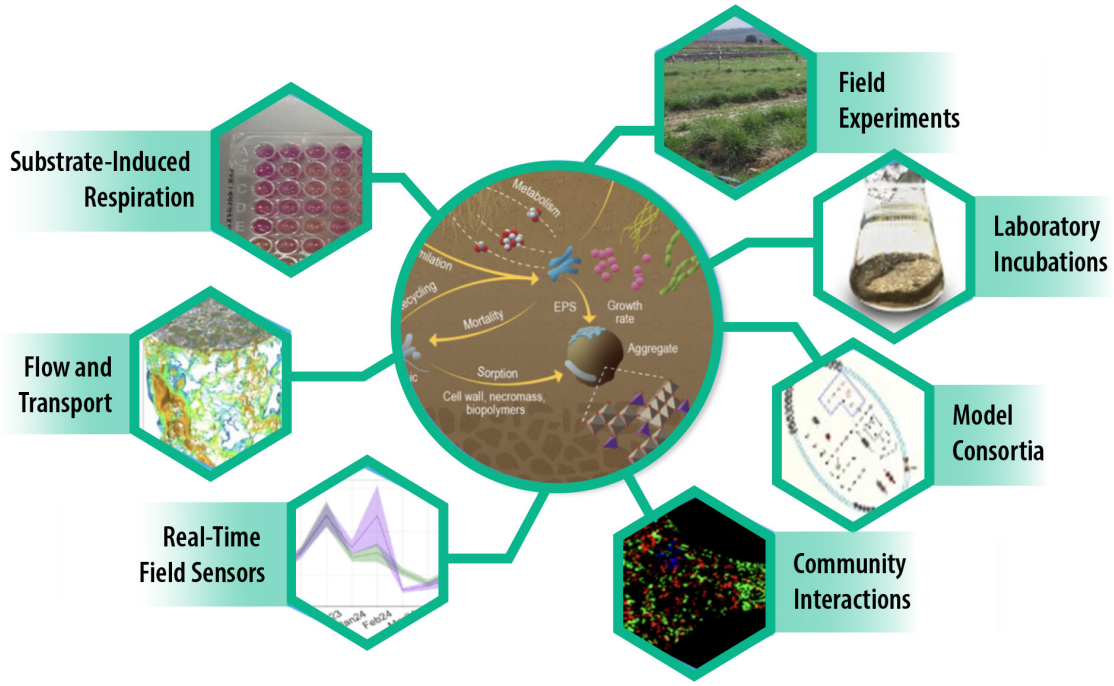


Fig. 2.2. Diverse Approaches to Predictive Phenomics. Biological data used to understand systems biology traverse scales of time, space, and complexity, ranging from milliseconds to millennia, from microns to miles, and from the molecular, cellular, and organismal levels to the ecosystem scale. Integrating this wealth of multimodal information requires data fusion and rigorous model interpretation. Multimodal representation learning methods and AI-driven biological data integration analysis through model pretraining, deep learning methods, and knowledge integration have the capacity to aggregate complex multimodal data across scales to discover new relationships currently hidden within existing complex biological datasets.

interpretability, and generalizability of biological simulations?

Impact

Improved multiscale biological simulations will transform the ability to understand and control biological processes across scales, from genome-level molecular interactions to large-scale systems, enabling precise simulations and targeted interventions in bioengineering and ecosystems. By integrating AI-driven multiscale modeling and digital twins, researchers can enhance predictive accuracy to optimize both laboratory and field experiments.

Target Activities

Integrate Cross-Scale Modeling of Biological Systems. AI methods such as multiplex network learning algorithms, multiresolution transformers, and reinforcement learning agents can integrate data across scales (Silver et al. 2021), from molecular interactions to phenotypic traits to field-level measurements (Alber et al. 2019). These models enable the discovery of how specific regulatory or metabolic pathways shape whole-organism performance or influence broader processes like soil nutrient flux, nitrogen fixation, and microbial competition. Mechanisms identified in model organisms, for example, can be traced across species and contextualized within community-scale simulations to assess how pathway rewiring affects system stability or output. Concrete applications include (1) predicting drought-induced carbon allocation in switchgrass roots, (2) modeling the influence of microbiome composition on bioenergy, biomaterials, or phytomining productivity, and (3) predicting how community-level metabolic networks in microbial consortia affect element cycling in soils. These

integrated models serve as the foundation for biologically grounded digital twins (Corral-Acero et al. 2020), guiding hypothesis generation, trait engineering, and targeted experimentation across DOE-relevant systems.

Innovate and Implement AI-Driven Experimental Design. New approaches that effectively quantify target molecules, pathways, and processes across scales of biological organization will enable new insights about the multiscale nature of systems biology. Predicting phenotypes and metaphenomes requires a focus on biomolecules, environmental factors, and emergent outcomes. The dynamic nature of systems biology calls for adaptive and iterative experimental platforms (Shahriari et al. 2016) that quantify systems at multiple temporal and spatial scales. Generative AI can accelerate model development (Sanchez-Lengeling and Aspuru-Guzik 2018) to fill gaps in sparse data and ensure imputed data are consistent with known biological constraints. Exascale computing can facilitate analysis of high-dimensional data and update models in response to rapidly generated experimental data. The interactive nature of AI–model–experiment research includes validating models with ground-truth data and experimental perturbations (Huang et al. 2016) to test the accuracy of predicted phenotypes (i.e., observable traits) within dynamic biological contexts.

Develop Novel Mathematical Models To Better Reflect the Distinctive Complexity of Biological Systems. Collaborative computational and biological research creates remarkable opportunities to analyze and interpret the high-dimensional data involved in discovering the molecular levers that drive ecosystem functions and responses to stresses and other perturbations (Brunton and Kutz 2022; Hoffmann, M. A., et al. 2022). Interpreting the function of unannotated genes, metabolites, and proteins under a range of abiotic conditions requires novel approaches that capture both instantaneous and long-term impacts (Riesselman et al. 2018). Computational methods that resolve combinatorial complexity and adapt to changing contexts will transform understanding of biological systems.

The reciprocal nature of organismal influence on and by the environment poses important challenges for mathematical representations and uncertainty quantification (Smith 2013). For example, plant–microbe interactions, soil conditions, and environmental factors all contribute to complex, nonlinear system behavior. However, biological systems science lacks clearly defined equations like physics-based models, increasing uncertainty (Karpatne et al. 2017). Biological systems often operate in a physical regime that cannot be perfectly described by continuous mathematical frameworks (e.g., essential molecules with less than one copy number per cell), requiring improved scalable algorithms that accommodate the stochasticity inherent in these systems (Raj and van Oudenaarden 2008). New algorithmic approaches and diverse computational platforms will be needed for efficient parallel processing and data handling and to enable AI models to adapt based on multimodal data (Avsec et al. 2021).

2.3 AI-Enabled Drivers for Experimental Systems

> **PRD 3:** Establish AI-enabled drivers for experimental systems as tools to understand and explore *de novo* design of biomolecules, metabolic pathways, and metabolic networks to extend the limits of nature's biochemical repertoire.

Rationale (Challenges and Opportunities)

Biological systems have a vast capacity to produce, manipulate, and efficiently separate numerous useful molecules using engineered proteins, pathways, and (multi)cellular processes. These biosystems range from the complex ecosystems needed to produce bioenergy feedstocks to the engineered strains used to convert them into a wide array of valuable bioproducts. However, the development of novel biodesign solutions is slowed by knowledge gaps in understanding complex biological functions and their interactions across spatiotemporal scales and environmental contexts, compounded by the massive protein, pathway, and bioprocess design space.

Interdisciplinary teams and cutting-edge facilities are needed to effectively use AI to predictably harness nature's biochemical repertoire. In *de novo* design, exascale computing will transform high-throughput

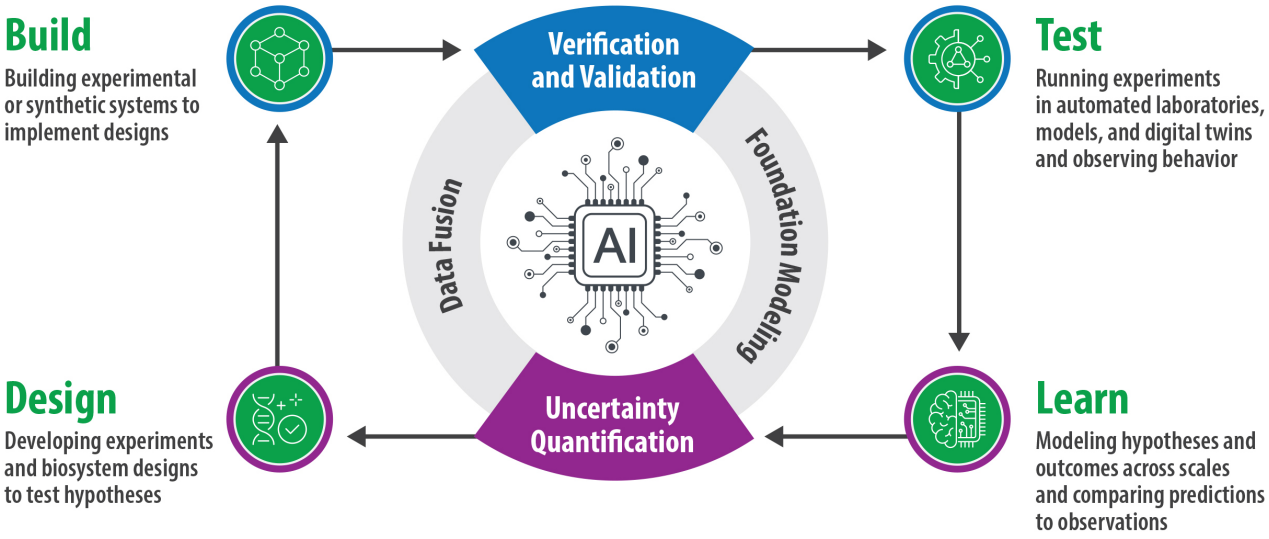


Fig. 2.3. Design-Build-Test-Learn (DBTL) Framework. AI is placed at the core of this DBTL cycle as an intelligent driver. AI models can propose hypotheses and guide experimental design for maximum information gain. Systems are built to challenge these hypotheses. Test strategies are AI-driven and adaptive, efficiently targeting measurement gaps. All resulting data is assimilated into a data backbone (e.g., BER Data Lakehouse or American Science Cloud), an uncertainty-aware knowledge-base that enables continuous model updates and systematic learning for the next iteration.

virtual screening of the massive chemical and biological space. AI algorithms can optimize experimental designs by controlling autonomous experiments, identifying critical knowledge gaps, and providing causal inferences from heterogeneous and high-dimensional biological data sources across both scale and complexity. Implementing this approach will require close partnerships among computational scientists, experts in applied math and computation, biologists, and ecologists (among others), along with access to HPC and cutting-edge biological and environmental research facilities.

Key Questions

- How can AI-driven digital twins enhance the design and optimization of biosystems, ensuring accurate uncertainty quantification and robust performance?
- How can autonomous experimentation, powered by advanced AI algorithms, improve the design and optimization of biosystems and enable more efficient and effective discovery processes?

Impact

Integrating AI-driven digital twins and autonomous experimentation into biosystems design will significantly enhance the ability to model, predict, and optimize biological processes. This synergistic combination of approaches will improve the speed and efficacy with which new biological insights are gleaned from experimental outputs while streamlining efforts to design new experiments to validate discoveries. Essentially, this work will improve the throughput and efficiency of the scientific Design-Build-Test-Learn (DBTL) cycle (see Fig. 2.3, this page). Standardized and automated workflows will greatly improve the utility of derived data.

Target Activities

Use AI To Achieve Both Experimental Tractability and Relevance. Digital twins have tremendous potential to serve as an integrative framework balancing trade-offs in tractability and relevance in biosystems design. This balance is critical because scale-up is one of the fundamental challenges in biology, whether

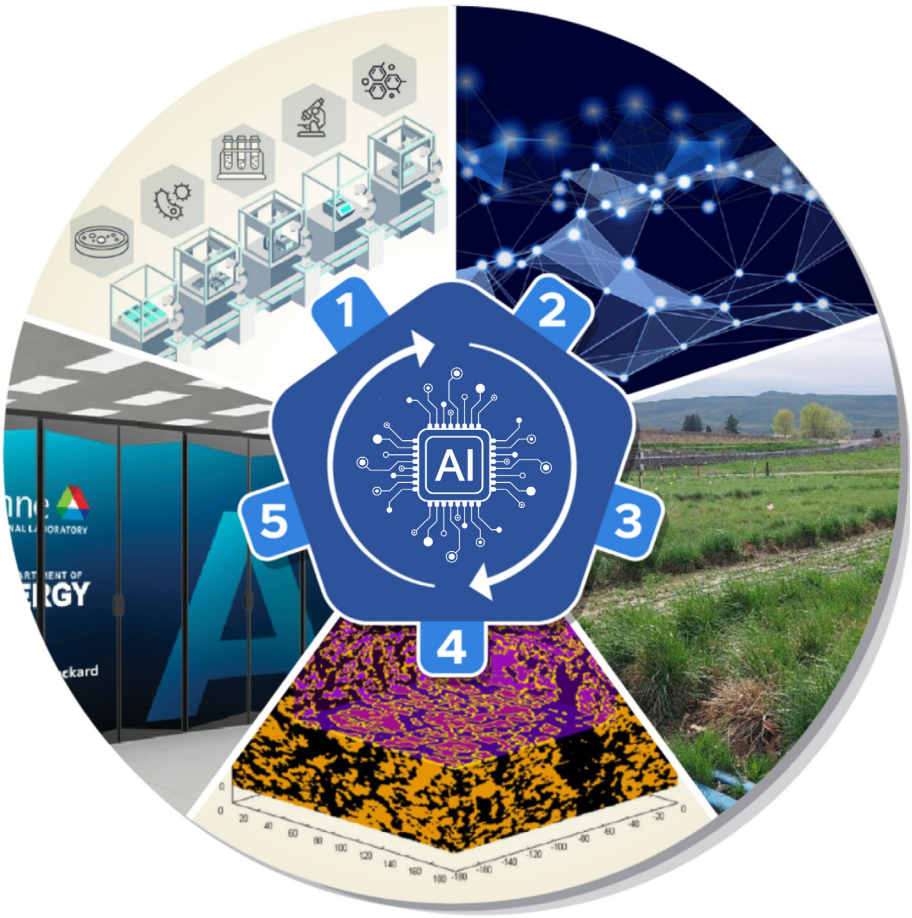


Fig. 2.4. Experiment Design with AI. AI provides unprecedented opportunities to integrate **(1)** automated workflows, **(2)** reduced-complexity models on the edge, and **(3)** real-time data from field sensors to develop **(4)** micron- to meter-scale digital twins of soil ecosystems that incorporate **(5)** novel algorithms and exascale computing.

going from a test tube to a fermenter or from a laboratory microbiome to the field.

Fermenter and field conditions are the end goal, but they are expensive, slow, complex, and often have many latent variables. Laboratory systems are typically inexpensive, fast, simple, and sufficiently controlled for mechanistic studies. Developing and integrating digital and physical twins spanning scales and complexity will create a powerful new capability to accelerate translation and generalization. For example, the development and use of fabricated ecosystems (e.g., EcoFABs, EcoPODs, and Soil Chips) will play a key role in the development, refinement, and control of digital twin experiments (Zengler et al. 2019). Given the commonality of underlying causal mechanisms such as gene regulation and metabolism (Karniadakis et al. 2021), mathematical and computational methods offer transferable utility across diverse biological systems. For instance, Bayesian neural networks excel at providing robust uncertainty quantification and interpretation, while graph neural networks can effectively identify complex relationships like correlations between metabolites and microbial interactions (Kwon et al. 2020).

Rapidly Enable Autonomous Experimentation Using AI. The incredible potential for AI design tools to be integrated with autonomous laboratories is a

unifying theme across diverse biological applications. This integration can vastly accelerate navigation of the biodesign experimental space using iterative DBTL cycles (Carbonell et al. 2018; see Fig. 2.3, p. 18, and Fig. 2.4, this page). AI-driven experimental design, monitoring, and control can enable much more efficient experimental designs, including adaptive experimentation (Burger et al. 2020).

Biologists typically perform a series of replicate treatment-control studies, where each study motivates the next. Autonomous experiments can use models (e.g., digital twins) and unreplicated data to initially explore experimental space and then rapidly converge on the areas with the largest uncertainty or the greatest potential for advancement [e.g., delving more deeply into mutations that display significant changes in phenotype (Stokes et al. 2020)]. These approaches can greatly improve experimental reproducibility, replicability, and productivity through standardization, optimization, and clear definition of all experimental parameters (NASEM 2019).

Apply Reasoning Models and Digital Twins To Hypothesize New Biological Constructs. Reasoning models offer a means of rapidly synthesizing existing biological knowledge to propose new potential capabilities of biological systems (e.g., novel enzymatic activities, novel pathways, novel microbial interactions, and new biology-based materials; Zhang et al. 2025). Reasoning models can generate hypotheses much faster than human scientists can experimentally test or even evaluate the hypotheses. Verification approaches are needed to support the effort to ground-truth hypotheses proposed by foundation models. For example, digital twins may be used to rapidly test hypotheses against current mechanistic or data-driven, evidence-based understanding of biological systems (Karpatne et al. 2017). This tactic could filter out unrealistic hypotheses and identify potential knowledge gaps exposed by a new proposed hypothesis, leading to the rapid design of more fruitful experiments.

Emphasize Explainable AI Capabilities To Advance Scientific Understanding. AI models in biology operate along a continuum that spans predictive, causal, and mechanistic reasoning (Sundararajan et al. 2017). Clarifying where a model falls on this spectrum is essential for ensuring that AI tools are aligned with DOE's mission to understand and manipulate biological processes for energy, biomaterials, and CMM.

Predictive models are designed to identify statistical patterns in data and use patterns to forecast future outcomes or classify biological states. These models, such as those used to predict protein structure, gene expression levels, or phenotypic traits from multiomic inputs, can achieve high accuracy but typically lack interpretability; they do not explain why a prediction is correct.

Causal models go beyond pattern recognition to infer directional relationships among variables, identifying which genes, pathways, or environmental factors influence others under specific conditions. In biological systems, causal inference can uncover the regulatory factors driving gene expression, the interactions shaping microbial communities, or the triggers of phenotypic shifts.

Mechanistic models, the most explanatory tier, aim to reconstruct the internal organization of biological systems. They identify modular structures—such as regulatory circuits, metabolic subnetworks, or signaling cascades—that collectively produce a functional outcome (Orth et al. 2010). Mechanistic models are closely aligned with experimental biology, as they offer interpretable, testable hypotheses about how biological function emerges from system structure.

While noncausal, nonmechanistic predictive AI models can enable effective design of biological systems, applications and advances from these models will ultimately plateau, as this kind of extrapolation can only be extended so far (Frazer et al. 2021; Peng et al. 2025). For example, protein language models (PLMs), purely predictive AI algorithms, are exceptionally good at predicting the types of mutations and variations that arise from natural evolution because that is the dominant variation found in protein sequence databases (see Fig. 2.5, p. 21). Without fine-tuning, PLMs fail to predict mutations that occur in adaptive laboratory evolution experiments because these mutations are a product of evolution in artificial, laboratory-created conditions.

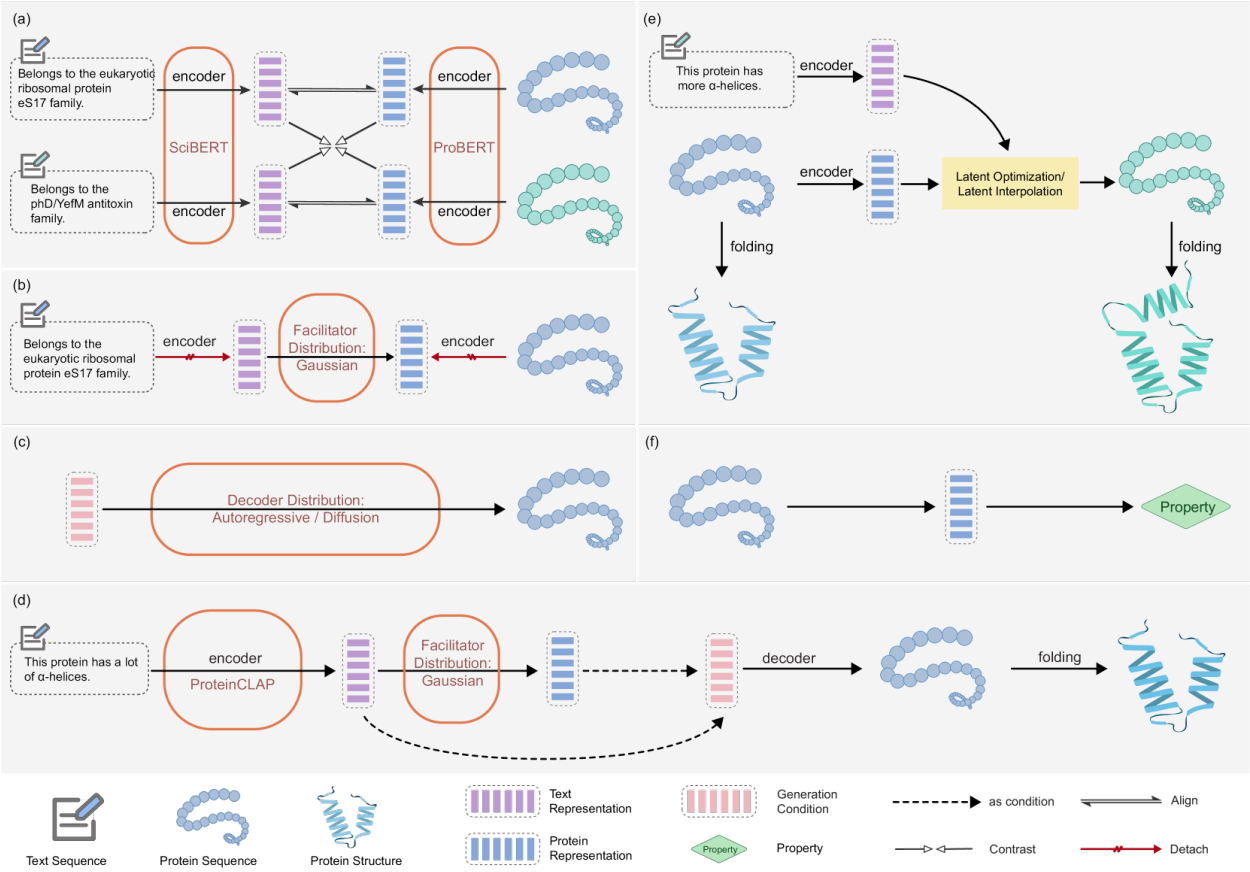


Fig. 2.5. Pipeline for the ProteinDT Pretraining Framework. Pretraining steps (A–C) and downstream tasks (D–F). **(A)** ProteinCLAP, a contrastive learning paradigm, aligns the representation spaces of the text and protein sequence modalities. **(B)** The ProteinFacilitator model augments the mapping from text sequence representation to protein sequence representation. **(C)** A protein sequence decoder generates protein sequences conditioned on the representations produced in the previous steps. **(D)** Downstream text-to-protein generation task. **(E)** Downstream text-guided protein editing task. **(F)** Downstream protein property prediction task. [Image republished from Liu, S., et al. 2025. "A Text-Guided Protein Design Framework," *Nature Machine Intelligence* **7**, 580–91. DOI:10.1038/s42256-025-01011-z.]

While PLMs are certainly excellent tools for protein design, their predictive nature limits the protein design space they can effectively explore. This limitation highlights the importance of developing explainable AI methods in close collaboration with scientists and laboratory facilities, thus taking advantage of both AI and human capabilities. Integrating digital twins and human-in-the-loop systems with experimental data sources is particularly powerful in this context because comparison and control of the two provide a robust framework to direct experiments toward key knowledge gaps.

2.4 Novel Algorithms for Genomics

> **PRD 4:** Develop algorithms to detect patterns in gene and genome organization within and across species to predict phenotypic plasticity.

Rationale (Challenges and Opportunities)

Deciphering the intricate relationships between genes, genomes, epigenomes, transcripts, proteins, metabolites, and phenotypes in diverse species is an algorithmic challenge complicated by massive and

complex datasets across communities and scales. Exascale computing can provide the necessary resources to train models with vast amounts of data and to optimize complex networks, leading to more efficient pathways and promising outcomes. Novel AI approaches integrating HPC with experiments are an opportunity to overcome the high dimensionality and sparsity of omics data.

Key Question

- How can novel AI approaches such as digital twins, foundation models, and other computational tools be integrated with omics research to discover, understand, and predict the molecular mechanisms in plants, microbes, and microbial communities that govern macroscale processes?

Impact

Advances in this space will greatly accelerate the development and integration of novel AI capabilities in omics research, enabling a deeper understanding of how genes, epigenetic marks, transcripts, proteins, and metabolites within plants, microbes, and microbial communities govern key emergent processes. Specifically, novel algorithms will transform understanding and control of the complex networks of interactions among the molecular entities that make up biological systems to operate together to produce cellular, microbiome, and even ecosystem-level behaviors.

Target Activities

Advance Biological Algorithms. Research is needed to advance network-theoretic frameworks that synergistically integrate multiplex networks and knowledge graphs with foundation models to understand multiscale biological interactions from molecular mechanisms to ecosystem dynamics. Unified network representations capable of harmonizing heterogeneous omics datasets—genomic, epigenomic, transcriptomic, proteomic, metabolomic, and microbiomic—can reveal emergent patterns otherwise obscured by data fragmentation (Wang, T., et al. 2021). The robust computational methodologies underpinning this ability need to be developed. These methods necessitate the implementation of specialized multiplex network learning architectures tailored for biological applications. Such architectures can simultaneously process and learn from both structured biological relationship data (e.g., protein–protein interaction networks, metabolic pathways, and microbial community interactions) and unstructured information sources [e.g., scientific literature, experimental narratives, and multiplex networks derived from large-scale omics datasets (Szklarczyk et al. 2025)].

Furthermore, innovative edge algorithms must be developed specifically for complex plant and microbial community networks, incorporating domain-specific biological constraints and leveraging transfer learning to overcome data sparsity while accurately predicting novel molecular interactions that govern community behavior and function (Lotfollahi et al. 2023). These computational advances will collectively transform the ability to model, predict, and ultimately manipulate biological systems across unprecedented scales of organization, from isolates (nanometers) to communities (micrometers) to full microbiomes (centimeters to meters), and across systems (kilometers). This work will lead to improved and more generalized network-based tools that are applicable to diverse data at multiple scales. Such tools will be more accessible and readily applicable to biologists, permitting analysis of many data combinations that currently lack good analytical approaches.

Develop New Models To Understand Evolution of Sequence Features. New models are needed to describe how molecular function changes with evolutionary, model-informed sequence and structure features (Brandes et al. 2013). Understanding molecular function in the context of proximal interacting functions with other evolving or coevolving molecules is critical for contextualizing this interaction within the physical and regulatory milieu of the cell (Green et al. 2021). Membrane and cell wall structure, improved estimation of transport (Almagro Armenteros et al. 2019), and the evolution of regulation and resource balancing across cellular systems need to be inferred so that knowledge from one organism in a given environment can be transferred to another in a different environment (Dalla-Torre et al. 2025).

Research needs include organized data of these types, evolutionary and physical models for function and interaction, and systems that can interpolate the transformation between better-studied molecules or organisms and less-studied ones.

Innovate Interpretable Biological Models. Significant methodological advances are needed to develop advanced attention-based mechanisms (Vaswani et al. 2017) for genomic sequence models that precisely identify and elucidate biologically significant motifs and regulatory elements, enabling transparent interpretation of deep learning predictions in molecular biology. Engineering advanced visualization frameworks that systematically map complex model decision boundaries and feature importance metrics to specific biological entities is crucial for establishing interpretable connections between computational predictions and underlying biological mechanisms.

Innovation is required in counterfactual explanation methodologies specifically tailored to biological systems. Such methodologies could simulate perturbation effects with statistical rigor (e.g., quantifying predicted expression changes following regulatory element modification; La Fleur et al. 2024), thus enabling hypothesis generation for experimental validation. These capabilities, if applied to data in BER facilities like the Environmental Molecular Sciences Laboratory (EMSL) and DOE Joint Genome Institute (JGI) or data repositories like KBase, NMDC, and the BER Data Lakehouse, could translate into powerful new hypothesis-driven experimental design frameworks, enhancing experiment productivity and improving the capacity to draw insights from complex data.

Engineer High-Throughput Hardware and Software. Developing large-scale computing frameworks optimized for the unique computational characteristics of biological multiplex network algorithms requires leveraging exascale computing architectures to process unprecedented volumes of interconnected biological data (Mammoliti et al. 2021; Acosta et al. 2022). Transformative research is necessary to develop containerized, reproducible AI workflow ecosystems—specifically designed for multiomics data processing—that seamlessly scale from personal computing environments to leadership-class supercomputing facilities while maintaining reproducibility across computational platforms. Advances are needed in memory-efficient algorithmic approaches for processing and analyzing omics-scale datasets (e.g., files encoding raw DNA reads from deep metagenome sequencing are around 300 to 600 gigabytes). Additionally, specialized hardware accelerators and optimization techniques must be designed for computationally intensive bioinformatics operations (e.g., sequence alignment, structural prediction, and phylogenetic inference), dramatically improving throughput while reducing energy consumption for large-scale biological data analysis. All these advances would be immediately applicable to improving the cyberinfrastructure of BER facilities and resources (e.g., JGI, EMSL, NMDC, KBase, and the Protein Data Bank). These advances would also be valuable for ongoing development in ASCR's Integrated Research Infrastructure and High Performance Data Facility projects, which will ultimately provide support for all BER facilities and research.

Use Agent-Based Modeling To Aid Integration of Multiomic and Environmental Data. Agent-based modeling approaches must be designed to establish quantitative bridges between molecular-level mechanisms and emergent community-level dynamics in microbial systems, thereby connecting genomic information to observable environmental phenomena through principled computational abstractions. Such methods would reduce the substantial lag that presently exists between experimental data generation and rich mechanistic analysis, as without AI infrastructure, these activities are generally performed by multiple expert parties in collaboration. With AI agents, these collaborations would still be required, but modelers would be able to provide experimental collaborators with agentic interfaces that enable them to use natural language to drive models and ask models questions about their data (Thirunavukarasu et al. 2023; Borghoff et al. 2025).

Develop Experimental Designs for Model Validation and AI-Driven Discovery. Collaborative teams need to conduct AI-informed laboratory and

field experiments that generate data for validating and calibrating predictive models. An iterative process involving the design and construction of experimental systems, followed by testing model predictions against empirical data, will refine both models and experiments to ensure reproducible results and accurate predictions (NASEM 2019). Integrating biological insights with computational approaches will also help identify complex biological mechanisms and processes that can benefit from AI-driven analysis and model development (Karpatne et al. 2017). In addition, AI can be leveraged to explore and uncover emergent biological behaviors.

Leverage Reinforcement Learning Agents for Mechanistic Discovery. Reinforcement learning (RL) agents are emerging as powerful tools for mechanistic reasoning in biological systems, as recently shown with protein–ligand interactions (Lee et al. 2025). Unlike traditional predictive models that passively learn from labeled data, RL agents actively explore large, multiplex biological networks to uncover causal and functional relationships (Yang et al. 2023). These agents simulate how perturbations (e.g., gene edits or environmental shifts) affect system behavior, traversing complex omics and interaction networks to identify regulatory circuits, metabolic pathways, or condition-specific subnetworks. Their ability to incorporate rewards based on biological plausibility, such as network coherence, literature alignment, or experimental validation, makes RL especially well-suited to guiding hypothesis generation, adaptive model refinement, and targeted experimental design (Liu, H., et al. 2022). When embedded within digital twins or AI-guided laboratory platforms, RL agents can iteratively prioritize interventions, optimize trait engineering strategies, and accelerate the discovery of transferable biological mechanisms (Khdoudi et al. 2024).

These algorithmic innovations can also be applied to develop mechanistic models of numerous biological systems important to the DOE mission (e.g., biomining in plants and microbiomes to accumulate and separate CMM; engineering more resilient bioenergy crops; and designing microbes to produce oils and other valuable byproducts or digest waste; Rylott and van der Ent 2025). By integrating root transcriptomics, metal transport pathways, and microbial-assisted metal solubilization, such models can guide the design of crops and consortia for CMM recovery, an emerging DOE priority (U.S. DOE 2023c).

Chapter 3

Data Generation for AI

3.1 Rationale (Challenges and Opportunities)

AlphaFold would not have been possible without the coordinated efforts of a large community of protein crystallographers who generated the Protein Data Bank (PDB; see sidebar, DOE Powers Discovery, p. 9). The National Institutes of Health Protein Structure Initiative also contributed significantly to this database of high-quality structures, which made training the AI system possible.

Effective development of AI tools for cells and ecosystems will similarly require vast amounts of new, standardized data that are of sufficient quality, resolution, and content. Generating such data requires community standards and close partnerships between computational scientists and experimentalists. DOE's national laboratory complex—with world-leading user facilities and deep domain expertise across application areas—is uniquely suited to provide high-quality, AI-ready data at the needed scale. Exascale computing will facilitate the integration of multiscale models, allowing AI to learn relationships and predict emergent properties that are impossible to capture with smaller computational resources. To address BER missions in biology and achieve the identified priority research directions (PRDs), these efforts should focus on the largest knowledge gaps. This focused collaboration will motivate the development of breakthrough technologies for measuring key parameters.

3.2 Impact

The coordinated effort of the ASCR and BER research communities to rapidly generate standardized data will greatly facilitate all proposed PRDs. This coordination includes developing new modalities for making key measurements to illuminate currently "dark" biological processes across scales. Critically, this effort also includes creating community standards, establishing new incentives, and enabling large multilaboratory experiments to generate data of sufficient quality and scale.

3.3 Target Activities

Large Coordinated Experiments To Generate Multiscale and Multimodal Data at Scale. Just as the physics community assembled around the quest to discover the Higgs boson (ATLAS Collaboration 2022), the ASCR and BER communities need to assemble around key organisms, ecosystems, and questions to rapidly generate necessary data across scales and modalities (Thompson et al. 2017). Having a central theme represents a new mode of experimentation where communities of scientists collaborate to fill data analysis gaps and support community models. This approach demands high-quality annotations, which must be accurate, consistent, and clearly defined. Tooling for capturing user expertise plays a major role in validating inter-annotator agreement and in establishing uncertainty scores (Dumitrache et al. 2020).

Drive Data Standardization with Recognition of Data Generators. New paradigms for individual

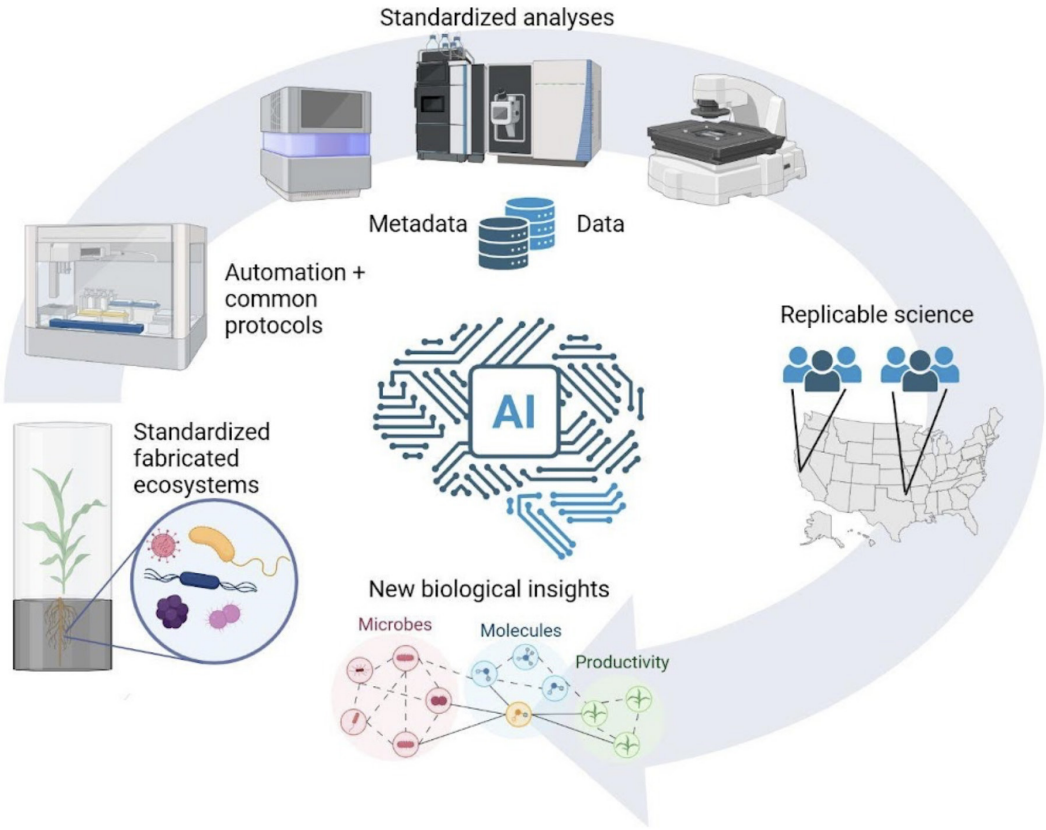


Fig. 3.1. Standardized Analyses. The diagram illustrates replicable experimental capabilities and community standards for data generation and deposition, which are required to enable integrative data analysis across time and laboratories.

contributions are needed to incentivize the development and acceptance of standardized experiments, data, and mathematical and computational tools (Data Citation Synthesis Group 2014; see Fig. 3.1, this page). This is especially true in the biological sciences, where scientists are evaluated based on the number, quality, and authorship position of their publications (Brand et al. 2015). Watermarking data and creating community leaderboards are innovative ways to track contributions that improve understanding and model accuracy (Choudhary et al. 2024; Gergov and Tsochev 2025; Rafi et al. 2025).

Community Standards for Data Generation and Deposition. Unknown data reproducibility and replicability are major challenges to using most existing biological and environmental studies for AI model training (Ball 2023). Few studies are replicated within the same laboratory; fewer still across multiple laboratories. This is a critical gap that must be addressed through community organization of replicate studies to identify and understand how uncontrolled or latent variables (e.g., protocol details, instrument type, and season) impact outcomes. These variables must be specified and measured to obtain reproducible results.

Replicate studies across multiple laboratories and locations will provide key insights into how sample size influences results, ultimately enabling AI-assisted study designs to generate high-confidence findings and useful data while facilitating improved use of automation (see Fig. 2.2, p. 16, and Fig. 3.2, p. 27). Addressing these challenges will require significant efforts in developing and disseminating standardized experimental protocols and resources, such as those being developed for fabricated ecosystems (Zengler et al. 2019; Novak et al. 2025).

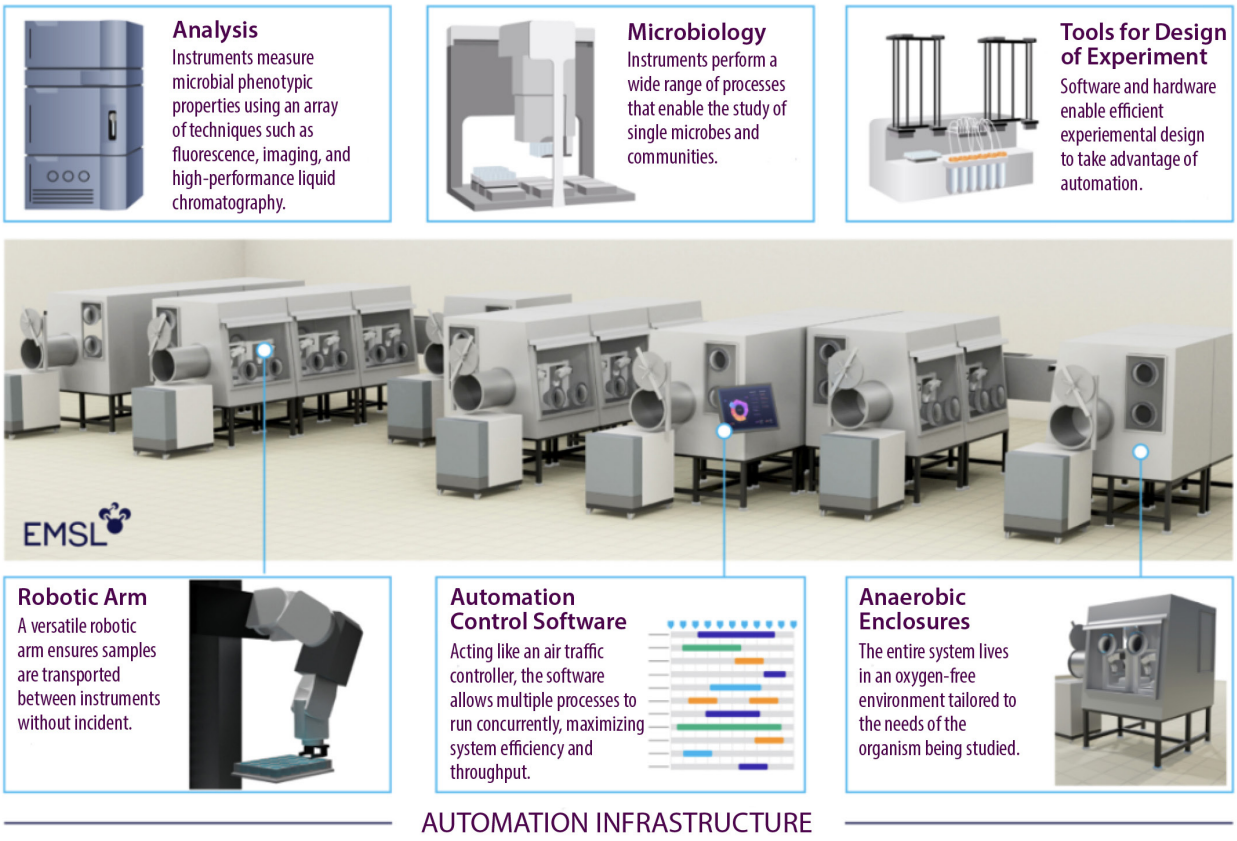


Fig. 3.2. Laboratory Automation. The Anaerobic Microbial Phenotyping Platform (AMP2) is one of many synergistic experimental user facilities supported by BER. AMP2 conducts anaerobic microbial phenotyping experiments by integrating complex devices and tools, such as robotic arms, sample transport, laboratory automation, software, and analytical instruments, all inside connected anaerobic chambers. Such studies will provide novel information about biological functions that enhance understanding, predictions, and control of complex bioeconomy-relevant processes. [Courtesy Environmental Molecular Sciences Laboratory]

Community efforts are also necessary for interoperable and standardized format adoption. Data should be stored in formats that are open, widely supported, and easy to parse. Example formats include:

- CSV, HDF5, and NETCDF for tabular and time series data
- PNG, TIFF, and OME-TIFF for images
- PDB, MMTF, mmCIF, and FASTQ for genomics
- JSON and YAML for metadata

Although not the focus of this workshop, considerations about domain ontologies for consistent terminology are important for team science and productivity.

AI Models To Focus Efforts on Biological Unknowns. Sequencing and multiomics have revealed the existence of a vast diversity of organisms, genes, and metabolites with unknown functions. For example, only 5% of microbes have been characterized. Generally, fewer than 50% of the genes in a given microbe have known functions (often it is far less), and less than 10% of the metabolites in a given sample can be identified (Hoffmann, M. A., et al. 2022; Vanni et al. 2022). Closing the gap on unknown functions is a grand challenge in biology.

AI tools can improve microbial isolation for subsequent characterization using high-throughput genetic and phenotyping activities (Liu, S., et al. 2022).

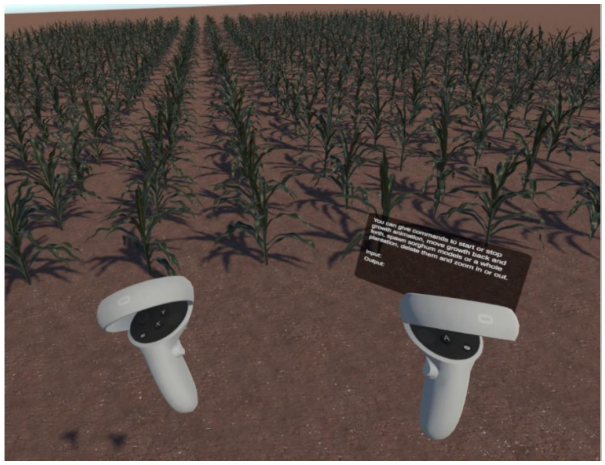


Fig. 3.3. Example of Virtual Reality Interaction. A photorealistic 3D rendering of a homogeneous sorghum field. [Courtesy NSF Grant 2417510 Collaborative Research IIBRP VR-Bio-Talk VR Voice-Controlled Visual Analytics Platform for Plant Digital Twins; B. Benes, V. Popescu, A. Magana, D. Pauli, and N. Merchant]

Complementing these efforts with *in situ* community editing and perturbation experiments is critical for testing gene and metabolite function in uncultivated microbes (Nethery et al. 2022).

State-of-the-art AI-driven methods (built on reasoning foundation models that utilize knowledge graphs) can systematically integrate and analyze diverse biological data, allowing for the prediction and exploration of unknown metabolites, genes, and microbes. This approach reveals new functional insights and even enables the conceptualization of entirely novel biological pathways or entities. Incorporating human-in-the-loop approaches and advanced visualization (see Fig. 3.3, this page) can further enhance the discovery process, allowing domain experts to iteratively refine AI outputs, prioritize plausible hypotheses, and guide targeted experimental validation. This human–AI synergy accelerates biological understanding and innovation, effectively transforming sparse insights into impactful discoveries (Prince et al. 2024).

Technologies To Measure Key Variables and Responses. The small scale of microbial interactions relative to the ability to measure and monitor systems (ideally noninvasively) leads to many knowledge gaps that impair efforts to apply AI to biological systems. Addressing this challenge requires the development and standardization of breakthrough technologies like quantum imaging to generate key data on micron-scale processes.

Benchmarking and Expert Review of AI-Driven Biological Models. As AI becomes increasingly central to biological discovery, rigorous benchmarking and expert-guided evaluation are essential to ensure that models are not only technically sound but also biologically meaningful (Marbach et al. 2012). Traditional machine learning metrics, such as Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall curve (AUPR), top-k precision, and calibration error, remain important for evaluating predictive performance. However, biological applications demand additional validation layers. Mechanistic models should be benchmarked against known pathways, gene–gene interactions, or regulatory circuits. Causal inference models should be compared to experimental perturbation data or inferred intervention effects (Bansal et al. 2022; Szklarczyk et al. 2023). Conservation across species, alignment with curated ontologies (Gene Ontology Consortium 2021), and the recovery of literature-supported relationships offer additional structure-aware metrics. Agentic reasoning should be assessed through consistency across agent instances, factual grounding, and clarity of mechanistic explanations (Jacovi and Goldberg 2020).

Critically, automated evaluation must be paired with human-in-the-loop expert review (Mosquiera-Rey et al. 2022). Domain experts play a central role in assessing the plausibility, novelty, and contextual relevance of AI-generated hypotheses. To facilitate this, AI systems should output interpretable intermediate representations—ranked mechanistic gene sets, subnetworks, or pathway narratives—accompanied by model confidence scores, provenance metadata, and natural language rationales. Integrating expert feedback into model refinement closes the loop between computational inference and experimental utility, aligning AI outputs with the goals of hypothesis generation, trait engineering, and biological insight.

Chapter 4

Crosscutting Approaches

ASCR supports work in AI, applied mathematics, computer science, and exascale computing. This chapter explores how those efforts may be applied in a crosscutting manner to enhance the biological research performed within BER's mission space (see Table 4.1, p. 30).

4.1 Novel Algorithms

Rationale (Challenges and Opportunities)

Biological processes interact with their environment to produce complex systems-level outcomes. Understanding and capturing these systems is a core DOE mission. Many ecosystem-scale processes are the result of interactions among multiple microbial community members, resulting in a community-level phenotype that is more than the sum of its parts (e.g., individual genes or genomes). For example, degrading complex organic feedstocks requires the cooperative hydrolytic capabilities of many individual microbial community members (Arnosti et al. 2021).

Most of these community members have only been measured in metagenomes, so they lack fully sequenced genomes or cultured representatives and have poor functional annotation. While AI applications in biology have primarily focused on deep representations of genes, proteins, or genomes (Knutson et al. 2022), their applicability to systems- or community-level processes is largely unexplored. Areas of sparsity in new data representations, dimension reduction, and uncertainty quantification need further exploration (see Fig. 4.1, p. 31). Efforts to apply AI to community-level processes raise the question of whether existing AI algorithms could be adapted to these tasks or whether only novel approaches, particularly those relying on exascale systems, are amenable to handling exponentially complex biological phenomena.

The limitations of current biological foundation models are another open question. These models are trained on data resources that are likely biased toward well-studied functions and organisms rather than representative systems. Creating generalizable models that capture true biological diversity will require novel mathematical and AI approaches to tackle data acquisition, usage, analysis, and model evaluation, as well as the identification of data gaps whose resolution will best aid model generalization and deployment in open environmental settings.

Key Questions

- What new mathematics, computer science, and computational sciences are needed to advance the analysis of complex genomic, microbial, and environmental data?
- How can the simulation of biological processes be advanced from the cellular scale to the reactor, biomaterial, crop field, or even ecosystem scale?
- Which algorithms must be developed to appropriately quantify and understand uncertainty?
- How can the amount of data necessary for training AI be defined?
- How should data be assessed?

Table 4.1. BER Challenges and Corresponding ASCR Research and Development

Crosscutting Focus Area	Example Biological Question	Computer Science/Math Focus
Novel Algorithms	How do individual microbe–metabolite interactions drive overall response and adaptation in complex microbiome systems?	• Develop energy-constrained sparse-learning algorithms inspired by bacterial networks (e.g., compressed sensing on graphs). • Create new reasoning models that reflect current biological reasoning approaches. • Develop mathematical approaches for modeling large language model efficiency.
Multiscale and Multimodal Modeling	Can scientists design an energy crop that will be resilient to dynamic unfavorable growth conditions encountered in nature?	• Couple agent-based root models with continuum computational fluid dynamics via surrogate mapping functions driven by machine learning.
Data Fusion	How can single-cell transcriptomes be integrated with bulk proteomics?	• Design exascale-scalable manifold-alignment methods with uncertainty estimates.
Foundation Models	How can regulatory motifs in *Arabidopsis* provide insights to support engineering of gene expression in bioenergy crops?	• Train multimodal large language models on federated genomic and phenotypic corpora, with domain-adaptation fine-tuning and guardrails. • Develop new models that explore graph learning approaches for inferring causal connections in data.
Digital Twins	How can researchers simulate engineered microbiome response to perturbations in growth conditions and individual genomes?	• Build hybrid physics–machine learning digital twins using partial differential equations/ordinary differential equation cosolvers accelerated by graph-based neural surrogates.
Verification and Validation	How can researchers ensure that predictions of optimal enzyme and pathway designs hold up during wet lab implementation?	• Implement out-of-distribution detection and conformal prediction uncertainty quantification layers across all machine learning pipelines. • Provide safety guardrails and monitoring processes for foundation models, agentic systems, and more.
Experiment Design and Automated Laboratories	How can researchers design and experimentally iterate toward optimal synthetic microbial consortia?	• Deploy Bayesian optimization and active-learning controllers in closed-loop robotic platforms.

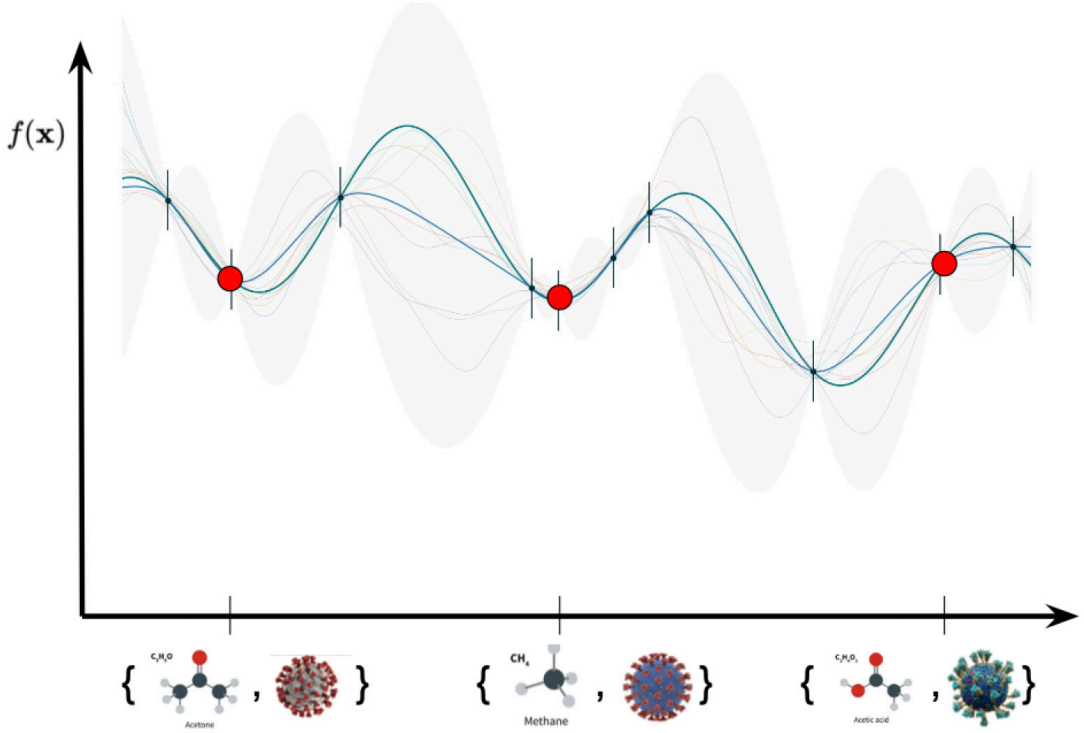


Fig. 4.1. Gaussian Process (GP) Regression Applied to Heterogeneous Biological Inputs. The plot illustrates GP-based function approximation, $f(x)$, designed to predict a biological or chemical property across a wide range of substances. The model is trained on known data (red dots) and generates a best-fit prediction (solid line) along with a measure of confidence (error bars). A key science challenge is how to compare and model entities that are fundamentally different, such as simple chemicals (e.g., acetone and methane) and complex viruses. To achieve accurate prediction in such a diverse dataset, upcoming science efforts should focus on developing new ways to describe these entities and measure their similarity to the model. [Courtesy University of California–Berkeley Center for Advanced Mathematics for Energy Research Applications, James Sethian, and Marcus Noack]

Impact

Developing new algorithms driven by advances in mathematics and computer science will play a prominent role in enabling the next generation of biological discovery. These innovations will facilitate the modeling and understanding of complex biological systems, improve the generalizability and interpretability of AI models in biology, and accelerate progress in energy and biotechnology challenges. Furthermore, these advances will drive the success of other crosscutting approaches.

Target Activities

Novel Algorithms for Community-Level and Systems Biology. AI approaches that can operate on sparse, high-dimensional, and incomplete data are needed to develop novel algorithms capable of modeling interactions among microbial communities and systems-level biological processes. These algorithms must move beyond the current focus on genes and genomes to address the complexity of biological systems at multiple organizational levels.

Bioinspired Computing Approaches. Developing biologically inspired computing systems that leverage principles such as parallel processing, distributed architectures, and adaptive learning for energy-efficient AI is a promising direction. By emulating the computational efficiency and adaptability of biological systems, new paradigms in AI—including sparse learning,

attention mechanisms, and energy-efficient training algorithms—can be realized.

Mathematical Innovations for Biological Data. Advances in mathematical frameworks are critical for analyzing and interpreting large-scale biological data. Innovations such as information geometry for biological network analysis, statistical models of collective biological computing, and optimization theory for energy-constrained learning systems will enable deeper insights into biological complexity. Additional needs include novel optimization algorithms for metabolic networks, advanced AI architectures for protein structure prediction, improved graph-theoretical approaches for modeling cellular networks, and new mathematical frameworks for capturing multiscale biological processes.

Advancements in partial differential equations (PDEs) for complex mechanisms, computational topology, manifold learning, and information theory could enable better modeling of biological systems. Furthermore, developments in high-performance computing (HPC) can accelerate biological simulations and analyses. These mathematical and computational advances will be increasingly important for engineering biological solutions to energy challenges.

Causal Inference and Handling Incomplete Data. To advance causal inference in biological systems, new techniques need to be developed, including topological, stochastic, and information theory–based methods for analyzing complex biological networks. Scalable computational approaches are essential, including specialized graph neural networks (Knutson et al. 2022) and Bayesian methods robust to biological noise and missing data (Noack and Ushizima 2023). Furthermore, innovations such as Bayesian Gaussian Process latent variable models (Ziaei et al. 2024) for dimensionality reduction and manifold learning—as well as optimization techniques utilizing parallel tempering on exascale machines—can address challenges related to incomplete data and scale.

High-Performance Computing for Biological Simulation. Leveraging advances in HPC will accelerate biological simulations and analyses, enabling the study of complex, multiscale biological phenomena. These computational advances are essential for transforming the massive amounts of data generated by modern experimental techniques into actionable scientific insights.

4.2 Multiscale and Multimodal Modeling

Rationale (Challenges and Opportunities)

Multiscale and multimodal modeling involves integrating multiple models operating at different scales and modalities, utilizing diverse data inputs to achieve comprehensive insights into biological systems. The current lack of sufficient high-quality data spanning all relevant scales and modalities is a fundamental challenge for this type of modeling. Highly instrumented laboratory and field ecosystems have the potential to provide the necessary data. However, there are several challenges in using multiple data sources, such as extracting and combining features from both structured information and unstructured data (e.g., images and videos). Recently, the sheer volume and heterogeneity of data have demanded new computational approaches to enhance the speed and scale of simulations (Cao and Gao 2022).

AI offers tremendous opportunities to meet these challenges and predict how molecules, cells, and organisms interact with each other and the environment over time. For example, AI can enable analysis of large omics data to better understand microbial phenotypes and community metaphenomes (Gao et al. 2022). When decoding biological systems, multimodal approaches can combine instrument data and simulations into inference networks that reveal gene regulatory connections responsible for physiological outcomes (Yang et al. 2021). Computational modeling incorporating statistical and mechanistic methods can identify key control points for microbiome engineering (Leggieri et al. 2021). Data from genomic, metabolomic, proteomic, and phenotypic sources can be fused to create comprehensive models that forecast biological responses and interactions (Singh et al. 2016; Mansoor et al. 2024).

Building faithful representations of relevant variables that communicate across scales is a critical problem in multiscale modeling. Addressing this issue is especially challenging in biology, where transition models between scales must be carefully formulated and robustly deployed (see Fig. 2.2, p. 16). New AI techniques have transformed the conventional Edisonian Design-Build-Test cycle into a multidimensional Design-Build-Test-Learn-Predict workflow, enabling the combination of multiscale and multimodal models that have significantly improved operational efficiency.

Key Questions

- What new mathematical and computational approaches are needed to bridge genome-based molecular-scale models with ecosystem-scale models to better understand biological processes across scales?
- How can multiscale modeling leverage sparse, hard-to-acquire data to generate meaningful and verifiable predictions about biological processes, especially in complex microbial and plant communities?
- When using data across modalities, can AI enable insights beyond correlation?

Impact

Advancements in multiscale and multimodal modeling, powered by AI, will enable deeper understanding of complex biological systems and their interactions across scales. These innovations will accelerate the discovery and engineering of target organisms, molecules, and metabolic pathways for desired outcomes such as improved crop yields, enhanced production of valuable compounds, and increased resistance to disease and environmental stress. Integrating AI-driven workflows will reduce costs, optimize resource use (Naveed et al. 2024), and sharpen the focus of field experiments (Singh et al. 2016; Gong et al. 2024; Mansoor et al. 2024; Zhang et al. 2024), ultimately facilitating enhanced bioproduct development and more efficient biotechnological processes.

Target Activities

Integration of Multiscale and Multimodal Data Using AI. AI offers unique opportunities for predicting how molecules, cells, and organisms interact with each other and the environment over time, enabling the analysis of large omics datasets to better understand microbial and plant phenotypes and community metaphenomes. Multimodal approaches can combine instrument data and simulations into inference networks that reveal gene regulatory connections responsible for physiological outcomes (Eissing et al. 2011; Deisboeck et al. 2014; Cao and Gao 2022; Loumeaud et al. 2024). Fusing data from genomic, metabolomic, proteomic, and phenotypic sources enables the creation of comprehensive models that forecast biological responses and interactions (Yang et al. 2021).

Computational Modeling and Microbiome Engineering. Computational modeling incorporating statistical and mechanistic methods can identify key control points for microbiome engineering (Gao et al. 2022). These approaches aid in the discovery of target organisms, molecules, and metabolic pathways that produce desired compounds or environmental feedbacks, supporting outcomes such as better-yielding crops, higher production of desired molecules, and greater resistance to disease and environmental stress.

AI-Enabled Simulation and Model Communication. AI-based simulation approaches such as surrogate models can replace comparatively expensive computational methods, while AI agents can automate the analysis of experimental data and refine and curate models. In addition, AI matching models generated from collected and measured data can communicate between highly accurate solvers at different scales, and AI co-scientists can assist in interpreting simulation results.

4.3 Data Fusion

Rationale (Challenges and Opportunities)

Advanced scientific computing, applied mathematics, and fundamental computer science underpin AI approaches that excel at integrating complex biological data, particularly the challenging multiomics and multimodal datasets that traditional mechanistic

approaches struggle to parse (Er et al. 2024). However, integrating this data relies upon robust data organization and standardized information associated with consistent sample IDs (U.S. DOE 2022a). Preparing "AI-ready" data requires careful attention to data sources, metadata, and domain knowledge prior to the development of encoding algorithms.

Once properly prepared, AI techniques powered by linear algebra, optimization theory, and graph algorithms can improve and accelerate integration by encoding diverse data types into a common vector format, enabling seamless association and analysis. Close collaboration between applied mathematicians, computer scientists and domain scientists is essential to ensure the resulting encodings have meaningful biological interpretations. This assessment process is likely to require the analysis of latent embedding spaces relative to different biological interpretations to guarantee meaningful separation of differently classed entities. Embedding space representations, obtained through high-performance parallel algorithms, could aid in understanding the impact of data uncertainty on biological conclusions, as these methods can identify variations that have the most impact on required biological interpretations.

Key Questions

- How will computational methods grounded in advanced applied mathematics and computer science enable the discovery of new behaviors, mechanisms, and designs of biological processes by, for example, extracting more information from available experimental data and coupling mechanistic insights to high-resolution imaging?
- What are the challenges and potential solutions for ensuring data interoperability and standardization when integrating high-resolution imaging and computational advances?
- How can AI algorithms capable of integrating multimodal data (e.g., high-resolution imaging, omics, and environmental metadata, including text data) be designed to derive insights into microbial phenotypes and their role in ecosystem resilience?

Impact

Advances in AI-driven data fusion will enable the integration and interpretation of complex, multiscale, and multimodal biological data, facilitating new discoveries in systems biology, environmental science, and biotechnology. Improved data fusion will accelerate the identification of causal relationships, enhance the predictive power of biological models, and support the design and engineering of biological systems for accurate outcomes. Developing robust, scalable, and interpretable AI methods will provide new tools to address critical challenges in understanding and manipulating complex biological systems, ultimately leading to more efficient and impactful scientific research.

Target Activities

Develop AI-Driven Data Integration and Fusion Methods. AI-driven data integration techniques, backed by scalable algorithms and HPC resources, can support the detection of causal relationships from data. Data integration enables interventional and counterfactual analysis, which can subsequently be integrated into mechanistic models (Pearl 2009). AI-driven data integration methods can leverage existing knowledge to predict unknown model parameters (e.g., growth rates and kinetic parameters) and aid in understanding the impact of data uncertainty on simulation parameterization, output, and interpretation (Schillings and Stuart 2017).

Key scientific goals of AI-driven data integration include understanding pore-scale soil–water interactions (Wang, Y. D., et al. 2021) and the impacts of plant and microbial phenotypes on emergent processes, such as biogeochemical fluxes, aggregate formation and turnover, and resistance or resilience to perturbation in flood and drought studies (Oikawa et al. 2024). Additionally, data fusion approaches aim to reveal how abiotic conditions influence biodiversity, biogeography, and future responses to environmental change. Another important role of data fusion is to integrate diverse data products to facilitate iterative design and engineering of plant, microbe, and microbial community systems (Arkin et al. 2018).

Advance Biostructure Recognition and Multimodal Data Analysis. To advance biostructure recognition from multiscale and multimodal data, novel, scalable computational methods that can leverage HPC resources need to be tailored to data from scientific experiments. These computational methods should couple AI with new mathematical frameworks and scalable algorithms to (1) reconstruct cell, soil, and protein structures from new and evolving complex imaging capabilities using scalable inverse techniques (Raissi et al. 2019); (2) seamlessly fuse data across multimodal and multiscale imaging techniques (Kalamkar and Geetha 2023); (3) extract feature vectors and compact representations for recognition and classification (Robitaille et al. 2022); and (4) build reference libraries of AI-ready data containing observed and measured biological information.

Develop and Apply Foundation Models and Generative AI. Current and emerging efforts in foundation models, large language models (LLMs), and vision transformers (ViTs; Dosovitskiy et al. 2021), many leveraging exascale machines, could enable rapid analysis and understanding of text, images, and videos (Hoffmann, J., et al. 2022; Truhn et al. 2024). Foundation models also have the potential to generate diverse data types, which could remediate studies hindered by data scarcity (Baek et al. 2021; Tunyasuvunakool et al. 2021). These new technologies are likely to enable more accurate identification of biomolecules, cellular structures, and organismal phenotypes.

4.4 Foundation Models

Rationale (Challenges and Opportunities)

Foundation models are a class of general-purpose AI models trained on massive unlabeled datasets and characterized by their scalable multimodal capabilities, since they often process and generate various forms of data. These models can be specialized and adapted to tasks based on domain-specific agents (see Fig. 4.2, p. 36) or on fine-tuning methods grounded in optimization theory and transfer learning (Zheng et al. 2025). For example, CACTUS (McNaughton et al. 2024) demonstrates how foundation models can become practical scientific assistants when wrapped in transparent, instrumented agents that enforce guardrails, expose intermediate reasoning, and seamlessly leverage HPC-hosted analytic workflows.

Beyond applications in natural language processing, foundation models using scalable HPC infrastructure to combine text, images, and audio are increasingly being developed for various scientific domains. With vast amounts of data available across fields such as chemistry, physics, and biology, these models are beginning to revolutionize those fields with new insights, capabilities, and even discoveries. Foundation models have shown the ability to identify patterns and relationships that may be too intricate or subtle for traditional computational methods, thereby pushing the boundaries of scientific knowledge.

Incorporating foundation models in biology holds immense potential for understanding protein and molecular properties. For instance, Functional Annotation of Proteins using Multimodal models (FAPM), a contrastive model linking natural language and protein sequence language, leverages both a pretrained protein sequence model and an LLM (Xiang et al. 2024). This allows FAPM to generate natural language labels for protein functions, including Gene Ontology terms and catalytic activity predictions. Such findings were broadly tested using public benchmarks (e.g., UniProt Knowledgebase's Swiss-Prot) to demonstrate FAPM's superior ability to understand protein properties compared to models relying only on sequence or structural data. Promising results from few-shot learning models (Zhou et al. 2024) using minimal wet laboratory data indicate there are imminent opportunities to understand complex biological phenomena at unprecedented resolution across modalities (for example, in aggregating information across diverse biological readouts from sequencing, multiomics approaches, structural data, and other experimental measurements).

The intrinsic opacity of foundation models is a critical obstacle to their deployment in scientific settings. Their billions of parameters, arranged in deep, multimodal architectures, resist model interpretability. Interpreting these models requires scalable methods that can attribute input modalities (e.g., sequence positions, image pixels, and spectral channels) to

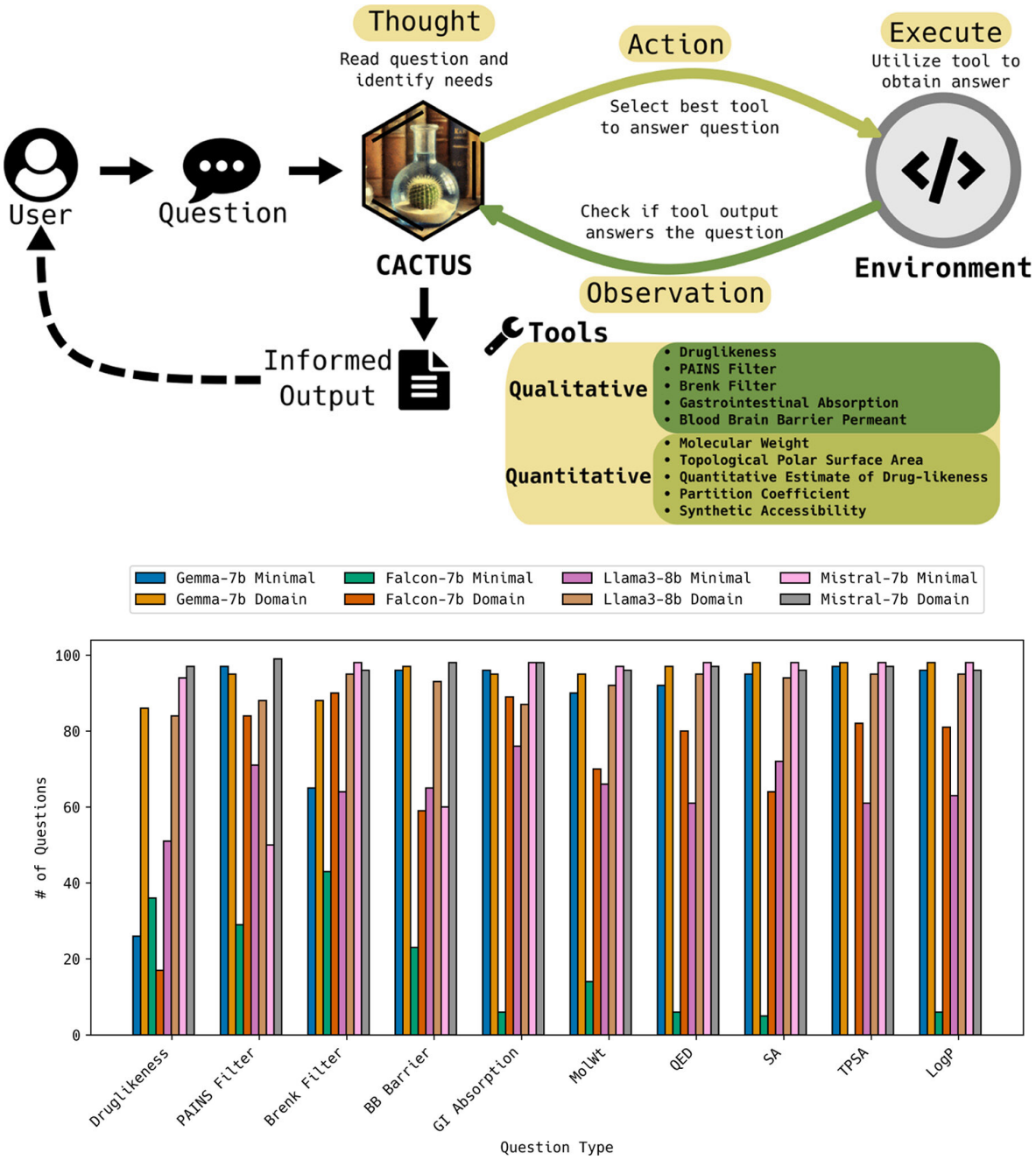


Fig. 4.2. Specializing 7B-Parameter Foundation Models with Cheminformatics Domain Benchmarking. Top panel: Agent workflow. CACTUS wraps any open-weight 7B (7 billion parameter) large language model from Hugging Face in a transparent Thought-Action-Observation loop, routing questions through a menu of qualitative and quantitative cheminformatics questions. Because reasoning, tool calls, and checks occur entirely at inference time, a general-purpose foundation model is domain-specialized without fine-tuning or reinforcement learning from human feedback, preserving model weights while adding traceable guardrails. **Bottom panel:** Benchmark results across chemistry questions spanning 10 property classes. The domain-prompt and tool orchestration boost accuracy of some models significantly over minimal prompting, confirming that (1) compact open models can reach near-expert performance when coupled with domain tools and (2) prompt-level adaptation alone delivers large gains, vital when fine-tuning is compute constrained. [Reprinted under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0) from McNaughton, A. D., et al. 2024. "CACTUS: Chemistry Agent Connecting Tool Usage to Science," *ACS Omega* **9**(46), 46563–73. DOI:10.1021/acsomega.4c08408.]

downstream predictions. Traditional interpretability techniques, such as gradient-based saliency, attention-score analysis, or SHAP value approximations (SHapley Additive exPlanations; Lundberg et al. 2020), must be rethought and highly parallelized to process massive datasets and model checkpoints simultaneously. HPC plays two roles in this space: (1) it provides the raw throughput for evaluating thousands of attribution queries in parallel and (2) supports the development of novel frameworks that reduce interpretability to tractable subproblems.

The fragmented training of foundation models means researchers have a limited understanding of how models are trained, how they are distributed, which data types the models support, and the purpose of the models themselves. Building on community efforts for foundation models is challenging because domain-specific datasets are siloed and common metadata standards are lacking. Key challenges in this context include defining the context of foundation models, identifying associated pretraining and fine-tuning tasks, developing scalable methods for model adaptation, and determining how users intend to interact with these models, particularly when using supercomputing resources.

Because foundation models span modalities, unified metrics rooted in information theory and statistical learning are necessary to quantify cross-modal feature importance and to ensure explanations generalize across data types. Without scalable, HPC-driven interpretability toolkits grounded in rigorous applied math, foundation models risk being labeled as black boxes rather than leveraged to their full scientific potential. Establishing robust guardrails is essential to ensure these models can be deployed in a reliable way. Guardrails help prevent misuse, mitigate dataset and model biases, and ensure the models' outputs are accurate and aligned with scientific objectives. Further considerations for accuracy and alignment are discussed in the Verification and Validation section (see p. 41).

Key Questions

- How can foundation models (including LLMs) be useful for fast analysis and discovery?
- How can foundation models be enhanced to move beyond quick assessments and support deeper reasoning?
- What new foundation models are necessary to capture microbial community processes and their interactions with their environment?
- Does DOE need to create new foundation models to avoid unknown corporate and foreign influences?
- Is it possible for DOE to create models with comparable performance, or should DOE focus on developing AI agents with controls for possible biases?

Impact

The advancement and responsible deployment of foundation models will transform scientific discovery by integrating and analyzing massive multimodal datasets across disciplines. These models will accelerate the identification of complex patterns and relationships, support new scientific insights, and facilitate breakthroughs in areas such as protein function prediction, molecular property analysis, and cross-modal data integration. By establishing robust standards, interpretability frameworks, and resource management strategies, the scientific community will be empowered to leverage foundation models in a reliable, transparent, and HPC-scalable manner, ultimately driving innovation and expanding the frontiers of knowledge.

Target Activities

Development and Specialization of Foundation Models for Science

- Advance the development, fine-tuning, and adaptation of foundation models for scientific domains—including biology, chemistry, and physics—by leveraging large, diverse, and multimodal datasets.

- Promote the creation of domain-specific agents and transparent, instrumented wrappers that enforce guardrails, expose intermediate reasoning, and leverage HPC-hosted analytic workflows.
- Facilitate the use of foundation models in tasks such as protein function annotation (Xiang et al. 2024), multimodal integration, and few-shot learning for biological discovery (Zhou et al. 2024).

Standardization, Metadata, and Community Collaboration

- Address fragmented foundation model training by developing and promoting common metadata standards, interoperable data formats, and open sharing of domain-specific datasets.
- Foster community efforts to define best practices for model pretraining, fine-tuning, distribution, and user interaction, particularly in the context of supercomputing resources.
- Ensure robust guardrails to establish model alignment with scientific objectives and mitigate misuse and bias.

HPC-Driven Interpretability and Model Transparency

- Develop scalable, HPC-enabled interpretability toolkits that can attribute input modalities to predictions across massive multimodal models.
- Innovate new frameworks, such as randomized linear algebra for low-rank approximation of activation subspaces and tensor decomposition to isolate task-specific latent factors, to reduce interpretability to tractable subproblems.
- Create unified, information theory–based metrics to quantify cross-modal feature importance and ensure generalizability of explanations across data types.

Resource Management and Model Access

- Develop strategies for efficient compute resource management, including graphics processing unit allocation, job scheduling, and resource-aware parallelism, for hosting and serving foundation models on research infrastructure.
- Balance the use of commercial application programming interface (API) services with open-source and publicly developed models to address concerns about cost, intellectual property exposure, and accessibility.
- Adapt research infrastructure to meet the growing demand for foundation model applications in scientific research.

4.5 Digital Twins

Rationale (Challenges and Opportunities)

Digital twins (i.e., virtual replicas of physical systems; see Fig. 4.3, p. 39) emerged in engineering but are now revolutionizing various areas of study (Fuller et al. 2020; NASEM 2024). A particularly promising application is in soil microbiome science. Digital twins' ability to process diverse data types in real time can address an urgent need for experimental and virtual models of soils, especially those surrounding plant roots (i.e., the rhizosphere). These models can bridge laboratory and field studies, rapidly improve feedstock crop performance under suboptimal growth conditions, and develop a molecular-level understanding of systems (Zhalnina et al. 2019).

At the ecosystem scale, hyperspectral imaging, automated rhizotron imaging, and real-time sensor data can inform digital ecosystem twins. Purpose-built microelectronic sensors codesigned for autonomous laboratory and field experiments can both inform and be informed by a digital ecosystem twin that captures micron-scale biogeochemical reactions responsible for ecosystem processes such as element cycling, plant–microbe exchange, and ecosystem productivity. Integrating these modalities through graph-based data fusion, manifold alignment, and *in situ* data compression can produce highly predictive simulations of plant growth, microbial metabolism, and environmental feedback.

AI modeling approaches—underpinned by scalable ML libraries, advanced data management, and

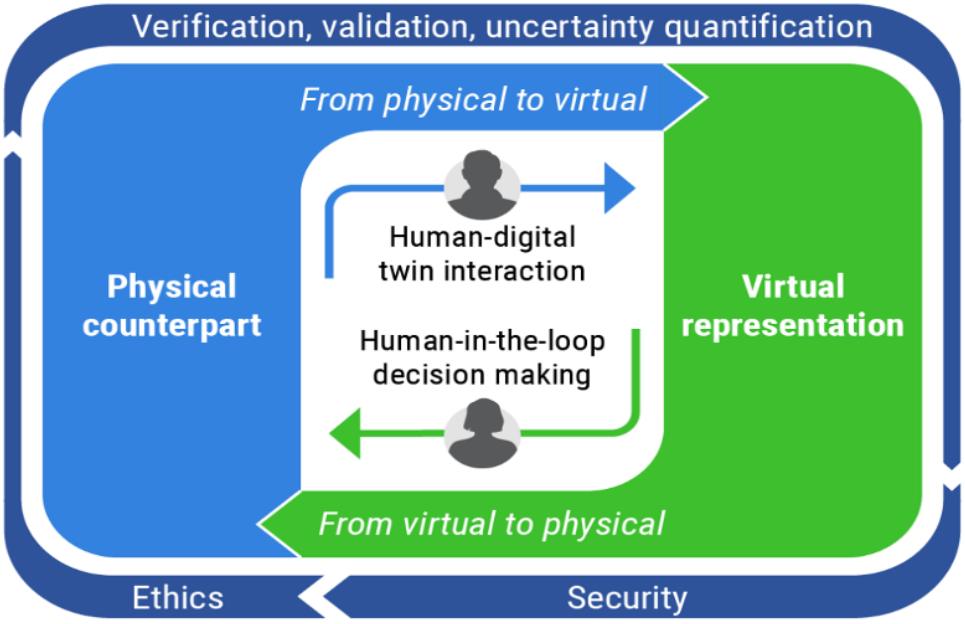


Fig. 4.3. Elements of the Digital Twin Ecosystem. Digital twins create a dynamic and intimate interaction among models, data, and decisions. The virtual representation evolves with the real-world biological (physical) counterpart, generating a feedback loop. [Image republished from NASEM. 2024. *Foundational Research Gaps and Future Directions for Digital Twins*. National Academies of Sciences, Engineering, and Medicine. National Academies Press, Washington, D.C., U.S. https://nap.nationalacademies.org/catalog/26894/foundational-research-gaps-and-future-directions-for-digital-twins]

codesign of hardware–software stacks—can identify key biological mechanisms driving productivity and stress tolerance, supporting the development of robust crop and feedstock systems. Deep learning techniques such as convolutional neural networks (Sordo et al. 2024) on HPC platforms can be used to analyze high-throughput plant and microbial phenotyping data, enabling the identification of traits associated with stress tolerance, microbial metabolism, and yield potential. Combining AI with mechanistic or data-driven models on HPC architectures allows researchers to simulate and predict plant–microbe–soil interactions under abiotic stresses (e.g., drought, flood, fire, and land management) using scalable parallel algorithms and resilient workflow engines.

Given the long duration of most plant studies, especially in the field, HPC-driven simulation pipelines can prioritize experiments to close key knowledge gaps by integrating data from both laboratory and field twins into experimental design. For example, digital twins can integrate data from laboratory and field twins to direct iterative experiments (see Fig. 4.4, p. 40) aimed at identifying the biological mechanisms driving field observations. To expedite digital twin development, reproducible and containerized experimental platforms of increasing biological, chemical, or physical complexity will be coupled with scalable computational workflows, empowering progressive translation from controlled systems to natural environments. This approach enables the construction of digital twins for highly controlled, well-defined systems, which can later be translated to more complex environments.

Key Questions

- Which new algorithms and scalable computational methods will enable digital twins to optimize biology experiments, generate and test hypotheses, and promote robust experimentation (e.g., reduced

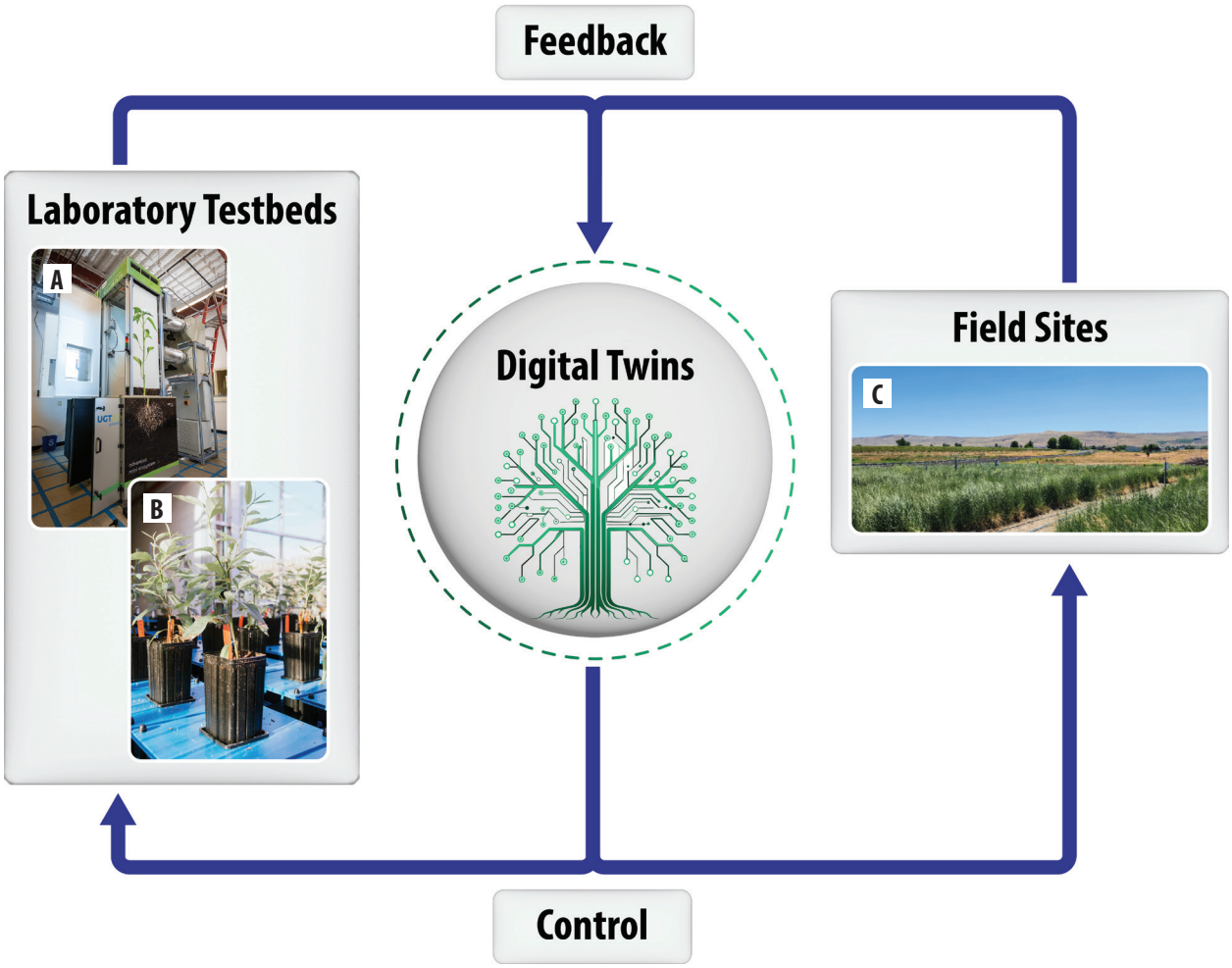


Fig. 4.4. Digital Twins Drive Experiments. Digital twins will enable effective, integrated experimentation and simulation through experimental feedback and model-driven control of both laboratory and field experiments. **(A)** EcoPOD at Lawrence Berkeley National Laboratory [Courtesy LBNL]. **(B)** Advanced Plant Phenotyping Laboratory at Oak Ridge National Laboratory [Courtesy ORNL]. **(C)** Pacific Northwest National Laboratory Phenotypic Responses experiment at a Washington State University field station [Courtesy PNNL].

risk of unintended consequences for microbial interventions)?

- How can upcoming exascale computational infrastructure and advanced software ecosystems impact the creation and execution of digital twins?

Impact

Integrating digital twins with AI and HPC will revolutionize the study and engineering of plant–microbe–soil systems, enabling real-time, predictive understanding of complex biological processes. These advances will accelerate the development of resilient crops and feedstocks, support novel biomanufacturing strategies, and facilitate the rapid translation of laboratory findings to field applications. Ultimately, digital twins will empower researchers to design, optimize, and scale biological systems for enhanced productivity and resilience.

Target Activities

Develop and Integrate Digital Twins for Microbial Communities and Ecosystems

- Advance the development of digital twins to understand and predict the behavior of microbial communities in complex environments, such as the soil microbiome, to transform the field of soil restoration.

Soil ecosystem digital twins could be used to understand the genomic and molecular basis of how soil microbiomes interact with plants to generate emergent functions such as organic matter decomposition and storage, stress resilience, plant growth promotion, and biomining of target elements. Digital twins encourage a learn-from-nature approach that supports the development of novel biomanufacturing technologies.

Apply AI and Digital Twins in Biomanufacturing and Biosystems Engineering

- Employ AI-enabled digital twins to simulate and optimize microbial metabolism for the production of biofuels, biomaterials, and other valuable compounds.

By understanding intricate microbial metabolic networks, researchers can engineer strains with improved performance and efficiency and develop chassis organisms—microbial hosts of genetic circuits—for biosensors, biomanufacturing, and environmental probiotic applications. Digital twins also provide predictive capabilities for scaling up optimized strains and microbiomes from the bench scale to production or field scales.

Leverage HPC and DOE Exascale Infrastructure

- Utilize DOE exascale infrastructure to accelerate microbiome engineering by integrating physical and AI models, creating predictive tools, and developing novel hypotheses for microbial interactions and gene, metabolite, and protein functions.
- Couple digital twin approaches with HPC-driven simulation pipelines to prioritize and design experiments, close knowledge gaps, and enable translation from controlled systems to natural environments.

4.6 Verification and Validation

Rationale (Challenges and Opportunities)

Verification and validation (V&V) in AI comprise systematic methodologies designed to rigorously assess the accuracy, reliability, and robustness of AI systems (Oberkampf and Roy 2010; U.S. DOE 2020; U.S. DOE 2023a). Verification confirms the AI workflow is implemented correctly, and validation (i.e., benchmarking) shows that its outputs represent reality within the intended domain. Methodologies include model-agnostic checks and model-specific methods that ensure AI predictions align with known biology and experimental data for continuous improvement (see Fig. 4.5, p. 42).

Together, V&V provide formal proofs or statistical evidence that the system meets explicit accuracy, reliability, and safety targets. These processes emphasize transparency in decision-making, clarity in model behaviors, and conformity to established scientific protocols (Oberkampf and Roy 2010). V&V protocols particularly focus on guardrails, explainability, uncertainty quantification, robustness against variations in data inputs, safety in diverse operational contexts, and accountability through detailed model documentation and validation procedures.

For example, Monte Carlo dropout can quantify uncertainty by sampling multiple predictions, and SHAP values provide interpretability by attributing prediction importance to input features (Lundberg et al. 2020). When it is not possible to ground-truth using first principles, theory, or experimental data (see Section 2.3: AI-Enabled Drivers for Experimental Systems, p. 17), comparing multiple independent models is critical, though this is acknowledged as a weaker form of V&V.

Despite advancements, significant challenges persist within V&V in biological contexts, particularly in validating AI models across different experimental conditions, managing rare or infrequent biological phenomena, and effectively integrating

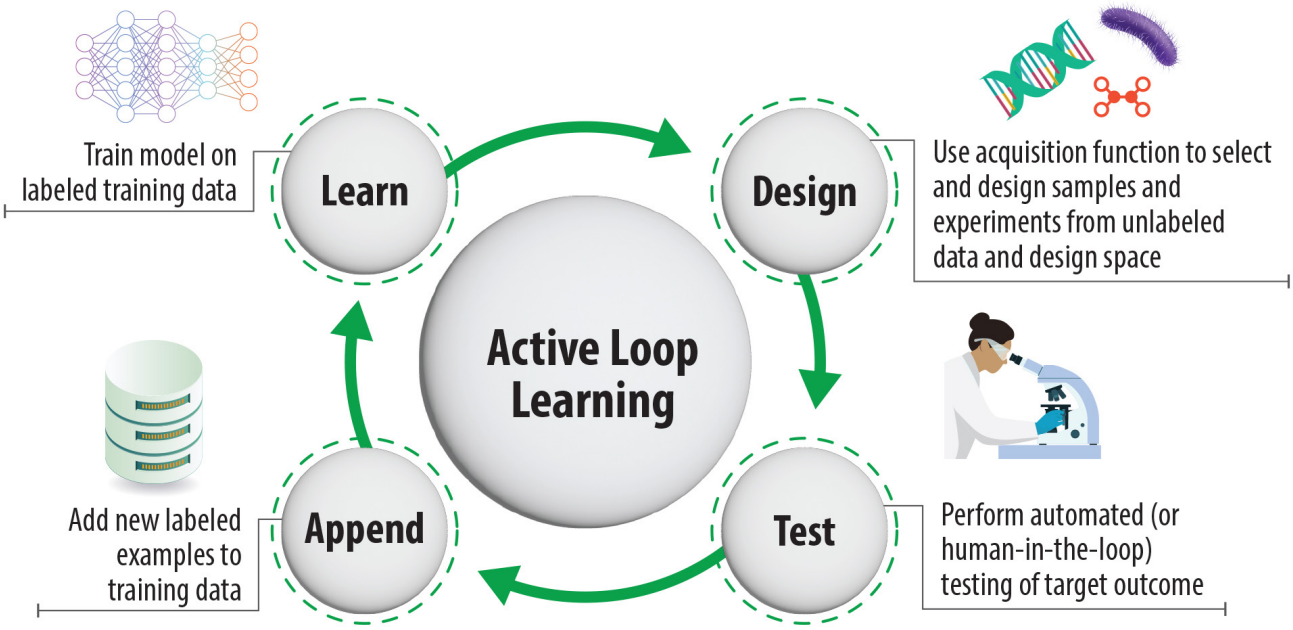


Fig. 4.5. Verification and Validation Within the Experimental Cycle. Incorporating rigorous assessment of AI predictions into the design and testing of hypotheses improves experimental outcomes, which reinforces the AI models' accuracy and reliability.

domain-specific expertise with AI-derived insights. Biological systems are inherently complex and variable, making reliable AI-based predictions particularly challenging. Effective V&V involves quantifying and explicitly representing the confidence and uncertainty associated with AI model predictions, thereby providing clarity on their reliability.

Incorporating interpretability within V&V enables researchers to understand the underlying biological mechanisms guiding AI-driven insights. For instance, AI-driven integration of multiomics data necessitates robust and scalable methods capable of handling heterogeneous data types, varying scales, missing values, biological noise, and technical artifacts. Rigorous V&V strategies must include extensive validation protocols, uncertainty quantification frameworks such as Bayesian inference, and adaptive methodologies tailored to diverse experimental scenarios, such as transfer learning approaches that validate models across varied datasets.

Scientific rigor demands reproducibility and thorough validation of AI-derived results versus direct observations (e.g., using digital and experimental twins to continuously benchmark model performance; see Fig. 4.4, p. 40). Robust V&V practices mandate transparent methodologies and comprehensive documentation, including version-controlled repositories for code; clearly standardized data formats; and detailed records of model architectures, hyperparameters, and training protocols. Transparency facilitates reproducibility and independent validation of findings. Experimental standardization and replication also facilitate those qualities, making them critical components of V&V.

In biological research—where field experiments can take over a year—erroneous predictions from AI systems can lead to significant setbacks. Hence, robust V&V protocols should incorporate systematic mechanisms for error detection, clearly defined uncertainty thresholds, methods for detecting out-of-distribution scenarios, and continuous validation against new experimental data, such as real-time anomaly detection algorithms, to proactively mitigate risks and ensure reliability.

Key Questions

- Which scalable AI methods or tools are critical for accurately simulating biological systems while rigorously incorporating and quantifying uncertainties?
- What specific V&V methods and metrics should be developed to ensure AI predictions are interpretable and robust for biologists?
- How can human-in-the-loop tools be designed to effectively integrate human expertise without creating bottlenecks?

Impact

Embedding rigorous V&V practices into AI-integrated biological research workflows will enhance scientific innovation, robustness, and reproducibility. These approaches will clearly communicate model limitations, quantify prediction confidence, and foster trust and engagement with the scientific community and public stakeholders. Ultimately, robust V&V will optimize resource allocation, enable reliable AI-driven discovery, and accelerate progress in complex biological research.

Target Activities

Develop and Implement Comprehensive V&V Protocols

- Develop robust V&V protocols that combine model-agnostic and model-specific checks, including mathematical and logistical consistency, numerical stability, and alignment with biological data.
- Incorporate uncertainty quantification methods (e.g., Monte Carlo dropout and Bayesian inference), interpretable feature selection, and out-of-distribution detection to ensure reliable and explainable AI predictions.

Integrate V&V into Scalable Workflows and HPC Platforms

- Leverage HPC platforms and scalable workflows to automate V&V processes, facilitate rigorous normalization, enable uncertainty-aware data integration, and support real-time anomaly detection.
- Promote the use of digital and experimental twins for continuous benchmarking and validation of AI models.

Promote Transparency, Reproducibility, and Documentation

- Establish best practices for transparent V&V methodologies, including version-controlled repositories, standardized data formats, and detailed documentation of model architectures, hyperparameters, and training protocols.
- Encourage experimental standardization and replication to support independent validation and reproducibility of findings.

Establish Adaptive and Domain-Aware V&V Strategies

- Develop adaptive V&V methodologies tailored to diverse experimental scenarios, such as transfer learning for cross-dataset validation, and approaches for handling rare or infrequent biological phenomena.
- Foster collaboration between domain experts and AI practitioners to integrate domain-specific knowledge into V&V processes.

4.7 Experiment Design and Automated Laboratories

Rationale (Challenges and Opportunities)

Experiment design is the process of planning and selecting the most effective methods of generating or acquiring data to reliably investigate and answer scientific questions, as well as to test, refine, and benchmark models (see Section 4.6: Verification and Validation, p. 41). With the increase in data from growing DOE instrument capabilities and other sources, choosing among potential experiment designs has become nontrivial, making expertise and intuition insufficient and suboptimal (U.S. DOE 2022c). The rise of HPC, AI, and automation introduces novel, human-in-the-loop approaches (e.g., machine teaching, interactive AI, and active learning) to guide experimental choices (Mosqueira-Rey et al. 2022).

Automated laboratories have revolutionized experimental capabilities by enabling high-throughput data generation, significantly increasing the speed and scale of scientific inquiry. These automated systems can execute complex experimental protocols with precision and consistency, generating vast amounts of data. For example, active learning models are capable of selecting the next set of experiments with high accuracy (Ding et al. 2024). Such approaches can balance the exploitation of current knowledge to achieve the learning objective with the exploration of the experimental space to accelerate discovery.

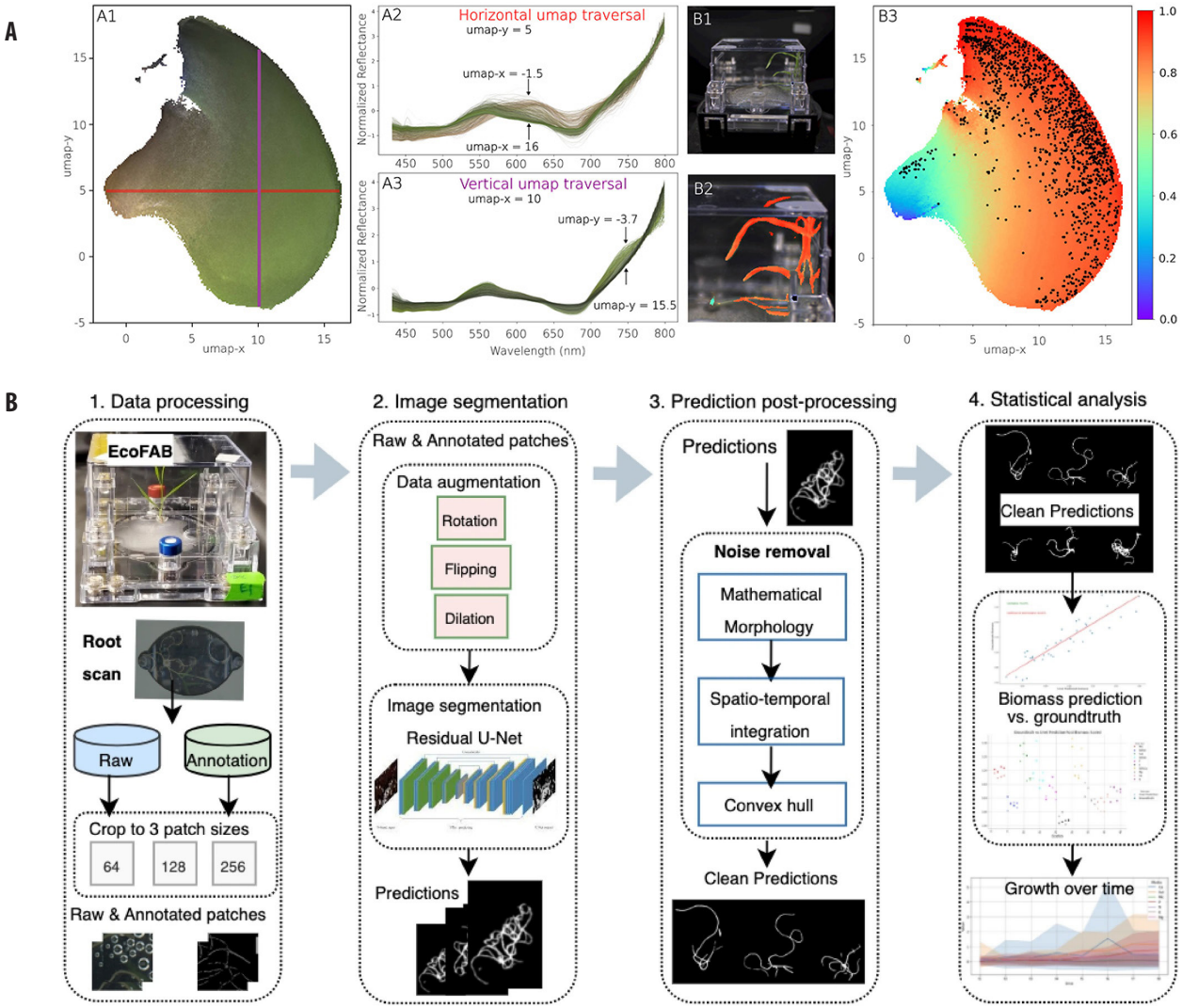
Building upon automated laboratories, autonomous experimentation aims to further enhance the research process by introducing AI-driven decision-making with HPC-backed data analysis workflows (see Fig. 4.6, p. 45). Autonomous experimentation systems, often referred to as "research robots" or "self-driving labs," can plan, execute, and evaluate experiments with minimal human intervention. These systems leverage AI algorithms to analyze experimental data in real time, learn from past results, and adapt future experiments, optimizing for specific research objectives.

Autonomous experiments require codifying biological design objectives into mathematically tractable optimization problems. This challenge is exacerbated by several factors. First, the objectives need to be transformed into the language of optimization: what is the autonomous experiment supposed to do? Should it home in on particularly interesting design choices or explore areas with little prior knowledge? Such decisions are difficult, and it is important to build mathematical descriptors that remove bias and systematically explore the design space. Second, the design space can be very large: building representations, energy functionals, probabilistic formulations, and efficient optimization methods are all important. These activities require state-of-the-art integration of mathematical ideas, scalable algorithms, partitioning and deployment in highly parallel compute environments, faithful representations of relevant biological formalisms, and close coupling to available data and domain knowledge (e.g., published literature). A scalable AI-driven system that meets these challenges holds tremendous potential.

Novel architectures, including specialized hardware and edge AI computing, offer exciting opportunities to advance autonomous experimentation. Integrating AI accelerators into experimental setups facilitates on-site computations, reducing latency and improving efficiency. Deploying AI capabilities directly at data sources enables real-time data processing and decision-making, allowing for rapid responses to dynamic experimental conditions. Experiment design, automated laboratories, and autonomous experimentation (Noack et al. 2020) are transformative approaches for investigating complex biological systems, particularly those with poorly characterized gene functions or community-level phenotypes. By integrating AI, HPC, and automation, researchers can develop computational approaches that ensure model generalization to diverse environmental conditions, facilitating rapid phenotyping and accelerating the discovery of biological processes.

Key Questions

- How can novel AI and scalable computational methods impact laboratory automation (e.g., robotics)?
- Can self-driving laboratories streamline analysis of complex multimodal data (e.g., genomic and phenomic data)?
- How can AI effectively incorporate prior scientific knowledge into data-driven modeling?
- What new AI tools that leverage exascale and edge computing will enable and accelerate real-time monitoring and feedback for plant, fungal, and microbial systems?
- What tools and algorithms are needed to advance AI experimental design?
- To what extent can AI-powered automated laboratories design and execute experiments to reach desired scientific discoveries?



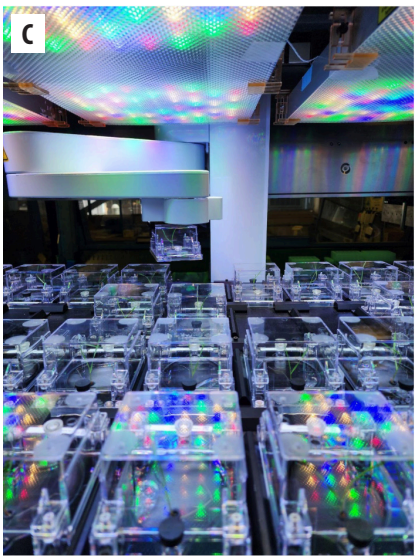


Fig. 4.6. Integrated AI with Automated Fabricated Ecosystems. (A) Application of Uniform Manifold Approximation and Projection (UMAP) to embed normalized spectra into a two-dimensional space for analysis of 3D hyperspectral imaging leaf data collected using EcoBOTs (see C panel). **(B)** Creation of RhizoNet, a computer vision system based on deep learning to automate root analysis from fabricated laboratory ecosystems acquired using the EcoBOT. **(C)** EcoBOT capabilities at Lawrence Berkeley National Laboratory. [(A) reprinted under a Creative Commons Attribution 4.0 International License (CC BY 4.0) from Zwart, P. H., et al. 2025. "Hyperspectral Segmentation of Plants in Fabricated Ecosystems," *Frontiers in High Performance Computing* **3**. DOI:10.3389/fhpcp.2025.1547340. (B) Reprinted under a Creative Commons Attribution 4.0 International License (CC BY 4.0) from Sordo, Z., et al. 2024. "RhizoNet Segments Plant Roots to Assess Biomass and Growth for Enabling Self-Driving Labs," *Scientific Reports* **14**, 12907. DOI:10.1038/s41598-024-63497-8. (C) Courtesy Lawrence Berkeley National Laboratory]

Impact

Integrating advanced experiment design, automated laboratories, and autonomous experimentation will transform biological research by enabling rapid, scalable, and reproducible investigation of complex systems. These approaches will accelerate scientific discovery, optimize resource allocation, and ensure robust model development and validation. Ultimately, the combination of AI, HPC, and automation will empower researchers to explore vast experimental landscapes, generate high-quality data, and drive innovation in biology and biotechnology.

Target Activities

Develop AI-Driven Experiment Design and Active Learning Approaches

- Develop and implement AI-driven experiment design strategies that combine human-in-the-loop methods (e.g., machine teaching, interactive AI, and active learning) with scalable HPC workflows to optimize data acquisition, model refinement, and hypothesis testing.

Advance Automated and Autonomous Laboratory Infrastructure

- Expand the deployment of automated and autonomous laboratory systems capable of executing high-throughput experiments, real-time data analysis, and adaptive experimental planning.
- Integrate AI accelerators and edge computing for on-site data processing and rapid decision-making.

Optimize Experimental Objective and Design Spaces

- Codify biological design objectives as mathematically tractable optimization problems to enable systematic exploration of large design spaces.
- Develop surrogate models, probabilistic frameworks, and efficient optimization algorithms to facilitate autonomous experimentation and discovery.

Integrate AI, HPC, and Automation for Rapid Biological Discovery

- Leverage the synergy of AI, HPC, and automation to accelerate phenotyping, model generalization, and the investigation of complex biological systems, including those with poorly characterized gene functions and community-level phenotypes.

Chapter 5

Concluding Remarks

Multidisciplinary discussions at the 2025 workshop on Envisioning Frontiers in AI and Computing for Biological Research identified four priority research directions (PRDs): Multimodal Data Assembly, Multiscale Biosystems Simulation, AI-Enabled Drivers for Experimental Systems, and Novel Algorithms for Genomics. These PRDs underscore the essential integration of advanced computational methodologies, such as deep learning, physics-informed modeling, scalable algorithms, and exascale computing platforms, to address complex biological challenges relevant to DOE missions.

While significant progress has been made, this workshop highlighted several areas requiring deeper computational specificity. Algorithmic innovation, explicit treatment of computational complexity with HPC, and rigorous verification and validation mechanisms are critical gaps. Addressing them will involve detailed exploration and development of specific computational methods such as graph neural networks, probabilistic inference, Bayesian optimization, and multiresolution modeling, all carefully tailored to biological contexts.

Many areas of biology continue to be data sparse. The DOE national laboratory system is uniquely positioned to generate the massive quantities of high-quality data needed to fill this gap by leveraging existing facilities and deep domain expertise. DOE user facilities are unmatched data generators because they operate multibillion-dollar scientific instruments that industry cannot afford, enabling researchers to conduct experiments under extreme conditions and at atomic resolution. The quality of this data is guaranteed by highly expert teams of scientists and engineers who conduct the experiments and operate the instruments. This specialized expertise ensures data meets rigorous quality standards and includes rich, standardized metadata, making the resulting datasets inherently AI ready and superior for training robust models. Unlike proprietary industry data, DOE's focus on basic research and open access creates massive public datasets for the entire scientific community.

Challenges and opportunities exist to improve throughput, accuracy, reproducibility, efficiency, and capability of experimental data generation platforms. Applying AI algorithms in this domain can improve the identification of knowledge gaps and guide the design of experimental campaigns that most effectively address such unknowns.

Advancing the computational biology frontier demands continued interdisciplinary collaboration, investment in computational infrastructure, and strategic alignment between computational scientists and biologists (see Table 5.1, p. 48). Applying the full potential of AI, ML, and computational sciences to biological research will drive transformative discoveries and enable unprecedented capabilities.

Table 5.1. Crosscutting Methodological Innovations in ASCR that Can Support BER Investigations

Focus Area	Biological Challenge	ASCR Methodological Innovation
Novel Algorithms	Identify systems-level emergent behavior from molecular rules	Sparse learning; bioinspired distributed architectures
Multiscale and Multimodal Modeling	Obtain mechanistic understanding of biological interactions across scales	Multiscale algorithms; exascale simulations; multimodal data integration and representation; agent-based modeling
Data Fusion	Fuse multiomics and imaging into causal models	Probabilistic models; exascale embedding and contrastive learning frameworks
Foundation Models	Generate and reason about hypothetical biomolecular functions	Federated multimodal large language models with fine-tuning pipelines and guardrails
Digital Twins	Conduct virtual crop, soil, and microbiome experiments	Hybrid physics–machine learning twin frameworks; surrogate machine learning cosolvers on exascale
Verification and Validation	Enable reliable prediction under biological variability	Conformal prediction; out-of-distribution detection; reproducible machine learning workflows
Experimental Design and Automated Laboratories	Incorporate closed-loop biodesign via robotics	Active learning; surrogate models; edge AI integration; hardware/software codesign; digital twins

Appendix A

Workshop Agenda

Day 1: February 4, 2025

8:00–9:00 a.m.	**Breakfast**
9:00–9:15 a.m.	**BER and ASCR Welcome** ***Speakers:*** Dorothy Koch, Associate Director, U.S. Department of Energy (DOE) Biological and Environmental Research (BER) program; Ceren Susut, Associate Director, DOE Advanced Scientific Computing Research (ASCR) program
9:15–9:30 a.m.	**Workshop Overview** ***Speakers:*** Daniela Ushizima (co-chair), Lawrence Berkeley National Laboratory; Christopher Henry (co-chair), Argonne National Laboratory
9:30–10:00 a.m.	**Foundation Models and Exascale Computing** ***Speaker:*** Rick Stevens, Argonne National Laboratory
10:00–10:15 a.m.	**Experiment Design with AI** ***Speaker:*** Kirsten Hofmockel, Pacific Northwest National Laboratory
10:15–10:30 a.m.	**Break**
10:30 a.m.–12:00 p.m.	**Breakout Sessions 1** ***Foundation Models:*** Three Groups ***Experiment Design:*** Three Groups
12:00–1:00 p.m.	**Lunch and Group Photo**
1:00–1:45 p.m.	**Morning Breakout Report-Outs: Foundation Models and Experiment Design**
1:45–2:00 p.m.	**Automated Labs/ Science** ***Speaker:*** Andrew Beam, Lila Sciences
2:00–2:30 p.m.	**Data Fusion** ***Speaker:*** David Baker, University of Washington
2:30–3:00 p.m.	**Automated Labs** ***Speaker:*** D.J. Kleinbaum, Emerald Cloud Lab
3:00–3:10 p.m.	**Break**
3:10–4:40 p.m.	**Breakout Sessions 2** ***Data Fusion:*** Three Groups ***Automated Labs:*** Three Groups
4:40–5:40 p.m.	**Afternoon Breakout Report-Outs: Data Fusion and Automated Labs**
5:40 p.m.	**Adjourn**

Day 2: February 5, 2025

8:00–9:00 a.m.	**Breakfast**
9:00–9:20 a.m.	**Day 1 Review and Day 2 Plan** ***Speakers:*** Christopher Henry and Daniela Ushizima
9:20–9:35 a.m.	**Digital Twins** ***Speaker:*** Shalin Mehta, Chan Zuckerberg Biohub
9:45–10:00 a.m.	**Digital Twins** ***Speaker:*** Jesse Tetreault, NVIDIA
10:00–10:15 a.m.	**Trustworthy AI** ***Speaker:*** Sergio Baranzini, University of California–San Francisco
10:15–10:30 a.m.	**Trustworthy AI** ***Speaker:*** Prasanna Balaprakash, Oak Ridge National Laboratory
10:30–10:35 a.m.	**Break**
10:45 a.m.–12:00 p.m.	**Breakout Sessions 3** ***Digital Twins:*** Three Groups ***Trustworthy AI:*** Three Groups
12:00–1:00 p.m.	**Lunch and Group Photo**
1:00–2:00 p.m.	**Morning Breakout Report-Outs: Digital Twins and Trustworthy AI**
2:00–2:15 p.m.	**Novel Algorithms** ***Speaker:*** Elebeoba May, University of Wisconsin–Madison
2:15–2:30 p.m.	**Novel Algorithms** ***Speaker:*** James Sethian, University of California–Berkeley
2:30–3:00 p.m.	**Multimodal Modeling at Exascale** ***Speaker:*** Arvind Ramanathan, Argonne National Laboratory
3:00–3:10 p.m.	**Break**
3:10–4:40 p.m.	**Breakout Sessions 4** ***Novel Algorithms:*** Three Groups ***Multiscale, Multimodal Modeling:*** Three Groups
4:40–5:40 p.m.	**Afternoon Breakout Report-Outs: Novel Algorithms and Multiscale Modeling**
5:40 p.m.	**Closing Remarks** ***Speakers:*** Margaret Lentz, ASCR; and Ramana Madupu, BER
6:00 p.m.	**Adjourn**

Day 3: February 6, 2025

8:00–9:00 a.m.	**Breakfast (all workshop participants who might be interested)**
9:00–10:30 a.m.	**Writing Session**
10:30–11:00 a.m.	**Break**
11:00 a.m.–12:00 p.m.	**Writing Session**
12:00 p.m.	**Adjourn**

Appendix B

Workshop Attendees

Co-Chairs

Daniela Ushizima, *Lawrence Berkeley National Laboratory*

Christopher Henry, *Argonne National Laboratory*

Program Committee

Prasanna Balaprakash, *Oak Ridge National Laboratory*

Ayan Biswas, *Los Alamos National Laboratory*

Adrienne Hoarfrost, *University of Georgia*

Kirsten Hofmockel, *Pacific Northwest National Laboratory*

Neeraj Kumar, *Pacific Northwest National Laboratory*

Arvind Ramanathan, *Argonne National Laboratory*

Trent Northen, *Lawrence Berkeley National Laboratory*

Strategic Committee

Margaret Lentz, *U.S. Department of Energy*

Ramana Madupu, *U.S. Department of Energy*

Todd Munson, *Argonne National Laboratory*

Workshop Attendees

Jacqueline Acres, *Whitman College*

Jonas Actor, *Sandia National Laboratories*

Francis Alexander, *Argonne National Laboratory*

Todd Anderson, *U.S. Department of Energy*

Sergio Baranzini, *University of California–San Francisco*

Paul Bayer, *U.S. Department of Energy*

Arunima Bhattacharjee, *Pacific Northwest National Laboratory*

Debsindhu Bhowmik, *Oak Ridge National Laboratory*

Aivett Bilbao, *Pacific Northwest National Laboratory*

Benjamin Bowen, *Lawrence Berkeley National Laboratory*

Benjamin Brown, *U.S. Department of Energy*

James Bruner, *Oak Ridge Institute for Science and Education*

Aydın Buluç, *Lawrence Berkeley National Laboratory*

William Cannon, *Pacific Northwest National Laboratory*

Romy Chakraborty, *Lawrence Berkeley National Laboratory*

Christine Chalk, *U.S. Department of Energy*

Tianlong Chen, *University of North Carolina–Chapel Hill*

Nicholas Chia, *Argonne National Laboratory*

Kriti Chopra, *Brookhaven National Laboratory*

Markus Covert, *Stanford University*

Kutter Craig, *Oak Ridge Institute for Science and Education*

Tanner Crowder, *U.S. Department of Energy*

Kevin Dalton, *SLAC National Accelerator Laboratory*

Paramvir Dehal, *Lawrence Berkeley National Laboratory*

Omar Demerdash, *Oak Ridge National Laboratory*

Sorin Draghici, *National Science Foundation*

Hal Finkel, *U.S. Department of Energy*

Ferdinando Fioretto, *University of Virginia*

Michael Fisher, *U.S. Department of Energy*

Marco Fornari, *U.S. Department of Energy*

Andrew Fowler, *Oak Ridge Institute for Science and Education*

Zachary Fox, *Oak Ridge National Laboratory*

Héctor García Martin, *Lawrence Berkeley National Laboratory*

Justin Hnilo, *U.S. Department of Energy*

Bin Hu, *Los Alamos National Laboratory*

Yunha Hwang, *Tatta Bio*

Daniel Jacobson, *Oak Ridge National Laboratory*

Paul Jensen, *University of Michigan*

Ravinder Kapoor, *U.S. Department of Energy*

Sagar Khare, *Rutgers University*

D.J. Kleinbaum, *Emerald Cloud Labs*

Carter Knutson, *Pacific Northwest National Laboratory*

Dorothy Koch, *U.S. Department of Energy*

Raga Krishnakumar, *Sandia National Laboratories*

Resham Kulkarni, *U.S. Department of Energy*

Yunqi Li, *Brookhaven National Laboratory*

Felice Lightstone, *Lawrence Livermore National Laboratory*

Pavel Lougovski, *U.S. Department of Energy*

Xiaoyi Lu, *University of California–Merced*

Sandeep Madireddy, *Argonne National Laboratory*

Costas Maranas, *The Pennsylvania State University*

Elebeoba May, *Wisconsin Institute of Discovery*

Jason McDermott, *Pacific Northwest National Laboratory*

Ambarish Nag, *National Renewable Energy Laboratory*

Peter Nugent, *Lawrence Berkeley National Laboratory*

Robinson Pino, *U.S. Department of Energy*

David Rabson, *U.S. Department of Energy*

Sridhar Raghavachari, *National Science Foundation*

Amit Roy-Chowdhury, *University of California–Riverside*

Ritimukta Sarangi, *SLAC National Accelerator Laboratory*

Vijayalakshmi Saravanan, *University of Texas–Tyler*

Gundolf Schenk, *University of California–San Francisco*

James Sethian, *University of California–Berkeley*

Sanna Sevanto, *Los Alamos National Laboratory*

Seth Steichen, *National Renewable Energy Laboratory*

Rick Stevens, *Argonne National Laboratory*

Ceren Susut, *U.S. Department of Energy*

Amy Swain, *U.S. Department of Energy*

Deneise Terry, *Oak Ridge Institute for Science and Education*

Jesse Tetreault, *NVIDIA*

Aeron Tynes Hammack, *Lawrence Berkeley National Laboratory*

Camilo Valdes, *Lawrence Livermore National Laboratory*

John Vant, *Oak Ridge National Laboratory*

Ming Wang, *University of California–Davis*

Bruce Warford, *Oak Ridge Associated Universities*

Pamela Weisenhorn, *Argonne National Laboratory*

Emma Westerman, *National Science Foundation*

Chenling Xu, *Lawrence Livermore National Laboratory*

Shinjae Yoo, *Brookhaven National Laboratory*

Byung-Jun Yoon, *Brookhaven National Laboratory*

Larry York, *Oak Ridge National Laboratory*

Karsten Zengler, *University of California–San Diego*

Petrus Zwart, *Lawrence Berkeley National Laboratory*

Appendix C
Glossary

AI co-scientist
An artificial intelligence system designed to act as a collaborative partner in scientific research, contributing substantively to hypothesis generation, experimental design, data analysis, interpretation, and discovery by integrating domain knowledge, reasoning, and adaptive learning in concert with human scientists.

anaerobic
A process, organism, or environment that occurs or exists in the absence of molecular oxygen.

biological dark matter
Biological molecules or organisms that are undetected or are detected but lack known functions.

causal inference
AI/statistical approaches designed to identify and understand the cause-and-effect relationships across data.

computational topology
Topology examines point sets and their invariants under continuous deformations, such as the number of connected components, holes, tunnels, or cavities. Computational topology deals with the complexity of topological problems and with the design of efficient algorithms for their solution in case these problems are tractable.

contrastive model
A machine learning model trained to learn representations by distinguishing between similar (positive) and dissimilar (negative) pairs of data, optimizing an objective function that increases similarity in the learned feature space for related inputs while maximizing separation for unrelated ones.

convolutional neural networks (CNN)
A class of deep, feedforward artificial neural networks designed to automatically and adaptively learn spatial hierarchies of features from structured data (such as images, sequences, or volumes) by applying convolutional operations, nonlinear activations, and pooling across multiple layers.

counterfactual explanation
An interpretable model output that identifies minimal changes to input features of a given instance that would alter the model's prediction to a specified desired outcome, thereby offering insight into the model's decision boundaries and causal behavior.

deep learning–based spectral analysis
The application of deep neural network architectures to interpret, model, or extract meaningful information from spectral data, such as those obtained from techniques including mass spectrometry, nuclear magnetic resonance, infrared spectroscopy, Raman spectroscopy, and ultraviolet–visible spectroscopy.

diffusion models
Generative models used primarily for image generation and other computer vision tasks. Diffusion-based neural networks are trained through deep learning to progressively "diffuse" samples with random noise, then reverse that diffusion process to generate high-quality images.

digital twin
A virtual representation or computational model of a physical object, system, or process designed to simulate real-world behaviors, interactions, and responses. By leveraging real-time data, advanced

simulations, and predictive analytics, digital twins allow users to monitor, analyze, optimize, and control their physical counterparts, enabling improved decision-making, experimentation, and forecasting in a controlled virtual environment.

edge algorithms
Computational methods designed for data processing and analysis directly on devices or sensors near data generation points, minimizing latency and bandwidth usage.

energy functionals
Mathematical constructs that assign a scalar energy value to a function or configuration of a physical system, typically representing the total energy (e.g., kinetic, potential, or free energy) as a function of fields, wavefunctions, or density distributions over space.

epigenetic
Heritable changes in gene expression that do not alter DNA sequences, typically involving chemical modifications like DNA methylation or histone modification.

exascale computing
Computation using computers capable of performing at least 1 exaFLOP (10^{18} floating point operations per second).

few-shot learning
A machine learning framework in which an AI model learns to make accurate predictions by training on a very small number of labeled examples.

foundation models
Deep learning models trained on vast datasets that can be applied across a wide range of use cases. Generative AI applications like large language models are common examples of foundation models.

global biogeochemical cycles
Integrated, planet-scale processes by which chemical elements and compounds are exchanged among the biosphere, atmosphere, hydrosphere, and geosphere, driven by biological, geological, and chemical mechanisms that regulate the composition and functioning of Earth's ecosystems.

global nutrient cycles
Large-scale, biogeochemical processes that govern the movement, transformation, and conservation of essential chemical elements (e.g., carbon, nitrogen, phosphorus, and sulfur) through the biosphere, atmosphere, hydrosphere, and geosphere, enabling the sustained productivity and regulation of Earth's ecosystems.

graph neural networks (GNNs)
Graph neural networks apply the predictive power of deep learning to rich data structures that depict objects and their relationships as points connected by lines in a graph.

hyperparameter optimization (HPO)
A mechanism for automatically exploring a search space of potential hyperparameters, building a series of models and comparing the models using metrics of interest.

in situ
Experiments or observations performed within the natural location or native context of a biological system, without removing the subject from its original environment or disrupting its structural or spatial organization.

isofunctional protein families
Groups of evolutionarily related proteins that, despite possible sequence divergence, catalyze the same biochemical reaction or perform the same molecular function across different organisms or contexts.

latent embedding spaces
In machine learning, a compressed representation of data points that preserves only essential features that inform the input data's underlying structure.

large language model (LLM)
A specialized type of machine learning model tailored for natural language processing tasks, including text generation. These models contain a large number of parameters and are typically trained using self-supervised techniques on extensive text datasets.

long molecular representations
Structured encodings of complex biological molecules (such as DNA, RNA, proteins, or metabolites) that capture detailed, extended information about their sequence, structure, modifications, or functional context across large spatial or informational scales.

manifold learning
A class of unsupervised estimators that seeks to describe datasets as low-dimensional manifolds embedded in high-dimensional spaces.

mass spectrometry
An analytical technique used to measure the mass-to-charge ratio (m/z) of ions.

metabolism
Describes how cells extract materials and energy from the environment and synthesize important byproducts.

metabolite
A small biological molecule produced by or involved in cellular metabolism.

metabolomics
The global profiling of small molecule metabolites in a biological sample (typically <1,500 daltons), providing insights into metabolic processes.

metaphenomes
Comprehensive sets of measurable phenotypic traits and functions expressed collectively by microbial communities (or other multiorganism systems) in a specific environment, emerging from the combined genetic potential (metagenomes) and environmental interactions of the constituent organisms.

microbiome
The collection of microorganisms (such as bacteria, fungi, viruses, and archaea) present in a specific environment, such as an animal gut or soil.

model-experiment-observation (ModEx)
An iterative research approach integrating computational modeling and experimental data to accelerate scientific discovery and validation.

multiscale
Analysis or modeling spanning multiple spatial or temporal scales, such as molecular to organism or milliseconds to years.

multimodal
Combining different data types or sensing modalities (e.g., images, text, and audio) to enhance analysis, prediction, or understanding.

multiomics
The integrative analysis of multiple omics data types (e.g., genomics, proteomics, and metabolomics) for comprehensive biological insight.

neural networks (NN)
Information processing paradigms inspired by the way biological neural systems process data.

out-of-distribution scenarios
Data points that fall outside the distribution of the training data for a model.

parallelized deep networks
Deep learning architectures whose training or inference processes are distributed across multiple computational units, such as central processing units, graphics processing units, or compute nodes, using parallel computing techniques to increase scalability, reduce runtime, and handle large-scale data or model sizes.

phenotyping
Process of observing, measuring, and analyzing an organism's traits or characteristics.

physics-informed AI/ML
AI/ML that seamlessly integrates data and mathematical physics models, even in partially understood, uncertain, and high-dimensional contexts.

probabilistic formulations
Mathematical framework in which phenomena, models, or hypotheses are expressed in terms of probability theory, enabling the representation of uncertainty, variability, and incomplete information through probability distributions over possible outcomes or parameters.

probabilistic inference
The process of calculating the conditional probability of a variable having a certain value, given specific evidence about other variables in a probabilistic model.

proteomics
The large-scale study of proteins, including their structures, functions, and interactions, within cells or organisms.

rhizosphere
Region of soil impacted by the presence of plant roots.

strong scaling
The efficiency of solving a fixed total program size or workload size with increasing numbers of workers.

surrogate representations
Simplified approximations of more complex, higher-order models. They are generally used to map input data to outputs when the actual relationship between the two is unknown or computationally expensive to evaluate.

threading
Small units of a computer program that can run independently, allowing the program to perform multiple tasks at the same time.

transcriptomics
The study of the complete set of RNA transcripts produced by the genome, providing insights into gene expression and regulation.

UniProt Knowledgebase/Swiss-Prot
A database providing high-quality, nonredundant protein sequence records with expert-reviewed functional annotations, including information on protein function, domain structure, post-translational modifications, variants, and protein–protein interactions. Swiss-Prot is a manually curated subsection of UniProtKB, comprised primarily of proteins with experimentally validated functions.

vectorization
A technique used to improve the performance of operations on data, especially large datasets, by processing multiple data points simultaneously using a single instruction, often reducing the use of for/while loops.

vision transformers (ViTs)
A transformer-like model that handles images for vision processing tasks.

Appendix D
References

Prior to the workshop, attendees were invited to submit position papers discussing key challenges and opportunities to formulate, implement, and apply AI/ML frameworks to biological systems relevant to BER's mission space. This community input shaped the workshop agenda, panelist discussions, and workshop report. These papers are registered together under DOI number 10.2172.2512398.

Abramson, J., et al. 2024. "Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3," *Nature* **630**, 493–500. DOI:10.1038/s41586-024-07487-w.

Acosta, J. N., et al. 2022. "Multimodal Biomedical AI," *Nature Medicine* **28**, 1773–84. DOI:10.1038/s41591-022-01981-2.

Alber, M., et al. 2019. "Integrating Machine Learning and Multiscale Modeling—Perspectives, Challenges, and Opportunities in the Biological, Biomedical, and Behavioral Sciences," *NPJ Digital Medicine* **2**, 115. DOI:10.1038/s41746-019-0193-y.

Almagro Armenteros, J. J., et al. 2019. "SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks," *Nature Biotechnology* **37**, 420–3. DOI:10.1038/s41587-019-0036-z.

Anderson, L. N., et al. 2025. "Computation Tools and Data Integration to Accelerate Vaccine Development; Challenges, Opportunities, and Future Directions, *Fronters in Immunology* **16**. DOI:10.3389/fimmu.2025.1502484.

Argelaguet, R., et al. 2020. "MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data," *Genome Biology* **21**, 111. DOI:10.1186/s13059-020-02015-1.

Arkin, A. P., et al. 2018. "KBase: The United States Department of Energy Systems Biology Knowledgebase," *Nature Biotechnology* **36**, 566–9. DOI:10.1038/nbt.4163.

Arnosti, C., et al. 2021. "The Biogeochemistry of Marine Polysaccharides: Sources, Inventories, and Bacterial Drivers of the Carbohydrate Cycle," *Annual Review of Marine Science* **13**, 81–108. DOI:10.1146/annurev-marine-032020-012810.

ASCAC. 2020. *ASCR@40: Highlights and Impacts of ASCR's Programs.* Advanced Scientific Computing Research Advisory Committee. https://doi.org/10.2172/1631812.

ASCAC. 2024. *2024 Advanced Scientific Computing Advisory Committee Facilities Subcommittee Recommendation.* Advanced Scientific Computing Research Advisory Committee. https://doi.org/10.2172/2370379.

ATLAS Collaboration. 2022. "A Detailed Map of Higgs Boson Interactions by the ATLAS Experiment Ten Years After the Discovery," *Nature* **607**, 52–59. DOI:10.1038/s41586-022-04893-w.

Avsec, Ž., et al. 2021. "Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions," *Nature Methods* **18,** 1196–1203. DOI:10.1038/s41592-021-01252-x.

Awan, M. G., et al. 2021. "Accelerating Large Scale *de novo* Metagenome Assembly using GPUs." In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.* New York, N.Y., U.S. pp. 1–11. DOI:10.5281/zenodo.5165333.

Baek, M., et al. 2021. "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network," *Science* **373**(6557), 871–6. DOI:10.1126/science.abj8754.

Ball, P. 2023. "Is AI Leading to a Reproducibility Crisis in Science?" *Nature* **624**, 22–5. DOI:10.1038/d41586-023-03817-6.

Ballard, J. L., et al. 2024. "Deep Learning-Based Approaches for Multi-omics Data Integration and Analysis," *BioData Mining* **17**, 38. DOI:10.1186/s13040-024-00391-z.

Bansal, P., et al. 2022. "Rhea, the Reaction Knowledgebase in 2022," *Nucleic Acids Research* **50**(D1), D693–700. DOI:10.1093/nar/gkab1016.

BERAC. 2024. *A Unified Data Infrastructure for Biological and Environmental Research: Report from the BER Advisory Committee*, DOE/SC-0214. Biological and Environmental Research Advisory Committee. https://doi.org/10.2172/2331276.

Berman, H. M., et al. 2000. "The Protein Data Bank," *Nucleic Acids Research* **28**(1), 235–42. DOI:10.1093/nar/28.1.235.

Borghoff, U. M., et al. 2025. "Human-Artificial Interaction in the Age of Agentic AI: A System-Theoretical Approach," *Frontiers in Human Dynamics* **7**. DOI:10.3389/fhumd.2025.1579166.

Brand, A., et al. 2015. "Beyond Authorship: Attribution, Contribution, Collaboration, and Credit," *Learned Publishing* **28**(2), 151–5. DOI:10.1087/20150211.

Brandes, N., et al. 2023. "Genome-Wide Prediction of Disease Variant Effects with a Deep Protein Language Model," *Nature Genetics* **55**, 1512–22. DOI:10.1038/s41588-023-01465-0.

Brunton, S. L. and J. N. Kutz. 2019. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, Cambridge, United Kingdom. DOI:10.1017/9781108380690.

Burger, B., et al. 2020. "A Mobile Robotic Chemist," *Nature* **583**, 237–41. DOI:10.1038/s41586-020-2442-2.

Burley, S. K. 2025. "Protein Data Bank: From Two Epidemics to the Global Pandemic to mRNA Vaccines and Paxlovid," *Current Opinion in Structural Biology* **90**. DOI:10.1016/j.sbi.2024.102954.

Cao, Z.-J., and G. Gao. 2022. "Multi-omics Single-Cell Data Integration and Regulatory Inference with Graph-Linked Embedding," *Nature Biotechnology* **40**, 1458–66. DOI:10.1038/s41587-022-01284-4.

Carbonell, P., et al. 2018. "An Automated Design-Build-Test-Learn Pipeline for Enhanced Microbial Production of Fine Chemicals," *Communications Biology* **1**, 66. DOI:10.1038/s42003-018-0076-9.

Choudhary, K., et al. 2024. "JARVIS-Leaderboard: A Large Scale Benchmark of Materials Design Methods," *NPJ Computational Materials* **10**, 93. DOI:10.1038/s41524-024-01259-w.

Clyde, A., et al. 2021. "High-Throughput Virtual Screening and Validation of a SARS-CoV-2 Main Protease Noncovalent Inhibitor," *Journal of Chemical Information and Modeling* **62**(1), 116–28. DOI:10.1021/acs.jcim.1c00851.

Corral-Acero, J., et al. 2020. "The 'Digital Twin' To Enable the Vision of Precision Cardiology," *European Heart Journal* **41**(48), 4556–64. DOI:10.1093/eurheartj/ehaa159.

Dalla-Torre, H., et al. 2025. "Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics," *Nature Methods* **22**, 287–97. DOI:10.1038/s41592-024-02523-z.

Data Citation Synthesis Group. 2014. *Joint Declaration of Data Citation Principles*. Ed. Martone, M. FORCE11, San Diego, Calif. DOI:10.25490/a97f-egyk.

Deisboeck, T. S., et al. 2014. "Multiscale Cancer Modeling," *Annual Review of Biomedical Engineering* **13**, 127–55. DOI:10.1146/annurev-bioeng-071910-124729.

Ding, K., et al. 2024. "Machine Learning-Guided Co-Optimization of Fitness and Diversity Facilitates Combinatorial Library Design in Enzyme Engineering," *Nature Communications* **15**, 6392. DOI:10.1038/s41467-024-50698-y.

Dosovitskiy, A., et al. 2021. "An Image is Worth 16 x 16 Words: Transformers for Image Recognition at Scale." In *Proceedings from the International Conference on Learning Representations.* Vienna, Austria. openreview.net/forum?id=YicbFdNTTy

Dumitrache, A., et al. 2020. "Empirical Methodology for Crowdsourcing Ground Truth," *Semantic Web* **12**(3), 403–21. DOI:10.3233/SW-200415.

ECP. "Mathematical Libraries." Accessed January 2025. The Exascale Computing Project. https://www.exascaleproject.org/research-group/math-libraries/

Eissing, T., et al. 2011. "A Computational Systems Biology Software Platform for Multiscale Modeling and Simulation: Integrating Whole-Body Physiology, Disease Biology, and Molecular Reaction Networks," *Frontiers in Physiology* **2**(4). DOI:10.3389/fphys.2011.00004.

Er, G. A., et al. 2024. "Multimodal Data Fusion Using Sparse Canonical Correlation Analysis and Cooperative Learning: A COVID-19 Cohort Study," *NPJ Digital Medicine* **7**, 117. DOI:10.1038/s41746-024-01128-2.

Frazer, J., et al. 2021. "Disease Variant Prediction with Deep Generative Models of Evolutionary Data," *Nature* **599**, 91–5. DOI:10.1038/s41586-021-04043-8.

Fuller, A., et al. 2020. "Digital Twin: Enabling Technologies, Challenges, and Open Research," *IEEE Access* **8**, 108952–71. DOI:10.1109/ACCESS.2020.2998358.

Gao, F., et al. 2022. "Artificial Intelligence in Omics," *Genomics, Proteomics & Bioinformatics* **20**(5), 811–3. DOI:10.1016/j.gpb.2023.01.002.

Gayoso, A., et al. 2021. "Joint Probabilistic Modeling of Single-Cell Multi-Omic Data with totalVI," *Nature Methods* **18**, 272–82. DOI:10.1038/s41592-020-01050-x.

Gene Ontology Consortium. 2021. "The Gene Ontology Resource: Enriching a Gold Mine," *Nucleic Acids Research* **49**(D1), D325–34. DOI:10.1093/nar/gkaa1113.

Gergov, I., and G. Tsochev. 2025. "Watermarking Fine-Tuning Datasets for Robust Provenance," *Applied Sciences* **15**(19), 10457. DOI:10.3390/app151910457.

Gong, X., et al. 2024. "Advancing Microbial Production through Artificial Intelligence-Aided Biology," *Biotechnology Advances* **74**. DOI:10.1016/j.biotechadv.2024.108399.

Green, A. G., et al. 2021. "Large-Scale Discovery of Protein Interactions at Residue Resolution Using Co-Evolution Calculated from Genomic Sequences," *Nature Communications* **12**, 1396. DOI:10.1038/s41467-021-21636-z.

Hoffmann, J., et al. 2022. "Training Compute-Optimal Large Language Models," *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2176. Curran Associates Inc., Red Hook, N.Y.

Hoffmann, M. A., et al. 2022. High-Confidence Structural Annotation of Metabolites Absent from Spectral Libraries," *Nature Biotechnology* **40**, 411–21. DOI:10.1038/s41587-021-01045-9.

Huang, P. S., et al. 2016. "The Coming of Age of *De Novo* Protein Design," *Nature* **537**, 320–7. DOI:10.1038/nature19946.

Jacovi, A., and Y. Goldberg. 2020. "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–205. Association for Computational Linguistics. DOI:10.18653/v1/2020.acl-main.386.

Jansson, J. K., and K. S. Hofmockel. 2020. "Soil Microbiomes and Climate Change," *Nature Reviews Microbiology* **18**, 35–46. DOI:10.1038/s41579-019-0265-7.

Ji, Y., et al. 2021. "DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome," *Bioinformatics* **37**(15), 2112–20. DOI:10.1093/bioinformatics/btab083.

Jinek, M., et al. 2012. "A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity," *Science* **337**(6096), 816–21. DOI:10.1126/science.1225829.

Jumper, J., et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* **596**, 583–9. DOI:10.1038/s41586-021-03819-2.

Kalamkar, S., and M. A. Geetha. 2023. "Multimodal Image Fusion: A Systematic Review," *Decision Analytics* Journal **9**, 100327. DOI:10.1016/j.dajour.2023.100327.

Karniadakis, G. E., et al. 2021. "Physics-Informed Machine Learning," *Nature Reviews Physics* **3**, 422–40. DOI:10.1038/s42254-021-00314-5.

Karpatne, A., et al. 2017. "Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data," *IEEE Transactions on Knowledge and Data Engineering* **29**(10), 2318–31. DOI:10.1109/TKDE.2017.2720168.

Khdoudi, A., et al. 2024. "A Deep-Reinforcement-Learning-Based Digital Twin for Manufacturing Process Optimization," *Systems* **12**(2), 38. DOI:10.3390/systems12020038.

Knight, C. G., et al. 2024. "Soil Microbiomes Show Consistent and Predictable Responses to Extreme Events," *Nature* **636**, 690–6. DOI:10.1038/s41586-024-08185-3.

Knutson, C., et al. 2022. "Decoding the Protein–Ligan Interactions using Parallel Graph Neural Networks," *Scientific Reports* **12**, 7624. DOI:10.1038/s41598-022-10418-2.

Kryshtafovych, A., et al. 2023. "Critical Assessments of Protein Structure Prediction(CASP)–Round XV," *Proteins* **91**(12), 1539–49. DOI:10.1002/prot.26617.

Kwon, Y., et al. 2020. "Uncertainty Quantification Using Bayesian Neural Networks in Classification: Application to Biomedical Image Segmentation," *Computational Statistics & Data Analysis* **142**, 106816. DOI:10.1016/j.csda.2019.106816.

La Fleur, A., et al. 2024. "Decoding Biology with Massively Parallel Reporter Assays and Machine Learning," *Genes & Development* **38**, 843–65. DOI:10.1101/gad.351800.124.

Lee, J., et al. 2025. "Beyond Rigid Docking: Deep Learning Approaches for Fully Flexible Protein–Ligand Interactions," *Briefings in Bioinformatics* **26**(5), bbaf454. DOI:10.1093/bib/bbaf454.

Leggieri, P. A., et al. 2021. "Integrating Systems and Synthetic Biology to Understand and Engineer Microbes," *Annual Review of Biomedical Engineering* **23**, 169–201. DOI:10.1146/annurev-bioeng-082120-022836.

Li, W., et al. 2024. "Knowledge-Guided Learning Methods for Integrative Analysis of Multi-omics Data," *Computational and Structural Biotechnology Journal* **23**, 1945–50. DOI:10.1016/j.csbj.2024.04.053.

Liu, H., et al. 2022. "Dynamic Knowledge Graph Reasoning Based on Deep Reinforcement Learning," *Knowledge-Based Systems* **241**, 108235. DOI:10.1016/j.knosys.2022.108235.

Liu, S., et al. 2022. "Opportunities and Challenges of Using Metagenomic Data To Bring Uncultured Microbes into Cultivation," *Microbiome* **10**, 76. DOI:10.1186/s40168-022-01272-5.

Liu, S., et al. 2025. "A Text-Guided Protein Design Framework," *Nature Machine Intelligence* **7**, 580–91. DOI:10.1038/s42256-025-01011-z.

Lookman, T., et al. 2019. "Active Learning in Materials Science with Emphasis on Adaptive Sampling Using Uncertainties for Targeted Design," *NPJ Computational Materials* **5**, 21. DOI:10.1038/s41524-019-0153-8.

Lotfollahi, M., et al. 2023. "Biologically Informed Deep Learning To Query Gene Programs in Single-Cell Atlases," *Nature Cell Biology* **25**, 337–50. DOI:10.1038/s41556-022-01072-x.

Loumeaud, A., et al. 2024. "Multiscale Mechanical Modeling of Skeletal Muscle: A Systemic Review of the Literature," *Journal of Medical and Biological Engineering* **44**, 337–56. DOI:10.1007/s40846-024-00879-3.

Lundberg, S. M., et al. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nature Machine Intelligence* **2**, 56–67. DOI:10.1038/s42256-019-0138-9.

Mammoliti, A., et al. 2021. "Orchestrating and Sharing Large Multimodal Data for Transparent and Reproducible Research," *Nature Communications* **12**, 5797. DOI:10.1038/s41467-021-25974-w.

Mansoor, S., et al. 2024. "Advance Computational Tools for Multiomics Data Learning," *Biotechnology Advances* **77**. DOI:10.1016/j.biotechadv.2024.108447.

Marbach, D., et al. 2012. "Wisdom of Crowds for Robust Gene Network Inference," *Nature Methods* **9**, 796–804. DOI:10.1038/nmeth.2016.

McNaughton, A. D., et al. 2024. "CACTUS: Chemistry Agent Connecting Tool Usage to Science," *ACS Omega* **9**(46), 46563–73. DOI:10.1021/acsomega.4c08408.

Mosqueira-Rey, E., et al. 2022. "Human-in-the-Loop Machine Learning: A State of the Art," *Artificial Intelligence Review* **56**, 3005–54. DOI:10.1007/s10462-022-10246-w.

Mukhtar, H., et al. 2022. "Digital Twins of the Soil Microbiome for Climate Mitigation," *Environments* **9**(3), 34. DOI:10.3390/environments9030034.

NASEM. 2019. *Reproducibility and Replicability in Science*. National Academies of Sciences, Engineering, and Medicine, Washington, D.C. The National Academies Press. DOI:10.17226/25303.

NASEM. 2024. *Foundational Research Gaps and Future Directions for Digital Twins.* National Academies of Sciences, Engineering, and Medicine. National Academies Press, Washington, D.C., U.S. https://nap.nationalacademies.org/catalog/26894/foundational-research-gaps-and-future-directions-for-digital-twins

NASEM. 2025. *The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations*. National Academies of Sciences, Engineering, and Medicine, Washington, D.C. The National Academies Press. DOI:10.17226/28868.

Naveed, M. H., et al. 2024. "Cellulosic Biomass Fermentation for Biofuel Production: Review of Artificial Intelligence Approaches," *Renewable and Sustainable Energy Reviews* **189**. DOI:10.1016/j.rser.2023.113906.

Nethery, M. A., et al. 2022. "CRISPR-Based Engineering of Phages for *In Situ* Bacterial Base Editing," *Proceedings of the National Academy of Sciences USA* **119**(46), e2206744119. DOI:10.1073/pnas.2206744119.

Noack, M. M., et al. 2020. "Autonomous Materials Discovery Driven by Gaussian Process Regression with Inhomogeneous Measurement Noise and Anisotropic Kernels," *Scientific Reports* **10**, 17663. DOI:10.1038/s41598-020-74394-1.

Noack, M., and D. Ushizima, Eds. 2024. *Methods and Applications of Autonomous Experimentation.* Chapman & Hall, Boca Raton, Fla. 444 pp.

Novak, V., et al. 2025. "Breaking the Reproducibility Barrier with Standardized Protocols for Plant–Microbiome Research," *PLoS Biology* **23**(9), e3003358. DOI:10.1371/journal.pbio.3003358.

NSCEB. 2025. *Charting the Future of Biotechnology: An Action Plan for American Security and Prosperity*. National Security Commission of Emerging Biotechnology (NSCEB). www.biotech.senate.gov

NVBL. "National Virtual Biotechnology Laboratory." Accessed January 2025. U.S. Department of Energy. https://science.osti.gov/nvbl.

Oberkampf, W. L., and C. J. Roy. 2010. *Verification and Validation in Scientific Computing*. Cambridge University Press, Cambridge, United Kingdom. DOI:10.1017/CBO9780511760396.

Oikawa, P. Y., et al. 2024. "A New Coupled Biogeochemical Modeling Approach Provides Accurate Predictions of Methane and Carbon Dioxide Fluxes Across Diverse Tidal Wetlands," *Journal of Geophysical Research*: *Biogeosciences* **129**(10), e2023JG007943. DOI:10.1029/2023JG007943.

One Big Beautiful Bill Act. 2025. H.R.1, 119th Cong. www.congress.gov/bill/119th-congress/house-bill/1/text

Orth, J. D., et al. 2010. "What is Flux Balance Analysis?" *Nature Biotechnology* **28**, 245–8 DOI:10.1038/nbt.1614.

Pearl, J. 2009. *Causality*. 2nd ed. Cambridge University Press, Cambridge, United Kingdom. DOI:10.1017/CBO9780511803161.

Peng, Y., et al. 2025. "Contrastive-Learning of Language Embedding and Biological Features for Cross Modality Encoding and Effector Prediction," *Nature Communications* **16**, 1299. DOI:10.1038/s41467-025-56526-1.

Prince, M. H., et al. 2024. "Opportunities for Retrieval and Tool Augmented Large Language Models in Scientific Facilities," *NPJ Computational Materials* **10**, 251. DOI:10.1038/s41524-024-01423-2.

Rafi, A. M., et al. 2025. "A Community Effort To Optimize Sequence-Based Deep Learning Models of Gene Regulation," *Nature Biotechnology* **43**, 1373–83. DOI:10.1038/s41587-024-02414-w.

Raissi, M., et al. 2019. "Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations," *Journal of Computational Physics* **378**, 686–707. DOI:10.1016/j.jcp.2018.10.045.

Raj, A., and A. van Oudenaarden. 2008. "Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences," *Cell* **135**(2), 216–26. DOI:10.1016/j.cell.2008.09.050.

Riesselman, A. J., et al. 2018. "Deep Generative Models of Genetic Variation Capture the Effects of Mutations," *Nature Methods* **15,** 816–22. DOI:10.1038/s41592-018-0138-4.

Robitaille, M. C., et al. 2022. "Self-Supervised Machine Learning for Live Cell Imagery Segmentation," *Communications Biology* **5**, 1162. DOI:10.1038/s42003-022-04117-x.

Sanchez-Lengeling, B., and A. Aspuru-Guzik. 2018. "Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering," *Science* **361**, 360–65. DOI:10.1126/science.aat2663.

Sasse, J., et al. 2018. "Feed Your Friends: Do Plant Exudates Shape the Root Microbiome?" *Trends in Plant Science* **23**(1), 25–41. DOI:10.1016/j.tplants.2017.09.003.

Schillings, C., and A. M. Stuart. 2017. "Analysis of the Ensemble Kalman Filter for Inverse Problems," *SIAM Journal on Numerical Analysis* **55**(3), 1264–90. DOI:10.1137/16M105959X.

Shahriari, B., et al. 2016. "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proceedings of the IEEE* **104**(1), 148–75. DOI:10.1109/JPROC.2015.2494218.

Silver, D., et al. 2021. "Reward is Enough," *Artificial Intelligence* **299**, 103535. DOI:10.1016/j.artint.2021.103535.

Singh, A., et al. 2016. "Machine Learning for High-Throughput Stress Phenotyping in Plants," *Trends in Plant Science* **21**(2), 110–24. DOI:10.1016/j.tplants.2015.10.015.

Smith, R. C. 2013. *Uncertainty Quantification: Theory, Implementation, and Application.* Society for Industrial and Applied Mathematics, Philadelphia, Pa. DOI:10.1137/1.9781611973228.

Sordo, Z., et al. 2024. "RhizoNet Segments Plant Roots to Assess Biomass and Growth for Enabling Self-Driving Labs," *Scientific Reports* **14**, 12907. DOI:10.1038/s41598-024-63497-8.

Stokes, J. M., et al. 2020. "A Deep Learning Approach to Antibiotic Discovery," *Cell* **180**(4), 688–702.e13. DOI:10.1016/j.cell.2020.01.021.

Sullivan, K. A., et al. 2024. "Analyses of GWAS Signal Using GRIN Identify Additional Genes Contributing to Suicidal Behavior," *Communications Biology* **7**, 1360. DOI:10.1038/s42003-024-06943-7.

Sundararajan, M., et al. 2017. "Axiomatic Attribution for Deep Networks," In *ICML'17: Proceedings of the 34th International Conference on Machine Learning —Volume 70,* 3319–28. JMLR.org. DOI:10.5555/3305890.3306024.

Szklarczyk, D., et al. 2023. "The STRING Database in 2023: Protein–Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest," *Nucleic Acids Research* **51**(D1), D638–46. DOI:10.1093/nar/gkac1000.

Szklarczyk, D., et al. 2025. "The STRING Database in 2025: Protein Networks with Directionality of Regulation," *Nucleic Acids Research* **53**(D1), D730–7. DOI:10.1093/nar/gkae1113.

Theodoris, C. V., et al. 2023. "Transfer Learning Enables Predictions in Network Biology," *Nature* **618**, 616–24. DOI:10.1038/s41586-023-06139-9.

Thirunavukarasu, A. J., et al. 2023. "Large Language Models in Medicine," *Nature Medicine* **29**, 1930–40. DOI:10.1038/s41591-023-02448-8.

Thompson, L., et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity," *Nature* **551**, 457–63. DOI:10.1038/nature24621.

Truhn, D., et al. 2024. "Large Language Models and Multimodal Foundation Models for Precision Oncology," *NPJ Precision Oncology* **8**, 72. DOI:10.1038/s41698-024-00573-2.

Tunyasuvunakool, K., et al. 2021. "Highly Accurate Protein Structure Prediction for the Human Proteome," *Nature* **596**, 590–6. DOI:10.1038/s41586-021-03828-1.

U.S. DOE. 2019. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*. U.S. Department of Energy Advanced Scientific Computing Research program. https://doi.org/10.2172/1478744.

U.S. DOE. 2020. *AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science,* Report No. ANL-20/17, 158802. U.S. Department of Energy. https://doi.org/10.2172/1604756.

U.S. DOE. 2022a. *Report for the ASCR Workshop on the Management and Storage of Scientific Data.* U.S. Department of Energy Office of Science Advanced Scientific Computing Research program. https://doi.org/10.2172/1845707

U.S. DOE. 2022b. *DOE National* Virtual *Biotechnology Laboratory Report on Rapid R&D Solutions to the COVID-19 Crisis*. U.S. Department of Energy Office of Science. https://doi.org/10.2172/1968213

U.S. DOE. 2022c. *Envisioning Science in 2050.* U.S. Department of Energy. https://doi.org/10.2172/1871683.

U.S. DOE. 2023a. *Advanced Research Directions on AI for Science, Energy, and Security: Report on Summer 2022 Workshops*. U.S. Department of Energy Office of Science and U.S. Department of Energy National Nuclear Security Administration. https://doi.org/10.2172/1986455

U.S. DOE. 2023b. *Artificial Intelligence and Machine Learning for Bioenergy Research: Opportunities* and *Challenges*, DOE/SC-0211. U.S. Department of Energy Office of Science and Office of Energy Efficiency and Renewable Energy. https://doi.org/10.2172/1968870.

U.S. DOE. 2023c. *Critical Materials Assessment*. U.S. Department of Energy. www.energy.gov/sites/default/files/2023-07/doe-critical-material-assessment_07312023.pdf

U.S. DOE. 2024. *Artificial Intelligence for the Methane Cycle*, DOE/SC-0213. U.S. Department of Energy Office of Science. https://doi.org/10.2172/2204972.

Ushizima, D., et al. 2021. "Deep Learning for Alzheimer's Disease: Mapping Large-Scale Histological Tau Protein for Neuroimaging Biomarker Validation," *Neuroimage* **248**. DOI:10.1016/j.neuroimage.2021.118790.

Vanni, C., et al. 2022. "Unifying the Known and Unknown Microbial Coding Sequence Space," *eLife* **11**, e67667. DOI:10.7554/eLife.67667.

Vaswani, A., et al. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems Vol. 30* . Eds. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc. DOI:10.5555/3294996.

Wang, T., et al. 2021. "MOGONET Integrates Multi-Omics Data Using Graph Convolutional Networks Allowing Patient Classification and Biomarker Identification," *Nature Communications* **12**, 3445. DOI:10.1038/s41467-021-23774-w.

Wang, Y. D., et al. 2021. "Deep Learning in Pore Scale Imaging and Modeling," *Earth-Science Reviews* **215**, 103555. DOI:10.1016/j.earscirev.2021.103555.

Wilkinson, M. D., et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* **3**, 160018. DOI:10.1038/sdata.2016.18.

Wilkinson, M. D., et al. 2019. "Evaluating FAIR Maturity Through a Scalable, Automated, Community-Governed Framework," *Scientific Data* **6**, 174. DOI:10.1038/s41597-019-0184-5.

Wu, P., et al. 2025. "Microbial Synthesis of Branched-Chain β,γ-diols from Amino Acid Metabolism," *Nature Communications* **16**, 4568. DOI:10.1038/s41467-025-59753-8.

Xiang, W., et al. 2024. "FAPM: Functional Annotation of Proteins using Multimodal Models beyond Structural Modeling," *Bioinformatics* **40**(12). DOI:10.1093/bioinformatics/btae680.

Xing, S., et al. 2023. "BUDDY: Molecular Discovery via Bottom-Up MS/MS Interrogation," *Nature Methods* **20**, 881–90. DOI:10.1038/s41592-023-01850-x.

Yang, D., et al. 2023. "Reinforcement Causal Structure Learning on Order Graph," *Proceedings of the AAAI Conference on Artificial Intelligence* **37**(9), 10737–44. DOI:10.1609/aaai.v37i9.26274.

Yang, K. D., et al. 2021. "Multi-Domain Translation between Single-Cell Imaging and Sequencing Data using Autoencoders," *Nature Communications* **12**, 31. DOI:10.1038/s41467-020-20249-2.

Yetgin, A. 2025. "Revolutionizing Multi-omics Analysis with Artificial Intelligence and Data Processing," *Quantitative Biology* **13**(3), e70002. DOI:10.1002/qub2.70002.

Yoon, J. H., et al. 2024. "Integrative Approach of Omics and Imaging Data to Discover New Insights for Understanding Brain Diseases," *Brain Communications* **6**(4). DOI:10.1093/braincomms/fcae265.

Yu, S., et al. 2024. "Two-Step Hyperparameter Optimization Method: Accelerating Hyperparameter Search by Using a Fraction of a Training Dataset," *Artificial Intelligence for the Earth Systems* **3**(1). DOI:10.1175/AIES-D-23-0013.1.

Zengler, K., et al. 2019. "EcoFABs: Advancing Microbiome Science Through Standardized Fabricated Ecosystems," *Nature Methods* **16**, 567–71. DOI:10.1038/s41592-019-0465-0.

Zhalnina, K., et al. 2018. "Need for Laboratory Ecosystems to Unravel the Structures and Functions of Soil Microbial Communities Mediated by Chemistry," *mBio* **9**(4). DOI:10.1128/mbio.01175-18.

Zhang, Y., et al. 2025. "Exploring the Role of Large Language Models in the Scientific Method: From Hypothesis to Discovery," *NPJ Artificial Intelligence* **1**, 14. DOI:10.1038/s44387-025-00019-5.

Zhang, Z., et al. 2024. "Integrating High-Throughput Phenotyping and Genome-Wide Association Studies for Enhanced Drought Resistance and Yield Production in Wheat," *New Phytologist* **243**(5), 1758–75. DOI:10.1111/nph.19942.

Zheng, H., et al. 2025. "Learning from Models Beyond Fine-Tuning," *Nature Machine Intelligence* **7**, 6–17. DOI:10.1038/s42256-024-00961-0.

Zhou, Z., et al. 2024. "Enhancing Efficiency of Protein Language Models with Minimal Wet-Lab Data through Few-Shot Learning," *Nature Communications* **15**, 5566. DOI:10.1038/s41467-024-49798-6.

Ziaei, N., et al. 2024. "A Bayesian Gaussian Process-Based Latent Discriminative Generative Decoder (LDGD) Model for High-Dimensional Data," *IEEE Access* **12**, 113314–35. DOI:10.1109/ACCESS.2024.3443646.

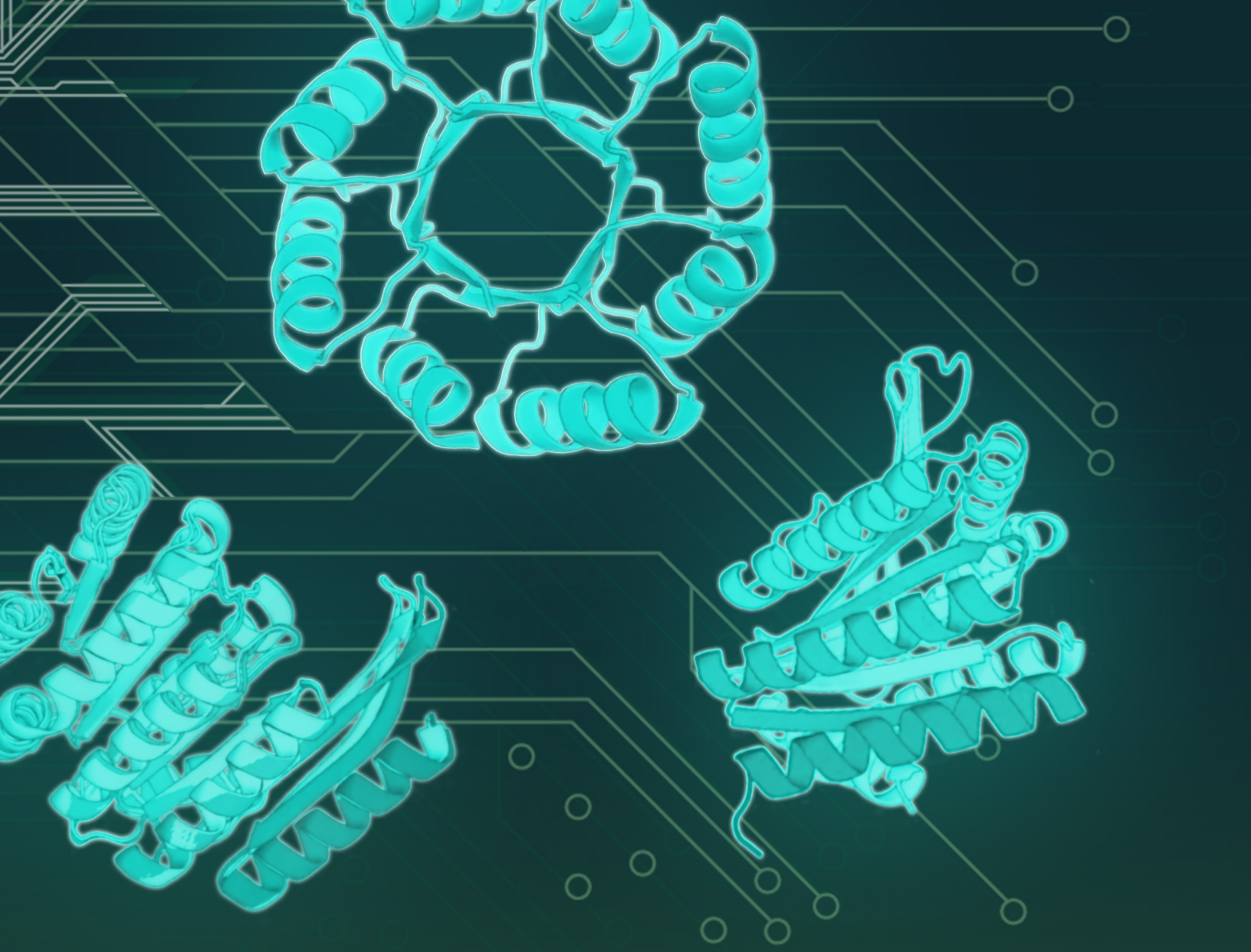
Zvyagin, M., et al. 2023. "GenSLMs: Genome-Scale Language Models Reveal SARS-Cov-2 Evolutionary Dynamics," *The International Journal of High Performance Computing Applications* **37**(6). DOI: 10.1177/10943420231201154.

Zwart, P. H., et al. 2025. "Hyperspectral Segmentation of Plants in Fabricated Ecosystems," *Frontiers in High Performance Computing* **3**. DOI:10.3389/fhpcp.2025.1547340.

Appendix E

Acronyms and Abbreviations

AI	artificial intelligence
ALCF	Argonne Leadership Computing Facility
API	application programming interface
ASCR	Advanced Scientific Computing Research program
AUPR	area under the precision-recall curve
AUROC	area under the receiver operating characteristic curve
BER	Biological and Environmental Research program
CASP	critical assessment of structure prediction
DOE	U.S. Department of Energy
EMSL	Environmental Molecular Sciences Laboratory
FAIR	findable, accessible, interoperable, reusable
FAPM	Functional Annotation of Proteins using Multimodal models
GPU	graphics processing unit
HPC	high-performance computing
JGI	DOE Joint Genome Institute
KBase	DOE Systems Biology Knowledgebase
LLM	large language model
ML	machine learning
ModEx	model-observation-experiment
NERSC	National Energy Research Scientific Computing Center
NMDC	National Microbiome Data Collaborative
NVBL	National Virtual Biotechnology Laboratory
OLCF	Oak Ridge Leadership Computing Facility
PDB	Protein Data Bank
PDE	partial differential equations
PLM	protein language model
PRD	priority research direction
RL	reinforcement learning
SC	U.S. Department of Energy Office of Science
SHAP	SHapley Additive exPlanations
UMAP	uniform manifold approximation and projection
V&V	verification and validation
ViT	vision transformer



U.S. DEPARTMENT of ENERGY | Office of Science