DOE Office of Science Workshop on

Envisioning Frontiers in AI and Computing for Biological Research

February 4–6, 2025

U.S. DEPARTMENT of ENERGY | Office of Science

PREPARED FOR THE U.S. DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Advanced Scientific Computing Research

Biological and Environmental Research

Transforming Bioscience with Innovative AI/ML

In February of 2025 a joint ASCR/BER workshop was held to identify key transformational research directions for understanding biology using artificial intelligence (AI), digital twins and high-performance (HPC) computational methods to facilitate scientific discovery and innovation in support of the Department of Energy mission. AI technologies offer exciting new groundbreaking methods to analyze large volumes of complex biological data, thereby greatly accelerating the ability to understand, predict, and design biological processes for beneficial purposes. In the laboratory, the bridging of AI-enabled automated experimental technologies, HPC and digital twins will provide potent tools for researchers to explore the fundamental nature of biology and harness its inherent metabolic potential for a variety of beneficial purposes. The focus of this workshop was on how high-performance computational methods can impact this objective by exploring digital twins, foundational models, and data-driven approaches with applications to advance automated laboratory experiments, modeling of complex living systems and engineering new functions into plants and microbial systems relevant to DOE mission. Workshop attendees with expertise in plant science, microbiology, mathematics, computer science, and AI assessed the current state of the science, trends, and AI challenges at the interface of plant and microbial systems biology and computational science to identify opportunities for high-impact research. This collaborative effort capitalized on ASCR's advancements in applied mathematics, computer science, and Exascale systems, and BER's expertise in basic genomics-enabled research on DOE relevant plant and microbial systems. The workshop culminated in four key priority research directions to guide future research and development within DOE Office of Science programs.

Priority Research Directions

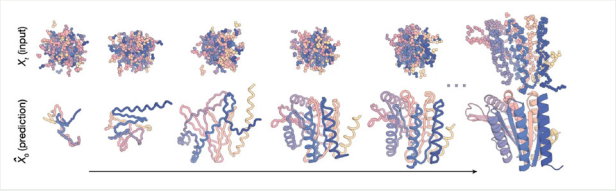
1. Advance novel computational approaches for data fusion to assemble multimodal data from disparate sources and link biological processes from molecules to the functional traits of organisms.

Key question: *What computational approaches can be developed to fuse complex biological data to enable discovery of new biological behaviors, mechanisms, and design principles, while also addressing data interoperability, noise, standardization, and quantification of uncertainty?*

Challenge: Biological data is often sparse, noisy, uncertain, and lacking in standardization. Also, our knowledge of molecular mechanisms and fundamental principles is incomplete. Limitations in measurement strategies across biological scales from molecular to organismal scales to improve this situation, even in model systems, is challenging. To address these challenges, we need (1) new integrative experimental, computational, and theoretical strategies informed by advances in AI, (2) reasoning models for data fusion, and (3) optimal experimental design strategies and verification approaches. These models require advances in scalability using HPC, automated labs and digital twins that enable reasoning models with experimental feedback to accelerate the generation of multiscale biological datasets.

Impact: By seeking ways to integrate diverse data streams and using advanced computational methods, we gain insights into the broader interplay of biological processes, enabling advancements in the understanding and design of biological systems.

SUCCESS STORY



In 2024, Dr. David Baker, working out of the University of Washington, was awarded the Nobel Prize in Chemistry (with two other colleagues) 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. This landmark achievement exemplifies how AI-driven approaches using DOE HPC systems, has advanced computational protein design and protein structure prediction leading to innovative applications in biotechnology, biomanufacturing, energy, agriculture, and medicine.

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.

Key questions: *What new scalable mathematical and computational approaches are required to bridge genome-based models with ecosystem-scale studies, ensuring consistency across scales while using multimodal data? How can AI-driven multiscale modeling be integrated with lab and field systems to enhance the accuracy, interpretability, and generalizability of biological simulations?*

Challenge: A core challenge in building mathematically consistent models of biological systems is the inherent dimensionality, including sparse, incomplete, and noisy observations and the coupling of spatial and temporal-scales. AI models are poised to bridge genome-based models and ecosystem-scale studies, however challenges in adapting DOE's HPC systems and algorithms for AI still need to be addressed. Specifically, AI-driven scale-bridging approaches are needed that can combine the strengths of Exascale platforms, reasoning and causal learning, and experimental design to address knowledge gaps. This requires advances in hardware/software co-design, integration of novel AI workflows, simulations, and experiments.

Impact: Mathematically grounded predictive models will transform our ability to understand and control biological processes across scales, enabling precise simulations and targeted interventions in strain engineering and ecosystem management.

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.

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 advance design and optimization of biosystems and bioprocesses?*

Challenge: Biological systems are a reservoir of vast metabolic potential to make virtually any biomolecule and biomaterial using proteins, pathways, and (multi)cellular processes. Unfortunately, there are vast knowledge gaps for understanding complex biological functions of living systems, their interactions across spatiotemporal scales and within environmental contexts. Also, the current lack of key datasets and the massive protein, pathway, and bioprocess design space, impedes the development of novel biosystems design solutions. Autonomous labs must be developed to efficiently navigate the biosystems design space. AI/ML tools (including domain-informed and physics-based models), digital twins, and scalable approaches for uncertainty quantification, must be developed and implemented to understand and take advantage of the enormous potential afforded within the biosystems design space.

Impact/Outcome: Integrating AI-driven digital twins and autonomous experimentation into biosystems design can enhance our ability to model and design biological processes, leading to accelerated discoveries in biotechnology. These capabilities will greatly enable efficient manipulation and design of biological systems.

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

Key question: *How can novel AI/ML approaches be integrated with genomic research to discover and manipulate molecular mechanisms in plants, microbes, and microbial communities?*

Challenge: The challenge lies in deciphering the intricate relationships between genes, genomes, proteins, metabolites, and observable traits in diverse species and their communities across scales. This requires the development of new AI paradigms, including foundational models, that can design and be directly integrated with experiments to overcome high dimensionality and sparsity of 'omics data. Furthermore, the development of scalable, HPC-ready strategies with uncertainty quantification interfaced with digital twins can help guide experimentation to infer nuanced relationships of interest.

Impact: Advances in AI driven multi-omics integration for multi-scale predictive modeling of genotype-phenotype-environment relationships will provide insights into how genes, proteins, and metabolites govern key emergent processes in biological systems of interest.

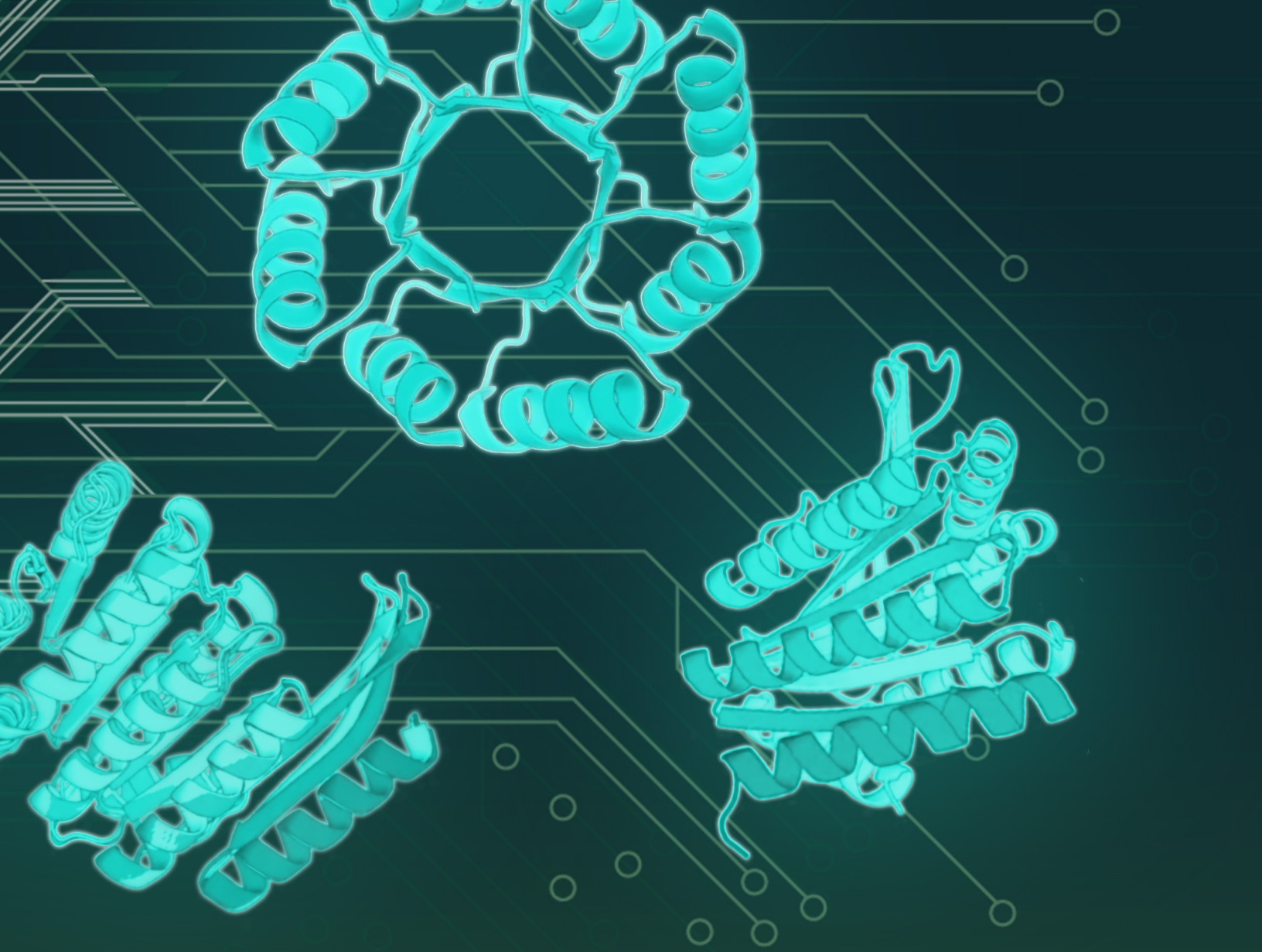


Image courtesy of Argonne National Laboratory

U.S. DEPARTMENT of ENERGY | Office of Science