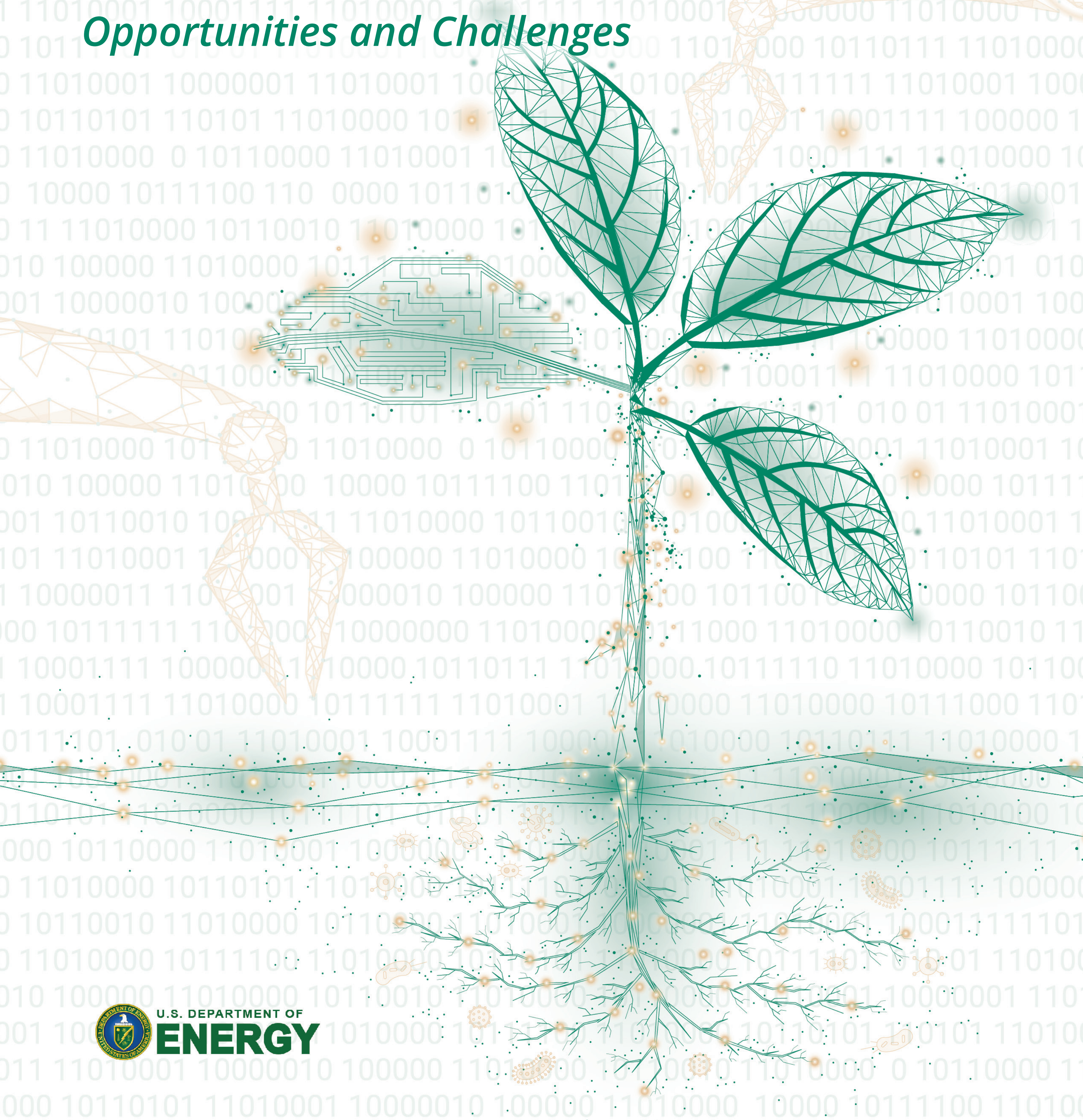
DOE/SC-0211

Artificial Intelligence and Machine Learning for Bioenergy Research

Opportunities and Challenges



U.S. DEPARTMENT OF ENERGY

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

August 23–25, 2022

Convened by
U.S. Department of Energy

Office of Science, Biological and Environmental Research Program and
Office of Energy Efficiency and Renewable Energy, Bioenergy Technologies Office

Organizing Committee

Huimin Zhao
(Chair)
University of Illinois,
Urbana-Champaign

Nathan Hillson
(co-Chair)
Lawrence Berkeley
National Laboratory

Kerstin Kleese van Dam
(co-Chair)
Brookhaven
National Laboratory

Deepti Tanjore
(co-Chair)
Lawrence Berkeley
National Laboratory

Resham Kulkarni
DOE Office of Science

R. Todd Anderson
DOE Office of Science

Jay Fitzgerald
DOE Office of Energy
Efficiency and
Renewable Energy

Gayle Bentley
DOE Office of
Energy Efficiency and
Renewable Energy

Wayne Kontur
DOE Office of Science

Ramana Madupu
DOE Office of Science

Pablo Rabinowicz
DOE Office of Science

About BER

The Biological and Environmental Research (BER) program supports transformative science and scientific user facilities examining complex biological, Earth, and environmental systems for clean energy and climate innovation. BER research seeks to understand the fundamental biological, biogeochemical, and physical principles needed to predict a continuum of processes occurring across scales, from molecules and genomes at the smallest scales to environmental and Earth system change at the largest scales. This research—conducted at universities, U.S. Department of Energy national laboratories, and research institutions across the country—is contributing to a future of reliable, resilient energy sources and evidence-based climate solutions.

About BETO

The Bioenergy Technologies Office (BETO) focuses on developing technologies that convert domestic lignocellulosic biomass (e.g., agricultural residues, forestry residues, dedicated energy crops) and waste resources (e.g., municipal solid wastes, animal manure, biosolids, plastic waste, biogas) into affordable biofuels and bioproducts that significantly reduce carbon emissions on a life-cycle basis (minimum of 70% decrease in greenhouse gases) as compared to equivalent petroleum-based products. These bioenergy technologies can enable a transition to a clean energy economy, create high-quality jobs, and support rural economies. Key to these activities is a focus on process techno-economics and life-cycle emissions, ensuring development of economically viable and environmentally friendly technologies.

This report is available at genomicscience.energy.gov/amber-ai-ml/

Suggested citation: U.S. DOE. 2023. *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.

DOE/SC-0211

Artificial Intelligence and Machine Learning for Bioenergy Research

Opportunities and Challenges

April 2023

U.S. DEPARTMENT OF
ENERGY

Contents

Executive Summary

The integration of artificial intelligence and machine learning (AI/ML) with automated experimentation, genomics, biosystems design, and bioprocessing technologies is poised to revolutionize scientific investigation and, particularly, bioenergy research. To identify the opportunities and challenges in this emerging research area, the U.S. Department of Energy's (DOE) Biological and Environmental Research program (BER) and Bioenergy Technologies Office (BETO) held a joint virtual workshop on AI/ML for Bioenergy Research (AMBER) on August 23–25, 2022 (see Appendix A: DOE Charge, p. 33). These interests have since been amplified in a September 2022 Executive Order, "*Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure U.S. Bioeconomy*," to promote a whole-of-government approach to biotechnology development (White House 2022).

Approximately 50 scientists with various backgrounds and expertise from academia, industry, and DOE national laboratories met to discuss the opportunities and challenges of AI/ML for bioenergy research. Workshop participants were tasked with assessing the potential for AI/ML and laboratory automation to advance biological understanding and engineering in general. They particularly examined how integrating AI/ML tools with laboratory automation could accelerate biosystems design and optimize biomanufacturing. Discussions included the data and computational infrastructure needed to augment biosystems design applications and the expertise and workforce development efforts urgently required to shift integrated systems toward bioenergy research more broadly.

Participants discussed many existing and future applications of AI/ML for biosystems design ranging from enzymes to plants and microbes, microbiomes, and bioprocess development. They also identified three key categories of scientific and technical opportunities and challenges: high-quality data, AI/ML algorithms, and laboratory automation.

Several main takeaways emerged from the workshop:

1. Numerous AI/ML and automated experimentation applications exist for a variety of DOE mission needs in energy and the environment.
2. Exemplary research grand challenges for which AI/ML could provide solutions include: building microbes and microbial communities to specifications, developing closed-loop autonomous design and control for biosystems design, and advancing scale-up and automation.
3. Lack of sufficient high-quality, annotated data hinders the development of AI/ML applications.
4. New and improved AI/ML tools are needed, particularly those meeting the specific needs of the BER and BETO research communities.
5. Trade-offs in performance, cost, and reliability exist between deploying commercially available versus building custom-developed instrumentation and software for automated or autonomous experimentation; translation of manual to automated or autonomous methods is often a nontrivial endeavor.
6. Training a new generation of young scientists who can develop and apply AI/ML tools is needed to solve long-standing scientific challenges in bioenergy research.

The integration of AI/ML tools and automated experimentation represents a new data-driven research paradigm complementary to the traditional hypothesis-driven research paradigm. This paradigm accelerates design and optimization of biological systems and processes for a variety of DOE mission needs in energy and the environment. The AMBER workshop broadly explored the potential of this new paradigm for bioenergy research, of particular interest to BER and BETO, and identified key challenges and opportunities that DOE can address in the coming years by leveraging its unique capabilities and resources.

1. Artificial Intelligence and Machine Learning Needs in Bioenergy Research

Thanks to recent advances in data science, synthetic biology, and laboratory automation, interest is growing in developing artificial intelligence (AI), machine learning (ML), and autonomous experimentation for broader genomics-based research and biotechnology applications. To explore the potential of AI/ML and automation in a bioenergy research paradigm, BER and BETO jointly organized the AI/ML for Bioenergy Research (AMBER) virtual workshop (see Appendix B: Workshop Agenda, p. 35). The meeting included four breakout sessions addressing a broad range of topics including microbiomes, plant-microbe interactions, bioprocess engineering, infrastructure for data and computing, outreach, and workforce development (see Appendix C: Breakout Session Assignments, p. 38).

The breakout groups reported similar AI/ML needs in their individual application spaces that fall under three pillars: high quality data, AI/ML algorithms, and laboratory automation (see Fig. 1.1, p. 2). The groups also identified two characteristics necessary for DOE bioenergy projects to succeed: transferability and human centricity. Addressing these needs and characteristics can help achieve the modeling and engineering of complex biological systems in specific application spaces in the bioenergy research paradigm. Application spaces comprise end-to-end pipelines at BER and BETO, from gene target identification and protein function prediction to scale-up science and distributed biomanufacturing.

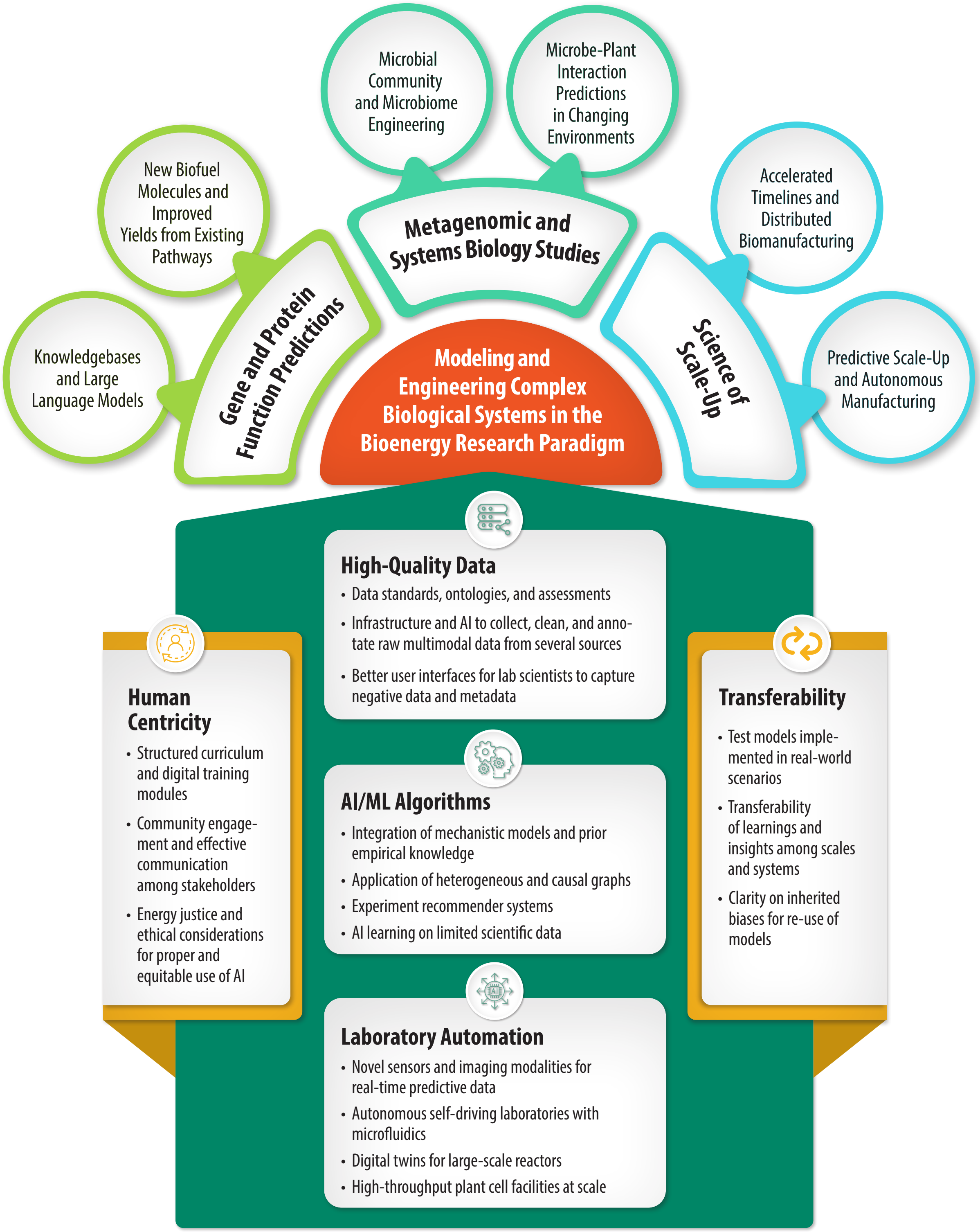
The needs identified in this report are specifically designed to address technical hurdles in the bioenergy research paradigm. For example, much of the automation and real-time bioreactor monitoring tools used for biofuel and bioproduct process development were originally designed for the pharmaceutical industry, which prioritizes time to market rather than titer, rates, and yield. Repurposing automation and computational tools from other industries may save development costs, but workshop participants (see Appendix D, p. 44) emphasized the need to identify inherent biases that accompany such tools.

Participants focused on bioenergy-specific topics such as "science of scale-up" for bioproduction to identify several needs specific to biosystems design and process development:

- Reducing risks in large-scale studies by developing transfer functions from lab-scale studies to substantially accelerate bioprocess development timelines.
- Developing autonomous bioprocessing in reactors to accelerate biofuel research at scale and manufacture vital (or critical) bioproducts through pandemics, during space travel, or on other planets (Berliner et al. 2022).
- Predicting gene and protein function to improve current strain engineering methods for biofuel and bioproduct production and populating large language models that substantially compress timelines in strain development.
- Designing AI/ML-enabled metagenomics and systems biology studies to help predict plant-microbe interactions on a warming planet and engineer soil microbial communities necessary to counter the impacts of climate change on crop yield.

Addressing these AI/ML needs will substantially improve the chances of delivering on BER and BETO strategic goals.

Workshop participants also identified AI/ML needs in the end-to-end process pipeline. To achieve distributed biomanufacturing, biorefineries should be equipped with computational tools that can continuously optimize processes based on upstream feedstock attributes, which can substantially impact downstream fermentation and separation yields. Petroleum refineries have long utilized nonlinear modeling to tune


Microbial Community and Microbiome Engineering
Microbe-Plant Interaction Predictions in Changing Environments
New Biofuel Molecules and Improved Yields from Existing Pathways
Accelerated Timelines and Distributed Biomanufacturing
Knowledgebases and Large Language Models
Predictive Scale-Up and Autonomous Manufacturing
Metagenomic and Systems Biology Studies
Gene and Protein Function Predictions
Science of Scale-Up
Modeling and Engineering Complex Biological Systems in the Bioenergy Research Paradigm

High-Quality Data

- Data standards, ontologies, and assessments
- Infrastructure and AI to collect, clean, and annotate raw multimodal data from several sources
- Better user interfaces for lab scientists to capture negative data and metadata

AI/ML Algorithms

- Integration of mechanistic models and prior empirical knowledge
- Application of heterogeneous and causal graphs
- Experiment recommender systems
- AI learning on limited scientific data

Laboratory Automation

- Novel sensors and imaging modalities for real-time predictive data
- Autonomous self-driving laboratories with microfluidics
- Digital twins for large-scale reactors
- High-throughput plant cell facilities at scale

Human Centricity

- Structured curriculum and digital training modules
- Community engagement and effective communication among stakeholders
- Energy justice and ethical considerations for proper and equitable use of AI

Transferability

- Test models implemented in real-world scenarios
- Transferability of learnings and insights among scales and systems
- Clarity on inherited biases for re-use of models

Fig. 1.1. Modeling and Engineering Complex Biological Systems in the Bioenergy Research Paradigm. Numerous outcomes (circles at top) can be realized by pursuing fundamental and applied artificial intelligence and machine learning (AI/ML) research and tool development specific to the bioenergy research paradigm, including high-quality data, AI/ML algorithms, and laboratory automation (green box at center). Successful projects will include transferability and human centricity features (yellow left and right boxes) which are fundamental to disruptive changes in the bioenergy field.

process conditions and fully convert each batch of crude oil. However, the practice produces yield variances in the pre-established suite of products (Hsu and Robinson 2006; Hu et al. 2002). In the biorefinery space, downstream unit operations should be chosen through AI/ML simulations based on prior knowledge to minimize product yield losses. DOE researchers should also explore next-generation feedstocks, including municipal wastes and C1 compounds like carbon monoxide, methane, carbon dioxide, and others.

Finally, participants identified researcher engagement with the community, especially through dynamic spokespeople, as a necessary endeavor. A large-scale biofoundry providing not only data for researchers to analyze and publish but also access to the community could democratize innovation in this field. Investment in large-scale facilities (e.g., high-throughput plant transformation and biomanufacturing facilities) can generate the data necessary to obtain high-fidelity AI/ML models. By ensuring that these models are openly available to transfer to industry in real-world scenarios and for training and other purposes, DOE can fundamentally impact bioenergy production and catalyze commercialization and knowledge sharing in this paradigm.

2. DOE’s Role in Advancing AI/ML

Numerous DOE workshops and reports have explored how AI/ML can advance science, with specific focus on where science can leverage industry and what science-specific needs the community must address. Another key question is whether DOE can and should play a specific role in advancing AI/ML given its unique capabilities. These reports (e.g., DOE 2020a,b; DOE 2022a,b) have identified a clear DOE niche in the AI/ML research space: integration of prior scientific knowledge into AI/ML solutions for problems at scale.

Knowledge integration should include not only data but also fundamental chemical, physical, and biological principles. These principles are key to achieving high-quality results, but industry has shown little interest in incorporating them into AI/ML solutions.

DOE can leverage its computational, experimental, and observational facilities to create large-scale scientific data collections that train AI/ML models for scientific discovery and extract underpinning scientific principles. During the COVID-19 pandemic, for example, DOE’s National Virtual Biotechnology Laboratory (NVBL) project developed AI/ML tools to screen for potential COVID-19 treatment compounds at a scale not achievable by industry (DOE 2021b). The project combined the world’s fastest computers with computational modeling, novel AI/ML models, and fundamental scientific knowledge.

2.1 Foundation Models for Complex Tasks

Foundation models are a recent AI/ML trend for addressing complex tasks. They are the Swiss Army knives of the AI/ML world and can self-train on extremely large-scale data minimized, or “tokenized,” to key characteristics (e.g., text, code, DNA, RNA, proteins, protocols, graphs, images tokenized as patches, waveforms tokenized as samples, robotic control sequences, and time-dependent data). The tokenized characteristics are then lightly customized and used by a single AI/ML model to tackle diverse tasks.

Foundation models can produce results at scale. For example, the Generative Pre-Trained Transformer 3 (GPT-3), an autoregressive language model with 175 billion parameters compared to GPT-2’s 1.5 billion (Brown et al. 2020), permits in-context learning. The model can be adapted to a downstream task simply by providing a prompt (i.e., a natural language description of the task)—an unanticipated emergent property for which GPT-3 was not trained. DeepMind’s Gato is another example of such a foundation model (Reed et al. 2022). It can perform over 600 multimodal complex tasks including engaging in a dialogue, playing video games, and controlling a robotic arm to stack blocks.

Many discovery processes in biology and life science research could be accelerated and enhanced with foundation models, such as complex autonomous experiments at scale that include sample preparation, design, and execution of broad field studies. Foundation models in this area could support tasks such as knowledge distillation from literature and tailor-made generation of sequences (e.g., nucleic acids, proteins, viruses, and microbes), small molecules, and research protocols.

DOE is well-positioned to develop foundation models for science due to its access to extremely large datasets; deep scientific knowledge and computing capabilities to train models; and expertise at creating large, multidisciplinary, mission-oriented teams. In accordance with DOE’s mission, this work could fuel scientific discovery in the research community and innovation in industry.

2.2 AI/ML-Based Surrogates for High-Performance Computing

Another recent trend in AI/ML for science is the development of AI/ML-based surrogates for high-performance computing (HPC). That is, replacing or augmenting computing-intensive kernels in HPC applications with machine-learned functions that compute the same function much faster.

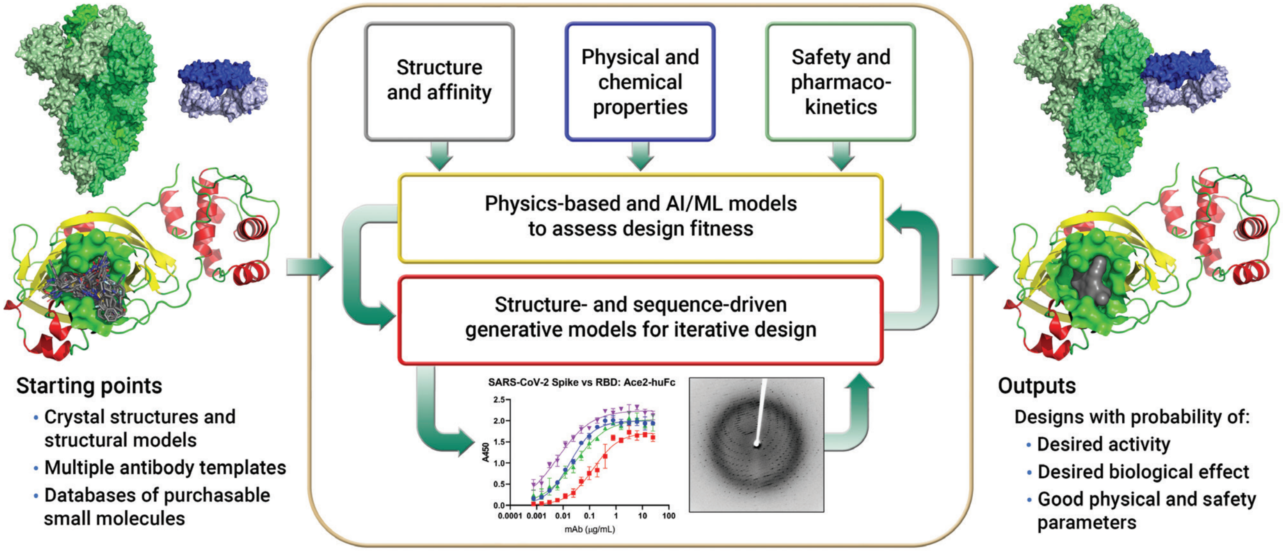

Structure and affinity
Physical and chemical properties
Safety and pharmaco-kinetics
Physics-based and AI/ML models to assess design fitness
Structure- and sequence-driven generative models for iterative design
SARS-CoV-2 Spike vs RBD: Ace2-huFc
2.5
2.0
1.5
1.0
0.5
0.0
A450
0.0001
0.001
0.01
0.1
1
10
100
mAb (µg/mL)
Starting points
• Crystal structures and structural models
• Multiple antibody templates
• Databases of purchasable small molecules
Outputs
Designs with probability of:
• Desired activity
• Desired biological effect
• Good physical and safety parameters

Fig. 2.1. Artificial Intelligence and Machine Learning Functions Can Speed High-Performance Computing. The National Virtual Biotechnology Laboratory project on molecular therapeutics created an integrated computational and experimental platform for designing COVID-19 therapeutics. [Courtesy Oak Ridge National Laboratory]

The approach has been demonstrated in many problem domains including physics, climate, computational fluid dynamics, molecular dynamics, drug docking, chemistry, density function theory, and others. For example, DeepDriveMD, a deep-learning-driven adaptive molecular simulator for protein folding, benefits from such augmentation, achieving speedups of >1,000 times to >100 million times (Lee et al. 2019). The approach was successfully applied during the NVBL project on molecular therapeutics to scan 100 billion molecules for potential suitability as COVID-19 treatments (Saadi et al. 2020; see Fig. 2.1, this page). Recently, hybrid AI/ML HPC solutions were replaced with end-to-end AI/ML, such as in the protein structure predictor AlphaFold (Jumper et al. 2021), achieving similar accuracy much faster.

2.3 Autonomous Control and Discovery in Experimentation

Finally, AI/ML for science is moving from automated experimentation to autonomous design, control, and discovery. Automated workflows simply complete pre-programmed steps, whereas autonomous experiments use AI/ML to make novel decisions based on experimental goals and real-time discoveries. AI/ML algorithms intelligently select new experiments based on current experimental results, creating a loop that explores scientific problems more quickly and efficiently than a human researcher.

Workshop participants presented several early examples of autonomous experiments relevant to BER and BETO research, primarily in the field of materials design and discovery (see Fig. 2.2, p. 7, and "Materials Discovery" sidebar, p. 8). Much can be learned and leveraged from existing materials design experiences and tools.

2.4 Data Quality and Computing Resources

Two critical components underpin novel AI/ML developments: the availability of large volumes of high-quality, annotated data and suitable computing and storage resources to effectively train and execute DOE-developed AI/ML models. The FAIR standards (go-fair.org/fair-principles/) make DOE data accessible for AI/ML training:

- **Findable:** Data should be findable by humans and computers.

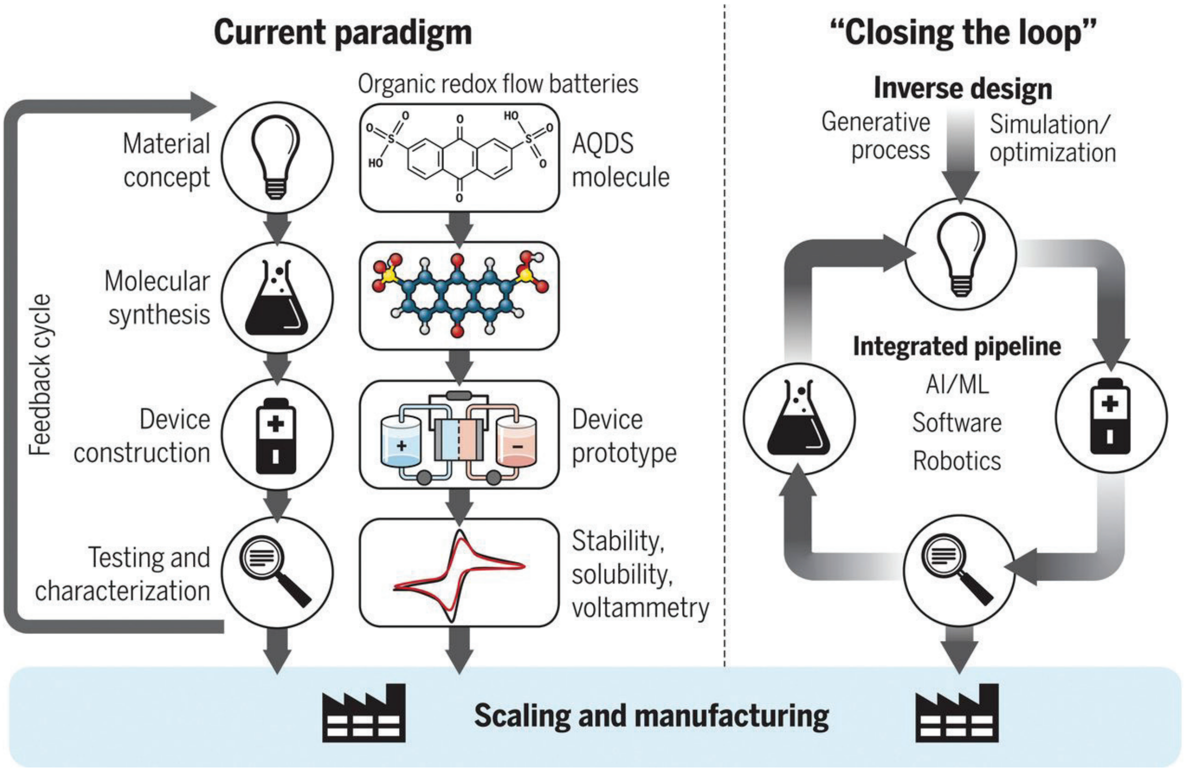


Fig. 2.2. High-Level Paradigm Comparisons for Material and Molecular Sciences. Redox flow batteries (left) exemplify the current paradigm. A closed-loop discovery process (right) utilizes inverse design and a tightly integrated workflow to enable faster identification, scale-up, and manufacturing. [From Sanchez-Lengeling, B., and A. Aspuru-Guzik. 2018. "Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering," ***Science*** **361**(6400), 360–65. Reprinted with permission from AAAS.]

- **Accessible:** Users and computers know how to access and use the data.
- **Interoperable:** Data needs to work with more than one application and workflow for analysis and integration.
- **Reusable:** Optimizing data reuse (the ultimate goal of FAIR) requires describing metadata and data well so they can be replicated and combined in different settings.

In addition to FAIR, data quality and actionability must also be considered.

Data Quality

- **Correctness**: Data collection is not error free; quality checks are needed.
- **Completeness**: Complete data collection may never be achieved, so ensuring data volume and coverage are sufficient for a given task is necessary, along with collecting both positive and negative experimental results and clearly identifying missing data.
- **Bias-Free**: Most data are biased due to the the type and manner collected. Biases must be identified and made explicit if they cannot be removed or corrected, including determining the source of bias, how strong it is, and whether it can be mitigated.

Data Actionability

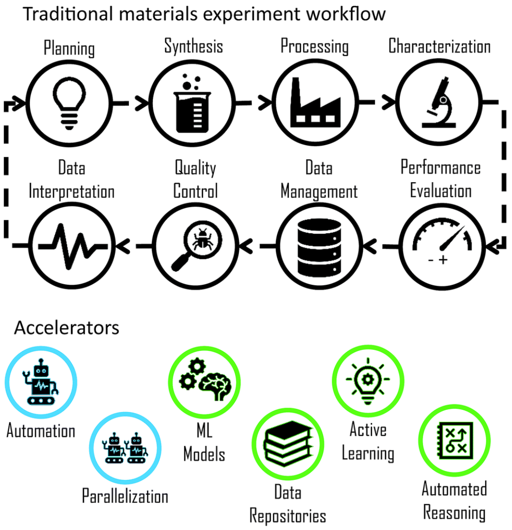
- **Reproducible**: Science is verifiable through reproducibility of results. Therefore, the data used to train AI/ML models and the methods used to create the data must also be reproducible. A key aspect to reproducibility is uncertainty quantification.

Continued on p. 10

Materials Discovery

The general workflow for materials discovery consists of synthesis, processing, characterization, and performance evaluation. These steps are traditionally executed sequentially, but automation and AI/ML methods have sped up the process by performing and evaluating many design loops in parallel while still building on respective outcomes (see figure). A 2019 review (Stein and Gregoire 2019) assessed the quantitative impact of different types of acceleration, such as automation, parallelization, ML models, data repositories, active learning, and automated reasoning, on traditional materials science discovery workflows.

Stein and Gregoire described the workflows used by four different research teams in terms of these components, including the level of automation introduced. To determine the speedup achieved by the automated discovery workflows, they compared the number of experiments that could be conducted in one pass with the number that a traditional experimental workflow would deliver—assumed to be one experiment per pass.



Experimental Materials Science Research Life Cycle. Overview of core research tasks with arrows indicating the cyclic execution of a traditional materials science experimental workflow (top). Acceleration of each task in a workflow can be obtained by incorporating acceleration techniques, as represented by six types of accelerators (bottom). [From Stein, H. S., and J. M. Gregoire. 2019. "Progress and Prospects for Accelerating Materials Science with Automated and Autonomous Workflows," *Chemical Sciences* **10**, 9640–9649. Reprinted under a Creative Commons license (CC BY 3.0).]

In Example A (see figure, p. 9), researchers, researchers optimized growth conditions for carbon nanotubes, achieving a speedup of 100 times. Learning was used to analyze prior experimental results and propose new experiments by automated robotics to optimize material combinations and growth conditions. The same methods could be applied, for example, to an autonomous bioreactor, the synthesis of biological samples, or growth conditions in a greenhouse or laboratory.

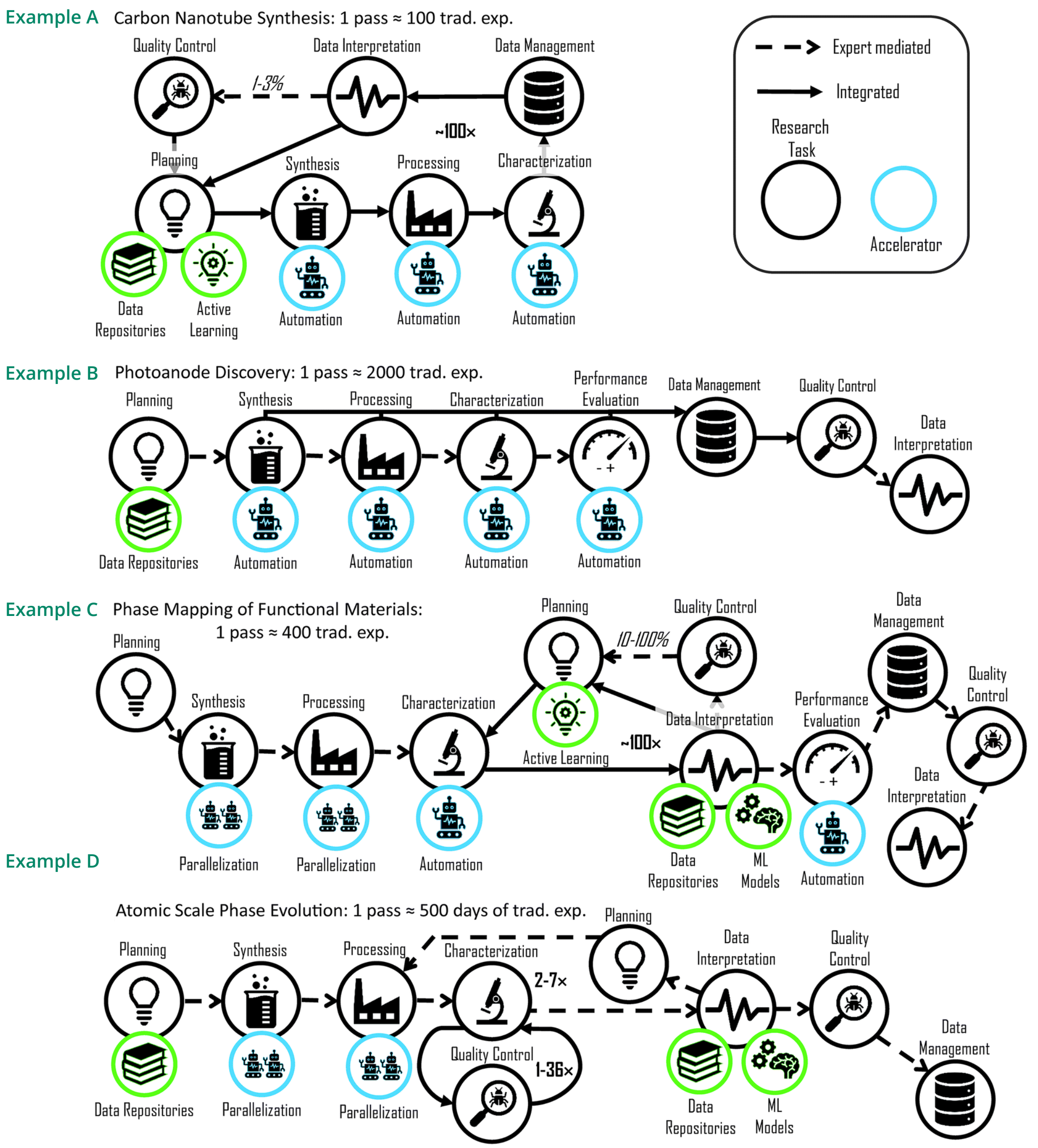
Example B represents a combinatorial exploration of research space like the NVBL molecular therapeutics project described in Fig. 2.1, p. 6. The team used large-scale automation to simultaneously operate on libraries of up to 2,000 samples, achieving a speedup of 2,000 times. To meaningfully design, steer, and evaluate experiments, the team selected high-value targets using computational screening of candidates. Experts determined the needed growth conditions. Results of the entire pipeline were captured, analyzed, and used to inform future experiments.

Example C describes a combinatorial research workflow (similar to example B) that achieved a speedup of 400 times using a combination of automation, parallelization, and expert-driven integration. Researchers added active learning to accelerate decisions on the best candidates to advance to the next step. Final characterization was further accelerated using real-time analysis and autonomous selection of the next best sample to screen.

Example D examined sample evolution. Instead of using a single bulk experiment, researchers used several smaller specialized experiments in parallel to evaluate sample stability and progression, replacing one large reactor with 36 custom nanometer-sized reactors. Key improvements in the autonomous workflow occurred in real-time monitoring and quality control. Results were compared to external sources using AI models. The team achieved a speedup of 500 times.

Continued on p. 9

Continued from p. 8



Workflow Diagrams of Accelerated Materials Experimentation Spanning a Range of Techniques, Strategies, and Research Goals. Various speedups in the discovery pipeline can be achieved as increasing levels of AI and automation are embedded and more processes become part of the design loop. Clear parallels with systems biology and synthetic biology workflows can be drawn. [From Stein, H. S., and J. M. Gregoire. 2019. "Progress and Prospects for Accelerating Materials Science with Automated and Autonomous Workflows," *Chemical Sciences* **10**, 9640–9649. Reprinted under a Creative Commons license (CC BY 3.0).]

Continued from p. 7

Using data with confidence requires knowing the level of uncertainty inherent in the data.

- **Provenance and Explainability**: The origin of data and results must be available to users in a form that enables assessment of the correctness and suitability of data and AI/ML tools for a task at hand. This information creates trust in the data for future use.
- **Range of Validity**: Metadata or other concepts can be used to clearly state boundaries regarding what purposes data can be used for and where it can or cannot be used (i.e., broad, limited).
- **Distilled**: Data are summarized, analyzed, and packaged for specific use cases.

An example of a FAIR data repository is the collaboration between two DOE projects: the National Microbiome Data Collaborative (NMDC; microbiomedata.org) and Benchmark Datasets and AI/ML Models with Queryable Metadata (ENDURABLE; crd.lbl.gov/divisions/amcr/computer-science-amcr/par/research/endurable/). ENDURABLE is establishing the means to provide AI/ML researchers with access to massive data repositories for developing AI/ML models to solve problems in microbiome science. More specifically, ENDURABLE is storing curated NMDC data and associated metadata and disseminating it for AI/ML research. These efforts, which follow FAIR principles, are making microbiome data useable to the deep learning community and catalyzing the development of AI/ML models for microbiome data science. Additionally, defining AI/ML tasks and their necessary data and metadata breaks down barriers to using AI/ML in microbiome data science and makes AI/ML research more broadly reproducible.

3. Assessing and Supporting Automated and Autonomous Experimentation

Laboratory automation broadly describes the process, or resultant systems, of replacing human operators with computational or robotic equivalents in a laboratory setting. Automation aims to reduce human tedium, repetitive stress, and strain injuries; lower labor and other costs (e.g., reagents through microfluidic or tiny droplet dispensers); increase throughput, reproducibility, and reliability; and extend lab operations toward a 24/7 schedule. Note that automation is not only physical but also relates to information and data processing. The advent of AI has enabled new forms of self-driving or autonomous experiments and laboratories that use AI and ML to define a research path as it progresses based on overall project goals and discoveries made during an experiment (Martin et al. 2023; Beal and Rogers 2020).

Considerable laboratory automation already exists within DOE-supported national laboratories and academic research facilities, including both off-the-shelf commercial and custom systems. When available, commercial systems are generally preferable because they often are more cost-effective, especially considering the exceedingly high costs of developing, supporting, and maintaining custom systems. However, custom systems might be needed when commercial systems are unavailable (e.g., during early phases of technology development or when the market size is too small to justify commercialization) or are insufficiently configurable, extendable, or accessible to meet given business needs.

Perhaps not surprisingly, DOE project workflows in national laboratories and academic institutions are executed across very heterogeneous robotic, instrument, equipment (e.g., mass spectrometry), and software platforms. Such workflows are especially prevalent in low technology readiness level (TRL) research and development environments where custom automation system components are common. Many of these workflows, which often constitute much of DOE's supported capabilities, have at their core nonautomated instrumentation that may be difficult to automate or integrate into automated workflows. Heterogeneity, at least in the absence of physical standards (e.g., labware dimensions) and informatic standards (e.g., data exchange formats), is often required to conduct bespoke research. However, heterogeneous platforms place heavy burdens on efforts to integrate, operate, support, and maintain workflow systems. In many cases, not all workflow operations can be automated, so optimizing collaborative workflow contributions from both humans and automated systems becomes important. These issues also relate to workforce development, which is essential to ensure that developers, operators, and maintainers of these hybrid human and automated systems achieve their performance potentials.

DOE project workflows span an automation gradient from fully manual (i.e., no automation) to semiautomated (i.e., mixtures of interlaced human and robotic/software operations) to fully automated (i.e., no human operations). In some, perhaps increasing, instances, these workflows have become fully self-driving (i.e., beyond full automation and autonomous iterative/cyclical workflows). Each of these automation and autonomy levels has a proper time and place for use with commercial and custom systems. In the context of this workshop, which emphasized (meta) data quantity and quality (e.g., reliability, reproducibility, and comparability), the more automated and autonomous a workflow, the better perhaps for supporting AI/ML-directed DOE science and technology development. However, autonomous workflows may not always be the best approach; cost, performance, and reliability trade-offs need to be evaluated in each case to decide on the best path forward. Future research not only should focus on new components that make workflows more automated or autonomous but also on methods to guarantee their quality, reliability, reproducibility, and explainability.

4. AI/ML Algorithms and Their Current Bioenergy Applications

Workshop participants discussed potential ways in which AI/ML approaches could enhance current applications in bioenergy. This chapter describes four areas underlying these research opportunities: fundamental challenges, process development, foundational AI/ML algorithms, and automated and autonomous experimentation.

4.1 Fundamental Challenges in Synthetic Biology and Biosystems Design

Biosystems design, or synthetic biology, aims to engineer biological systems that have novel or improved functions for basic and applied biological research. Quantitatively and predictively engineering these systems—including enzymes; pathways; and whole genomes of microorganisms, plants, and microbial communities (microbiomes)—is overwhelmingly challenging due to their intricate connectivity and complexity. AI/ML advancements that enable computers to learn automatically from experience have emerged in recent years as potentially powerful tools to address this challenge (Carbonell et al. 2019; Volk et al. 2020). This section highlights four examples of AI/ML-enabled biosystems design and their development challenges to illustrate the status of the field: (1) enzyme engineering, (2) pathway and metabolic engineering, (3) plant engineering, (4) and microbiome engineering. For details about specific AI/ML tools, see Section 4.3: Foundational AI/ML Algorithms for Bioenergy Research, p. 19).

Enzyme Engineering

Enzyme engineering aims to improve enzyme phenotypes desirable for biotechnological, industrial, and scientific applications (Yang et al. 2019). Directed evolution is one of the most widely used and successful tools (Wang et al. 2021). Despite its success, directed evolution is time intensive, labor intensive (Yang et al. 2019), and inefficient because beneficial variants are rare and the possible variant space is enormous (Hie and Yang 2022).

Recently, the research community has increasingly applied AI/ML to facilitate enzyme engineering (Wittmann et al. 2021; Li et al. 2019). Compared to traditional directed evolution, AI/ML-assisted directed evolution can be more efficient in locating beneficial variants with considerably fewer experiments (see Fig. 4.1, p. 14). For example, scientists developed a deep learning framework called ECNet (evolutionary context-integrated neural network) to accurately predict variant fitness (Luo et al. 2021). Additionally, researchers used an ML algorithm called upper confidence bound (UCB) to explore a model's uncertainty region and simultaneously sample the region with high fitness (Greenhalgh et al. 2021). UCB is an iterative process that repeatedly trains the model with experimentally determined variant-fitness data and makes predictions of new variants for follow-on screening. In another example, researchers developed an *in silico* directed evolution workflow based on Markov chain Monte Carlo to engineer green fluorescent protein and TEM-1 beta lactamase (Biswas et al. 2021).

Pathway and Metabolic Engineering

Researchers have effectively used AI/ML to improve the production of fuels, chemicals, and materials in only a few design-build-test-learn (DBTL) iterations and to analyze data to predict new biological interactions or characterize component parts (Volk et al. 2022; see Fig. 4.2, p. 15). For example, in just three rounds, the BioAutomata platform improved lycopene production by 77% compared to random screening (HamediRad et al. 2019). A related platform, the Automated Recommendation Tool (Radivojević et al. 2020), demonstrated improved design predictions

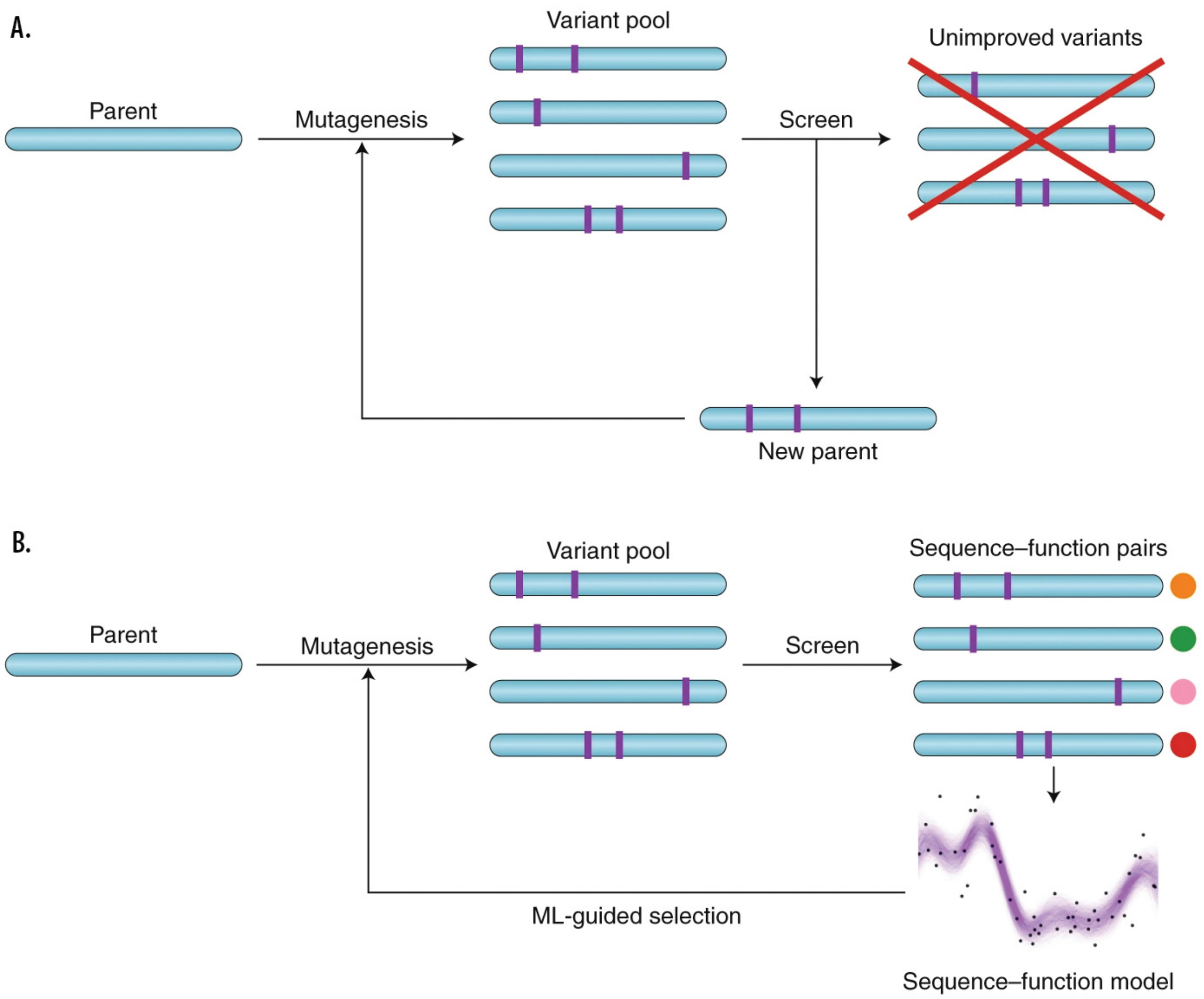


Fig. 4.1. Comparison of Traditional Directed Evolution and Machine Learning (ML)–Assisted Directed Evolution. Traditional directed evolution (A) uses iterative cycles of diversity generation and screening to find improved variants and discard information from unimproved variants. ML methods (B) use the data collected in each round of directed evolution to choose which mutations to test in the next round. Careful choice of which mutations to test decreases the screening burden and improves outcomes. [Reprinted with permission from Springer Nature from Yang, K. K., et al. 2019. "Machine-Learning-Guided Directed Evolution for Protein Engineering," *Nature Methods* **16**, 687–94.]

for fatty acids, and a subsequent study combined genome-scale models with AI/ML to overproduce tryptophan (Zhang et al. 2020). A recent example used sequence information and cell sorting to characterize all promoters in the yeast *Saccharomyces cerevisiae*, creating a model that, in principle, could enable promoter design in an engineered pathway (Vaishnav et al. 2022). Another example linking genotype to phenotype used a set of kinase knockouts to predict the yeast metabolome under different knockout settings (Zelezniak et al. 2018).

Plant Engineering

Crop domestication and traits could be improved by addressing a fundamental challenge in plant biology: understanding how the vast cis-regulatory DNA sequences that surround genes control gene expression. To advance this understanding, various supervised AI/ML models have been trained on good-quality functional genomics data (e.g., chromatin accessibility and transcription factor binding). For example, inspired by recent progress in zero-shot

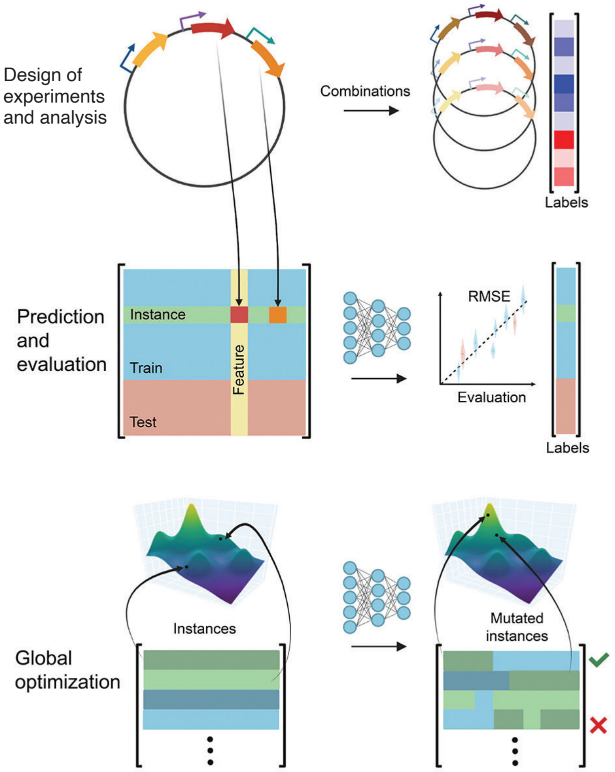


Fig. 4.2. A Standard Workflow that Integrates Machine Learning (ML) with Metabolic Engineering. First, a library of variants is constructed and analyzed by assigning labels to each variant. In this example, labels are assumed to be titers associated with a pathway on the plasmid. Then, pathway data is converted to a data matrix where an ML model is trained to make predictions based on the reserved test data. New high variants predicted to perform well are then recommended for future design. [Reprinted with permission from Volk, M. J., et al. 2022. "Metabolic Engineering: Methodologies and Applications," *Chemical Reviews* (special section). ©2022 American Chemical Society.]

fitness prediction of protein variants from global language models, Benegas et al. (2022) reported the first zero-shot noncoding variant effect predictor trained on the genomic sequence of *Arabidopsis thaliana*. Because this AI/ML model is trained using only one genome in an unsupervised manner, it can be easily transferred to any plant genome for predicting variant-effect fitness and improving crop traits.

Additionally, genome-editing tools, such as the CRISPR-Cas system in plants, have enabled DNA sequence manipulation, helping inform yield improvement and increase stress tolerance. However, relating phenotypic outcomes to genomic features remains a huge challenge. Cheng et al. (2021) developed an evolutionarily informed AI/ML approach to predict nitrogen use efficiency both within and across species. van Dijk et al. (2021) discusses, among other topics, increased efforts in using computer vision for plant phenotyping, implementing ML for plant-pathogen interactions, and identifying metabolic pathways.

Microbiome Engineering

Precise microbiome engineering requires accurately understanding community-level interaction edges between different microbial species. However, experimental discovery of such edges becomes impractical in naturally occurring microbiomes, as pairwise combinations become prohibitively large. Recent designs of AI/ML-based approaches employ different classes of models to predict community interactions and thus community networks. To predict new interaction edges in a microbiome context, these models leverage data (e.g., community interactions) from previous experiments or prior knowledgebases and generally understood features of specific microbial species. For example, DiMucci et al. (2018) developed a random forest model capable of predicting previously unknown pairwise interactions in microbial communities. The study found that for moderately sized communities (i.e., 20 species), training the model on just 5% of experimentally confirmed interactions was sufficient to predict the remaining 95% of interactions with 80% accuracy. Scientists used the model to rank the importance of each feature in the trait-level feature vectors and thus identify the most important traits governing interaction edge presence in a community. These results provided not only interpretability to the model's predictions but also a framework for hypothesis generation of the mechanisms by which organisms interact within a community.

Achieving precision microbiome engineering requires highly refined prior knowledge of both the community interactions and predictive capabilities of community performance on modifying any microbe's function. So far, scientists have largely applied AI/ML techniques to the former category. A prospective application path could involve using AI/ML models to design better microbiome engineering strategies for achieving a desired function. A recent review addresses the role of ML in microbiome-related research (Hernández Medina et al. 2022).

4.2 Process Development

Scale-up process development is often seen as more of an art than a science (Humphrey 1998; Stocks 2013). Lab-scale tests in microtiter plates are subject to capillary effects and therefore do not represent scale-up performance that involves turbulence and heterogeneities. Large-scale reactor studies can be resource intensive, limiting the number of process development studies that scientists can perform. Statistical design-of-experiments approaches can help achieve statistical confidence but only in a limited portion of the vast multiparametric experimental design space. Low representation of experimental space, along with minimal replication, leads to restricted understanding of scale-up performance, thereby increasing risk of failure during commercial production. Scientists and engineers often rely on intuition to counter unanticipated events (Crater and Lievense 2018), and this empirical practice explains the apt perception of scale-up process development as an art. To review this issue tangibly, breakout groups focused on fermentation and particularly discussed the challenges of scaling up fermentation processes in bioreactors, identifying research needs and opportunities related to (1) data availability, (2) imaging and autonomous bioreactors, (3) bioreactor digital twins, (4) downstream processing, and (5) scale-up science.

Data Availability

Data generated from fermentation campaigns are acquired from several sensors and often include inadequately annotated metadata. This lack of annotation leads to poor understanding of all interactions among multiple variables. Consequently, campaign results are largely unactionable. Data availability alone can be a challenge, as few online and real-time measurements are currently accessible with off-the-shelf equipment.

Most researchers use online dissolved oxygen and off-gas concentration measurements to estimate product titers, rates, and yields; they then conduct offline chromatography-based studies to validate these estimates. However, data acquisition rates from chromatography and other studies are very slow. Furthermore, the costs associated with data acquisition from multiple analytical equipment (offline and online) lead to low data volumes insufficient for analysis with a full suite of AI/ML methods. While data can be collected from multiple fermentation campaigns across several scale-up facilities, very few public facilities can offer such data volumes. Most data from fermentation campaigns is located with companies in proprietary forms, thus making data sharing difficult. Ultimately, data acquisition, curation, and sharing in bioprocessing is a challenge that has received very little attention, even though data forms the backbone of AI/ML applications.

Data at the laboratory scale (<10 mL) is more readily available because academic and industrial institutions can now collect data using microfluidics and microtiter plates. However, insights cannot yet be transferred confidently from the lab to bench scale (0.25 L to 10 L) or pilot scale (1,000+ L) for new target-host combinations.

Similar challenges exist with transferability of different production systems. While literature is available for a few processes and production strains (e.g., penicillin production and *Escherichia coli*–based processes), many other industrial processes conducted with noncanonical microorganisms are not widely published. Such data will be necessary to transfer insights from well-studied strains to other systems (e.g., studying bacterial processes to help inform fungal fermentations). Knowledge transferability from lab to pilot scales and among microorganisms is still an ambitious goal that, at present, is primarily impeded by resource limitations in bench- and pilot-scale testing.

Imaging and Autonomous Bioreactors

Fermentation is a critical unit operation in generating biofuels through biological pathways that depend on microbial catalysts to convert sugars and intermediates from agricultural and other waste feedstocks. Due to an explosion of tools in the past decade, synthetic biologists can rapidly engineer microbial hosts and generate several strain variations with improved productivity at the milliliter scale in shake flasks and well plates. However, researchers face challenges in predicting these strains' performance in bioreactors, even at the slightly larger scale of 2 L (Wehrs et al. 2019). As they grow and survive in bioreactors, microbial hosts within the same culture undergo both genotypical and phenotypical changes that often lead to lower productivity due to fitness-related mutations. Cell viability tests, sequencing, and omics are direct indicators of microbial cellular health but are only available *post hoc*. Developing novel real-time monitoring methods is essential for understanding single-cell level changes in culture while in process.

Real-time imaging, such as feature extraction from single cells, is a long-standing approach to assess physiological heterogeneity in mammalian cells (Bevan et al. 2019). Spectroscopy also has the potential to deliver real-time chemical information. Near-infrared spectroscopy at 2 nm resolution can provide information on bond rotation changes within one degree and bond length changes of 0.01Å. Such data can differentiate stereoisomers, a capability that state-of-the-art mass spectrometry measurements cannot provide. Information from high-resolution infrared spectroscopy (near and far) enables AI/ML models to learn signatures of distinct cell states and phenotypes.

Raman measurements, quantum sensors, and other modalities are also of interest in providing novel data that can substantially boost the outcome from AI/ML methods for fermentation processes. Imaging-based modalities can be coupled to large investments in chromatography, mass spectrometry, and transcriptomics, for example, to obtain mechanistic insights into bioprocesses. In principle, inverting AI-derived signatures of cell states should be possible to discover the chemistry that underlies predictive power. Exploring this frontier could enable the passive, nondestructive monitoring of sample biochemistry. Specifically, learning to invert the spectral signature into chemistry, genes, and other relevant information will generate previously unavailable insights. Imaging microbial hosts could substantially enhance understanding of industrial-scale mutations, which could improve scale-up challenges in biofuel and biochemical production processes.

Additionally, new imaging-based datasets will be essential to develop self-driving bioreactors with fully automated process control and minimal human interventions for all modalities (i.e., host-product combinations). Such bioreactors are paramount to removing the bottleneck in bench-scale process development capacity. AI infrastructure is needed to learn from data-driven models, suggest operational perturbations, and accelerate the pace of process optimization.

The performance of AI-based control in predicting extreme events and other foundational problems will lead to exciting new scientific studies and improved bioproduction processes. At the bench scale, tolerance for a failed experiment is high, but the cost of a failed commercial production campaign due to extreme events can be prohibitive for any company or research institution. AI-based methods are very suitable for predicting such events when signal-to-noise regimes are infinite, especially in reinforcement learning settings. However, in biology, signal-to-noise is quite low and onerous, requiring the development of fundamental new paradigms of specific objective functions, including new non-Markovian formulations. Sensor development to minimize noise must go hand-in-hand with algorithm development.

Finally, development is needed for controls that enable nonexpert users to operate bioreactors for process development to maximize titers, rates, and yields of bioenergy molecules. Such efforts can have a far-reaching impact. For example, applications in defense and space travel, where austere environments demand automated production of food, fuel, and medicines via fermentation, will need controls that users can operate without the help of expert process engineers.

Bioreactor Digital Twins

Although the pharmaceutical industry already applies mechanistic modeling for bioreactor studies based on digital twins, or virtual models designed to accurately represent a physical object or process, few such studies exist in the biofuel and biochemical domains.

One reason for this disparity is that large bioreactors (10,000+ L) are required for biofuel production compared to smaller reactors (~1,000 L) used to produce pharmaceutical ingredients. Spatial heterogeneity in process conditions occurs across the height of a large-scale bioreactor due to the water column's weight on bottom layers, which experience higher pressures, higher oxygen and carbon dioxide concentrations, and possibly lower glucose concentrations. Through computational fluid dynamics coupled with metabolic modeling, researchers have shown that heterogeneity in process conditions in large-scale bioreactors impacts microbial cultures and their productivity (Haringa et al. 2018). In many cases worldwide, only global, single-point samples are taken during a process, leading to limited information on local process performance. Novel sensors and data streams from different parts of a large bioreactor can help describe spatial heterogeneity and develop digital twins for simulations that can minimize experimental testing in large bioreactors.

Downstream Processing

Downstream processing (DSP) refers to one or multiple unit operations performed on a fermentation culture after it exits a bioreactor. DSP could involve a one-step centrifugation or filtration process or a multistep serial process that includes cell disruption, extraction, and purification, or evaporation and drying. The state-of-the-art approach to develop a DSP suitable for a particular molecule involves testing several unit operations in series and parallel. DSP development requires large amounts of fermentation broth (i.e., at least tens of liters but more typically hundreds) and, although expensive, is essential for establishing an end-to-end process at large scales prior to commercialization. Researchers can use AI/ML approaches—in combination with chemical, rheological, and other physical properties of the fermentation culture and product—to predict the performance of individual and combinations of DSP unit operations as well as their optimal operating conditions. Such solutions can substantially reduce both the cost and time needed to identify optimal process pathways, a task currently conducted using an empirical, trial-based approach.

Finally, most DSP unit operations used in current biofuel and biochemical production chains were developed for other industries, such as pharmaceuticals and food. Substantial innovation is needed to develop unique separation equipment for biofuels and bioproducts, and AI/ML approaches may help identify previously unconsidered methods for desirable molecules.

Scale-Up Science

To date, lessons learned from the pharmaceutical industry have guided industrial bioreactor scaling and development despite many differences between pharmaceutical and biofuel production. For example, biofuels must be manufactured at massively higher quantities compared to vaccines and other medicines (i.e., millions of gallons versus thousands of kilograms). Also, contaminated biofuels, unlike pharmaceuticals, are salvageable because the batch ultimately can be purified enough to burn in an engine. Additionally, scientists can engineer synthetic and natural microbial communities to generate biofuels, especially communities associated with waste and second-generation feedstocks (e.g., molasses and bagasse; Senne de Oliveira Lino et al.). Finally, while the pharmaceutical industry is typically focused on the filing time for drug approval from the U.S. Food and Drug Administration, the biofuels industry is often working to maximize titers, rates, and yields to attain economic viability.

The biofuels industry needs much higher capacity for bench-scale bioreactor studies and could substantially benefit from sharing specific lessons learned about biofuel processes. However, with no opportunity to publish such process-based knowledge, lessons learned are often shared orally. A central knowledge repository and a self-driving and digital twin approach that enables conducting multiple experiments in a single bioreactor could maximize resources and commercialization prospects. Biofuel-centric sensor and tool development, along with AI/ML applications, can also lead to radical improvements in bioprocessing.

4.3 Foundational AI/ML Algorithms for Bioenergy Research

Some of the many AI/ML tools (i.e., development frameworks and models) broadly used today stem from industry developments. However, the quality of available AI/ML solutions varies widely not only in robustness, reproducibility, and explainability but also in applicability to scientific challenges. Industry tool suites that are often good for general tasks require scientific knowledge integration to produce acceptable results in research settings. For applications in the BER mission space, scientists are using AI/ML for increasingly complex applications, and AI/ML-accelerated data and image analysis is becoming standard in various scenarios across the BER community.

The following sections describe six examples of new, more complex AI/ML opportunities and research needs in bioenergy: (1) matching AI/ML models to problems of interest, (2) merging AI/ML predictive capabilities with mechanistic insight, (3) overcoming the limited data problem, (4) integrating data from various resources, (5) quantifying the predictive capacity of AI/ML models, and (6) developing generally applicable large language models and foundation models. These examples also highlight future challenges that could be addressed with AI/ML approaches.

Matching AI/ML Models to Problems of Interest

Choosing an AI/ML model for the problem of interest depends on multiple factors, such as the nature of labels or output, the number of data points available for training the AI/ML model, and the type of input. Depending on the labels, ML models fall into three primary categories: classification, regression, and clustering. If the labels for training the AI/ML model are not available, clustering can help find similarities between data points. Small datasets typically restrict model choice to traditional ML models (e.g., ridge regression, support vector machines, and random forest). However, larger quantities of data allow for the consideration of deep neural networks. Recently, Greener et al. (2022) developed a guide to ML for biologists (see Fig. 4.3, p. 20).

While AI/ML seems ideal for the scale and complexity of synthetic biology problems, limited data availability is a critical bottleneck to developing bigger and better AI/ML models. A new paradigm in addressing the problem of limited data is manifold learning or, in other words, feature engineering. This approach enables representation of complex, high-dimensional data in low dimensions while capturing problem-specific information and reducing unnecessary noise. This outcome can be achieved using techniques from unsupervised learning such as autoencoders and training a low-parametrized traditional AI/ML model on these low-dimensional data representations.

In the case of highly parameterized AI/ML models, such as deep neural networks, model architecture can influence prediction capabilities. The simplest neural network architecture is multilayer perceptron in which layers of artificial neurons are arranged in a fully connected fashion. Input types require different model architectures. For example, convolutional neural networks (CNNs) can capture local spatial structures and are most often used for image-like data. One major application of CNN is to identify or predict subcellular organization and cell fate using microscopy images.

Graph convolutional networks (GCN) are applied in tasks involving entities connected by defined relationships or interactions. GCNs update node properties in the network by combining predictions from all neighboring nodes. They therefore are better suited for graph-structured biological data, such as molecules (composed of atoms and bonds) and gene–gene interaction networks (composed of genes and interactions). Alternatively, recurrent neural networks like long short-term memory are more suited for sequential biological data, such as time series prediction and protein function or structure prediction.

Language models from natural language processing (NLP) provide another framework to encode sequential data in biology. NLP models, for example, could treat protein sequences as sentences in a foreign language and only make a viable variant or meaningful

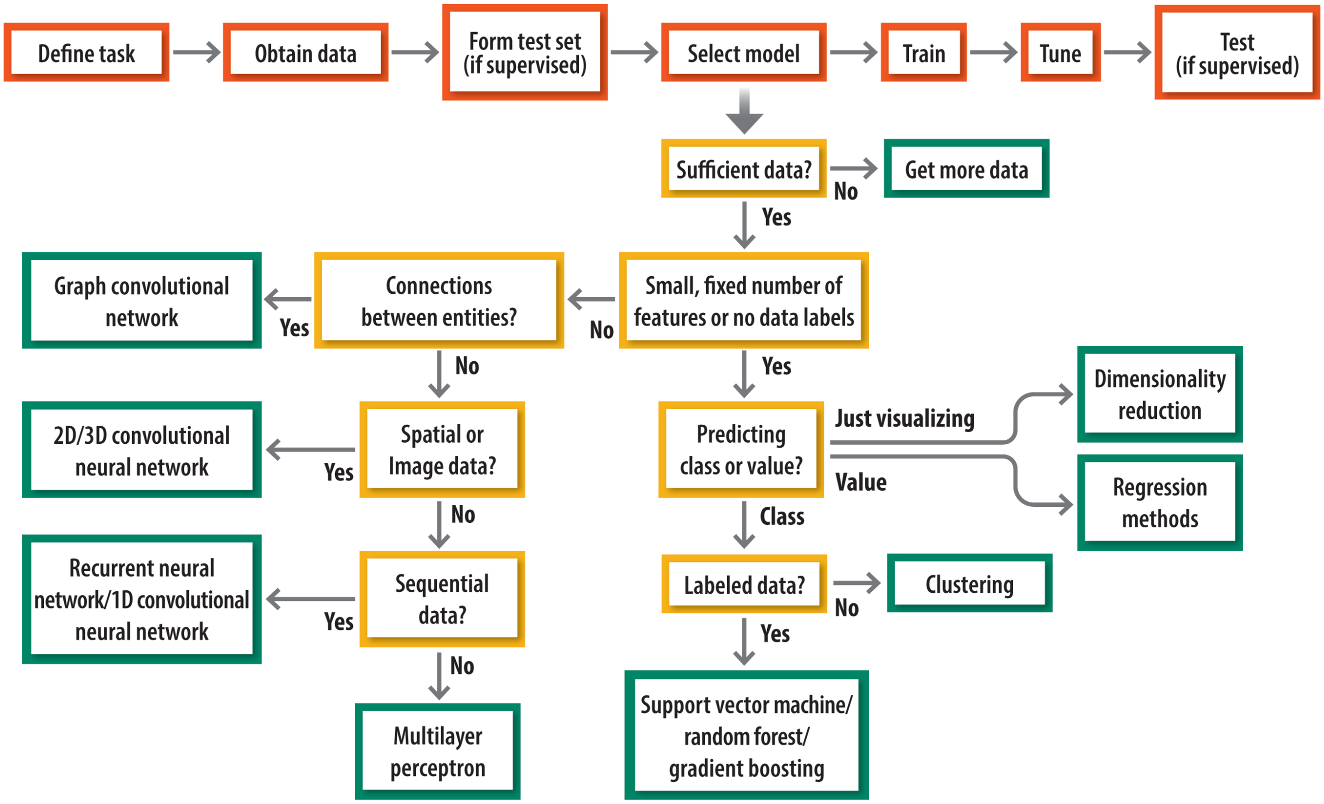

Define task
Obtain data
Form test set (if supervised)
Select model
Train
Tune
Test (if supervised)
Sufficient data?
No
Get more data
Yes
Small, fixed number of features or no data labels
No
Connections between entities?
Yes
Graph convolutional network
No
Spatial or Image data?
Yes
2D/3D convolutional neural network
No
Sequential data?
Yes
Recurrent neural network/1D convolutional neural network
No
Multilayer perceptron
Yes
Predicting class or value?
Just visualizing
Dimensionality reduction
Value
Regression methods
Class
Labeled data?
No
Clustering
Yes
Support vector machine/ random forest/ gradient boosting

Fig. 4.3. Flowchart Summarizing How to Select a Machine Learning (ML) Model. The overall procedure for training an ML method is shown along the top. A decision tree to assist researchers in selecting a model is below. However, a simple overview such as this cannot cover every case. For example, the number of data points required for ML to become applicable depends on the model being used and the number of features available for each data point, with more features requiring more data points. Deep learning models that work on unlabeled data also exist. [Reprinted with permission from Greener, G., et al. 2022. "A Guide to Machine Learning for Biologists," *Nature Reviews Molecular Cell Biology* **23**(1), 40–55.]

sentence when amino acids are put in a certain order. For instance, Transformer, a state-of-the-art model in NLP, tracks relationships in sequential data like words in a sentence, thereby learning context and meaning. As such, the model can perform translation tasks (e.g., translating an enzyme to the substrate it can catalyze).

Merging AI/ML Predictive Capabilities with Mechanistic Insight

While AI/ML models are known for their predictive capabilities, their inner logic is difficult to interpret and thus obstructs scientific understanding of biological insights or mechanisms. However, advances in the field of interpretable AI/ML enable important patterns and features underlying an AI/ML model to be identified using sensitivity analysis, saliency, and attention-based methods. Additionally, genome-scale metabolic models (GEMs) provide features that can merge AI/ML predictive capabilities with mechanistic insight. GEMs are designed to satisfy known biological constraints on metabolism, such as reaction stoichiometry, mass conservation, gene-product-reaction encoding, and nutrient environment. As a result, GEM-derived features are biologically feasible and can be used to discriminate and interpret differences between phenotypic states. One example of a GEM-based ML framework is the Metabolic Allele Classifier (MAC), which takes the genome sequence of a particular tuberculosis strain as its input and classifies strains as either resistant or susceptible to a specific antibiotic (Kavvas et al. 2020). As MAC provides an allele-parameterized form of flux balance analysis, statistical tests between antibiotic-specific resistance and susceptible strains can provide a biochemical interpretation of the genotype-phenotype map.

Overcoming the Limited Data Problem in Bioenergy Research

Although AI/ML can greatly benefit synthetic biology, it also has some limitations. One major challenge is that AI/ML is notoriously data hungry (Hsu et al. 2022). Training accurate AI/ML models generally requires sufficient training data, yet biological data can be limited by the difficulty and expense of experiments and data acquisition, which consequently hinder the training effectiveness of AI/ML models.

Recently, several studies aimed at building AI/ML models that leverage fewer data points (Wittmann et al. 2021) revealed that the limited data problem can potentially be solved using a generative model (Madani et al. 2020) or a "low-N" model, which relies on a low number of training data points (Hsu et al. 2022). Generative models create new samples following a distribution, making full use of the unlabeled information abundant in biology. For example, Madani et al. (2020) successfully applied generative models to *de novo* protein design and introduced a protein language model termed ProGen. Trained on billions of protein sequences, ProGen can generate protein sequences with controllable features (e.g., function). Data-efficient low-N models offer another potential solution to data limitations. A low-N study by Hsu et al. (2022) successfully trained a linear regression model tasked to predict protein variant effect using as few as 48 variants.

Integrating Data from Various Resources

Data integration from different types is often an empirical task that requires testing to find the highest-performing model for the specific biological objective (Kim et al. 2016; Nguyen and Wang 2020; Zampieri et al. 2019). Integration techniques include multimodal ML in which (1) a learned function at various stages of the learning pipeline brings various data streams together (Culley et al. 2020) or (2) data is mapped to an intermediate data structure that hypothetically represents the underlying biological ontology (Cho et al. 2016; Ma et al. 2018). Additionally, enforcing constraints such as mass balance between reaction and product in a chemical reaction can encode biophysical information into neural networks (Wang et al. 2022). Other approaches can encode domain knowledge into physics-inspired neural networks where loss functions are designed to optimize a domain-specific property (Ji et al. 2021). Encoding domain knowledge and mechanistic knowledge directly into AI/ML models is a more promising pursuit than either parallel mechanistic and AI/ML models joined at the final stage for prediction or mechanistic models used to generate features for AI/ML models.

Quantifying Predictive Capacity of AI/ML Models

Quantifying the overall predictive performance of AI/ML models requires multiple metrics. However, trade-offs between metrics are typical; for instance, optimizing mean squared error might result in lower correlation. Additionally, optimizing loss functions might drastically increase overall model complexity. For example, Salis et al. (2009) used a linear model instead of a popular nonlinear model like an artificial neural network to create the promoter calculator that provides an explainable mode for RNA polymerase binding and transcription. Since metrics can bias data processing and model development, evaluation metrics must be considered at the origin of project planning. Ultimately, the engineering goal is tightly bound to evaluation metrics. As a best practice, a wide variety of balanced metrics should be reported to enable future developers to benchmark against previous results for the same or similar tasks. When working on an unprecedented learning task, metrics should be compared to mechanistic models that can make similar comparisons.

Developing Generally Applicable Large Language Models and Foundation Models

Recently emerging as the preeminent strategy for scaling AI/ML model capabilities, large language models (LLMs), foundation models, and their underlying technologies have quickly revolutionized NLP and computer vision. In less than 4 years, LLMs have grown more than a thousandfold. Current models—namely, Open AI's Generative Pre-trained Transformer 3 (GPT-3), Google's Language Model for Dialogue Applications (LaMDA) and Pathways Language Model (PaLM), and Google subsidiary DeepMind's Gopher—take in terabytes of data to train hundreds of billions of parameters. At these scales, LLMs have demonstrated unprecedented, and often uncanny, capabilities not only in language generation quality but also understanding and reasoning about the knowledge they ingest. These capabilities and the further potential of LLMs pose an important opportunity to drive and accelerate systems and synthetic biology research. Thanks largely to their flexibility in digesting different data sources (e.g., text, images, signals, and spectra) at tremendous scales, LLMs and their derivatives have achieved state-of-the-art results not only in general language and image tasks but also in biological literature parsing. Examples include Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT), DNA sequence analysis (DNABERT), gene regulatory analysis (GeneBERT), protein structure prediction (AlphaFold2), and others. These models have also been extraordinarily successful in multitask and multimodal applications, including the construction of sparse expert models such as Switch Transformers and Google's Generalist Language Model (GLaM), which may scale to trillions of parameters.

4.4 Automated and Autonomous Experimentation

AI/ML has profoundly impacted automation. Production arrays and test robotic platforms are now available for automating DBTL cycles, and computational control over the build and test steps enables the development of closed-loop systems performing AI/ML-driven scientific experiments. This capability can significantly reduce the combinatorial complexity of a given problem, optimizing systems faster than traditional methods.

Preliminary efforts in self-driving microfluidics laboratories for systematic titer, rate, and yield improvements in synthetic biology have produced a semiautomated process that leverages a droplet-based microfluidic system to enable CRISPR-based gene editing and high-throughput screening on a chip (Iwai et al. 2022). CRISPR-based engineering demonstrated the system's capabilities in two test cases: (1) function disruption of the galactokinase gene (*galK*) in *E. coli* and (2) targeted engineering of the glutamine synthetase gene (*glnA*) and the blue-pigment synthetase gene (*bpsA*) to improve indigoidine production in *E. coli*.

Newer autonomous experiments go beyond pure process automation. While automated experiments follow a predefined plan, autonomous experiments instead adapt and suggest new experimental pathways based on developments during an experiment. These early autonomous experiments are supported by several key technologies, such as (1) optimal experimental design, which determines the next best step given a set of experimental goals; (2) decision-making under uncertainty, which assesses and expresses the underpinning AI/ML model's confidence in the available information used for decision-making; and (3) unsupervised and reinforcement learning, which provides a means for continuous learning throughout the experiment with potential system and user feedback to improve the outcome. A few examples of successful implementations of first prototypes include BioAutomata for synthetic biology (see sidebar, BioAutomata: A Self-Driving Biofoundry for Biosystems Design, p. 23) and Brookhaven National Laboratory's National Synchrotron Light Source II (NSLS-II) for materials discovery and physics (BNL 2021). An even more recent demonstration showed the interaction between two beamlines at NSLS-II via AI/ML, each informing the other of further explorable regions of interest during parallel experiments. The approach has shown great promise not only in making experiments more efficient but also in significantly accelerating discovery. The developed principles could be equally applied to biological experiments with great impact.

BioAutomata: A Self-Driving Biofoundry for Biosystems Design

A fully integrated biofoundry, BioAutomata enables closed-loop design and optimization of biological systems (HamediRad et al. 2019). After setting initial parameters, designing the sequence space of variable regions (e.g., promoter variants in a combinatorial pathway assembly), and defining objective functions, BioAutomata selects which experiments expect the highest yield improvements. It then performs those experiments, generates and learns from data, and updates its predictive model based on newly presented evidence. Using this new information, BioAutomata decides which experiments to perform next to reach a user's goal while simultaneously working to minimize experiments conducted and project costs.

BioAutomata: An Integrated Robotic System for Autonomous Experimentation Driven by Artificial Intelligence and Machine Learning (AI/ML). The platform was developed by researchers from DOE's Center for Advanced Bioenergy and Bioproducts Innovation (CABBI). [Courtesy CABBI]

5. Opportunities and Challenges

AI/ML and automated experimentation present key opportunities and challenges for advancing biological understanding and engineering, especially in bioenergy, biosystems design, and biomanufacturing. These opportunities intersect three research grand challenges and require addressing gaps in (1) experimental, data, and computing infrastructure; (2) various bioenergy applications of AI/ML, such as genotype to phenotype prediction, biosystems design, and bioprocessing; and (3) education, training, and workforce development. By leveraging its unique capabilities and resources, DOE is well-positioned to realize AI/ML-based opportunities for missions in energy and environment.

5.1 Science Challenges

Three exemplary research grand challenges could benefit from AI/ML solutions.

1. **Microbes and Microbial Communities Built to Specifications.** AI/ML could design genomes with predefined properties for specific environments, ensuring that genomes perform as expected. A key needed novel foundational AI/ML capability is the use of specifications, theory, and experiments to learn new biology.

2. **Closed Loop Autonomous Design and Control for Biosystems Design.** Autonomous, self-driving experiments that optimize facility resources, reduce the number of experiments, and limit redundant data collection are poised to profoundly transform experimentation by searching for new experimental parameters and settings and targeting new biological materials and processes. Key needed novel foundational AI/ML capabilities are human-in-the-loop hypothesis creation and testing, the ability to work with few data, and delivery of trustworthy solutions.

3. **Bioprocess Scale-Up and Automation.** Translating research progress into industrial bioengineering innovation requires scale-up of experiments at multiple scales (e.g., from microtiter plates to flasks to large and highly mixed bioreactors instrumented with comparable measurements for integration). Key needed capabilities are good-quality data and metadata, available and accessible data and computing at scale, the ability to work with few data, and digital twins for guidance.

Similar AI/ML needs exist for all three grand challenges:

- Massive, annotated datasets or, alternatively, the ability to learn from few data.
- Predictive capabilities to foresee outcomes.
- Trustworthy AI/ML.
- Effective collaboration between humans and AI/ML (e.g., in developing and testing hypotheses).
- AI/ML models capable of executing on small edge devices as part of larger complex workflows.
- Availability of objective-based decision-making under uncertainty (i.e., optimal experimental design).
- Efficient and effective capture of expert knowledge in AI/ML models.
- Digital twins to design, accompany, guide, and inform complex experiments.
- Ontologies in addition to AI/ML for supporting technologies (e.g., robotics and sensors) and AI/ML-specific computing infrastructure, metadata, and data standards.

5.2 Technology Gaps

Breakout participants discussed challenges in experimental, data, and computing infrastructure relevant to BER and BETO application areas and AI/ML. These challenges included concerns about automation, standardization, and data quality; underdevelopment and use of certain methodologies and tools; communication and technology gaps between biological and computational domains; and current funding models' lack of coordination at scale, which impedes risk-taking, agility, and innovation.

Laboratory Automation

Tension and trade-offs exist between deployment of commercial, off-the-shelf instrumentation and software versus custom development (or retrofitting). Desirable systems are encompassing and stable yet flexible enough to be adapted to changing needs. Several representative systems and environments would require hybrid commercial and custom components, including AI/ML algorithm–directed robotic (as well as microfluidic) systems capable of collecting large amounts of diverse data in a nondestructive manner, to guide AI/ML systems that explore bioprocess optimization space in bioreactors.

Automation gradients also pose challenges. Translating manual to automated methods is often nontrivial, but an additional challenge for semimanual workflows is the difficulty in achieving enough contiguous automated method coverage to avoid frequent interleaving of manual and automated steps.

Data Infrastructure

Important challenges in data infrastructure include data exchange standardization, data quality, and data privacy (especially commercial). Integrating heterogeneous software, data, and automation across vendors and developers is difficult, partly due to a lack of standardized metadata formats, vocabularies, and syntaxes. Likewise, ontologies may be too static as foundational models evolve over time. However, dynamic data ontologies and exchange standards can create interface-breaking changes, so these systems must have clear change management processes, proper migration procedures, and ongoing contact with stakeholders to maintain stability.

A key challenge in the development of capable AI/ML models for scientific discovery is the need for very large, high-quality datasets suitable for the research questions at hand. Data quality matters more than the amount because high-quality negative data is required for ML and model development. Given the necessary scale and coverage of these datasets to avoid gaps or undesirable biases, small research efforts no longer have the capacity to create them. Instead, automated coordinated campaigns are needed, enabled by changes in experimental design. Autonomous experiments and laboratories can play key roles in these campaigns, but they present their own implementation challenges. An additional data-related challenge is how to build models on top of a foundation of private (e.g., company-owned) primary data and make the trained models available to the public without revealing the primary data and creating issues with intellectual property or copyrights.

Digital Twins

Building predictive digital twins (or using AI/ML to automatically develop them) requires new types of tools to reliably predict nonintuitive targets and produce more accurate multiscale modeling of biological systems and processes. A big challenge in metabolic engineering, for example, is researchers' tendency to rely on gene targets reported in the literature. Powerful computational tools, whether biophysical, ML, or a combination, could predict nonintuitive targets for metabolic engineering (e.g., genes of unknown function or not reported in the literature) that could significantly benefit system performance. However, the reliability of these predictions, either actual or perceived, has not surpassed the threshold needed to secure resources. Perceptions of reliability and risk may be confounded by metabolic engineers if, for example, they do not understand how biophysical or ML models work and thus cannot assess risk versus reward.

Technology Adoption

Computing infrastructure-related challenges include barriers to technology adoption and establishment of benchmarks that encourage adoption. As technologies show success, their adoption increases. However, early success can lead to algorithm fatigue. In such cases, the continuous release of new algorithms that are not necessarily improved based on community-accepted metrics contributes to the reluctance to adopt new ones.

Benchmarks must therefore be established to test new algorithms, models, and methods and a leaderboard established to publicize tool performance based on the benchmarks. The frequent gap between benchmark performance and usefulness for novel scientific discovery must also be addressed.

Interdisciplinary Communication

Similar challenges affect interdisciplinary communication and the accessibility of AI/ML capabilities and data repositories to nonexperts. Many are related to interactions between experimentalist and computational teams and ensuring that each understands the parlance of the other. Part of the solution could be to improve and simplify software and automation user interfaces or low-code environments to make a subset of AI/ML capabilities accessible (e.g., 20% of the functionality that will satisfy 80% of needs) while simultaneously preventing nonexperts from misusing the tools or misinterpreting their results. Similarly, the substantial effort required to enter data into repository systems (e.g., laboratory information management systems and electronic notebooks) could be minimized by developing AI/ML-guided methods to capture metadata or extract raw instrumental data for deposit into AI/ML-serving data systems.

Risk Appetite

The reluctance to take risks, which is difficult to evaluate, presents an inertial challenge to innovation, such as in metabolic engineering efforts that would benefit from pursuing nonintuitive targets. Researchers in resource-constrained environments are less likely to take risks, often developing small, bespoke infrastructures without coordination with other groups, rendering them difficult to maintain and support. Such factors produce incremental improvements that are more evolutionary than revolutionary.

Efforts to foster higher-level coordination across different endeavors and provide the resources needed to encourage innovation and risk-taking will help develop core solutions to big common challenges. Overarching organizations could achieve this by helping knit together experimental, data, and computing infrastructure.

A related challenge for researchers developing technologies and capabilities, as opposed to those pursing scientific goals, is how to prepare specific and quantitative milestone-driven plans that deliver accountability to funders without a commitment to prescribed features or specifications that would impede agility and responsiveness to users' unanticipated and often changing needs. One possible approach to this challenge, which is compatible with user-centered design practices, is to prescribe procedural milestones (e.g., conduct customer discovery interviews and develop features prioritized by these interviews).

Autonomous Experiments and Laboratories

Many research directions in DOE's mission space (e.g., systems biology and biosystems design) will require moving toward autonomous laboratories driven by AI/ML. These capabilities will enable researchers to execute high-throughput experiments (e.g., microbial systems from growth through omics data generation and molecular structure determination) and leverage high-throughput robotic laboratories capable of producing datasets on the order of 100,000 data points in run times of a few days. Autonomous infrastructure must be rapidly configurable for different organisms, instruments, experiments, and protocols. Scientists will require the ability to define growth goals and quality criteria directly to robotic AI/ML control systems, enabling generation of high-throughput, population-scale, multiple-omics data types. Below are several exemplary unmet needs.

Capture Expert Knowledge to Drive Autonomous Experiments and Laboratories. For AI/ML-enabled bioenergy research, establishing high-throughput facilities for plant cell experiments at scale is important. Nearly all existing large-scale robotic laboratories were developed for microbial or mammalian cell systems, so hardware and software are needed to address the unique challenges of plant biotechnology. Plant biotechnology expertise (e.g., cell transformation and growth) must be transferred to AI/ML systems that will drive robots.

Establish Environmentally Hardy Technology for Field-Scale Autonomous Experiments and Laboratories. Automated field sites capable of carrying out large-scale field trials autonomously, as well as autonomous greenhouse facilities, must be established. Commercial solutions (e.g., from the agriculture community) must be leveraged as much as possible to enable remote surveys of plant health, growth conditions, stress tolerance, and optimal growth conditions within environmental constraints. In the research space, development of a species-agnostic platform may be a required investment for planting systems and plant identification tracking systems (e.g., 100,000 plantlets → field → assays → data). In addition, there is a need to integrate AI/ML into small sensors, small robots, large field equipment, and remote sensing.

Address Increased Complexity Due to Scale for Autonomous Experiments and Production. Many differences exist between current pharmaceutical paradigms and those suitable for biofuels and bioproducts. BETO applications often require translation to scale and reliability. Needs include massively increased product quantities, more concern about scaling, less concern about batch contamination, and more attention to the potential benefits of including feedstock-associated microbes. Unlike the pharmaceutical industry, the biofuels and bioproducts community faces key questions that involve (1) determining how to investigate microbial communities that thrive on feedstocks and with other microbial communities that feedstocks introduce into a process and (2) understanding the evolution of microbial communities in large-scale bioreactors. Accelerating biofuels and bioproducts research requires identifying which scales must be measured to predict behavior and detecting emergent, divergent phenomena and their causes. These fundamental capabilities will enable autonomous production at scale.

Facilitate Training of AI/ML Models for Bioenergy Scenarios. The advanced AI/ML research community requires data and computing infrastructure capabilities beyond those available today, specifically:

- **Data Archives** — Needs include new services from existing data archives, such as AI/ML-ready data sufficiently labeled and cleaned for immediate use in AI/ML training. Users require tools to identify gaps in existing data collections and automatically initiate additional data collection campaigns of varying sizes to fill those gaps, linking to computing and experimental facilities and projects.
- **Computing Resources** — Training large-scale AI/ML models requires sufficient computing resources.
- **New AI/ML Training Infrastructure** — Scientists need capabilities for combining or leveraging large, potentially multidisciplinary data collections that support AI/ML designs, development, and training. Data should be permanently available on AI/ML-specific computing hardware for medium timescales of months to years (depending on the project or campaign) with the ability to share data and models with others.
- **Integrative Technology Test Labs** — Many of the described scenarios rely on integrating AI/ML with robotics, sensors, or domain-specific edge devices. Test laboratories would benefit these scenarios by enabling researchers to rapidly integrate and test solutions, particularly for complex experiments.

AI/ML Challenges

AI/ML challenges identified in the workshop fall into three principal needs: (1) new or improved AI/ML methods, (2) tools that meet the specific needs of BER and BETO research communities, and (3) industry partnerships.

Opportunities for autonomous operation of experiments, laboratories, and field sites cannot be fully met with existing AI/ML models and methods. New research is needed to enable the bold science grand challenges outlined by workshop participants. Several science communities have conducted first-of-their-kind autonomous experiments, but they remain limited in scope. Fully autonomous complex experiments, laboratory operations, and field sites will require AI/ML models for hundreds to thousands of different tasks for controlling every aspect of the work, including devising hypotheses, predicting results, and choosing paths that optimize scientific outcomes based on set research goals. Some of these needed capabilities exist today, but others, like models for hypothesis generation and translation of scientific goals, are in their infancy at best. An even bigger challenge is customizing so many models to widely varying experiments; the training data and available workforce are currently insufficient to accomplish this customization.

Foundation models could provide a pathway to autonomous operation of experiments, laboratories, and field sites, but they are still untried in scientific research. Furthermore, even the most advanced foundation models can currently perform only 600 different tasks, which may be insufficient in a bioenergy research paradigm. Given their proven potential for easy customization once trained and their general versatility, foundation models warrant more research to explore how they can be used to support autonomous operation.

In existing automated or basic autonomous experiments, the inclusion of specialized scientific knowledge is key to successful AI/ML solutions. This knowledge goes beyond what is included in data and requires the explicit incorporation of foundational scientific principles and tacit expert knowledge (e.g., for sample preparation). To scale up such efforts, the community must develop methods for knowledge capture that are more efficient, more robust, and better directed.

AI/ML-ready data will be key to advancing the impact of AI/ML in the field of biology. BER, with its significant enabling infrastructure of experimental and computational facilities, could create data ready for AI/ML training at the needed scale (e.g., to facilitate training of new foundation models). Clear definitions, benchmarks, and standards will be needed to determine which metadata, provenance, data preparation, and other techniques will enable data to be easily utilized in AI/ML training without resource-intensive data identification, cleaning, or preparation.

However, gathering sufficient training data will be impossible in some situations. Consequently, new AI/ML models that can cope with little training data must be developed, along with methods to create more information-rich training input. This active research field requires more investments and robust testing of new methods.

Many settings in autonomous experimentation and observation will require embedding AI/ML models in edge computing devices with limited computational capabilities (e.g., instruments, sensors, and drones). This integration will require developing compact AI/ML models that have capabilities similar to larger models but use less computing power. Furthermore, many edge devices, particularly those in the field, have limited network connectivity, making upgrades and retraining difficult. New methods are needed to enable unsupervised *in situ* retraining of these compact AI/ML models while maintaining their quality and correctness.

Finally, BER- and BETO-supported biologists must be able to trust the AI/ML solutions they wish to deploy. Therefore, developing robust, reproducible, and explainable models is important. All three qualities are being actively researched, but significant progress is needed to achieve easy-to-use, standardized approaches that can be integrated broadly in all AI/ML solutions.

5.3 Application Challenges

AI/ML could help address BER and BETO application areas facing scientific and technical challenges. Some of these challenges are specific, while others are more general and crosscutting.

General challenges include associating genotype with phenotype and elucidating relationships between biological components. High-throughput phenotypic measurements are only possible in very few cases, and substantial genotype data is often not available for the systems seeking phenotype prediction. Forward phenotype prediction may uncover missing biological knowledge, whereas inverse design (i.e., designing to achieve a particular specification) uncovers design principles and additional gaps in biology and is perhaps a more compelling goal. Approaches are needed that use integrated datasets to elucidate biological relationships leading to testable hypotheses.

Domain-specific challenges include:

- Engineering enzymes with specified kinetics, substrate specificity, and other properties.
- Engineering microbial communities *in situ* and at scale for pathogen protection; carbon sequestration; community stability; carbon dioxide (CO_2) and water cracking; nitrogen fixation; stress protection (e.g., drought); nutrient mobilization; hydrogen, phosphate, and carbon cycling; and closed-loop, energy-efficient regenerative bioproduction using waste or C1 feedstocks.
- Characterizing microbial gene-environment interactions effectively, which is prerequisite to foundational understanding, prediction, and inverse design, including microenvironments within bioreactors.
- Improving bioreactor performance for bioprocess scale-up. Optimizing fermentations could be pursued through a variety of approaches, including changing feedstock injection sites, redesigning vessels to improve mixing and mass transfer, adding sensing methods that are easier to interpret and act upon, and leveraging microbial population heterogeneity to improve process robustness.
- Scaling bioprocess information to better inform laboratory-scale experiments since many bioprocess developers lack regular access to pilot, let alone industrial, scales.
- Engineering plants to: create specific transcriptional levels or transcriptional circuits, implement metabolic pathways within chloroplasts, enhance stress resistance (e.g., to pathogens, drought, or other environmental factors), redesign photosynthesis, achieve nitrogen-fixing endosymbiosis, improve crops for animal agriculture, achieve a foundational understanding of plant development including interactions at the tissue level, predict how a plant will perform under different environmental conditions, sequester CO_2, and integrate plant biology with climate and economic modeling.

5.4 Workforce Development, Diversity, Outreach, and Social Responsibility

Data-driven design of biological systems represents an emerging paradigm in basic and applied biological research. Currently, few bioenergy researchers are well-versed in both AI/ML and biology. Therefore, encouraging and facilitating research collaborations between computational scientists and biologists is important, along with training a new generation of scientists to develop and apply AI/ML tools to long-standing scientific challenges in bioenergy research.

Workshop participants discussed community development issues related to education, outreach, workforce development, partnerships with other funding agencies, and social responsibilities and ethics (see Appendix C: Breakout Session Assignments, p. 38). The session opened with two talks: “Applying a Human-Centered Design Approach for AI/ML Educational Outreach” and “Charting a Course for a Resilient and Competitive Future—Bioeconomy Strategy Engagement and Recommendations for Training.” The talks provided valuable recommendations from related efforts and potential leverageable strategies, including the Task Force on Synthetic Biology and the Bioeconomy at Schmidt Futures. Key takeaways compiled from participant breakout groups are described below.

Workforce Development

Participants were deeply engaged in conversations on education, training, and workforce development. Similar key themes emerged in multiple breakout groups, including building an interdisciplinary workforce, ensuring workforce diversity and inclusiveness, and creatively engaging the public. The breakout group on workforce development discussed how to train the workforce to become synergistic with AI/ML and how to encourage established, senior professionals to welcome new technology. Two main challenges that emerged from discussions were competition to retain data scientists and engineers and the need for an interdisciplinary workforce.

Competing with the private sector for a highly trained workforce is another challenge potentially attributable to salary, impact, job stability, and professional growth opportunities. Industry may be perceived as having more impact than government and academia and as being more related to real-world challenges as, for example, the AI/ML industry increasingly publishes open-source tools. Other distinctions between industry and academia are the current incentive mechanisms for principal investigators and faculty and what might be most appropriate for data scientists and engineers.

Another primary topic of discussion was creating an interdisciplinary workforce. Most graduate students and postdocs currently working in bioenergy research are unfamiliar with AI/ML algorithms and automation and the kinds of problems AI/ML can solve. A lack of understanding of different types of model-derived knowledge can limit researchers moving into ML-driven bioenergy research from non-ML fields. Conversely, graduate students and postdocs who are experts in AI/ML are not necessarily familiar with the major scientific challenges related to bioenergy research.

Overcoming these knowledge and skill gaps and training a new generation of scientists well-versed in both AI/ML and biology require making AI/ML more accessible to biological experimentalists and recruiting talented computational scientists to bioenergy research. DOE can create targeted opportunities for collaboration and cross-training. Examples include hackathons, CASP-like competitions, kaggle competitions, protein design competitions, or "protein-paloozas" to encourage collaboration between biologists and computer scientists. Participants also discussed training opportunities to overcome knowledge and skill gaps, including 1- to 2-year postbaccalaureate programs, flexible certificates similar to MBA programs, and internship opportunities. In addition, effort is needed to excite researchers about AI/ML potential and motivate them to enter their data into databases. Finally, communication of common standards among researchers is needed to ensure proper and rigorous model validation standards.

Strategies to build a more interdisciplinary workforce involve: (1) increasing incentives for interdisciplinary research and team-based science, including funding mechanisms for computational and experimental scientists; (2) building more transdisciplinary research centers and training opportunities; and (3) gaining recognition from funders. To engage computer scientists with bioenergy research, participants suggested more cross-training opportunities, additional funding mechanisms for data scientists, and increased recognition of deep knowledge in computational and engineering areas.

Diversity

Workforce breakout group participants also focused on ways to ensure diversity and inclusiveness in the developing workforce, noting that diversity is especially poor in computational areas. Multiple strategies to increase diversity include: (1) adopting better, more inclusive hiring practices; (2) developing mechanisms to support partnerships with minority-serving institutions (MSIs); (3) offering more targeted summer research opportunities for undergraduate students from MSIs and historically black colleges and universities; (4) ensuring diverse representation at meetings and workshops; (5) connecting with rural communities, including land grant universities and extension scientists; and (6) creating additional internship opportunities.

Outreach

Four main themes emerged from the outreach breakout group, several of which synergize with the diversity discussion. First, early inclusion of biology and computational science into students' education paths is important. More programs that ease student access to curricula (e.g., industry internships and iCLEM) are needed at the high school and college levels. Early introduction to interdisciplinary training between computational science and biology will enable greater access to AI/ML in bioenergy research. Additionally, creative activities like AlphaFold games will engage and broaden participation.

Second, participants discussed the lack of student funding for nonmedical biotechnology training programs. Biology-related internships are difficult to scale and are comparatively expensive. A potential solution is to create engaging virtual programs that provide students with hands-on experience when in-person opportunities are unavailable.

Third, participants recommended using better metrics to incentivize outreach efforts, suggesting the importance of tracking science impact in ways other than just number of publications. Number of citations, such as for datasets, could be another metric of success. Faculty could also support additional mentees or help promote the field in other ways.

Finally, participants discussed the implications of public misconceptions of bioenergy research, such as genetically modified organisms (GMOs). Using spokespeople to help communicate the benefits of nonmedical biotech would help dispel myths. Social media platforms are good avenues to disseminate interesting work in the field.

Social Responsibility and Ethics

The social responsibility and ethics breakout participants also discussed misconceptions and disinformation in the public domain, along with potential outreach efforts to better inform people about the field and publicly funded efforts. An emphasis on critical thinking and risk-benefit trade-off analysis could be accomplished by (1) creating training materials and activities that bridge AI/ML, biosecurity, and research communities; (2) ensuring that materials resonate with the community (i.e., what is the impact on health and employment); (3) increasing researcher engagement with professional creatives (e.g., writers) to develop interesting and accurate narratives; (4) providing small supplemental funding mechanisms to attract professional creatives to produce new forms of content; and (5) including social and ethical considerations in agency reports for policy-makers.

Participants also discussed the importance of diversity, equity, inclusion, and accessibility (DEIA) in workforce development. At all educational levels (vocational, undergraduate, and graduate), programs and training opportunities are needed that combine biology with computer science or applied mathematics. One approach that can scale quickly is a hybrid training environment that provides most materials online (e.g., videos, standard operating procedures, and quizzes) while also offering hands-on experience. In addition, industry engagement is key to building new, more inclusive collaboration models such as internships and visiting positions.

When considering DEIA, participants also explored how to identify real versus perceived risk in applying AI/ML to biotechnology. Three suggestions were discussed. First, mechanisms should be developed to screen potentially harmful cases early. Identifying risks and risk tolerance at an early stage can constrain research but defining risks and mitigation strategies is critically important. Considering the potential dual use of technologies might also be helpful. Second, biosecurity tools (e.g., environmental sequencing for pathogen detection) should be developed and used to assess ecological risks of environmentally deployed GMOs and engage regulators and the public on rational risk analysis and governance. Finally, transparency, care, and caution should be pillars that guide the impacts of biotechnology applications. This is especially critical in ethically dubious applications such as explosives or narcotics production.

Finally, this breakout group considered ways to equitably distribute the benefits and risks of technology. One key suggestion was to distribute biomanufacturing, especially jobs, where the raw materials are sourced (e.g., in the agricultural Midwest). Also critical for equitable benefit distribution is considering the geography of next-generation feedstocks, including agriculture, forestry, municipal wastes, and C1s. This model will require both capital infrastructure build-out and local workforce training for biomanufacturing jobs.

Appendix A

DOE Charge

Department of Energy
Washington, D.C.

Recent advances in computing and data analytics have resulted in powerful artificial intelligence and machine learning (AI/ML) techniques with significant potential for use in biotechnology and broader genomics-based research. These techniques combined with advances in automation in the laboratory offer the ability to rapidly accelerate the design and optimization of biological systems and processes for a variety of DOE mission needs in energy and the environment.

A previous joint BER-BETO workshop in April 2021, "*Designing for Deep Decarbonization: Accelerating the U.S. Bioeconomy*," identified several areas within the transportation, industry, and agricultural sectors within the U.S. economy where advances in biotechnology were poised to make significant contributions (DOE 2021a). The Biological and Environmental Research (BER) program within DOE's Office of Science (SC) and the Bioenergy Technology Office (BETO) within DOE's Office of Energy Efficiency and Renewable Energy (EERE) have an interest in accelerating the pace of development and transition of biotechnology solutions out to industry as part of an overall strategy to promote a globally competitive U.S. bioeconomy.

Building on the previous workshop, the AI/ML for Bioenergy Research (AMBER) workshop should explore the integration of AI/ML techniques within genomics-enabled basic and applied science and biodesign for optimization of biological systems and processes (a fully automated laboratory system to accelerate iterative design-build-test-learn systems) and toward advancing biomanufacturing. The use of AI/ML within an automated laboratory affords the ability for iterative learning that builds on previous data collection and characteristics of the chassis organism (microbe or plant) to accelerate the optimization and design of new metabolic processes for the production of desirable products and/or new functions. The pairing of AI/ML techniques with automated instrumentation could lead to significant improvements not only in the more rapid design and optimization of engineered organisms but also, if applied broadly, in the potential to change scientific investigation in general.

The AMBER workshop should specifically focus on the broader scientific potential and immediate applications of integrated AI/ML systems with automation in the laboratory. Workshop participants should be tasked with assessing the potential for AI/ML systems to advance the understanding of biology in general, how integration of AI/ML techniques with automation in the laboratory could accelerate the design of biological systems and optimize biomanufacturing, what data and compute infrastructure would be needed in such systems, and what expertise and workforce development efforts would be needed to shift toward these systems within the broader science. We anticipate the recruitment of a broad and diverse group of participants that would bring multidisciplinary expertise and knowledge to this effort. The participants would have expertise with applications in AI/ML (related to areas of genomics, protein/structure prediction, imaging, synthetic biology, lab automation and bioprocess development), data resource needs, and expertise in the areas of plant and microbial systems (with experts outside these systems to be invited to bring in additional perspectives).

As the growth of more sophisticated AI/ML models, fueled by the availability of ever larger datasets within the DOE infrastructure, enables more automated analyses through the use of robotics, the stage is set for potentially game-changing approaches to scientific investigation. The integration of AI/ML techniques with automated experimentation offers powerful new approaches to research that not only take better advantage of previous research results and data but iteratively build on and learn from new information generated within these envisioned approaches to science. In addition to experimental design approaches, AI/ML can have immediate impacts on bioprocess design for biomanufacturing as it is industrially practiced through more precise control of conditions in bioreactors. These discussions are very important across both SC and EERE programs as DOE looks to take advantage of breakthroughs in data science within a scientific complex rich in computational and experimental capabilities.

We are excited about the workshop attendees sharing their thoughts, expertise, and imagination during the AMBER workshop discussions and look forward to exciting times ahead for biological science and biotechnology development where DOE plays an important and leading role.

Sincerely,

R. Todd Anderson

Director, Biological Systems Science Division
Biological and Environmental Research program
Office of Science

Jay Fitzgerald

Chief Scientist
Bioenergy Technologies Office
Office of Energy Efficiency and Renewable Energy

Appendix B

Workshop Agenda

August 23–25, 2022

All times Eastern

August 23, 2022

Session 1: Workshop Goals and Introduction to Artificial Intelligence/Machine Learning

12:00 p.m.	**Welcome and Opening**	
	Welcome	Huimin Zhao (University of Illinois, Urbana-Champaign)
	Opening Remarks, Motivation, Background	R. Todd Anderson (U.S. Department of Energy), Jay Fitzgerald (U.S. Department of Energy)
	Objectives and Structure	Huimin Zhao
12:30 p.m.	**Introduction into AI for Biology**	**Moderator:** R. Todd Anderson
	Advanced Research Directions in AI for Science, Energy, and Security	Rick Stevens (Argonne National Laboratory)
1:10 p.m.	**Break**	
1:35 p.m.	**Opportunities and Challenges in Emerging AI/ML-enabled Bioenergy Research**	
	Short Talks and Panel Discussion	**Moderator:** Kerstin Kleese van Dam (Brookhaven National Laboratory)
	Structuring Data for Statistical Learning	Kjiersten Fagnan (DOE Joint Genome Institute)
	Opportunities and Challenges in Emerging AI/ML-Enabled Bioenergy	Dmitry Grapov (Amyris)
	Recent Advances of AI for Biology and Biotechnology	Marinka Zitnik (Harvard University)
2:35 p.m.	**Elevator Pitch Presentations**	**Moderator:** Nathan Hillson (Lawrence Berkeley National Laboratory)
4:05 p.m.	**Concluding Notes from Day 1**	Huimin Zhao

August 24, 2022

Session 2: Defining Focus on Applications of AI/ML for Bioenergy Research

12:00 p.m.	**Opening Remarks**	Huimin Zhao, Deepti Tanjore (Lawrence Berkeley National Laboratory)
12:05 p.m.	**Presentations and Q&A**	**Moderator:** Deepti Tanjore
	AI- and XAI-Driven Systems Biology	Daniel Jacobson (Oak Ridge National Laboratory)
	Taking the Cellular Perspective: A Multiscale, Computation-Driven Approach to Bioprocess Design, Operation, and Optimization	Cees Haringa (Delft University of Technology)
12:35 p.m.	**Breakout Groups**	**Moderator:** Huimin Zhao
	2-1 AI/ML Applications – Biology (Microbe/Microbiome)	
	2-2 AI/ML Applications – Biology (Plant)	
	2-3 AI/ML Applications – Biodesign (Microbe/Microbiome)	
	2-4 AI/ML Applications – Biodesign (Plant)	
	2-5 AI/ML Applications – Process	
2:05 p.m.	**Break**	
2:30 p.m.	**Report Out**	**Moderator:** Deepti Tanjore

Session 3: AI/ML Approaches to Meet Bioenergy Research Needs

3:15 p.m.	**Opening Remarks**	Deepti Tanjore
3:20 p.m.	**Presentations and Q&A**	**Moderator:** Kerstin Kleese van Dam
	Mathematically-Based AI/ML to Guide and Analyze Experiments: Autonomous Self-Driving Labs, Complex Inversion, and Reconstruction from Limited Scientific Data	James Sethian (University of California, Berkeley)
	Integrated Mechanistic and AI/ML Approach for Bioenergy	Frank Alexander (Brookhaven National Laboratory)
3:50 p.m.	**Breakout Groups**	**Moderator:** Huimin Zhao
	3-1 AI/ML Approaches	
	3-2 AI/ML Approaches	
	3-3 AI/ML Approaches	
	3-4 AI/ML Approaches	
	3-5 AI/ML Approaches	
5:20 p.m.	**Break**	
5:45 p.m.	**Report Outs**	**Moderator:** Kerstin Kleese van Dam
6:30 p.m.	**Concluding Notes Day 2**	Kerstin Kleese van Dam

August 25, 2022

Session 4: Data and Compute Infrastructure Needed

12:00 p.m.	**Opening Remarks**	Nathan Hillson
12:05 p.m.	**Presentations and Q&A**	**Moderator:** Nathan Hillson
	Data and Compute Infrastructure Needed for AI/ML in Bioenergy Research	Héctor García Martin (Lawrence Berkeley National Laboratory)
	ML for CRISPR Genome Editing: A Case Study for Enhanced Methods in Agricultural Genetics	Matt Hudson (University of Illinois, Urbana-Champaign)
12:35 p.m.	**Breakout Groups**	**Moderator:** Kerstin Kleese van Dam
	4-1 Data and Compute Infrastructure – Large-Scale Experimental Facilities	
	4-2 Data and Compute Infrastructure – Automation	
	4-3 Data and Compute Infrastructure – Laboratory-Based Research	
	4-4 Data and Compute Infrastructure – Computational Science	
	4-5 Data and Compute Infrastructure – Biological System Design and Control	
2:05 p.m.	**Preparation for Report Out**	
2:30 p.m.	**Report Outs**	**Moderator:** Kerstin Kleese van Dam

Session 5: Community Development Including Outreach, Engagement, and Training

3:15 p.m.	**Opening Remarks**	Huimin Zhao
3:20 p.m.	**Presentations and Q&A**	**Moderator:** Huimin Zhao
	Applying a Human-Centered Design Approach for AI/ML Educational Outreach	Rachel Switzky (University of Illinois, Urbana-Champaign)
	Charting a Course for a Resilient and Competitive Future: Bioeconomy Strategy Engagement and Recommendations for Training	Mary Maxon (Schmidt Futures)
4:10 p.m.	**Breakout Groups**	**Moderator:** Huimin Zhao
	5-1 Community Development – Education	
	5-2 Community Development – Outreach	
	5-3 Community Development – Workforce Development	
	5-4 Community Development – Partnerships with Other Funding Agencies	
	5-5 Community Development – Social Responsibilities/Ethics	
5:10 p.m.	**Break**	
5:35 p.m.	**Report Outs**	**Moderator:** Huimin Zhao
6:05 p.m.	**Concluding Notes Day 3**	Huimin Zhao
6:10 p.m.	**Workshop Co-chairs Meeting**	

Appendix C

Breakout Session Assignments

Session 2: AI/ML Applications

2-1 Biology: Microbe/Microbiome

Adam Arkin, group leader
Lawrence Berkeley National Laboratory

Emiley Eloe-Fadrosh
DOE Joint Genome Institute

Kjiersten Fagnan
DOE Joint Genome Institute

Ee-Been Goh
Zymergen, Inc.

Kerstin Kleese van Dam
Brookhaven National Laboratory

Shinjae Yoo
Brookhaven National Laboratory

Mary Maxon
Schmidt Futures

Arvind Ramanathan
Argonne National Laboratory

Rick Stevens
Argonne National Laboratory

Dawn Adin, observer
U.S. Department of Energy

Wayne Kontur, observer
U.S. Department of Energy

2-2 Biology: Plant

Sue Rhee, group leader
Carnegie Institution for Science

Arti Singh, group leader
Iowa State University

Kristofer Bouchard
Lawrence Berkeley National Laboratory

Mary J. Dunlop
Boston University

Daniel Jacobson
Oak Ridge National Laboratory

Lee Ann McCue
Pacific Northwest National Laboratory

Carlos Soto
Brookhaven National Laboratory

Marinka Zitnik
Harvard University

Resham Kulkarni, observer
U.S. Department of Energy

Ramana Madupu, observer
U.S. Department of Energy

Catherine Ronning, observer
U.S. Department of Energy

2-3 Biodesign: Microbe/Microbiome

Héctor García Martin, group leader
Lawrence Berkeley National Laboratory

Dmitry Grapov
Amyris

Lydia Kavraki
Rice University

Nina Lin
University of Michigan

Christopher Long
Ginkgo Bioworks, Inc.

Costas Maranas
The Pennsylvania State University

Chris Mungall
Lawrence Berkeley National Laboratory

Peter St. John
National Renewable Energy Laboratory

Huimin Zhao
University of Illinois, Urbana-Champaign

R. Todd Anderson, observer
U.S. Department of Energy

Boris Wawrik, observer
U.S. Department of Energy

2-4 Biodesign: Plant

Shin-Han Shiu, group leader
Michigan State University

Frank Alexander
Brookhaven National Laboratory

Qun Liu
Brookhaven National Laboratory

Baskar Ganapathysubramanian
Iowa State University

Nathan Hillson
Lawrence Berkeley National Laboratory

James Sethian
University of California, Berkeley

Matthew Hudson
University of Illinois, Urbana-Champaign

Rachel Switzky
University of Illinois, Urbana-Champaign

Neeraj Kumar
Pacific Northwest National Laboratory

Pablo Rabinowicz, observer
U.S. Department of Energy

Amy Swain, observer
U.S. Department of Energy

2-5 Process

Cees Haringa, group leader
Delft University of Technology

Gyorgy Babnigg
Argonne National Laboratory

Ben Brown
Lawrence Berkeley National Laboratory

Deepti Tanjore
Lawrence Berkeley National Laboratory

Corey Hudson
Sandia National Laboratories

Adam Perer
Carnegie Mellon University

Gina Tourassi
Oak Ridge National Laboratory

Bobbie-Jo Webb-Robertson
Pacific Northwest National Laboratory

Gayle Bentley, observer
U.S. Department of Energy

Jay Fitzgerald, observer
U.S. Department of Energy

Paul Sammak, observer
U.S. Department of Energy

Session 3: AI/ML Approaches

3-1 AI/ML Approaches

Arvind Ramanathan, group leader
Argonne National Laboratory

Adam Arkin
Lawrence Berkeley National Laboratory

Emiley Eloe-Fadrosh
DOE Joint Genome Institute

Kjiersten Fagnan
DOE Joint Genome Institute

Ee-Been Goh
Zymergen, Inc.

Kerstin Kleese van Dam
Brookhaven National Laboratory

Mary Maxon
Schmidt Futures

Rick Stevens
Argonne National Laboratory

Shinjae Yoo
Brookhaven National Laboratory

Dawn Adin, observer
U.S. Department of Energy

Wayne Kontur, observer
U.S. Department of Energy

3-2 AI/ML Approaches

Carlos Soto, group leader
Brookhaven National Laboratory

Kristofer Bouchard
Lawrence Berkeley National Laboratory

Mary J. Dunlop
Boston University

Daniel Jacobson
Oak Ridge National Laboratory

Lee Ann McCue
Pacific Northwest National Laboratory

Sue Rhee
Carnegie Institution for Science

Arti Singh
Iowa State University

Marinka Zitnik
Harvard University

Resham Kulkarni, observer
U.S. Department of Energy

Ramana Madupu, observer
U.S. Department of Energy

Catherine Ronning, observer
U.S. Department of Energy

3-3 AI/ML Approaches

Costas Maranas, group leader
The Pennsylvania State University

Héctor García Martin
Lawrence Berkeley National Laboratory

Chris Mungall
Lawrence Berkeley National Laboratory

Dmitry Grapov
Amyris

Lydia Kavraki
Rice University

Nina Lin
University of Michigan

Christopher Long
Ginkgo Bioworks, Inc.

Peter St. John
National Renewable Energy Laboratory

Huimin Zhao
University of Illinois, Urbana-Champaign

R. Todd Anderson, observer
U.S. Department of Energy

Boris Wawrik, observer
U.S. Department of Energy

3-4 AI/ML Approaches

Neeraj Kumar, group leader
Pacific Northwest National Laboratory

Frank Alexander
Brookhaven National Laboratory

Qun Liu
Brookhaven National Laboratory

Baskar Ganapathysubramanian
Iowa State University

Nathan Hillson
Lawrence Berkeley National Laboratory

James Sethian
University of California, Berkeley

Matthew Hudson
University of Illinois, Urbana-Champaign

Rachel Switzky
University of Illinois, Urbana-Champaign

Shin-Han Shiu
Michigan State University

Pablo Rabinowicz, observer
U.S. Department of Energy

Amy Swain, observer
U.S. Department of Energy

3-5 AI/ML Approaches

Ben Brown, group leader
Lawrence Berkeley National Laboratory

Gyorgy Babnigg
Argonne National Laboratory

Cees Haringa
Delft University of Technology

Corey Hudson
Sandia National Laboratories

Adam Perer
Carnegie Mellon University

Deepti Tanjore
Lawrence Berkeley National Laboratory

Gina Tourassi
Oak Ridge National Laboratory

Bobbie-Jo Webb-Robertson
Pacific Northwest National Laboratory

Gayle Bentley, observer
U.S. Department of Energy

Jay Fitzgerald, observer
U.S. Department of Energy

Paul Sammak, observer
U.S. Department of Energy

Session 4: Data and Compute Infrastructure

4-1 Large-Scale Experimental Facilities

Lee Ann McCue, group leader
Pacific Northwest National Laboratory

Matthew Hudson
University of Illinois, Urbana-Champaign

Rachel Switzky
University of Illinois, Urbana-Champaign

Daniel Jacobson
Oak Ridge National Laboratory

Gina Tourassi
Oak Ridge National Laboratory

Kerstin Kleese van Dam
Brookhaven National Laboratory

James Sethian
University of California, Berkeley

Arti Singh
Iowa State University

Rick Stevens
Argonne National Laboratory

R. Todd Anderson, observer
U.S. Department of Energy

Amy Swain, observer
U.S. Department of Energy

4-2 Automation

Ben Brown, group leader
Lawrence Berkeley National Laboratory

Shinjae Yoo, group leader
Brookhaven National Laboratory

Gyorgy Babnigg
Argonne National Laboratory

Emiley Eloe-Fadrosh
DOE Joint Genome Institute

Qun Liu
Brookhaven National Laboratory

Ee-Been Goh
Zymergen, Inc.

Dmitry Grapov
Amyris

Mary Maxon
Schmidt Futures

Arvind Ramanathan
Argonne National Laboratory

Resham Kulkarni, observer
U.S. Department of Energy

Boris Wawrik, observer
U.S. Department of Energy

4-3 Laboratory-Based Research

Bobbie-Jo Webb-Robertson, group leader
Pacific Northwest National Laboratory

Mary J. Dunlop
Boston University

Corey Hudson
Sandia National Laboratories

Costas Maranas
The Pennsylvania State University

Adam Perer
Carnegie Mellon University

Sue Rhee
Carnegie Institution for Science

Shin-Han Shiu
Michigan State University

Huimin Zhao
University of Illinois, Urbana-Champaign

Marinka Zitnik
Harvard University

Gayle Bentley, observer
U.S. Department of Energy

Catherine Ronning, observer
U.S. Department of Energy

4-4 Computational Science

Chris Mungall, group leader
Lawrence Berkeley National Laboratory

Frank Alexander
Brookhaven National Laboratory

Carlos Soto
Brookhaven National Laboratory

Kjiersten Fagnan
DOE Joint Genome Institute

Baskar Ganapathysubramanian
Iowa State University

Nathan Hillson
Lawrence Berkeley National Laboratory

Lydia Kavraki
Rice University

Peter St. John
National Renewable Energy Laboratory

Dawn Adin, observer
U.S. Department of Energy

Wayne Kontur, observer
U.S. Department of Energy

Ramana Madupu, observer
U.S. Department of Energy

Paul Sammak, observer
U.S. Department of Energy

4-5 Biological System Design and Control

Héctor García Martin, group leader
Lawrence Berkeley National Laboratory

Adam Arkin
Lawrence Berkeley National Laboratory

Kristofer Bouchard
Lawrence Berkeley National Laboratory

Deepti Tanjore
Lawrence Berkeley National Laboratory

Cees Haringa
Delft University of Technology

Neeraj Kumar
Pacific Northwest National Laboratory

Nina Lin
University of Michigan

Christopher Long
Ginkgo Bioworks, Inc.

Jay Fitzgerald, observer
U.S. Department of Energy

Pablo Rabinowicz, observer
U.S. Department of Energy

Session 5: Community Development

5-1 Education

Rachel Switzky, group leader
University of Illinois, Urbana-Champaign

Emiley Eloe-Fadrosh
DOE Joint Genome Institute

Matthew Hudson
University of Illinois, Urbana-Champaign

Daniel Jacobson
Oak Ridge National Laboratory

Gina Tourassi
Oak Ridge National Laboratory

Kerstin Kleese van Dam
Brookhaven National Laboratory

Lee Ann McCue
Pacific Northwest National Laboratory

James Sethian
University of California, Berkeley

Arti Singh
Iowa State University

Rick Stevens
Argonne National Laboratory

R. Todd Anderson, observer
U.S. Department of Energy

Amy Swain, observer
U.S. Department of Energy

5-2 Outreach

Ee-Been Goh, group leader
Zymergen, Inc.

Gyorgy Babnigg
Argonne National Laboratory

Arvind Ramanathan
Argonne National Laboratory

Ben Brown
Lawrence Berkeley National Laboratory

Dmitry Grapov
Amyris

Qun Liu
Brookhaven National Laboratory

Mary Maxon
Schmidt Futures

Shinjae Yoo
Brookhaven National Laboratory

Resham Kulkarni, observer
U.S. Department of Energy

Boris Wawrik
U.S. Department of Energy

5-3 Workforce Development

Chris Mungall, group leader
Lawrence Berkeley National Laboratory

Frank Alexander
Brookhaven National Laboratory

Mary J. Dunlop
Boston University

Corey Hudson
Sandia National Laboratories

Costas Maranas
The Pennsylvania State University

Shin-Han Shiu
Michigan State University

Sue Rhee
Carnegie Institution for Science

Bobbie-Jo Webb-Robertson
Pacific Northwest National Laboratory

Huimin Zhao
University of Illinois, Urbana-Champaign

Marinka Zitnik
Harvard University

Gayle Bentley, observer
U.S. Department of Energy

Catherine Ronning, observer
U.S. Department of Energy

5-4 Partnerships with Other Funding Agencies

Kjiersten Fagnan, group leader
DOE Joint Genome Institute

Baskar Ganapathysubramanian
Iowa State University

Nathan Hillson
Lawrence Berkeley National Laboratory

Lydia Kavraki
Rice University

Adam Perer
Carnegie Mellon University

Carlos Soto
Brookhaven National Laboratory

Peter St. John
National Renewable Energy Laboratory

Wayne Kontur, observer
U.S. Department of Energy

Ramana Madupu, observer
U.S. Department of Energy

Paul Sammak, observer
U.S. Department of Energy

5-5 Social Responsibilities/Ethics

Nina Lin, group leader
University of Michigan

Adam Arkin
Lawrence Berkeley National Laboratory

Kristofer Bouchard
Lawrence Berkeley National Laboratory

Héctor García Martin
Lawrence Berkeley National Laboratory

Deepti Tanjore
Lawrence Berkeley National Laboratory

Cees Haringa
Delft University of Technology

Neeraj Kumar
Pacific Northwest National Laboratory

Christopher Long
Ginkgo Bioworks, Inc.

Dawn Adin, observer
U.S. Department of Energy

Jay Fitzgerald, observer
U.S. Department of Energy

Pablo Rabinowicz, observer
U.S. Department of Energy

Appendix D

Workshop Participants and Position Papers

Organizing Committee

Huimin Zhao (Chair)
University of Illinois,
Urbana-Champaign

Nathan Hillson (Co-Chair)
Lawrence Berkeley National
Laboratory

Kerstin Kleese van Dam (Co-Chair)
Brookhaven National Laboratory

Deepti Tanjore (Co-Chair)
Lawrence Berkeley National
Laboratory

Resham Kulkarni
DOE Office of Science

R. Todd Anderson
DOE Office of Science

Jay Fitzgerald
DOE Office of Energy Efficiency and
Renewable Energy

Gayle Bentley
DOE Office of Energy Efficiency and
Renewable Energy

Wayne Kontur
DOE Office of Science

Ramana Madupu
DOE Office of Science

Pablo Rabinowicz
DOE Office of Science

Participants

Frank Alexander
Brookhaven National Laboratory

Adam Arkin
Lawrence Berkeley National Laboratory

Gyorgy Babnigg
Argonne National Laboratory

Kristofer Bouchard
Lawrence Berkeley National Laboratory

Ben Brown
Lawrence Berkeley National Laboratory

Mary J. Dunlop
Boston University

Emiley Eloe-Fadrosh
DOE Joint Genome Institute

Kjiersten Fagnan
DOE Joint Genome Institute

Baskar Ganapathysubramanian
Iowa State University

Héctor García Martin
Lawrence Berkeley National Laboratory

Ee-Been Goh
Zymergen, Inc.

Dmitry Grapov
Amyris

Cees Haringa
Delft University of Technology

Corey Hudson
Sandia National Laboratories

Matthew Hudson
University of Illinois, Urbana-Champaign

Daniel Jacobson
Oak Ridge National Laboratory

Lydia Kavraki
Rice University

Neeraj Kumar
Pacific Northwest National Laboratory

Nina Lin
University of Michigan

Qun Liu
Brookhaven National Laboratory

Christopher Long
Ginkgo Bioworks, Inc.

Costas Maranas
The Pennsylvania State University

Mary Maxon
Schmidt Futures

Lee Ann McCue
Pacific Northwest National Laboratory

Chris Mungall
Lawrence Berkeley National Laboratory

Adam Perer
Carnegie Mellon University

Arvind Ramanathan
Argonne National Laboratory

Sue Rhee
Carnegie Institution for Science

James Sethian
University of California, Berkeley

Shin-Han Shiu
Michigan State University

Arti Singh
Iowa State University

Carlos Soto
Brookhaven National Laboratory

Peter St. John
National Renewable Energy Laboratory

Rick Stevens
Argonne National Laboratory

Rachel Switzky
University of Illinois, Urbana-Champaign

Gina Tourassi
Oak Ridge National Laboratory

Bobbie-Jo Webb-Robertson
Pacific Northwest National Laboratory

Shinjae Yoo
Brookhaven National Laboratory

Marinka Zitnik
Harvard University

Position Papers

“The Data, Computing, and Experimental Infrastructures Needed to Integrate AI/ML Approaches into Biological Research: Developing Self-Driving Laboratories for AI-Driven Science,” by Gyorgy Babnigg, Casey Stone, Doga Ozgulbas, Rafael Vescovi, Dion Antonopoulos, Arvind Ramanathan, Thomas Brettin.

“Microfluidics Self-Driving Labs for Systematic TRY Improvement in Synthetic Biology,” by Héctor García Martin.

“AI for Crop Improvement with Reduced Input for Future Environments,” by Matthew Hudson.

“State of the Art Bioreactor Operation: Self-Driving Bioreactor Adapting with Biological Changes Position Paper,” by Deepti Tanjore.

“Image Analysis for Quantifying Biofuel Production in Single Cells,” by Mary J. Dunlop.

“AI/ML for Integration of Multi-Scale and Multi-Modal Bioimaging Data for Bioenergy Research,” by Qun Liu, Xiao Zhang, Yuewei Lin.

“Large Language Model (LLM)-Enabled Knowledge Base for Systems & Synthetic Biology,” by Carlos Soto.

“Biological Applications of Optimal Prior Development and Transfer Learning,” by Francis J. Alexander.

“AI-Enabled Data Integration and Fusion for Optimal Bioenergy/Bio-product Design,” by Arvind Ramanathan, Alexander Brace, Thomas Brettin, Austin Clyde, Gautham Dharuman, Ian Foster, Michael Irvin, Carla Mann, Priyanka Setty, Rick Stevens, Max Zvyagin.

Appendix E

References

Beal, J., and M. Rogers. 2020. "Levels of Autonomy in Synthetic Biology Engineering," *Molecular Systems Biology* **16**(12), e10019. DOI:10.15252/msb.202010019.

Benegas, G., et al. 2022. "DNA Language Models are Powerful Zero-Shot Predictors of Non-Coding Variant Effects," *bioRxiv*, preprint. DOI:10.1101/2022.08.22.504706.

Berliner, A. J., et al. 2022. "Space Bioprocess Engineering on the Horizon," *Communications Engineering* **1**(13). DOI:10.1038/s44172-022-00012-9.

Bevan, N., et al. 2019. "Quantifying Cell Subsets and Heterogeneity in Living Cultures Using Real Time Live-Cell Analysis," *Cancer Research* **79**(13) Supplement, 2156. DOI:10.1158/1538-7445.AM2019-2156.

Biswas, S., et al. 2021. "Low-*N* Protein Engineering with Data-Efficient Deep Learning," *Nature Methods* **18**, 389–396. DOI:10.1038/s41592-021-01100-y.

BNL. 2021. "Physics on Autopilot: Brookhaven National Lab Applies AI to Make Big Experiments Autonomous," Brookhaven National Laboratory. 10 November 2021. bnl.gov/newsroom/news.php?a=219206

Brown, T. B., et. al. 2020. "Language Models are Few Shot Learners," *arXiv*, 205.14165v4, preprint. DOI:10.48550/arXiv.2005.14165.

Carbonell, P., et al. 2019. "Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation," *ACS Synthetic Biology* **8**(7), 1474–1477. DOI:10.1021/acssynbio.8b00540.

Cheng, C.-Y., et al. 2021. "Evolutionarily Informed Machine Learning Enhances the Power of Predictive Gene-to-Phenotype Relationships," *Nature Communications* **12**, 5627. DOI:10.1038/s41467-021-25893-w.

Crater, J. S., and J. C. Lievense. 2018. "Scale-Up of Industrial Microbial Processes," *FEMS Microbiology Letters* **365**(13), fny138. DOI:10.1093/femsle/fny138.

Cho, H., et al. 2016. "Compact Integration of Multi-Network Topology for Functional Analysis of Genes," *Cell Systems* **3**(6), 540-548.e5. DOI:10.1016/j.cels.2016.10.017.

Culley, C., et al. 2020. "A Mechanism-Aware and Multiomic Machine-Learning Pipeline Characterizes Yeast Cell Growth," *PNAS* **117**(31), 18869–18879. DOI:10.1073/pnas.2002959117.

Dijk, A. D. J. van, et al. 2021. "Machine Learning in Plant Science and Plant Breeding," *iScience* **24**(1), 101890. DOI:10.1016/j.isci.2020.101890.

DiMucci, D., et al. 2018. "Machine Learning Reveals Missing Edges and Putative Interaction Mechanisms in Microbial Ecosystem Networks," *mSystems* **3**(5), e00181-18. DOI:10.1128/mSystems.00181-18.

GFISCO. 2019. "FAIR Principles," GO FAIR International Support and Coordination Office. go-fair.org/fair-principles/

Greener, J. G., et al. 2022. "A Guide to Machine Learning for Biologists," *Nature Reviews Molecular Cell Biology* **23**(1), 40–55. DOI:10.1038/s41580-021-00407-0.

Greenhalgh, J. C., et al. 2021. "Machine Learning-Guided Acyl-ACP Reductase Engineering for Improved *In Vivo* Fatty Alcohol Production," *Nature Communications* **12**(5825). DOI:10.1038/s41467-021-25831-w.

HamediRad, M., et al. 2019. "Towards a Fully-Automated Algorithm-Driven Platform for Biosystems Design," *Nature Communications* **10**(5150). DOI:10.1038/s41467-019-13189-z.

Haringa, C., et al. 2018. "Computational Fluid Dynamics Simulation of an Industrial *P. chrysogenum* Fermentation with a Coupled 9-Pool Metabolic Model: Towards Rational Scale-Down and Design Optimization," *Chemical Engineering Science* **175**, 12–24. DOI:10.1016/j.ces.2017.09.020.

Hernández Medina, R., et al., et al. 2022. "Machine Learning and Deep Learning Applications in Microbiome Research," *ISME Communications* **2**(98). DOI:10.1038/s43705-022-00182-9.

Hie, B. L., and K. K. Yang. 2022. "Adaptive Machine Learning for Protein Engineering," *Current Opinion in Structural Biology* **72**, 145–152. DOI:10.1016/j.sbi.2021.11.002.

Hsu, C., et al. 2022. "Learning Protein Fitness Models from Evolutionary and Assay-Labeled Data," *Nature Biotechnology* **40**, 1114–1122. DOI:10.1038/s41587-021-01146-5.

Hsu, C. S., and P. R. Robinson, Eds. 2006. *Practical Advances in Petroleum Processing*, Springer-Verlag, New York. DOI:10.1007/978-0-387-25789-1.

Hu, S., et al. 2002. "Combine Molecular Modeling with Optimization to Stretch Refinery Operation," *Industrial and Engineering Chemistry Research* **41**(4), 825–841. DOI:10.1021/ie0010215.

Humphrey, A. 1998. "Shake Flask to Fermentor: What Have We Learned?" *Biotechnology Progress* **14**(1), 3–7. DOI:10.1021/bp970130k.

Iwai, K., et al. 2022. "Scalable and Automated CRISPR-Based Strain Engineering Using Droplet Microfluidics," *Microsystems and Nanoengineering* **8**(31). DOI:10.1038/s41378-022-00357-3.

Ji, W., et al. 2021. "Stiff-PINN: Physics-Informed Neural Network for Stiff Chemical Kinetics," *The Journal of Physical Chemistry A* **125**, 8098–8106. DOI:10.1021/acs.jpca.1c05102.

Jumper, J., et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* **596**, 583–589. DOI:10.1038/s41586-021-03819-2.

Kavvas, E. S., et al. 2020. "A Biochemically-Interpretable Machine Learning Classifier for Microbial GWAS," *Nature Communications* **11**, 2580. DOI:10.1038/s41467-020-16310-9.

Kim, M., et al. 2016. "Multi-Omics Integration Accurately Predicts Cellular State in Unexplored Conditions for *Escherichia coli*," *Nature Communications* **7**, 13090. DOI:10.1038/ncomms13090.

Lee, H., et al. 2019. "DeepDriveMD: Deep-Learning Driven Adaptive Molecular Simulations for Protein Folding," *arXiv*, preprint. DOI:10.48550/arXiv.1909.07817.

Li, G., et al. 2019. "Can Machine Learning Revolutionize Directed Evolution of Selective Enzymes?" *Advanced Synthesis and Catalysis* **361**, 2377–2386. DOI:10.1002/adsc.201900149.

Luo, Y., et al. 2021. "ECNet Is an Evolutionary Context-Integrated Deep Learning Framework for Protein Engineering," *Nature Communications* **12**, 5743. DOI:10.1038/s41467-021-25976-8.

Ma, J., et al. 2018. "Using Deep Learning to Model the Hierarchical Structure and Function of a Cell," *Nature Methods* **15**, 290–298. DOI:10.1038/nmeth.4627.

Madani, A., et al. 2020. "ProGen: Language Modeling for Protein Generation," *arXiv*, preprint. DOI:10.1101/2020.03.07.982272.

Martin, H. G., et al. 2023. "Perspectives for Self-Driving Labs in Synthetic Biology," *Current Opinion in Biotechnology* **79**, 102881. DOI:10.1016/j.copbio.2022.102881.

Nguyen, N. D., and D. Wang. 2020. "Multiview Learning for Understanding Functional Multiomics," *PLOS Computational Biology* **16**, e1007677. DOI:10.1371/journal.pcbi.1007677.

Radivojević, T., et al. 2020. "A Machine Learning Automated Recommendation Tool for Synthetic Biology," *Nature Communications* **11**, 4879. DOI:10.1038/s41467-020-18008-4.

Reed, S., et al. 2022. "A Generalist Agent," *arXiv*, preprint. DOI:10.48550/arXiv.2205.06175.

Saadi, A. al-, et al. 2020. "IMPECCABLE: Integrated Modeling Pipeline for COVID Cure by Assessing Better Leads," *arXiv*, preprint. DOI:10.48550/arXiv.2010.06574.

Salis, H. M., et al. 2009. "Automated Design of Synthetic Ribosome Binding Sites to Control Protein Expression," *Nature Biotechnology* **27**, 946–950. DOI:10.1038/nbt.1568.

Sanchez-Lengeling, B., and A. Aspuru-Guzik. 2018. "Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering" *Science* **361**(6400), 360–365. DOI:10.1126/science.aat2663.

Senne de Oliveira Lino, F., et al. 2021. "Complex Yeast–Bacteria Interactions Affect the Yield of Industrial Ethanol Fermentation," *Nature Communications* **12**, 1498. DOI:10.1038/s41467-021-21844-7.

Stein, H. S., and J. M. Gregoire. 2019. "Progress and Prospects for Accelerating Materials Science with Automated and Autonomous Workflows," *Chemical Science* **10**, 9640–9649. DOI:10.1039/C9SC03766G.

Stocks, S. M., 2013. "Ch. 7: Industrial Enzyme Production for the Food and Beverage Industries: Process Scale Up and Scale Down." In *Microbial Production of Food Ingredients, Enzymes and Nutraceuticals*. Eds. McNeil, B., et al. Woodhead Publishing, Cambridge, Mass. elsevier.com/books/microbial-production-of-food-ingredients-enzymes-and-nutraceuticals/mcneil/978-0-85709-343-1

U.S. DOE. 2022a. *AI@DOE: Interim Executive Report*. U.S. Department of Energy Office of Science. DOI:10.2172/1872103.

U.S. DOE. 2022b. *Artificial Intelligence for Earth System Predictability (AI4ESP): 2021 Workshop Report*. U.S. Department of Energy Office of Science. publications.anl.gov/anlpubs/2022/09/177828.pdf

U.S. DOE. 2021a. *Designing for Deep Decarbonization: Accelerating the U.S. Bioeconomy Workshop Report*. U.S. Department of Energy Office of Energy Efficiency and Renewable Energy and Office of Science. biosciences.lbl.gov/wp-content/uploads/2021/12/21-BAO-3054-Designing-the-Bioeconomy-for-Deep-Decarbonization-Report_v5.pdf

U.S. DOE. 2021b. *National Virtual Biotechnology Laboratory: Report on Rapid R&D Solutions to the COVID-19 Crisis*. U.S. Department of Energy Office of Science. science.osti.gov/-/media/nvbl/pdf/NVBL_Technical_Report.pdf

U.S. DOE. 2020a. *AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science*. U.S. Department of Energy Office of Science. publications.anl.gov/anlpubs/2020/03/158802.pdf

U.S. DOE. 2020b. *Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee Report on AI/ML, Data-Intensive Science and High-Performance Computing*. U.S. Department of Energy Office of Science. science.osti.gov/-/media/ascr/ascac/pdf/meetings/202009/AI4Sci-ASCAC_202009.pdf

Vaishnav, E. D., et al. 2022. "The Evolution, Evolvability and Engineering of Gene Regulatory DNA," *Nature* **603**, 455–463. DOI:10.1038/s41586-022-04506-6.

Volk, M. J., et al. 2022. "Metabolic Engineering: Methodologies and Applications," *Chemical Reviews*. DOI:10.1021/acs.chemrev.2c00403. In review.

Volk, M. J., et al. 2020. "Biosystems Design by Machine Learning," *ACS Synthetic Biology* **9**(7), 1514–1533. DOI:10.1021/acssynbio.0c00129.

Wang, H., et al. 2022. "Chemical-Reaction-Aware Molecule Representation Learning," *International Conference on Learning Relations*. openreview.net/forum?id=6sh3pIzKS-

Wang, Y., et al. 2021. "Directed Evolution: Methodologies and Applications," *Chemical Reviews* **121**(20), 12384–12444. DOI:10.1021/acs.chemrev.1c00260.

Wehrs, M., et al., 2019. "Engineering Robust Production Microbes for Large-Scale Cultivation," *Trends in Microbiology* **27**(6), 524–537. DOI:10.1016/j.tim.2019.01.006.

White House. 2022. *Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure U.S. Bioeconomy*. whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/

Wittmann, B. J., et al. 2021. "Advances in Machine Learning for Directed Evolution," *Current Opinion in Structural Biology* **69**, 11–18. DOI:10.1016/j.sbi.2021.01.008.

Yang, K. K., et al. 2019. "Machine-Learning-Guided Directed Evolution for Protein Engineering," *Nature Methods* **16**, 687–694. DOI:10.1038/s41592-019-0496-6.

Zampieri, G., et al. 2019. "Machine and Deep Learning Meet Genome-Scale Metabolic Modeling," *PLOS Computational Biology* **15**, e1007084. DOI:10.1371/journal.pcbi.1007084.

Zelezniak, A., et al. 2018. "Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts," *Cell Systems* **7**(3), 269-283.e6. DOI:10.1016/j.cels.2018.08.001.

Zhang, J., et al. 2020. "Combining Mechanistic and Machine Learning Models for Predictive Engineering and Optimization of Tryptophan Metabolism," *Nature Communications* **11,** 4880. DOI:10.1038/s41467-020-17910-1.

Appendix F

Acronyms and Abbreviations

AI	artificial intelligence
AMBER	Artificial Intelligence and Machine Learning for Bioenergy Research
BER	DOE Biological and Environmental Research program
BETO	DOE Bioenergy Technologies Office
BioBERT	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
bpsA	blue-pigment synthetase gene
C1 compounds	one-carbon molecules
CABBI	DOE's Center for Advanced Bioenergy and Bioproducts Innovation
Cas	CRISPR-associated protein
CNN	convolutional neural network
$\mathbf{CO_2}$	carbon dioxide
COVID-19	coronavirus disease 2019
CRISPR	clustered regularly interspaced short palindromic repeats
DBTL	design-build-test-learn cycle
DEIA	diversity, equity, inclusion, and accessibility
DOE	U.S. Department of Energy
DNABERT	Bidirectional Encoder Representations from Transformers for DNA sequence analysis
DSP	downstream processing
ECNet	evolutionary context-integrated neural network
EERE	DOE Office of Energy Efficiency and Renewable Energy
ENDURABLE	Benchmark Datasets and AI/ML Models with Queryable Metadata
FAIR	findable, accessible, interoperable, reusable
galK	galactokinase gene
GCN	graph convolutional network
GEM	genome-scale metabolic model
GeneBERT	Bidirectional Encoder Representations from Transformers for gene regulatory analysis
GLaM	Google's Generalist Language Model
***gln*A**	glutamine synthetase gene
GMO	genetically modified organism
GPT	Generative Pre-Trained Transformer
HPC	high-performance computing
iCLEM	Introductory College-Level Experience in Microbiology
LaMDA	Google's Language Model for Dialogue Applications
LLM	large language model
MAC	Metabolic Allele Classifier
ML	machine learning
MSI	minority-serving institution
NLP	natural language processing
NMDC	National Microbiome Data Collaborative
NSLS-II	Brookhaven National Laboratory's National Synchrotron Light Source II
NVBL	National Virtual Biotechnology Laboratory
PaLM	Google's Pathways Language Model
SC	DOE Office of Science
TRL	technology readiness level
UCB	upper confidence bound