EBRC Engineering Biology Research Consortium

Response to DOE's Request for Information on Partnerships for Transformational AI Models

The Engineering Biology Research Consortium (EBRC) is a nonprofit, public-private partnership that brings together scientists, engineers, and industry leaders to advance the field of engineering biology to address national and global needs. EBRC's members include experts from over 90 universities and research institutes, alongside leaders from more than 25 companies, philanthropies, and other organizations. Working closely with partners across the engineering biology ecosystem, EBRC focuses on four key areas: Research Roadmapping, Policy & International Engagement, Education, and Security.

1. How should DOE best mobilize National Laboratories to partner with industry sectors within the United States to form a public-private consortium to curate the scientific data of the DOE across the National Laboratory complex so that the data is structured, cleaned, and preprocessed in a way that makes it suitable for use in AI models? How can DOE anonymize and desensitize data and/or make use of privacy-preserving AI training methods to enable AI model development using sensitive or proprietary data?

DOE should partner with the Center for AI Standards and Innovation (CAISI) and other expert stakeholders from academia and industry to develop AI-ready data standards. High-quality, reproducible, and interoperable data is critical for developing next-generation AI tools and models. In order to make use of the existing data within DOE and ensure that new data is AI-ready, clear and robust standards are necessary. To accomplish this, DOE should partner with NIST's CAISI to identify and engage experts in data science and AI model training from across academia and industry to develop these standards. These standards should include specifications for:

- Provenance
- Uncertainty metrics
- Rich metadata
- Domain-specific annotations
- Common ontologies
- Formatting
- Performance metrics

DOE should centralize as much data as possible while creating federated infrastructure for sensitive or proprietary data. In order for the data housed across all of DOE's National Laboratories to be maximally useful, a proportion of non-sensitive DOE data should be identified by engaging stakeholders within the National Laboratories and incorporated into a centralized repository that is accessible to verified academic and private-sector researchers.. Centralization greatly improves the usability of the data and efficiency of its use, removing any latency associated with retrieving data from multiple sources. Sensitive, classified, and otherwise excluded DOE data as well as privately-owned proprietary data should be incorporated into a federated infrastructure managed by the DOE and made available through a monitored managed access system. Federation greatly limits the complexity of models that can be trained on such a system. However, a federated network of data is advantageous for governance, tracking provenance, distribution of maintenance cost and labor, and overall security. By federating rather than centralizing, the institutions that originally generated the data maintain control over their own data, which allows each institution to ensure compliance with legal constraints and maintain security around sensitive data. Centralization also increases the risk of single-point security breaches. Therefore, DOE should create infrastructure for both centralized and federating data to maintain security and compliance but enable broader integration and utility.

DOE should coordinate with CAISI to develop risk-based categorization standards and privacy systems for sensitive data and models. While DOE should prioritize making data as accessible and open as possible, some data will require additional governance, particularly with respect to sensitive or proprietary data. Therefore, DOE should develop standards for performing risk-based categorization of data into discrete hazard levels, with an emphasis on dual-use potential and private personal health information. Each level should be connected to specific governance and risk mitigation mechanisms, such as tiered access control, encryption, secure compute environments, and cybersecurity standards. DOE should draw on existing best practices for securing these data, such as Crypt4GH for genomic data encryption. Additionally, DOE should develop mechanisms to support model-to-data training methods that would support model training on sensitive data without requiring data egress outside of a secured environment.

DOE should explore incentive structures and privacy systems that would motivate private companies to contribute data and/or models for use within the consortium. To maximize the mutual benefit of a public-private partnership, DOE should explore mechanisms by which private companies could be incentivized to contribute proprietary data and/or models to the consortium. A major concern for most private companies will be protecting their proprietary information from competitors. Therefore, DOE should explore mechanisms for federated learning that would allow private companies to retain sovereignty over their data by locally training models. DOE could incentivize participation by private companies by giving priority access to new models to companies that contribute their data. Similar structures could be constructed for incentivizing companies to share models by granting advance access to data.

2. How should DOE best structure the public-private consortium to enable activities across a range of scientific and technical disciplines, including partnerships with industry, to develop self-improving AI models for science and engineering using DOE's data, potentially in combination with data from other partners? Specific, related questions for consideration include but are not limited to:

   DOE should prioritize implementing general-purpose AI models for agentic AI workflows that enable self-improving model training and chaining tools together. Advances from leading developers like FutureHouse and Google have shown that the most effective way to combine general-purpose AI models with specialized models is by implementing general-purpose language and reasoning models as an agent that can call other specialized tools or models through APIs. This approach is advantageous over joint training or fine tuning general-purpose models because it lends itself to greater modularity, does not require multi-modal training, and limits the overall amount of training required. DOE could utilize existing APIs like FutureHouse's Aviary or utilize Model Context Protocol.

DOE should prioritize implementing self-improving AI models for biomanufacturing applications. The most effective use cases for self-improving AI models will be those that can be tested and evaluated through automated high-throughput experimentation. Automated high-throughput experimentation will enable generation of large amounts of data through experimentation that is more easily standardized and reproducible. These data can then be used iteratively to fine-tune models and to inform subsequent rounds of experimentation, both of which could also be automated. A critical component of this automated system for self-improving AI models will be designing assays that can be integrated into the automated experimentation platform for evaluating the performance of the model. While computational benchmarks of performance will also be valuable, performance on physical benchmarks should be prioritized and form the basis for improving the capabilities of the model. **Biomanufacturing is an area where methods for high-throughput experimentation are already well-established and could benefit greatly from self-improving AI models.** This would enable advances and breakthroughs in key areas such as novel fuel sources, biocatalysis, and novel materials among others.

3. How should DOE best provide AI models to the scientific community through programs and infrastructure making use of cloud technologies to accelerate innovation in discovery science and engineering for new energy technologies?

   DOE should create infrastructure to support a low-cost cloud computing environment that provides centralized access to an AI model toolbox that is connected to a federated network of AI-ready data. Some of the major barriers to the use and adoption of AI models and tools in scientific research are the lack of access to high-performance computing (HPC) infrastructure and the expertise needed to implement these technologies. While many well-resourced research institutions have dedicated HPC facilities and staff, smaller institutions would benefit from low-cost access to HPC infrastructure and personnel. Therefore, DOE should develop a low-cost cloud computing environment that could provide access to a toolbox of AI models and tools. **Priority should be given to researchers at institutions that lack HPC infrastructure.** This environment should also provide access to the data curated and generated by the consortium.

   DOE should continue to support existing training programs and create new training opportunities with a focus on expanding computational literacy with experimentalists and placing computational scientists within experimental research groups. Workforce development is a critical component for accelerating innovation at the intersection of AI and the life sciences. Current programs within DOE, like the Computational Science Graduate Fellowship, have seen great success in training more than 425 students in applying high-performance computing to science and engineering challenges. However, a critical gap remains in expanding the use and application of computational tools within experimental research groups, in particular with researchers at smaller institutions. To facilitate greater use and application of computational tools, DOE should create opportunities that support interdisciplinary training, within and outside DOE facilities. These opportunities should give experimentalists an opportunity to learn practical application-focused computational skills, like Python for data analysis, interpretation of model outputs, recognizing model limitations and failure modes, and basic command line skills. Similarly, opportunities should be created for computational scientists to work in experimental research groups, where they can identify application challenges, improve the usability of their tools, and provide advice for data generation, standardization, and formatting for use in model training.