Driving AIxBio Innovation Through Data and Standardization

June 2024

Introduction

If the United States wants to remain a leader in artificial intelligence (AI) and biotechnology (AIxBio), it must treat biological data as a strategic asset to support the next phase of AIxBio models. These models will rely on biological data sets of unprecedented scale, likely generated through high-throughput lab automation and new experimental methods. Biological data enable the use of AIxBio models, but advances in AIxBio are limited by the availability of appropriate and usable data.[1] Additionally, data standardization would enable the United States to combine data from across its robust and diverse life science ecosystem to further advance AIxBio and maximize its potential benefits. This white paper describes considerations for generating and standardizing biological data to support continued AIxBio research, development, and application.

> ***Biological data:*** Facts, statistics, or other pieces of information about the composition, characteristics, and behavior of organisms and their component parts.

Advancing AIxBio with biological data

Artificial intelligence models learn from patterns in the data on which they are trained. Typically, the more data the AI model learns from, the more accurate, precise, and nuanced the model is. Having sufficient biological data is important to ensure AIxBio models are available for use across different applications, as AIxBio models are likely to only succeed in applications in which the appropriate training data are available. Ensuring appropriate training data exist will require consideration of both the type of biological data (e.g., genomics or proteomics) and the quantity of data. For example, if an AIxBio model were intended to predict something about humans, the model would benefit from biological data from many different humans, but also many data points from each of those humans.

Applications of AI are limited by the existence of appropriate data. But unlike applications of AI where general internet data is abundant, biological data is currently a bottleneck for AIxBio. Biotechnology advances, such as DNA synthesis and synthetic biology, combined with robotics and lab automation could enable biological data generation at unprecedented scale.[2] The U.S. public and private sectors could strengthen investments in generating biological data to create a strategic asset that would advance AIxBio and unlock new potential within the biotechnology field. However, data generation alone may not be enough; collaboration among the U.S. government, industry, and academia to develop data standards will ensure that data are usable and searchable.

Data acquisition

Researchers can acquire biological data in multiple ways:

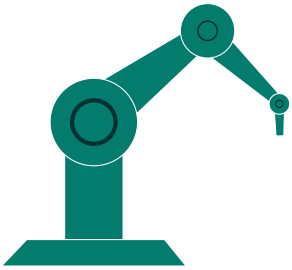
Curate existing data

Researchers identify and reuse existing, public data to fit their needs. This method is relatively low cost and does not require additional laboratory experimentation. It is, however, limited by the fact that the data may not perfectly suit the needs of the researcher. For example, data may exist from multiple academic laboratories, but each laboratory used different analytical instrumentation making it challenging to combine the data. This process is currently time-consuming because data require careful review and often reformatting. Researchers can retrieve and reformat data from various databases to create datasets suitable for training machine learning models.[3]

Manual laboratory measurements

Researchers perform experiments in life sciences laboratories to generate new data. Researchers conducting manual experiments have the most flexibility compared to other methods because of their full control and knowledge of what and how experiments are conducted.

Humans introduce a source of variation in the data: there can be person-to-person variation (e.g., each person correctly performs an action, but slightly differently) and individual variation (e.g., a single person correctly performs an action multiple times, but slightly differently each time). For example, the angle at which a liquid is pipetted (pictured above) could influence the amount of liquid used in an experiment which could influence the results and generated data.



High-throughput lab automation

Specially designed, purpose-built laboratories can leverage advances in robotics and automation to generate large amounts of data quickly. However, these laboratories require significant capital expenditures to establish and expertise to operate. Laboratory robots have distinct advantages, including the ability to conduct experiments with improved accuracy and precision, and require fewer resources to collect a single data point. Robotic systems can also move samples between different locations in the lab, including to measurement devices that generate data. Automated labs can also upload the data produced from the measurement devices directly to software that oversees the lab. Updating an automated lab to generate a new type of data can require purchasing new equipment, changing the laboratory layout, and optimizing experimental protocols. Automated labs are therefore less flexible than manual human measurement when adapting to generating additional types of data.

New experimental methods

Researchers are developing new laboratory methods to collect more data than ever before. These methods are a result of combining (1) biotechnology to create tens of thousands (or more) variations in a single test tube and (2) advances in measurement instrumentation that enable measuring all the variants at once.[4] Currently, executing these methods requires researchers with diverse sets of expertise and resource-intensive optimization. Their use must be considered on a case-by-case basis because the methods are not applicable to all types of biological data.

Why is data standardization important?

Standardizing the way in which data are recorded, reported, and stored in databases is necessary to maximize the full potential of biological data and AIxBio. Standardization also enables the researcher to search for, find, and use data resources more efficiently. Ideally, standardization enables the combined use of different biological datasets, regardless of the process used for generating each one. Through standardization, newly generated data will likely be more amenable for training AIxBio models, providing more opportunities for innovation.

What could be standardized?

Many aspects of AIxBio data would benefit from standardization. These include standardizing *what* data and metadata are reported, as well as *how* the data are reported and stored for access after generation. Additionally, the way in which the data and metadata are described could be standardized.

Sources

1 Notin et al. "Machine learning for functional protein design"
2 McKinsey. "From bench to bedside: Transforming R&D labs through automation"
3 Ruffolo et al. "Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences"
4 A-Alpha Bio. https://www.aalphabio.com/

For any questions about this white paper, or related work at the National Security Commission on Emerging Biotechnology, please contact us at ideas@biotech.senate.gov.

Staff at the National Security Commission on Emerging Biotechnology authored this paper with input from the expert Commissioners. The content and recommendations of this white paper do not necessarily represent positions officially adopted by the Commission.