Congressional Research Service
Informing the legislative debate since 1914

Artificial Intelligence in the Biological Sciences: Uses, Safety, Security, and Oversight

November 22, 2023

Congressional Research Service
https://crsreports.congress.gov
R47849

Congressional Research Service
Informing the legislative debate since 1914

Artificial Intelligence in the Biological Sciences: Uses, Safety, Security, and Oversight

SUMMARY

R47849

November 22, 2023

Todd Kuiken
Analyst in Science and Technology Policy

Artificial intelligence (AI) is a term generally thought of as computerized systems that work and react in ways commonly thought to require intelligence. AI technologies, methodologies, and applications can be used throughout the biological sciences and biology R&D, including in engineering biology (e.g., the application of engineering principles and the use of systematic design tools to reprogram cellular systems for a specific functional output). This has enabled research and development (R&D) advances across multiple application areas and industries. For example, AI can be used to analyze genomic data (e.g., DNA sequences) to determine the genetic basis of a particular trait and potentially uncover genetic markers linked with those traits. It has also been used in combination with biological design tools to aid in characterizing proteins (e.g., 3-D structure) and for designing new chemical structures that can enable specific medical applications, including for drug discovery. AI can also be used across the scientific R&D process, including the design of laboratory experiments, protocols to run certain laboratory equipment, and other "de-skilling" aspects of scientific research. The convergence of AI and other technologies associated with biology can lower technical and knowledge barriers and increase the number of actors with certain capabilities. These capabilities have potential for beneficial uses while at the same time raising certain biosafety and biosecurity concerns. For example, some have argued that using AI for biological design can be repurposed or misused to potentially produce biological and chemical compounds of concern.

Both AI and engineering biology are multidisciplinary fields that build on advances in multiple scientific disciplines and technical developments that each have associated benefits and risks. As they converge, those benefits and risks may compound and create unique uncertainties, challenge governance systems, and, in some instances, raise new biosafety and biosecurity concerns. The U.S. Intelligence Community's 2023 Annual Threat Assessment stated that the fields of AI and biotechnology are "being developed and are proliferating faster than companies and governments can shape norms, protect privacy, and prevent dangerous outcomes." AI's use in biology sits within a broader debate on how best to manage biosafety and biosecurity, including whether that use may lower technical and knowledge barriers and increase the likelihood of malign use. The potential for the use of AI in conjunction with biological design and other types of scientific research and experimentation has prompted recommendations from various groups on how to prevent the misuse of AI applications in biology and science more broadly. These include keeping a human "in the loop," controlling access to DNA sequences and synthesis capabilities, and governance mechanisms that restrict, or monitor, who can access certain AI applications and biological design tools. Numerous bills introduced in the 118th Congress address the risks, benefits, and strategic competitiveness of AI more generally, while H.R. 4704, S. 2399, and S. 2346 focus more specifically on AI and biological threats. Other policy considerations for Congress could include the following:

- Whether AI, and its use in biology and other scientific fields, should be regulated broadly across all use cases and areas of development, regulated on a case-by-case basis focused on particular application areas or end-use products, or whether current biosafety and biosecurity oversight frameworks are sufficient.
- Whether a broad risk management approach examining the R&D system as a whole is most appropriate, including at what stage oversight is warranted (e.g., basic research stage, prototype stage, or prior to release of a final product) and what entity might conduct such oversight (e.g., a federal agency or a system of self-governance incentives).
- Whether to provide an agency (or agencies) authority to conduct oversight of various aspects of AI, create a new agency with authority over AI, or authorize agencies under current law to enforce certain oversight responsibilities. Authorizing agencies under current law may require additional coordination amongst federal agencies in order to establish oversight responsibilities. Other types of self-governance incentives could also be examined.

Oversight and governance of the use of AI in the biological sciences involves potential benefits and risks associated with the biological design capabilities of AI models, as well as the feasibility of producing AI generated designs that pose possible biosecurity risks. Regulating AI broadly, or the use of AI in one area of R&D, could impact other areas of R&D in specific or unintended ways. For example, limiting access to AI models, restricting the types of data an AI model may be trained on, or limiting the capabilities an AI model is allowed to execute, could each impact a biological design tool's capability.

Contents

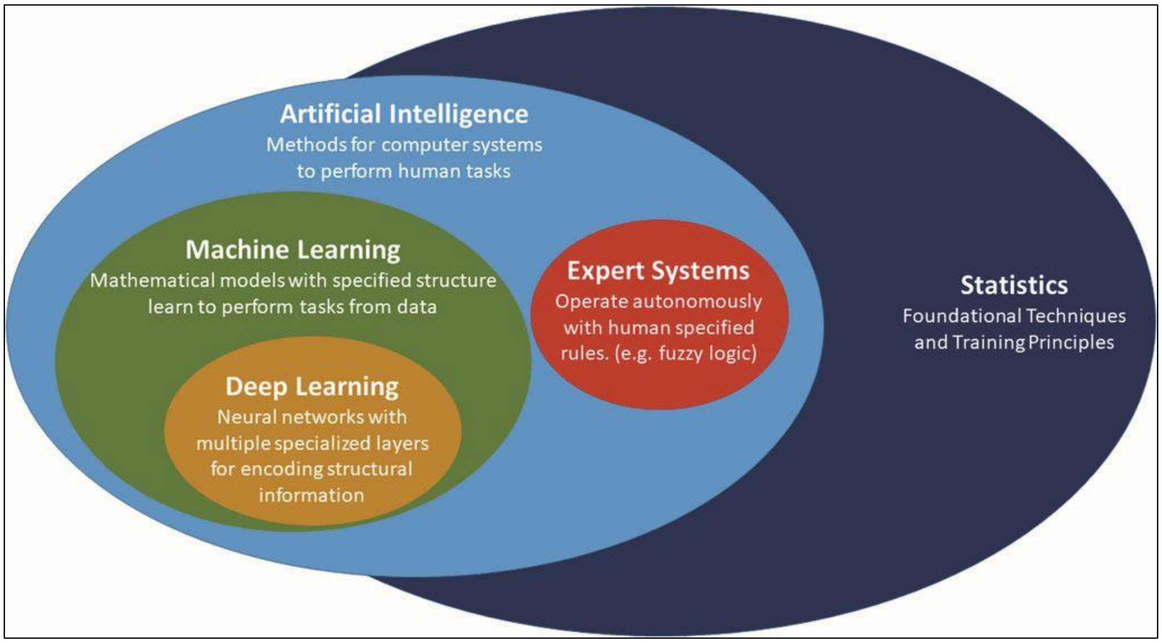
Figures

Contacts

Introduction[1]

Artificial Intelligence (AI) technologies, methodologies, and applications can be used throughout the biological sciences and biology R&D. AI can generally be thought of as a multidisciplinary field that studies and develops machines that work and react in ways commonly thought to require intelligence, such as the ability to learn, solve problems, and achieve goals under uncertain and varying conditions, with varying levels of autonomy. AI is not one thing; rather, AI systems can encompass a range of technologies, methodologies, and applications, such as natural language processing, robotics, and facial recognition. Common terms used in the field of AI include machine learning (ML), deep learning (DL), and neural networks (NN). ML is an umbrella term for training computer systems to solve complex problems normally requiring human intelligence, such as predictions, classification, and association, using a large amount of data, computer algorithms, statistical models, and other computing techniques.[2] DL is a subset of ML, using techniques such as multi-layer neural networks to improve the performance and accuracy of problem-solving, for example, in image recognition. An NN represents a set of algorithms and techniques that process data and extract information at multiple layers of computational nodes, based on the idea of how neurons in a nervous system signal to one another. It can be as simple as a single layer with a few nodes, or complex, containing many layers with millions of nodes. See **Figure 1**.

Figure 1. Relationship Between AI, ML, DL, Expert Systems, and Statistics


Artificial Intelligence
Methods for computer systems to perform human tasks
Machine Learning
Mathematical models with specified structure learn to perform tasks from data
Deep Learning
Neural networks with multiple specialized layers for encoding structural information
Expert Systems
Operate autonomously with human specified rules. (e.g. fuzzy logic)
Statistics
Foundational Techniques and Training Principles

Source: Sue Ellen Haupt, David John Gagne, and William W. Hsieh, et al., "The History and Practice of AI in the Environmental Sciences," *Bulletin of the American Meteorological Society*, vol. 103, no. 5 (2022). © American Meteorological Society. Used with permission.

[1] For a general introduction to AI and policy considerations, see CRS Report R47644, *Artificial Intelligence: Overview, Recent Advances, and Considerations for the 118th Congress*, by Laurie A. Harris.

[2] Adapted from Erik Brynjolfsson, Tom Mitchell, and Daniel Rock, "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?," AEA Papers and Proceedings, vol. 108 (May 1, 2018), pp. 43-47, https://dspace.mit.edu/bitstream/handle/1721.1/120302/pandp.20181019.pdf.

Machine learning has been used in the biological sciences, and science in general, for some time.[3] AI's ability to process large amounts of raw, unstructured data (e.g., DNA sequence data) has reduced the time and cost to conduct certain experiments in biology, enabled others types of experiments that previously were unattainable,[4] and contributed to the broader field of engineering biology. There is no universally agreed definition of engineering biology, but it is generally thought of as the application of engineering principles and the use of systematic design tools to enable the reprogramming of cellular systems at the genetic level for a specific functional output.[5] This approach has increased human ability to make direct changes at the cellular level and create novel genetic material (e.g., DNA and RNA) to obtain specific functions.

Engineering biology is a multidisciplinary field that leverages a broad set of sciences and technologies.[6] Additionally, engineering biology relies on and builds upon advances in other fields such as nanotechnology, robotics, and, increasingly, artificial intelligence (see **Figure 2**).[7] These technologies enable engineering biology researchers to read and write DNA. For example, sequencing technologies "read" DNA, while gene synthesis technologies can take sequence data and "write" DNA, turning the data into physical material which can then be designed or engineered for different purposes. Other terms such as synthetic biology, biotechnology, and biological sciences are used interchangeably with engineering biology in this report when discussed in relation to AI.

Policy concerns have been raised that the use of AI in the biological sciences could be repurposed or misused. For example, as part of a recent study, an AI model for drug development was retrained to design molecules for toxicity instead of designing against them. The study reported that in less than six hours the AI model generated 40,000 molecules that scored within their desired threshold of toxicity and bioactivity, which included the nerve agent VX, other known chemical warfare agents, and new molecules that were predicted to have even higher toxicities.[8]

This report focuses on the use of AI in the biological sciences, particularly engineering biology and biological design, as well as some aspects of laboratory automation and design. Biological design tools can be described as the tools and methods that enable the design and understanding of biological processes (e.g., DNA sequences/synthesis or the design of novel organisms). Such capabilities have potential positive impacts but at the same time raise certain biosafety and biosecurity concerns. AI may also have impacts on other fields of science, but those are beyond the scope of this report.[9] The report aims to inform Congress as it considers potential oversight of AI, either generally, or for specific scientific application areas.

[3] Joe G. Greener, Shaun M. Kandathil, and Lewis Moffat, et al., "A Guide to Machine Learning for Biologists," *Nature Reviews Molecular Cell Biology*, vol. 23 (2022).

[4] Abhaya Bhardwaj, Shristi Kishore, and Dhananjay K. Pandey, "Artificial Intelligence in Biological Sciences," *Life*, vol. 12, no. 1430 (2022).

[5] CRS Report R47265, *Synthetic/Engineering Biology: Issues for Congress*, by Todd Kuiken.

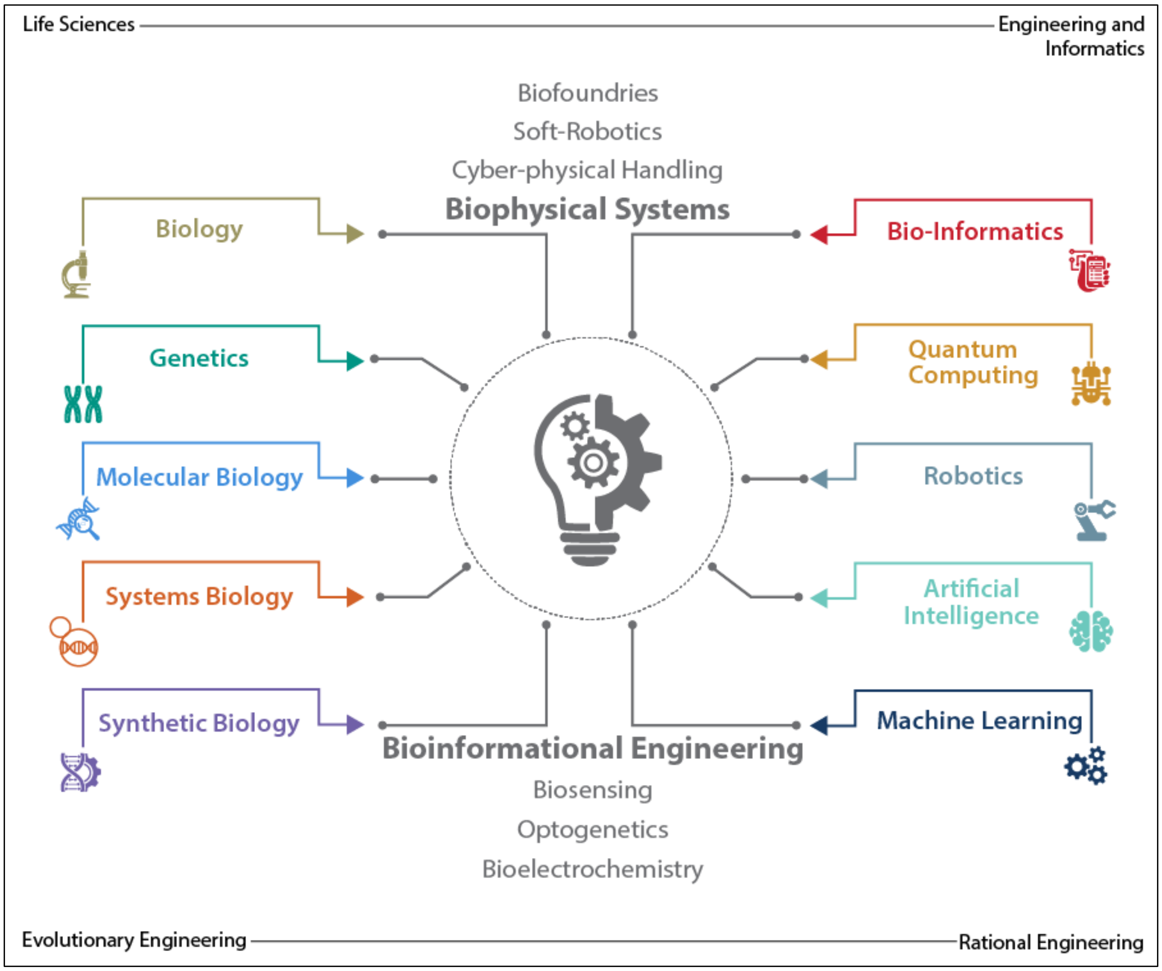
[6] Kent Redford, Thomas Brooks, and Nicholas Macfarlane, et al., *Genetic Frontiers for Conservation: An Assessment of Synthetic Biology and Biodiversity Conservation*, IUCN, Technical Assessment, Gland, Switzerland, 2019. National Academies of Sciences, Engineering, and Medicine, *Biodefense in the Age of Synthetic Biology*, Washington, DC, 2018, https://doi.org/10.17226/24890.

[7] Arlindo L. Oliveira, "Biotechnology, Big Data and Artificial Intelligence," *Biotechnology Journal*, vol. 14, no. 1800613 (2019).

[8] Fabio Urbina, Filippa Lentzos, and Cédric Invernizzi, et al., "Dual Use of Artificial-Intelligence-Powered Drug Discovery," *Nature Machine Intelligence*, vol. 4 (2022).

[9] For examples on how AI is impacting other areas of science, see "How Scientists Are Using Artificial Intelligence," *The Economist*, 2023, https://www.economist.com/science-and-technology/2023/09/13/how-scientists-are-using-artificial-intelligence.

Figure 2. Associated Sciences and Technologies Enabling Engineering Biology



Source: Thomas A. Dixon, Paul S. Freemont, and Richard A. Johnson, et al., "A Global Forum on Synthetic Biology: The Need for International Engagement," *Nature Communications*, vol. 13, no. 3516 (2022).

Note: Machine learning is typically considered a subfield of AI, but is a specific tool used in engineering biology.

Convergence of AI and Biological Design

AI and the biological sciences have increasingly converged, each building upon the other's capabilities to enable new research and development across multiple areas (see **Figure 3**).[10] The CEO and cofounder of DeepMind, an AI subsidiary of Google, recently said of biology:

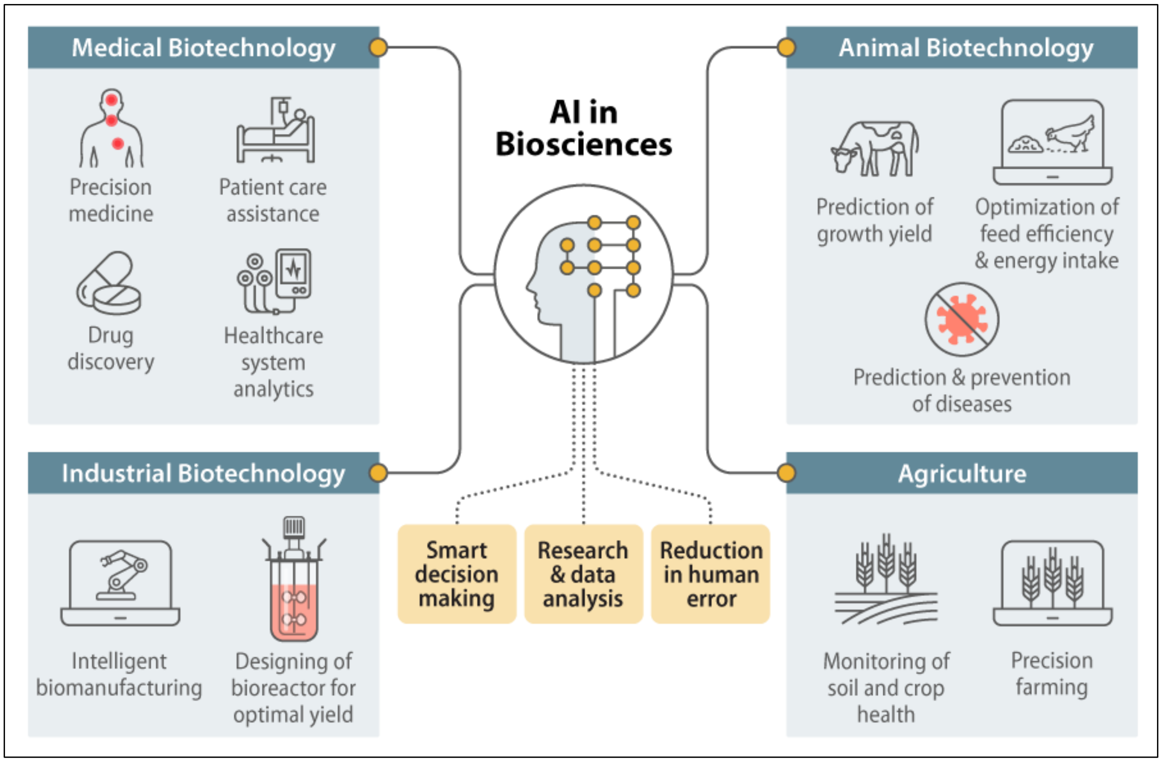
> At its most fundamental level, I think biology can be thought of as an information processing system, albeit an extraordinarily complex and dynamic one. Just as mathematics

[10] Abhaya Bhardwaj, Shristi Kishore, and Dhananjay K. Pandey, "Artificial Intelligence in Biological Sciences," *Life*, vol. 12, no. 1430 (2022).

> turned out to be the right description language for physics, biology may turn out to be the perfect type of regime for the application of AI.[11]

Both AI and engineering biology are multidisciplinary fields that rely on and build upon advances in other scientific disciplines and technology fields, such as nanotechnology, robotics, and increasingly big data (e.g., genetic sequence data). Each of these fields is itself a convergence of multiple sciences and technologies, so their impacts can combine to create new capabilities and concerns.[12] For example, each technology and scientific field in **Figure 2** has individual benefits and risks. As they converge, those benefits and risks can layer upon one another or create unique opportunities as well as uncertainties that may challenge governance systems, and, in some instances, raise new biosafety and biosecurity concerns not present in the individual fields considered separately.

Figure 3. Example Application Areas Where AI and Biological Sciences Converge


Medical Biotechnology
Precision medicine
Patient care assistance
Drug discovery
Healthcare system analytics
AI in Biosciences
Animal Biotechnology
Prediction of growth yield
Optimization of feed efficiency & energy intake
Prediction & prevention of diseases
Industrial Biotechnology
Intelligent biomanufacturing
Designing of bioreactor for optimal yield
Smart decision making
Research & data analysis
Reduction in human error
Agriculture
Monitoring of soil and crop health
Precision farming

Source: CRS, adapted from Abhaya Bhardwaj, Shristi Kishore, and Dhananjay K. Pandey, "Artificial Intelligence in Biological Sciences," *Life*, vol. 12, no. 1430 (2022).

The U.S. Intelligence Community's 2023 Annual Threat Assessment stated that the fields of AI and biotechnology are "being developed and are proliferating faster than companies and governments can shape norms, protect privacy, and prevent dangerous outcomes."[13] The report went on to identify genomic sequence data as a particular area of interest, pointing towards efforts by countries, universities, and private companies that have created, or are creating, centralized

[11] Demis Hassabis, quoted in Rob Toews, "The Next Frontier for Large Language Models Is Biology," *Forbes*, July 16, 2023, https://www.forbes.com/sites/robtoews/2023/07/16/the-next-frontier-for-large-language-models-is-biology/.

[12] John T. O'Brien and Cassidy Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology," *Health Security*, vol. 18, no. 3 (2020).

[13] Office of the Director of National Intelligence, *Annual Threat Assessment of the U.S. Intelligence Community*, 2023, https://www.dni.gov/files/ODNI/documents/assessments/ATA-2023-Unclassified-Report.pdf.

databases to collect, store, process, and analyze genetic data. The report further identified China's efforts to collect U.S. health and genomic data through its acquisitions of and investments in U.S. companies, as well as through cyberattacks. This analysis followed a 2021 assessment by the National Counterintelligence and Security Center suggesting that China understands that the collection and analysis of large genomic data sets from diverse populations can help foster new medical discoveries and cures with substantial commercial value, and can advance its AI and precision medicine industries.[14]

Importance of AI Related to Genetic Sequence Data and Biological Design

Some experts have suggested that AI is essential to analyzing the exponential growth of genetic sequence data.[15] According to the National Security Commission on Artificial Intelligence, "AI will be essential to fully understanding how genetic code interacts with biological processes."[16] AI has the ability to process large amounts of biological data (e.g., genetic sequence data), which can come from different biological sources. That capability is needed to understand complex biological systems[17] and has become a necessity for certain biological research, such as engineering biology.[18] For example, researchers can use AI to analyze genomic data sets to determine the genetic basis of a particular trait and potentially uncover genetic markers linked with that trait.[19] Different types of biological data can be utilized by AI and biological design tools: sequence data, molecular structure data, image data, time-series, and omics data.[20] See **Figure 4**. A limiting factor, however, is the quality and quantity of the biological data (e.g., DNA sequences) that the AI system is trained on.[21] For example, accurate identification of a particular species based on DNA requires reference sequences of sufficient quality to exist and be available. Individual databases have varying standards that govern access, as well as the type and quality of information they contain. The design, management, quality standards, and data protocols for

[14] National Counterintelligence and Security Center, *China's Collection of Genomic and Other Healthcare Data from America: Risks to Privacy and U.S. Economic and National Security*, 2021, https://www.dni.gov/files/NCSC/documents/SafeguardingOurFuture/NCSC_China_Genomics_Fact_Sheet_2021revision20210203.pdf.

[15] Zulema Udaondo, "Big Data and Computational Advancements for Next Generation of Microbial Biotechnology," *Microbial Biotechnology*, vol. 15, no. 1 (2022); Wardah S. Alharbi and Mamoon Rashid, "A Review of Deep Learning Applications in Human Genomics Using Next-Generation Sequencing Data," *Human Genomics*, vol. 16, no. 26 (2022).

[16] National Security Commission on Artificial Intelligence, *Final Report*, 2021, https://www.nscai.gov/2021-final-report/.

[17] Abhaya Bhardwaj, Shristi Kishore, and Dhananjay K. Pandey, "Artificial Intelligence in Biological Sciences," *Life*, vol. 12, no. 1430 (2022).

[18] Arlindo L. Oliveira, "Biotechnology, Big Data and Artificial Intelligence," *Biotechnology Journal*, vol. 14 (2019).

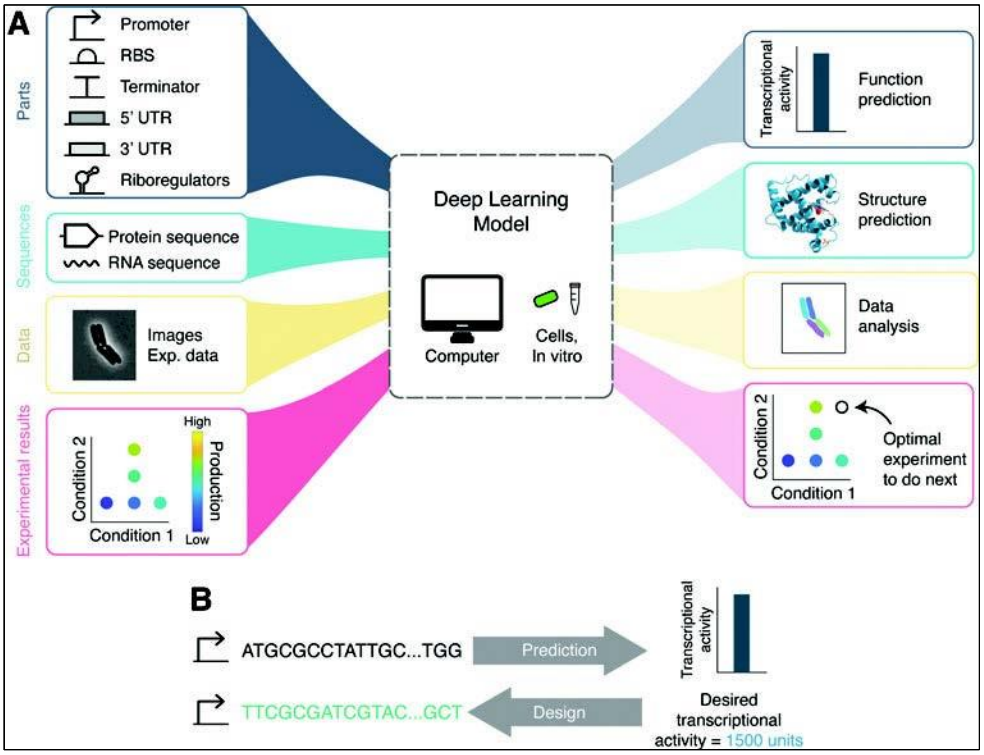
[19] Abhaya Bhardwaj, Shristi Kishore, and Dhananjay K. Pandey, "Artificial Intelligence in Biological Sciences," *Life*, vol. 12, no. 1430 (2022). A trait is a characteristic of an organism that is expressed by genes.

[20] William A.V. Beardall, Guy-Bart Stan, and Mary J. Dunlop, "Deep Learning Concepts and Applications for Synthetic Biology," *GEN Biotechnology*, vol. 1, no. 4 (2022). Omics data refers to various disciplines in biology whose names end in the suffix -omics, such as genomics, proteomics, metabolomics, etc.

[21] Fausto Artico, Arthur L Edge III, and Kyle Langham, "The Future of Artificial Intelligence for the BioTech Big Data Landscape," *Current Opinion in Biotechnology*, vol. 76 (2022). Training a model refers to providing a model with data to learn from, often called a training dataset. After a model is trained to recognize patterns from one dataset, some models can be provided with new data and still recognize patterns or predict results. Fine-tuning a model refers to training a previously trained model on new data, or otherwise adjusting an existing model. See CRS Report R47569, *Generative Artificial Intelligence and Data Privacy: A Primer*, by Kristen E. Busch.

reference databases can therefore affect the utility of a particular DNA sequence and its use with AI.[22]

Figure 4. Examples of AI Used in Engineering Biology



Source: William A.V. Beardall, Guy-Bart Stan, and Mary J. Dunlop, "Deep Learning Concepts and Applications for Synthetic Biology," *GEN Biotechnology*, vol. 1, no. 4 (2022).

Notes: "Parts" is a term used in engineering biology to describe a segment, or sequence, of DNA that encodes for a specific biology function.

Sequencing technologies have evolved rapidly, making it possible to sequence entire genomes more efficiently and at lower cost.[23] Gene synthesis technologies can take sequence data and "write" DNA, turning it into physical material which can then be designed or engineered for different purposes. The technologies enabling the ability to both read and write DNA is fundamental to engineering biology.[24]

DNA sequences, including human DNA sequences,[25] are stored in databases, many of which are publicly funded and freely accessible, while others are privately held. The volume of genetic sequence data has grown exponentially as sequencing technology has evolved (see **Figure 5** and

[22] CRS In Focus IF12285, *eDNA/eRNA: Scientific Value in What's Left Behind*, coordinated by Todd Kuiken.

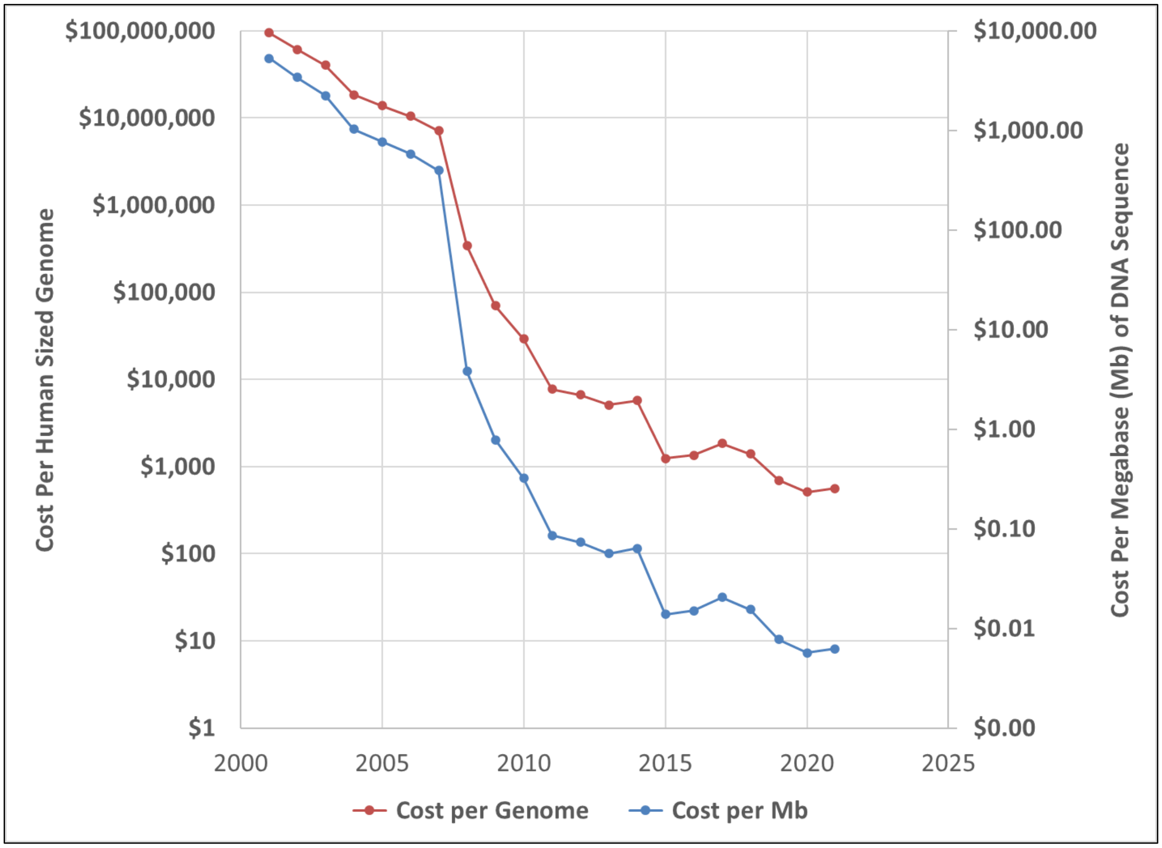
[23] For additional analysis on genetic sequence data see CRS In Focus IF12356, *Digital Biology: Implications of Genetic Sequencing*, by Todd Kuiken, and CRS In Focus IF12285, *eDNA/eRNA: Scientific Value in What's Left Behind*, coordinated by Todd Kuiken.

[24] DNA is made up of two linked strands that form a double helix. These strands are connected by base pairs consisting of the following bases: Adenine (A), Thymone (T), Cytosine (C), and Guanine (G). A connects with T and C connects with G.

[25] Genetic information is considered health information and protected under the Health Insurance Portability and Accountability Act (HIPAA). For additional information on genetic research, considerations with the use and misuse of genetic information, and relevant laws see CRS Report RL34584, *The Genetic Information Nondiscrimination Act of 2008 (GINA)*, by Amanda K. Sarata.

Figure 6). One analysis identified more than 1,700 databases incorporating data on genomics, protein sequences, protein structures, plants, and metabolic pathways, among others.[26] For example, the Protein Data Bank, a U.S.-funded data center, contains more than a terabyte of three-dimensional structure data for biological molecules, including proteins, DNA, and RNA.[27] While the Protein Data Bank is an open-source public database, other databases can be proprietary. For example, Gingko Bioworks claims to have more than 2 billion protein sequences in its proprietary database.[28] Public research groups can also produce large amounts of genetic sequence data. The Broad Institute claims to produce roughly 500 terabases (trillion bases) of genomic data per month.[29] There is great potential value in the aggregate volume of genetic datasets that can be collectively mined to discover and characterize relationships among genes.

Figure 5. Cost of DNA Sequencing over Time



Source: CRS analysis of data from Kris A. Wetterstrand, *DNA Sequencing Costs: Data*, National Institutes of Health, National Human Genome Research Institute, 2023, https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

Notes: The Y-axis is in logarithmic scale. A megabase (Mb) is a unit of measurement for DNA. One megabase = 1 million bases. The cost per genome assumes a genome size of 3,000 Mb.

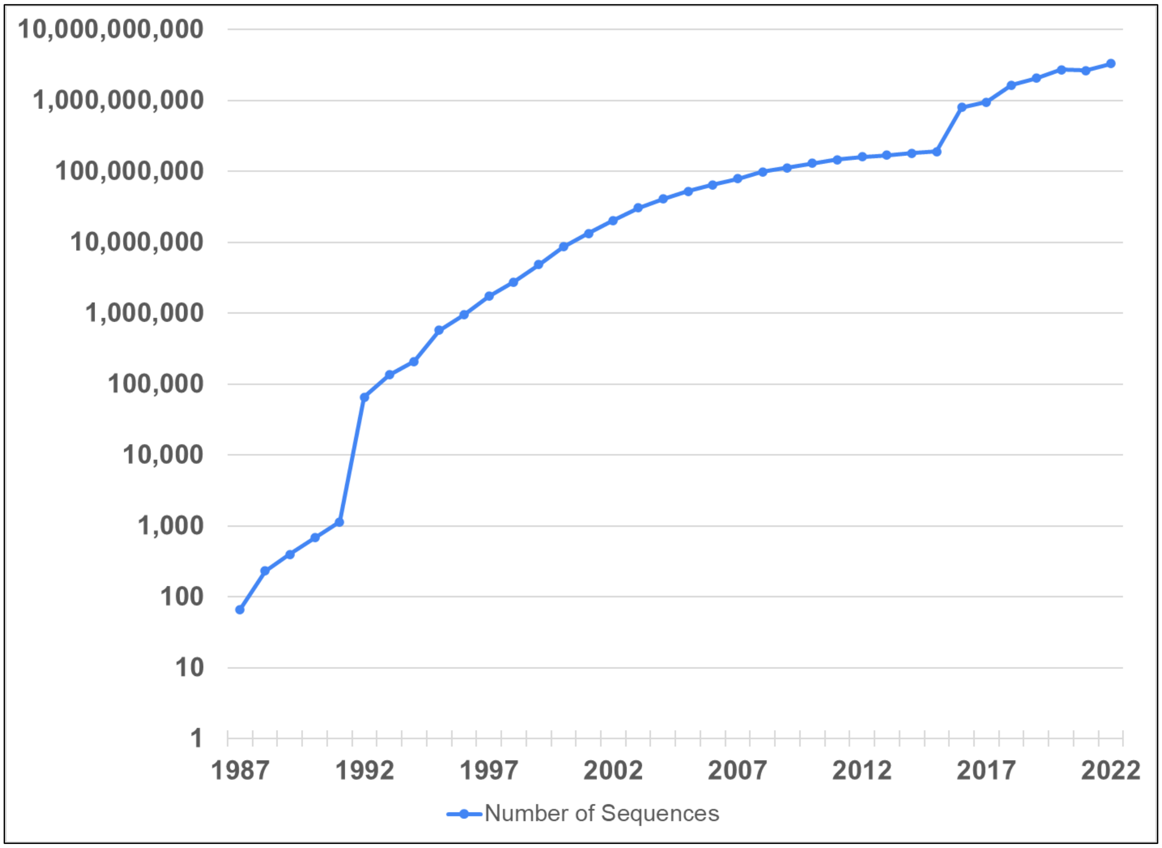
[26] Arlindo L. Oliveira, "Biotechnology, Big Data and Artificial Intelligence," *Biotechnology Journal*, vol. 14 (2019).

[27] RCSB Protein Data Bank, *About RCSB PDB: Enabling Breakthroughs in Scientific and Biomedical Research and Education*, https://www.rcsb.org/.

[28] Ginkgo Bioworks, "Google and Ginkgo: Foundry-Scale Data Meets AI," press release, 2023, https://www.ginkgobioworks.com/2023/08/29/google-and-ginkgo-foundry-scale-data-meets-ai/.

[29] Broad Institute, *Genomics*, 2023, https://www.broadinstitute.org/genomics.

Figure 6. Growth of Sequences in the International Nucleotide Sequence Database Collaboration



Source: CRS analysis of data from the International Nucleotide Sequence Database Collaboration (INSDC).

Notes: The Y-axis is in logarithmic scale. INSDC includes sequence data from the DNA Data Bank of Japan, the European Nucleotide Archive, and GenBank, the National Institutes of Health genetic sequence database.

The U.S. National Security Commission on Artificial Intelligence recommended that the U.S. fund and prioritize the development of a biobank containing a "wide range of high-quality biological and genetic data sets securely accessible by researchers." [30] The Commission further noted that the establishment of such a database containing a broad range of human, animal, and plant genomes would enhance and democratize biotechnology innovations and facilitate new levels of AI-enabled analysis of genetic data while also reducing U.S. researchers' reliance on foreign access to genomic databases for research.[31]

The availability of genetic data to train AI models, as well as decisions about the selection of genetic data for that purpose, can introduce bias. For example, training an AI model on datasets that emphasize or omit certain genetic traits can affect how that information is used and the types of applications developed, potentially privileging or disadvantaging certain populations. In addition, access to data and to AI models themselves may impact communities of differing socioeconomic status or other factors unequally.

[30] National Security Commission on Artificial Intelligence, *Final Report*, 2021, https://www.nscai.gov/2021-final-report/.

[31] Ibid. For additional information on federal initiatives related to DNA databases see CRS In Focus IF12285, *eDNA/eRNA: Scientific Value in What's Left Behind*, coordinated by Todd Kuiken.

AI-Enabled Biological Design Tools

Proteins are responsible for nearly every task of cellular life. Each has a unique amino acid sequence, which are the building blocks of proteins, and a complex three-dimensional physical structure.[32] Predicting the three-dimensional structure of proteins has been the focus of research for more than 50 years, with one of the first research papers on the subject published in 1973.[33] According to one paper, the structures of around 100,000 proteins have been deciphered; however, this is a fraction of the estimated billions of known protein sequences.[34] Despite the increase in genomic sequence data, scientists' ability to determine protein structures and reproduce them experimentally has remained a challenge. Understanding protein structures and how to design new proteins allows researchers to control a protein's stability and biochemical properties that can enable specific medical applications, including drug discovery.[35] The ability to predict protein structures can also impact other areas of biochemical engineering, including the identification of key protein residues and better estimates of optimal kinetic parameters, which are important for drug development.[36]

In 2021, an AI-based program called AlphaFold demonstrated the ability to predict protein structures even when no similar structures are known. The AlphaFold researchers also demonstrated that its predicted structures were accurate when compared to the experimental protein structure (i.e., physical material).[37] The AI system underpinning AlphaFold incorporated both physical and biological knowledge about protein structures and was trained on the Protein Data Bank, a U.S. federally funded open-source data center.[38] Another study in 2022 suggested that AI techniques, specifically deep network hallucination,[39] can be used to design new protein structures that are different from the sequences and structures of naturally occurring proteins.[40]

Research into designing new proteins typically involves protein structures being predicted based on previous protein structures. These predictions then have to be experimentally verified for accuracy (i.e., predicted proteins have to be synthesized and tested, typically in the bacterium *E. coli*).[41] Most such predictions in the past have not been accurate, and the protein in question could

[32] Clare O'Connor and Jill U. Adams, "Proteins Are Responsible for a Diverse Range of Structural and Catalytic Functions in Cells," in *Essentials of Cell Biology* (Cambridge, MA: NPG Education, 2010).

[33] Christian B. Anfinsen, "Principles That Govern the Folding of Protein Chains," *Science*, vol. 181, no. 4096 (1973).

[34] John Jumper, Richard Evans, and Alexander Pritzel, et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature*, vol. 596 (2021).

[35] Damiano Sgarbossa, Umberto Lupo, and Anne-Florence, "Generative Power of a Protein Language Model Trained on Multiple Sequence Alignments," *eLIfe*, vol. 12, no. e79854 (2023).

[36] Zulema Udaondo, "Big Data and Computational Advancements for Next Generation of Microbial Biotechnology," *Microbial Biotechnology*, vol. 15, no. 1 (2022). Note: Protein residues are individual amino acids that are linked by peptide bonds in the protein chain. Understanding kinetic parameters can be used for the characterization of disease processes and for monitoring treatment effects.

[37] John Jumper, Richard Evans, and Alexander Pritzel, et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature*, vol. 596 (2021).

[38] RCSB Protein Data Bank, *About RCSB PDB: Enabling Breakthroughs in Scientific and Biomedical Research and Education*, https://www.rcsb.org/.

[39] Hallucination is a term sometimes used to describe the case when an AI model perceives patterns or objects that are nonexistent or imperceptible to human observers. See IBM, "What Are AI Hallucinations?," https://www.ibm.com/topics/ai-hallucinations.

[40] B. I. M. Wicky et al., "Hallucinating Symmetric Protein Assemblies," *Science*, vol. 378 (2022).

[41] *E. coli* is a bacterium that is used in laboratory experiments to test foreign DNA and their protein products.

not function within a living organism.[42] Recent studies suggest that AI-generated protein structures may be more accurate, which could accelerate protein engineering by reducing the number of inaccurately predicted proteins that need to be tested.

While these and other advances show promise, some argue that determining how to efficiently research and design protein structures with specific functions remains a challenge.[43] Novel AI-generated molecules will still need to be physically synthesized and evaluated to determine whether they are of any practical use or value beyond the theoretical possibility that was predicted by the AI model.[44]

AI-Enabled Laboratory Capabilities

AI-based computer models trained on biological data sets and laboratory experimental procedures could be used to design experiments to optimize specific conditions (e.g., an organism's growth rate). In many engineering and science applications, experiments are conducted and empirical models are developed with the objective of improving specific responses of interest.[45] These optimization strategies could be combined with generative AI systems,[46] including AI large language models (LLMs) to write computer code instructing laboratory robots to fulfill experimental parameters and run experiments.[47] LLMs are a subset of generative AI and are characterized as "large" due, in part, to the vast amount of data necessary for training the model to learn the rules of language.[48] This approach could eventually include the use of individual desk-top DNA synthesizers to synthesize DNA under the direction of the AI output.[49] For example, one study used LLMs, including GPT-4,[50] to generate robotic scripts from written instructions, which were able to program laboratory robots to complete experiments. The study proposed that "if LLMs can simplify the translation of experimental processes and instructions into precise robotic movements, then the cost of educating life science researchers to use robots

[42] John Jumper, Richard Evans, and Alexander Pritzel, et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature*, vol. 596 (2021).

[43] John Ingraham, Max Baranov, and Zak Costello, et al., *Illuminating Protein Space with a Programmable Generative Model*, Generate Biomedicines, 2022.

[44] W. Patrick Walters and Mark Murcko, "Assessing the Impact of Generative AI on Medicinal Chemistry," *Nature Biotechnology*, vol. 38 (2020).

[45] National Institute of Standards and Technology, *How do you optimize a process*, https://www.itl.nist.gov/div898/handbook/pri/section5/pri53.htm.

[46] Note: Generative AI systems can generate new content—such as text, images, and videos—through learning patterns from data. See CRS Report R47569, *Generative Artificial Intelligence and Data Privacy: A Primer*, by Kristen E. Busch.

[47] Abhaya Bhardwaj, Shristi Kishore, and Dhananjay K. Pandey, "Artificial Intelligence in Biological Sciences," *Life*, vol. 12, no. 1430 (2022); Adam C. Dama, Kevin S. Kim, and Danielle M. Leyva, et al., "BacterAI Maps Microbial Metabolism Without Prior Knowledge," *Nature Microbiology*, vol. 8 (2023); Takashi Inagaki, Akari Kato, and Koichi Takahashi, et al., "LLMs Can Generate Robotic Scripts from Goal-Oriented Instructions in Biological Laboratory Automation," *arxiv (pre-print)*, 2023, https://arxiv.org/abs/2304.10267; Daniil A. Boiko, Robert MacKnight, and Gabe Gomes, "Emergent Autonomous Scientific Research Capabilities of Large Language Models," *arxiv (preprint)*, 2023, see https://arxiv.org/abs/2304.05332.

[48] CRS Report R47569, *Generative Artificial Intelligence and Data Privacy: A Primer*, by Kristen E. Busch.

[49] Sarah R. Carter, Jaime M. Yassif, and Christopher R. Isaac, *Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance*, Nuclear Threat Initiative, 2023, https://www.nti.org/analysis/articles/benchtop-dna-synthesis-devices-capabilities-biosecurity-implications-and-governance/.

[50] GPT-4 is an AI chatbot developed by a company called OpenAI underpinned by an LLM.

can be substantially reduced."[51] The preliminary results suggest that LLMs can interpret and execute tasks at a level similar to a graduate student working under the guidance of a professor. However, generating scripts based on longer instructions remains a challenge. Another study attempted to develop a system, using LLMs, that could be used for autonomous design, planning, and execution of scientific experiments and concluded that the system "demonstrates exceptional reasoning and experimental design capabilities, effectively addressing complex problems and generating high-quality code."[52]

One commentary stated that utilizing AI to help standardize certain processes through automation may help enhance the reproducibility and predictability of biological experiments.[53] Some have argued that AI may reduce the opportunity for human creativity in science, while others suggest that AI instead will free scientists from some of the repetitive tasks of scientific research and allow more time for critical thinking.[54]

Policy Considerations

AI's use in biology and the broader scientific enterprise could raise policy concerns for Congress, particularly regarding whether additional oversight may be warranted. There have been numerous bills introduced in the 118th Congress that address the risks, benefits, and strategic competitiveness of AI more generally.[55] In addition, in July 2023, two bills were introduced that focused more specifically on AI and biological threats.

- H.R. 4704 and S. 2399, "Artificial Intelligence and Biosecurity Risk Assessment Act," would require the Department of Health and Human Services (HHS) Assistant Secretary for Preparedness and Response to "conduct risk assessments and implement strategic initiatives or activities to address whether technical advancements in artificial intelligence, such as open-source artificial intelligence models and large language models, can be used intentionally or unintentionally to develop novel pathogens, viruses, bioweapons, or chemical weapons."
- S. 2346, "Strategy for Public Health Preparedness and Response to Artificial Intelligence Threats," would require the Secretary of HHS to develop a strategy for public health preparedness and response to artificial intelligence threats, and for other purposes.

Potential Biosafety and Biosecurity Concerns

Some have argued that using GenAI for biological design can be repurposed or misused, raising certain biosafety and biosecurity concerns.[56] As part of a "thought experiment" that may have

[51] Takashi Inagaki, Akari Kato, and Koichi Takahashi, et al., "LLMs Can Generate Robotic Scripts from Goal-Oriented Instructions in Biological Laboratory Automation," *arxiv (pre-print)*, 2023, https://arxiv.org/abs/2304.10267.

[52] Daniil A. Boiko, Robert MacKnight, and Gabe Gomes, "Emergent Autonomous Scientific Research Capabilities of Large Language Models," *arxiv (preprint)*, 2023, see https://arxiv.org/abs/2304.05332.

[53] Serina L. Robinson, "Artificial Intelligence for Microbial Biotechnology; Beyond the Hype," *Microbial Biotechnology*, vol. 15, no. 1 (2021).

[54] Ibid.

[55] CRS Report R47644, *Artificial Intelligence: Overview, Recent Advances, and Considerations for the 118th Congress*, by Laurie A. Harris.

[56] Sean Ekins, Maximilian Brackmann, and Cedric Invernizzi, et al., "Generative Artificial Intelligence-Assisted Protein Design Must Consider Repurposing Potential," *GEN Biotechnology*, vol. 2, no. 4 (2023).

implications for the Chemical and Biological Weapons Conventions,[57] Collaborations Pharmaceuticals, Inc., retrained its AI model for drug development[58] to design molecules for toxicity instead of designing against them.[59] The AI model normally penalizes predicted toxicity and rewards predicted target activity. By inverting this logic, the AI model was guided to reward both toxicity and bioactivity. The AI model was then trained with a public database of drug-like molecules and their bioactivities. The study reported that in less than six hours the AI model generated 40,000 molecules that scored within their desired threshold of toxicity and bioactivity, which included the nerve agent VX, other known chemical warfare agents, and new molecules that were predicted to have even higher toxicities. These new molecules were not included in the datasets the AI model was trained on. While the AI model was able to generate designs for these new molecules, the company did not attempt to actually synthesize and produce them.

AI's potential ability to produce molecules that have similar properties but different chemical structures or DNA sequences have raised concerns. Specifically, that current detection methods, and screening guidelines that use certain DNA sequences or other features as identifiers, may not be able to identify these new toxins of concern. For example, ricin is a plant protein toxin that is listed on the U.S. Federal Select Agent List.[60] Researchers claim that using AI, they could "take the active site of ricin, computationally remove the surrounding stabilizing protein structure, and ask a generative algorithm to hallucinate and imprint a new proteinaceous structure around the active site."[61] This would, in theory, remove the DNA sequence component that is screened by DNA synthesis service providers[62] and other control mechanisms (e.g., the Chemical and Biological Weapons Conventions or the U.S. Federal Select Agent Program). Note, however, that using AI to identify and design novel molecules, or DNA sequences, does not mean that the molecule could actually be produced, or a functioning living organism based on AI-generated molecules or DNA sequences. Additional scientific, technical, and other challenges would have to be overcome, such as synthesizing the DNA and getting it to function in a living organism as designed, to realize this biosecurity risk.[63]

Laboratory Automation and De-Skilling Concerns

While laboratory automation has already impacted certain research operations (e.g., liquid handling robots), the use of AI to automate certain procedures that used to require hands-on, tacit knowledge could expand the pool of individuals who can participate in the biological sciences. While this could be a positive development in some respects, it could also potentially increase biosafety and biosecurity concerns.[64] One study examined whether such a system would provide instructions and capabilities to design certain illicit drug compounds and chemical weapons

[57] Spiez Convergence Conference, 2021, https://www.spiezlab.admin.ch/en/home/meta/refconvergence.html.

[58] Fabio Urbina, Christopher Lowden, and Christopher Culberson, et al., "MegaSyn: Integrating Generative Molecule Design, Automated Analog Designer and Synthetic Viability Prediction," *ChemRxiv (Preprint)*, 2021.

[59] Fabio Urbina, Filippa Lentzos, and Cédric Invernizzi, et al., "Dual Use of Artificial-Intelligence-Powered Drug Discovery," *Nature Machine Intelligence*, vol. 4 (2022).

[60] The Department of Health and Human Services (HHS) and U.S. Department of Agriculture (USDA) Select Agent Program regulates the possession, use, and transfer of select agents and toxins deemed by HHS or USDA to pose a severe threat to public health and safety, based on a set of criteria. See http://www.selectagents.gov/.

[61] Sean Ekins, Maximilian Brackmann, and Cedric Invernizzi, et al., "Generative Artificial Intelligence-Assisted Protein Design Must Consider Repurposing Potential," *GEN Biotechnology*, vol. 2, no. 4 (2023).

[62] International Gene Synthesis Consortium, *International Gene Synthesis Consortium Updates Screening Protocols for Synthetic DNA Products and Services*, https://genesynthesisconsortium.org/.

[63] John T. O'Brien and Cassidy Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology," *Health Security*, vol. 18, no. 3 (2020).

[64] Ibid.

agents. While the model refused to return results for certain illicit compounds, the authors suggest this could be overcome by altering the terminology, prompting the authors to call for restrictions, particularly a mechanism that would prevent the model from moving forward without the approval of a "human" who can review and approve certain experiments.[65]

Wider access to AI, cloud labs,[66] genetic sequence information, and gene synthesis capabilities, including desktop DNA synthesizers,[67] have raised other biosafety and biosecurity concerns—such as who should be able to access these capabilities and what limits might be placed on synthesis capabilities. The convergence of AI and other technologies associated with biology can lower technical and knowledge barriers and increase the number of actors with certain capabilities.[68] AI can enable experimentation and design of biological systems, particularly the design and synthesis of DNA sequences, without the user necessarily understanding what those biological functions represent or do.

A recent study reported that, in a course at the Massachusetts Institute of Technology (MIT), students were asked to investigate whether AI chatbots could provide information that could potentially be used to cause a pandemic. Students were able to use chatbots that suggested potential pandemic pathogens, explained how they can be generated from synthetic DNA using reverse genetics, supplied the names of DNA synthesis companies that are unlikely to screen orders, identified detailed protocols and how to troubleshoot them, and recommended that anyone lacking the skills to perform reverse genetics engage a core facility or contract research organization.[69] The authors of the study argued that increased access to AI can exacerbate the "de-skilling" of biological research and that certain restrictions and oversight may be necessary when considered alongside other advances, such as desktop synthesizers and cloud-based laboratories. Such restrictions include evaluations of, and controlled access to, LLMs, as well as screenings to evaluate what a particular person may be attempting to develop, order, or use. Others may counter that access to information does not necessarily imply the scientific and technical ability to convert that information into actual harm, in this case the creation of a virus capable of causing a pandemic.

Automation and "de-skilling" could present increased biosafety and biosecurity concerns. How severe, or whether these concerns are true in actual practice, is an open debate[70] and an area that Congress may wish to examine. For example, if the output of an LLM poses certain biosafety and biosecurity risks, should there be restrictions on access to the data (e.g., certain DNA sequences

[65] Daniil A. Boiko, Robert MacKnight, and Gabe Gomes, "Emergent Autonomous Scientific Research Capabilities of Large Language Models," *arxiv (preprint)*, 2023, see https://arxiv.org/abs/2304.05332.

[66] Cloud labs are highly automated laboratories that allow researchers to design and run experiments remotely by sending "instructions" to a contracted facility via the web. Experiments can include sample preparation, bioassays, workflows for DNA synthesis, and methods for imaging and detection, as well as AI-driven drug discovery and testing. "What Are Cloud Labs?," The Biologist, https://www.rsb.org.uk/biologist-features/the-biologist-s-guide-to-cloud-labs.

[67] Sarah R. Carter, Jaime M. Yassif, and Christopher R. Isaac, *Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance*, Nuclear Threat Initiative, 2023, https://www.nti.org/analysis/articles/benchtop-dna-synthesis-devices-capabilities-biosecurity-implications-and-governance/.

[68] John T. O'Brien and Cassidy Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology," *Health Security*, vol. 18, no. 3 (2020).

[69] Emily H. Soice, Rafael Rocha, and Kimberlee Cordova, et al., "Can Large Language Models Democratize Access to Dual-Use Biotechnology?," *arxiv*, 2023, https://arxiv.org/abs/2306.03809.

[70] Sarah R. Carter, Nicole E. Wheeler , and Sabrina Chwalek, et al., *The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe*, Nuclear Threat Initiative , 2023, https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/; Helena, *Biosecurity in the Age of AI*, 2023, https://www.helenabiosecurity.org/.

of concern) on which LLMs are trained? Restricting access to certain DNA sequences may be difficult since there are a number of databases currently available globally, a large body of existing published scientific literature containing sequence data along with research results (a commonly accepted practice in the global research community), and a number of LLMs that have already been trained on data from existing databases and literature. There may also be concern that restricting data that LLMs are trained on could imperil potentially beneficial uses or create incentives for malicious actors to move certain LLMs to the dark web.

Another option could be restricting access to the LLMs themselves. Developers of the LLMs could require certain registration, authentication or screening requirements for users to access the LLM, or particular features of the LLM, that pose potential biosafety and biosecurity concerns. Additional reporting requirements could also be implemented that identified certain queries or outputs of the LLM that raised concerns, which potentially could be referred to a federal oversight or security agency.

Another concern may be the purposes of LLM outputs and what they are being used in conjunction with. For example, if an LLM is aiding in biological design then the output, or the mechanisms to produce the output, may be an area where certain restrictions could be implemented, if found necessary. As discussed below, certain aspects of this concern are addressed via DNA synthesis screening. However, if certain biological design tools become ubiquitous (e.g., desk-top DNA synthesizers), Congress may examine whether access to these tools should be restricted or whether certain aspects of these tools could be constrained. For example, could the data or biological design features of a desk-top synthesizer be restricted[71] to prevent certain capabilities of producing a biosafety or biosecurity hazard?

Biosafety and Biosecurity Oversight

AI's use in biology and the broader scientific enterprise sits within a debate on how best to manage laboratory biosafety and biosecurity.[72] Many of the concerns that have been raised around the convergence of AI and biology stem from information hazards. Information hazards are risks that arise from the dissemination or the potential dissemination of true information that may cause harm or enable some agent to cause harm.[73] Both AI and certain aspects of biological research can contain information hazards, either separately or collectively. Addressing these issues has led some experts to call for controlled access to certain materials, technologies, and equipment, as well as to develop international norms.[74] Some oversight mechanisms are already in place through U.S. laws and other guidance related to federally funded research,[75] as well as industry best practices that include DNA synthesis screening protocols.[76] For example, certain screening guidelines have been implemented through industry standards and other requirements for

[71] Sarah R. Carter, Jaime M. Yassif, and Christopher R. Isaac, *Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance*, Nuclear Threat Initiative, 2023, https://www.nti.org/analysis/articles/benchtop-dna-synthesis-devices-capabilities-biosecurity-implications-and-governance/.

[72] CRS Report R47695, *U.S. Oversight of Laboratory Biosafety and Biosecurity: Current Policies, Recommended Reforms, and Options for Congress*, by Todd Kuiken.

[73] Nick Bostrom, "Information Hazards: A Typology of Potential Harms from Knowledge," *Review of Contemporary Philosophy*, vol. 10 (2011).

[74] Helena, *Biosecurity in the Age of AI*, 2023, https://www.helenabiosecurity.org/.

[75] CRS Report R47695, *U.S. Oversight of Laboratory Biosafety and Biosecurity: Current Policies, Recommended Reforms, and Options for Congress*, by Todd Kuiken.

[76] International Gene Synthesis Consortium, "Harmonized Screening Protocol V2," 2017, https://genesynthesisconsortium.org/wp-content/uploads/IGSCHarmonizedProtocol11-21-17.pdf.

federally funded research to address biosafety and biosecurity risks associated with DNA synthesis, along with other mechanisms that may control access to certain equipment or capabilities.

Currently, the International Gene Synthesis Consortium's (IGSC's) Harmonized Screening Protocol specifies that synthetic gene sequence orders be screened against the IGSC's Regulated Pathogen Database, which contains sequences and organisms subject to regulatory control or licensing.[77] HHS developed its *Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids* to align with "Providers' and Customers' existing protocols and business practices, to be implemented without unnecessary cost, and to minimize any negative impacts on the conduct of research and business operations."[78] On October 30, 2023, the White House released Executive Order 14110, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,* which requires federal agencies to "establish a framework, incorporating, as appropriate, existing United States Government guidance, to encourage providers of synthetic nucleic acid sequences to implement comprehensive, scalable, and verifiable synthetic nucleic acid procurement screening mechanisms, including standards and recommended incentives."[79]

Questions Congress could face in this area include whether these types of voluntary, or agency-directed, screening programs should be applied to data that AI models are trained on, whether additional oversight focused on AI's use in biology and the broader scientific enterprise is necessary, and if so, whether agencies have expertise and capacity to implement such a program, or whether current oversight systems are sufficient.

Broad-Based vs. Case-by-Case Oversight

Some have identified the importance of distinguishing the risks associated with the use of AI models to generate experimental procedures from the risks associated with the use of AI-enabled biological design tools.[80] They argue that doing so may be useful when exploring potential governance options. This could prove difficult as such biology-focused AI models and biological design tools are often developed in tandem.[81] An example of this was described in an announcement by Google Cloud and Ginkgo Bioworks, an engineering biology company, of their plans to build a "generative AI platform for engineering biology and for biosecurity."[82] The announcement described the convergence of biological data (which includes genomic data), AI, and biodesign tools, in which biological data is designed and structured for use with the AI tools and integrated into a biofoundry. A biofoundry is a facility that provides integrated infrastructure (combination of biology, computer-aided design, robotics, and engineering principles) enabling

[77] Ibid.

[78] Administration for Strategic Preparedness and Response, *Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids,* U.S. Department of Health & Human Services, 2023, https://aspr.hhs.gov/legal/synna/Documents/SynNA-Guidance-2023.pdf.

[79] Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," 88 *Federal Register* 75191-75226 (2023).

[80] Jonas B. Sandbrink, "Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools," preprint, https://arxiv.org/abs/2306.13952.

[81] Anna Nowogrodzki, "The Automatic-Design Tools That Are Changing Synthetic Biology," *Nature,* vol. 564 (2018).

[82] Ginkgo Bioworks, "Google and Ginkgo: Foundry-Scale Data Meets AI," press release, 2023, https://www.ginkgobioworks.com/2023/08/29/google-and-ginkgo-foundry-scale-data-meets-ai/.

the rapid design, construction, and testing of engineered organisms for biotechnology applications and research.[83]

The potential capabilities of AI used in conjunction with biological design and other types of scientific research and experimentation have prompted recommendations on how to prevent the misuse of AI applications in biology and science more broadly. Those recommendations include keeping a human "in the loop," controlling access to DNA sequences and synthesis capabilities, and other governance mechanisms that restrict, or monitor, who can access certain AI applications and biological design tools.[84]

Existing and proposed oversight mechanisms take different approaches, such as controlling access to a particular technology or dataset, or focusing on a particular step in the R&D process (e.g., basic research, prototypes, or release of a final product).

If further regulation or oversight in this area were deemed necessary, potential options for Congress could include

- whether AI, and its use in biology and other R&D applications, should be regulated broadly across all use cases and areas of development; or
- whether to design oversight mechanisms on a case-by-case basis (e.g., limiting access to certain tools or models), focused on particular application areas (e.g., drug development), or based on particular end-use products (e.g., certain laboratory equipment).

Additional regulation and oversight could impact other areas of R&D in specific or unanticipated ways. For example, limiting access to an AI model, restricting the types of data on which an AI model is trained, or limiting the capabilities an AI model is allowed to execute could each individually impact a biological design tool's capability. Limiting AI's access to certain DNA datasets based on particular biosecurity concerns (e.g., pathogenicity) could prevent an AI model from designing certain toxins of concern. It might also, however, prevent the AI model from predicting molecular structures that could enable the development of new drug candidates.

Structured Access

Some experts have called for structured access to certain AI capabilities while others suggest a gradient approach.[85] In a structured access approach, the developer of an AI model maintains control over how the AI system can be used, modified, and reproduced.[86] Structured access could also be implemented for other research tools in the R&D process (e.g., desktop DNA

[83] CRS Report R47265, *Synthetic/Engineering Biology: Issues for Congress*, by Todd Kuiken. Nathan Hillson, Mark Caddick, and Yizhi Cai, et al., "Building a Global Alliance of Biofoundries," *Nature Communications*, vol. 10, no. 2040 (2019).

[84] Fabio Urbina, Filippa Lentzos, and Cedric Invernizzi, et al., "Preventing AI From Creating Biochemical Threats," *Journal of Chemical Information and Modeling*, vol. 63 (2023). Helena, *Biosecurity in the Age of AI*, 2023, https://www.helenabiosecurity.org/.

[85] Irene Solaiman, "Generative AI Systems Aren't Just Open or Closed Source," *Wired*, 2023, https://www.wired.com/story/generative-ai-systems-arent-just-open-or-closed-source/.

[86] Helena, *Biosecurity in the Age of AI*, 2023, https://www.helenabiosecurity.org/. Note: There are two broad categories of structured access: (1) use controls, which govern the direct use of the AI system (who, what, when, where, why, and how); and (2) modification and reproduction controls, which prevent the user from altering the AI system or building their own version in a way that circumvents use controls.

synthesizers),[87] for example, by controlling the acquisition of the device itself.[88] This would not necessarily impact the AI component of the R&D process. At one university where a team of researchers was proposing to develop an AI model to predict the toxicity of chemicals and materials, its research ethics and society review panel suggested that the team think about ways to control the distribution of the software, model, and outputs in order to reduce potential misuse.[89] This type of review and oversight is similar to how the United States addresses some aspects of biosafety and biosecurity, with research institutions responsible for implementation of certain guidelines as a condition for receiving federal funding.[90] Congress may examine whether federal agencies have the expertise, capacity, and authority to both develop and issue guidance requiring institutions, or other actors, who receive federal funding to conduct certain biosafety and biosecurity reviews of AI enabled R&D applications. Congress may also examine whether this type of biosafety/biosecurity review resulting in potential restricted access can be accomplished by granting authority to agencies to conduct such reviews, regardless of whether an actor receives federal funding or not.

Issues Pertaining to Genetic Sequence Data and Databases

The U.S. National Security Commission on Artificial Intelligence recommended that the U.S. fund and prioritize the development of a biobank containing a "wide range of high-quality biological and genetic data sets securely accessible by researchers." The United States currently has a number of such databases, including GenBank.[91] However, they are distributed across different agencies, contain different sources and types of sequence data, and are designed for different uses.[92] There are other sequence databases which are publicly funded and freely accessible, and others that are privately held. Some experts have raised privacy and security concerns around these databases in terms of access and potential use of the sequence data.[93]

Potential options for Congress in this area may include

- whether the United States should develop a sequence database strategy coordinating activities across the U.S. government in relation to the collection, storage, maintenance, and access to its sequence databases;
- whether certain privacy, biosafety, and biosecurity issues pertaining to access and use of sequence databases are adequately addressed, particularly related to the convergence of AI and biological design; and
- whether these issues should be addressed regarding sequence databases which are publicly funded and freely accessible or privately held.

[87] Sarah R. Carter, Jaime M. Yassif, and Christopher R. Isaac, *Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance*, Nuclear Threat Initiative, 2023, https://www.nti.org/analysis/articles/benchtop-dna-synthesis-devices-capabilities-biosecurity-implications-and-governance/.

[88] Helena, *Biosecurity in the Age of AI*, 2023, https://www.helenabiosecurity.org/.

[89] Sadasivan Shankar and Richard N. Zare, "The Perils of Machine Learning in Designing New Chemicals and Materials," *Nature Machine Intelligence*, vol. 4 (2022).

[90] CRS Report R47695, *U.S. Oversight of Laboratory Biosafety and Biosecurity: Current Policies, Recommended Reforms, and Options for Congress*, by Todd Kuiken.

[91] NIH, GenBank, https://www.ncbi.nlm.nih.gov/genbank/.

[92] CRS In Focus IF12285, *eDNA/eRNA: Scientific Value in What's Left Behind*, coordinated by Todd Kuiken.

[93] Saadia Arshad, Junaid Arshad, and Muhammad Mubashir Khan, et al., "Analysis of Security and Privacy Challenges for DNA-Genomics Applications and Databases," *Journal of Biomedical Informatics*, vol. 119 (2021).

Other Policy Considerations

A general policy consideration may be whether a broader risk management approach[94] examining the R&D system as a whole is more appropriate than addressing individual components of the R&D process. This would include at what stage oversight is warranted (e.g., basic research, prototyping, or prior to release of a final product) and who might conduct such oversight (e.g., by a federal agency or through self-governance incentives). For example, would a federal agency have the authority to conduct oversight or could oversight be achieved through self-governance incentives? This could be similar to the operation of institutional review boards[95] or other industry-led programs such as the International Gene Synthesis Consortium,[96] which conducts screening for DNA synthesis orders.

Congressional options for addressing federal oversight could include whether to provide an agency (or agencies) authority to conduct oversight of various aspects of AI, create a new agency with authority over AI, or authorize agencies under current law to enforce certain oversight responsibilities. Authorizing agencies under current law may require additional coordination among federal agencies in order to establish oversight responsibilities. While it was established via an executive order, this type of coordination among agencies would be similar to how the *Coordinated Framework for Regulation of Biotechnology* established mechanisms for three federal agencies (USDA, FDA, and EPA) to share responsibility for regulating the products of biotechnology.[97]

Potential oversight and governance of the use of AI in the biological sciences both domestically and internationally, either generally or focused on particular use cases or areas of development, may have unpredicted societal and economic impacts, both positive and negative, affecting U.S. strategic competitiveness. Such concerns could include how potential U.S. oversight and governance influences, or is influenced by, policy actions taken elsewhere; whether international harmonization may be needed, and, if so, through what mechanisms (e.g., the United Nations, international standards bodies, trade policy); and how to engage in those negotiations.

Author Information

Todd Kuiken
Analyst in Science and Technology Policy

[94] In January 2023 the National Institute of Standards and Technology released its *AI Risk Management Framework to better manage risks to individuals, organizations, and society associated with AI*. See https://doi.org/10.6028/NIST.AI.100-1.

[95] U.S. Food and Drug Administration, *Institutional Review Boards Frequently Asked Questions*, 2019, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/institutional-review-boards-frequently-asked-questions.

[96] International Gene Synthesis Consortium, “Harmonized Screening Protocol V2,” 2017, https://genesynthesisconsortium.org/wp-content/uploads/IGSCHarmonizedProtocol11-21-17.pdf.

[97] Executive Office of the President (EOP), Office of Science and Technology Policy, “Coordinated Framework for Regulation of Biotechnology,” 51 *Federal Register* 23302, June 26, 1986. For additional information on the Coordinated Framework see CRS Report R46737, *Agricultural Biotechnology: Overview, Regulation, and Selected Policy Issues*, by Renée Johnson.

Disclaimer

This document was prepared by the Congressional Research Service (CRS). CRS serves as nonpartisan shared staff to congressional committees and Members of Congress. It operates solely at the behest of and under the direction of Congress. Information in a CRS Report should not be relied upon for purposes other than public understanding of information that has been provided by CRS to Members of Congress in connection with CRS's institutional role. CRS Reports, as a work of the United States Government, are not subject to copyright protection in the United States. Any CRS Report may be reproduced and distributed in its entirety without permission from CRS. However, as a CRS Report may include copyrighted images or material from a third party, you may need to obtain the permission of the copyright holder if you wish to copy or otherwise use copyrighted material.