AIxBio Horizon Scan

Introduction

The AIxBio field stands at a critical juncture where rapid capability advances are outpacing governance frameworks and safety measures. While technical progress continues to accelerate through improved integration of AI with biological tools, fundamental challenges around data quality, biological complexity, and the inability to extrapolate beyond training data remain significant constraints. The next 18 months will likely prove pivotal in determining whether voluntary safety practices by AI companies, emerging evaluation frameworks, and international coordination efforts can keep pace with technological development. There is currently a very robust debate about open-sourcing AI models, where safeguards are unlikely to be robust enough to prevent misuse. As these powerful capabilities become more accessible through improved interfaces and automation, ensuring their beneficial use while mitigating risks will require sustained attention from policymakers, researchers, and civil society stakeholders alike.

To address critical governance needs, this horizon scan examines the rapidly evolving landscape of capabilities at the convergence of AI and the life sciences (AIxBio capabilities), and it is conducted under the auspices of the AIxBio Global Forum. The goal of this horizon scan is to outline anticipated trends in technology and capability development over the next three years. This has a wide range of potential applications, including supporting well-informed conversations about AIxBio risks and risk-reduction interventions. In addition to incorporating published studies and information, this analysis draws on interviews and engagement with 21 experts across academia, industry, and policy, which took place in late 2025. This report synthesizes expert insights to provide a comprehensive overview of current AIxBio capabilities, identifies important trends shaping the field's development, and presents predictions for capabilities expected to emerge over the next three years.

Current Capabilities

The current AIxBio landscape is characterized by significant advances in both the capabilities of large language models (LLMs) and specialized biological AI tools, as well as the integration of LLMs with biological research workflows.

LLM Advances

Large language models are now demonstrating expert-level performance in biological domains. According to the Virology Capabilities Test, a benchmark designed to measure how well large language models can answer expert-level virology questions, LLMs developed by OpenAI, Google, and Anthropic now outperform human experts in their virology-related areas of expertise. Methods like Retrieval Augmented Generation (RAG) can enhance capabilities of LLMs in specific areas of expertise by retrieving relevant documents from external knowledge sources—for example by using the internet to access journal articles on a particular organism or cell signaling pathway— and then using that retrieved information to inform more accurate and detailed responses. This technique can increase the capabilities of LLMs for life-science relevant applications, for example by pulling up laboratory protocols and helping a scientist understand how to modify an experimental procedure for the organism they are researching.

While these advanced capabilities offer significant potential benefits, they also raise significant biosafety and biosecurity concerns. LLMs are still prone to hallucinations, where they present inaccurate or incomplete information as factual; hallucinations would be particularly concerning when the models are sharing information about biosafety protocols, such as proper containment or disinfection procedures. Several leading labs have also determined that their un-safeguarded models may be able to help a novice create a biological weapon by providing step-by-step information that was historically only known by a small group of experts. Leading LLM labs have implemented safety systems designed to prevent this misuse and are continuously updating their safety systems.

Biological AI Tool Advances

Specialized biological AI tools are advancing rapidly in protein science. For example, AlphaProteo and continued iterations of AlphaFold, which predict how proteins fold and interact—a core problem in biology—have advanced protein design and structure prediction capabilities. This allows scientists to determine the three-dimensional shape of a protein and even how mutations may change its shape and function. AlphaFold 3 can now reliably model protein-protein interactions and complexes made up of proteins and other types of molecules, enabling the design of new therapeutics. An open-source competitor, Boltz-1, was released six months after AlphaFold 3 and trained using a quarter of the compute yet has comparable capabilities, showing just how quickly cutting edge, proprietary tools can be re-created and shared widely, enabling rapid democratization of these technologies.

Genomic AI models, which represent another category of biological AI tools, are pushing the boundaries of what's possible in biological design. Another example of rapidly advancing biological AI tools is Evo. Evo stands out as the current cutting edge for a life-science language model and is capable of modeling entire genomes for a subset of organisms with small genomes. Using a DNA sequence as a starting prompt, the model can complete the sequence, designing a novel protein, or even a complete genome in some cases. Similar to early natural language models, the tool requires guidance and filtering to produce meaningful outputs, but it is beginning to show preliminary capabilities of novel virus design. Coupled with additional fine-tuning, prompts, and filters specific to a virus that infects bacteria, the Evo team was able to use the model to generate 302 novel versions of the virus, 16 of which were viable.

Integration of LLMs and Biological AI Tools and Laboratory Robotics

The integration between LLMs and biological tools has improved markedly. For example, the MP4 and Pinal models use natural-language prompts for protein generation, which can allow for more intuitive use by a broader group of people who don't necessarily have deep expertise in protein design or command-line tool usage. However, at this time, generated enzymes are not always functional and are often less efficient than naturally occurring enzymes.

Natural language interfaces for control of laboratory robotics are further developed than those for protein design. Both proprietary software tools and non-specialized language models can now be used to generate code for liquid-handling robots, enabling more people to automate wet-lab work. Currently, laboratory robotics systems are not wide-spread, and even advanced laboratories typically require physical assistance by a human operator.

Current Key Players

Monitoring current key players can allow for understanding of where current technologies stand. In this dynamic space, many academic labs and new commercial players regularly announce advances. Major contributors to recent progress in tools at the intersection of AI and the life sciences include Arc Institute (developer of Evo), Google DeepMind (AlphaProteo and AlphaFold evolution), Stanford University (including Biomni), FutureHouse (advancing agent-based biological research), and commercial entities like Benchling, which has rolled out AI assistants embedding LLMs into research workflows. NVIDIA's BioNeMo framework has facilitated easier training and deployment of biological models, while companies like Profluent (OpenCRISPR) and EvolutionaryScale (ESM3), which was just acquired by the Chan-Zuckerberg Initiative, have made significant contributions to gene editing and protein design capabilities.

Important Trends

Integration of LLMs with biological research tools

The integration of LLMs with biological research tools represents perhaps the most transformative development. LLMs are designed to predict the next word in a sequence. AI agents use an LLM as their "brain," have a memory system, and can use a set of tools to complete a task. Agent-based systems can now debug their own work, iterate on experimental designs, and orchestrate complex workflows. Agents using biological AI tools and laboratory robotics may allow for both testing and optimization of AI-designed molecules as well as biological discovery. Agents can work towards more complex goals and independently decide and execute the next step, enabling more autonomous research processes. Model Context Protocols, which connect AI assistants to the data systems, will further allow integration of biological tools, datasets, and other models as part of AI agent workflows.

Automation and laboratory integration

Automation and laboratory integration will also be critical for setting the pace of AIxBio capabilities. Cloud labs and automated experimental platforms are dramatically lowering barriers to experimentation, removing the need for years of hands-on training from current experts, and may allow for fully automated AI-driven biology research without humans in the loop in the next decade, meaning systems will need deliberately designed safeguards and can't rely on a person flagging suspicious or malicious work. AI-driven systems may be able to generate hypotheses, design experiments to test them, run them via robotic systems, and use the data to generate more hypotheses. The convergence of AI design tools with robotic laboratory systems enables rapid iteration cycles, with some facilities already capable of synthesizing and testing hundreds of thousands of AI-generated molecules weekly.

This automation trend is closely linked to data generation improvements, as data quality is a persistent bottleneck.

Data as a bottleneck

Data collection presents a significant rate-limiting step for developing more powerful biological AI models that make higher fidelity predictions. Large amounts of additional data will need to be collected to significantly advance these models beyond current capabilities.

The data bottleneck is particularly acute for edge cases and rare biological phenomena that are most relevant for both breakthrough discoveries and potential misuse scenarios. Training datasets are inevitably biased toward well-studied model organisms, common proteins, and frequently investigated pathways, leaving AI models poorly equipped to make accurate predictions about novel or understudied biological systems. Addressing this bottleneck will require not only generating more data, but also strategic investments in generating high-quality data for critical but currently underexplored areas of biological space.

In addition to data quantity, there are significant data quality challenges including issues of standardization, reproducibility, and experimental design. Many biological datasets suffer from batch effects, inconsistent measurement protocols, and insufficient metadata documentation, making it difficult for AI models to extract reliable patterns. Furthermore, biological data often comes from disparate sources—academic labs, pharmaceutical companies, and biotech startups—each using different experimental conditions, organisms, and measurement techniques. This heterogeneity creates significant obstacles for training robust AI models that can generalize across contexts. As a result, even when large volumes of biological data exist, much of it may be too noisy or poorly annotated to effectively train next-generation AI systems.

Progress at different levels of biological scale: molecular, cellular and population scale

Potential progress may be unlocked by work at the right biological scale – whether that's at the level of individual molecules, cellular pathways, whole cells, tissues, organisms, populations, or even entire

ecosystems. Different scales present different opportunities for AI advancement. At the molecular level, biophysics-based AI models have shown promise in drug discovery and may enable capabilities beyond what current training data would suggest, such as designing enzymes with entirely novel functions by extrapolating from underlying physical principles.

At the cellular level, the concept of "virtual cells" is beginning to emerge as a potentially powerful predictive tool. Early, simple versions of this category of models are built using perturbation-based datasets, where cells are chemically or genetically disrupted and then analyzed through imaging, RNA sequencing, proteomics, or other comprehensive measurements. The resulting data helps create models that can predict how cells will respond to different interventions and can help find reciprocal similarities between diseases and therapies. Data quality remains a critical bottleneck – advances in single-cell measurement technologies will provide better training data for these models.

Monitoring both the resolution of available biological data and the predictive capabilities of models operating at different scales will be important for understanding future AIxBio progress.

Future Predictions

6 to 18 Months

In the next 6-18 months, there will likely be **significant improvements in agent-biological tool integration**, including tools that will help with bioinformatics and data processing, although agents for design of new proteins and organisms may be less developed in this time frame.

These agent systems can dramatically accelerate the pace of research by automating tasks that currently require specialized computational expertise, such as analyzing genomic sequences or identifying patterns in complex datasets. This acceleration could enable smaller research teams to accomplish what previously required large, well-funded laboratories, potentially democratizing access to advanced biological research capabilities for both beneficial applications like drug discovery and concerning applications like pathogen engineering.

The proliferation of automated laboratory platforms with LLM integration is expected to accelerate, alongside much better automated labs—towards fully automated rooms with zero human requirements. Removing human operators from the laboratory means that current oversight mechanisms which rely on trained scientists recognizing potentially dangerous work will no longer function. A fully automated system could theoretically execute a complete biological design-build-test cycle without any human review, raising fundamental questions about how to implement safety checks and responsible-use policies when there are no human decision-points in the workflow.

This time frame will also see the release of **second wave protein and genome language models**, which will be able to produce effective proteins and whole genomes with higher fidelity, reducing the amount of *in vitro* testing that will be necessary. Higher fidelity predictions mean researchers can move more quickly from computational design to functional biological products, dramatically shortening development timelines. While this could accelerate beneficial applications like enzyme design for industrial processes or therapeutic protein development, it also lowers the barrier between having an

idea for a dangerous biological agent and actually creating a functional version, as fewer rounds of trial-and-error testing would be required.

Wider deployment of protein binder design tools will allow for greater application to a wider range of beneficial and malicious uses. Protein binders—molecules that attach to specific target proteins—are foundational to both medicine and potential weapons. Therapeutically, they can block disease-causing proteins or deliver drugs to specific cells. However, the same capability could be used to design molecules that interfere with critical immune functions or enhance the infectivity of pathogens. As these tools become more widely available and easier to use, the technical barrier to creating such molecules drops significantly.

This timeframe will likely also see the emergence of **closed-loop discovery systems capable of orchestrating automatic synthesis, testing, and refinement cycles** for predicted proteins or genomes. These closed-loop systems represent a qualitative shift in research capability: rather than requiring human scientists to interpret results and design follow-up experiments, AI systems will autonomously iterate through design-test-refine cycles, potentially running hundreds or thousands of experiments in the time it would take a human team to complete a handful. This could unlock scientific discoveries at unprecedented speed but also means that dangerous capabilities could be developed and optimized without meaningful human oversight or opportunity for intervention.

Additionally, commercial biological AI tools will likely gain better **natural language interfaces**, making sophisticated biological tools accessible to users with limited technical expertise. This accessibility shift is double-edged: it could enable researchers in resource-limited settings to contribute to beneficial biological research and could accelerate innovation in fields like agriculture and environmental science. However, it also means that individuals without formal training in biosafety or biosecurity—who might not recognize when their work poses risks—could potentially use these tools to design dangerous biological agents, even unintentionally.

2 to 5 Years

Looking ahead 2-5 years, it is likely that **LLMs will be able to orchestrate biological workflows that meet or exceed the level of human experts,** with sophisticated agent tool-use becoming routine. For example, these automated systems will have high efficiency in designing and carrying out experiments, both designing the appropriate experiments and carrying them out at scale. This will likely be accompanied by the emergence of biological programming languages where researchers can specify desired biological functions, and AI systems will design and use automated lab systems to produce the corresponding genetic circuits, proteins, or organisms.

In addition, the **integration of experimental feedback into model training** will mature, where models will explicitly seek and integrate feedback of data for edge cases, potentially enabling things like protein design tools that can teach themselves how to predict function in a novel area.

However, incremental development towards new functions may not always be reasonable when the leap from known function to desired function is too great; progress will likely remain constrained by the fundamental challenge of extrapolation beyond training data.

Open Questions

Experts interviewed had differing opinions on several key issues:

Impact of genomic language models: The current impact and trajectory of genomic language models like Evo divided participants, with some viewing them as revolutionary while others found them underwhelming relative to expectations. Current genomic language models seem to have similar accuracy in some areas as original AlphaFold, but the full scope of their abilities, or inabilities, has not yet been shown.

Data vs algorithmic constraints: A second open question is the role of data versus algorithmic improvements in driving progress, with some arguing that "we have all the data points we need, we just need something smart enough to connect the dots," while others insist that current biological data is too messy and incomplete for AI to make transformative leaps.

Specialized tools vs foundation models and fine-tuning: Benchmarking tools are emerging to evaluate AI performance in viral research. The EVEREST benchmark, a testing framework that evaluates how well AI models can predict the effects of viral mutations, has emerged as an important tool for viral mutation effect prediction. The original publication showed that narrow, purpose-built tools designed for specific, virus-related work generally outperform larger foundation models like Evo2, which are built to generalize across biology. However, the authors did not fine-tune any of the foundation models on viral sequences and it is possible that fine-tuned versions of general purpose models would perform better.

Geopolitical trajectories: There are also important open questions regarding regulatory and geopolitical factors, including changes in both research funding and oversight. While some participants expressed optimism about responsible AI practices continuing voluntarily among major AI companies, others voiced significant pessimism about continued implementation of any safety requirements. particularly given recent political changes as well as increased competition among nations to develop the most advanced AI capabilities.

Frontier labs like Anthropic, OpenAI, and Google DeepMind have prioritized ensuring their models have responsible safeguards around biological issues, but these are not required and there are no universal standards for what safeguards should be implemented or what capabilities should be restricted.

There is increasing concern among scientists and biosecurity experts that model evaluations, used to determine model capabilities and safeguard effectiveness, may be subject to export controls, making it difficult to properly assess current capabilities globally.

Recent reductions in U.S. Government life science research funding, coupled with large private sector investments in AI and AIxBio development, may change the rate of AIxBio capability advancement and/or the center of gravity of the work—potentially shifting innovation hubs from academia to industry or out of the United States, which may require new strategies and networks to properly monitor.