Communication

A Call for Built-in Biosecurity Safeguards for Generative AI Tools

Mengdi Wang *, Zaixi Zhang *, Amrit Singh Bedi, Stephanie Guerra, Sheng Lin-Gibson, Le Cong, Souradip Chakraborty, Yuanhao Qu, Jian Ma, Eric Xing, George Church

Posted Date: 26 March 2025

doi: 10.20944/preprints202503.1761.v1

Keywords: generative AI; biosecurity

 doi:10.20944/preprints202503.1761.v1

Article

A Call for Built-In Biosecurity Safeguards for Generative AI Tools

Mengdi Wang [1,*], Zaixi Zhang [1], Amrit Singh Bedi [2], Alvaro Velasquez [3], Stephanie Guerra [4], Sheng Lin-Gibson [4], Le Cong [5], Souradip Chakraborty [6], Megan Blewett [7], Yuanhao Qu [5], Jian Ma [8], Eric Xing [8], George Church [9]

[1] Princeton University, NJ, USA
[2] University of Central Florida, FL, USA
[3] ADefense Advanced Research Projects Agency, USA
[4] National Institute of Standards and Technology, MD, USA
[5] Stanford University, CA, USA
[6] University of Maryland, MD, USA
[7] Iris Medicine, USA
[8] Carnegie Mellon University, PA, USA
[9] Harvard University, MA, USA
* Correspondence: mengdiw@princeton.edu

Abstract: The rapid adoption of generative AI (GenAI) in biotechnology offers immense potential but also raises serious safety concerns. AI models for protein engineering, genome editing, and molecular synthesis can be misused to enhance viral virulence, design toxins, or modify human embryos, while ethical and policy discussions lag behind technological advances. This Correspondence calls for proactive, built-in, AI-native safeguards within GenAI tools. With more research and development, emerging AI safety technologies—watermarking, alignment, anti-jailbreak methods, and unlearning—can complement governance policies and provide scalable biosecurity solutions. We also stress the global community's role in researching, developing, testing, and implementing these measures to ensure the responsible GenAI deployment in biotechnology.

Keywords: generative AI; biosecurity

1. Biosecurity Threats of GenAI in Biosciences

GenAI is changing biotechnology research, accelerating drug discovery, protein design, and synthetic biology. It also enhances biomedical imaging, personalized medicine, and lab automation, enabling faster and more efficient scientific advancements. However, these breakthroughs have also raised biosecurity concerns, prompting policy and community discussions [1–4].

The power of GenAI lies in its ability to generalize from known data to the unknown. Deep generative models can predict novel biological molecules that may not resemble existing genome sequences or proteins. This capability introduces dual-use risks and serious biosecurity threats—such models could potentially bypass the established safety screening mechanisms used by nucleic acid synthesis providers[5], which presently rely on database matching to identify sequences of concerns[6]. AI-driven tools could be misused to engineer pathogens, toxins, or destabilizing biomolecules, while AI science agents could amplify risks by automating experimental designs[7].

The research community has recognized biosecurity dangers for over twenty years[8], but AI amplifies and accelerates them. Baker and Church warned that "protein-design technology is vulnerable to misuse for producing dangerous biological agents" and "gene sequence and synthesis data should be collected and stored in repositories that are only queried in emergencies"[1]. Further, a community of scientists signed on to a set of guiding principles in Fall 2024 to "ensure that this technology develops in a responsible and trustworthy manner and that it is safe, secure, and beneficial for all". The creators of the genome foundation model Evo acknowledged its dual-use potential,

 doi:10.20944/preprints202503.1761.v1

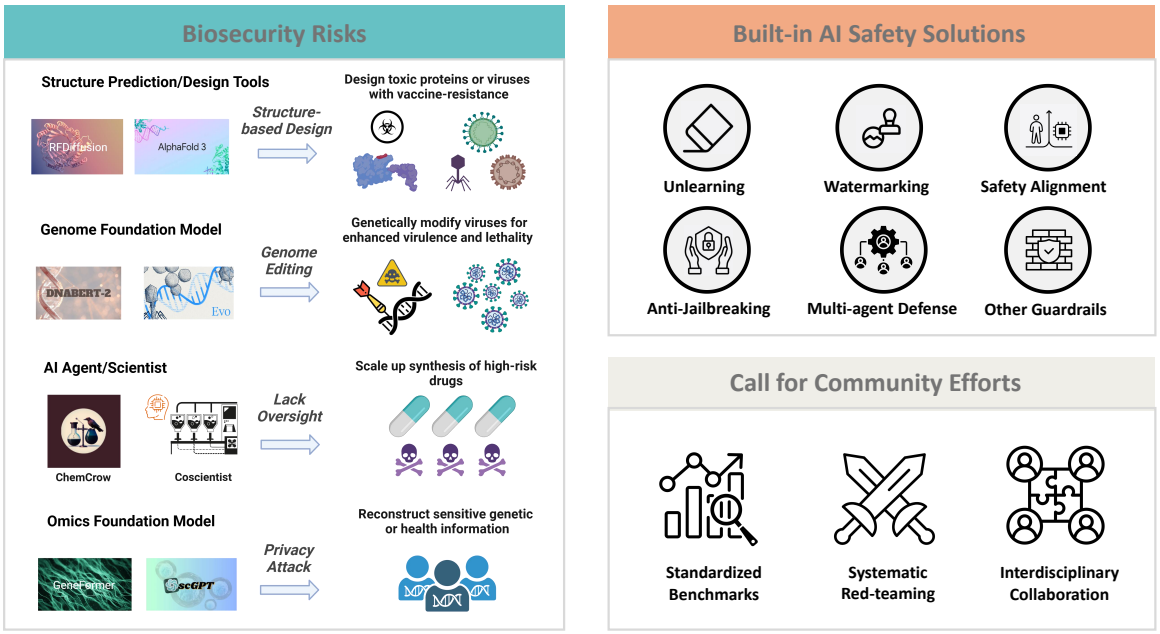

Figure 1. Generative AI-driven biosecurity challenges and emerging AI safeguard technologies. AI foundation models can generate novel biologics—including protein structures, DNA/RNA sequences, and biomarkers—beyond known data, potentially bypassing traditional safety screening. This heightens dual-use risks and privacy concerns, while AI agents further amplify threats by automating research and experimentation. To mitigate these risks, we advocate for built-in AI safeguards such as unlearning, watermarking, safety alignment, and AI agent defenses. Addressing these challenges requires community-wide efforts, including benchmarking, red-teaming, and interdisciplinary collaboration to ensure the responsible deployment of AI in the bioeconomy.

stating that while it could aid therapeutic discovery, it might also facilitate the development of harmful synthetic microorganisms[9]. Similarly, the developers of CRISPR-GPT[10] raised concerns about AI-driven gene editing being misused for modifying viruses or human embryos.

General-purpose AI safety has garnered considerable attention from researchers [4], yet AI biosecurity remains a largely underexplored area. The challenges posed by GenAI in biotechnology are unique due to the dual-use nature of biotechnological applications, the high stakes of genetic and biological manipulation, and a critical lack of cross-disciplinary expertise in both AI safety and biosciences. The complex, domain-specific nature of biotechnological research, combined with limited awareness of its potential risks, further exacerbates these vulnerabilities.

2. Call for Built-In AI Safety Solutions for Biosecurity

Today, many AI developments in biotechnology remain unsafeguarded, exposing serious risks. Closing this gap demands immediate, coordinated action—integrating technical safeguards, fostering global collaboration, and enacting robust policies to ensure responsible innovation in biotechnology.

Technical, built-in safeguards are one approach that could potentially mitigate the misuse risk of AI tools trained on biology. Such safeguards must be proactive, scalable, and effective in countering dual-use risks and malicious exploitation, without significant eroding the beneficial performance of the model. A number of emerging AI safety technologies hold promise but require further research and developments for appropriate biosecurity applications. First, watermarking – the embedding of imperceptible patterns within AI-generated biological designs – enables reliable tracing and auditing of the generated outputs. For example, FoldMark[11] applies watermarking to protein generative models like AlphaFold and RFDiffusion via embedding up to 32-bit identity tracing codes in the model's output, ensuring traceability in AI-designed proteins. Second, safety alignment can train models to avoid generating harmful responses when prompted with malicious queries. Alignment of foundation models is typically achieved via model-level finetuning or training-free controlled decoding. For instance, preference-optimized language models can be aligned to avoid generating pathogenic DNA

 doi:10.20944/preprints202503.1761.v1

sequences[12], preventing AI from inadvertently assisting in the design of biological threats. Third, removing specific harmful or private knowledge from pre-trained AI models through unlearning prevents them from generating dangerous biological constructs. For instance, if a model has been trained to optimize toxin synthesis, targeted unlearning can erase this capability while preserving its utility for beneficial applications[13]. Fourth, AI systems, such as large language models, must be robust against users' attempts to bypass safety restrictions. Anti-jailbreak may require strong reasoning abilities of large language models to accurately infer the intention of malicious users. In biosecurity, this involves training models to recognize and reject prompts that attempt to exploit weaknesses in AI-driven protein/DNA synthesis tools. Fifth, integrating autonomous AI agents into safety frameworks enables real-time monitoring and rapid response to anomalous behavior. For example, the agent defense layer in CRISPR-GPT[10] filters illegal queries or issues warnings when users attempt to generate hazardous biological sequences. Multiple AI agents can collaborate to cross-verify outputs, detect emerging threats, and enforce corrective measures, ensuring that any potential misuse is promptly contained. Lastly, cryptographic technologies have the potential to integrate AI safeguards into remote devices[14], ensuring unbreakable links between safety screening and physical synthesis. Together, these approaches form a framework for technical biosecurity safeguards that complement governance and screening policies.

While awareness of biosecurity risks in this field is growing, we are still in the early stages. The necessary AI technologies remain largely conceptual and under-developed, and current models lack protections. Adding built-in AI safeguards could raise costs and reduce performance – an important tradeoff that must be analyzed through the development and integration of these approaches. Community efforts are urgently needed to develop these safeguards and assess their impact on the bioeconomy.

3. Call for Community Efforts

Beyond technological advancements, we call for the development of standardized benchmarks and systematic red-teaming practices to evaluate and improve AI safety measures. For example, AI-driven benchmarks for identifying unsafe genome sequences are essential. These benchmarks shall be dynamic, and they will enable practitioners to move beyond traditional database matching toward proactive risk prediction. Additionally, robust biosafety prompt benchmarks should be developed to evaluate large language models' responses to misuse requests. These benchmarks should cover diverse biological domains and threat scenarios, reflecting real-world challenges and ensuring practical, reliable defenses.

Red teaming, involving adversarial testing by experts, is crucial for uncovering vulnerabilities in AI systems. Simulated attacks or misuse attempts during training help strengthen model resilience. We advocate for community-driven red teaming efforts that engage interdisciplinary researchers to simulate potential misuse cases and improve model robustness. Establishing a shared repository of test cases and threat models would accelerate learning and response capabilities across biotechnology and AI safety.

The integration of GenAI into biotechnology demands urgent collaboration between AI researchers, scientists, and security experts to preempt dual-use risks. We argue that built-in technical guardrails—spanning model-level constraints, decoding filters, and agent-level defenses—may be a critical approach to ensuring AI tools, such as protein designers or DNA synthesizers, cannot be co-opted for harm. To operationalize this vision, we advocate for three priorities: advancing AI safety research tailored to biology foundation models, establishing standardized risk assessments for AI-bio tools, and developing global monitoring systems to detect emerging threats. By improving our understanding, development, and integration of built-in safeguards in AI-bio tools alongside governance strategies, the biotechnology community can harness the benefits of GenAI while mitigating its biosecurity risks.

 doi:10.20944/preprints202503.1761.v1

Disclaimer: Certain tools and software are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the tools and software identified are necessarily the best available for the purpose.

Conflicts of Interest: Z.Z., A.S.B., A.V., S.G., S.L., S.C., M.B., and J.M. have no competing interests. E.X. has equity in GenBio AI. G.C. has biotech patents and equity in companies: arep.med.harvard.edu/t/. M.W., L.C., Y.Q. invented some of the technologies mentioned in the paper, with patent applications filed by Princeton University and Stanford University. L.C. is a Scientific Advisor to Acrobat Genomics and Arbor Biotechnologies.

References

1. Baker, D.; Church, G. Protein design meets biosecurity, 2024.
2. Bloomfield, D.; Pannu, J.; Zhu, A.W.; Ng, M.Y.; Lewis, A.; Bendavid, E.; Asch, S.M.; Hernandez-Boussard, T.; Cicero, A.; Inglesby, T. Ai and biosecurity: The need for governance. *Science* **2024**, *385*, 831–833.
3. Blau, W.; Cerf, V.G.; Enriquez, J.; Francisco, J.S.; Gasser, U.; Gray, M.L.; Greaves, M.; Grosz, B.J.; Jamieson, K.H.; Haug, G.H.; et al. Protecting scientific integrity in an age of generative AI, 2024.
4. Bengio, Y.; Mindermann, S.; Privitera, D. International AI Safety Report 2025 **2025**.
5. Wittmann, B.J.; Alexanian, T.; Bartling, C.; Beal, J.; Clore, A.; Diggans, J.; Flyangolts, K.; Gemler, B.T.; Mitchell, T.; Murphy, S.T.; et al. Toward AI-Resilient Screening of Nucleic Acid Synthesis Orders: Process, Results, and Recommendations. *bioRxiv* **2024**, pp. 2024–12.
6. Office of Science and Technology Policy. Framework for Nucleic Acid Synthesis Screening. Technical report, Office of Science and Technology Policy (OSTP), Executive Office of the President, 2024. Hosted by the Administration for Strategic Preparedness and Response (ASPR).
7. Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578.
8. Church, G. Synthetic Biohazard Non-proliferation Proposal. https://arep.med.harvard.edu/SBP/Church_Biohazard04c.htm, 2004. Accessed: March 19, 2025.
9. Nguyen, E.; Poli, M.; Durrant, M.G.; Kang, B.; Katrekar, D.; Li, D.B.; Bartie, L.J.; Thomas, A.W.; King, S.H.; Brixi, G.; et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* **2024**, *386*, eado9336.
10. Huang, K.; Qu, Y.; Cousins, H.; Johnson, W.A.; Yin, D.; Shah, M.; Zhou, D.; Altman, R.; Wang, M.; Cong, L. Crispr-GPT: An LLM agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021* **2024**.
11. Zhang, Z.; Jin, R.; Fu, K.; Cong, L.; Zitnik, M.; Wang, M. FoldMark: Protecting Protein Generative Models with Watermarking. *bioRxiv* **2024**.
12. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* **2023**, *36*, 53728–53741.
13. Liu, S.; Yao, Y.; Jia, J.; Casper, S.; Baracaldo, N.; Hase, P.; Yao, Y.; Liu, C.Y.; Xu, X.; Li, H.; et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence* **2025**, pp. 1–14.
14. SecureDNA Foundation. SecureDNA: Free, Secure DNA Synthesis Screening Platform. https://securedna.org/, 2025. Accessed: March 19, 2025.