https://doi.org/10.1038/s41587-025-02650-8

A call for built-in biosecurity safeguards for generative AI tools

Check for updates

The rapid adoption of generative artificial intelligence (AI) 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, and ethical and policy discussions lag behind technological advances. This Correspondence calls for proactive, built-in and AI-native safeguards within generative AI 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 deployment of generative AI in biotechnology (Fig. 1).

Biosecurity threats of generative AI in the biosciences

Generative AI is changing biotechnology research, and accelerating drug discovery, protein design and synthetic biology. It also enhances biomedical imaging, personalized medicine and laboratory automation, which enables faster and more efficient scientific advancements. However, these breakthroughs have also raised biosecurity concerns, which has prompted policy and community discussions[1–4].

The power of generative AI lies in its ability to generalize from known data to the unknown. Deep generative models can predict novel biological molecules that might 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, and 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]. The creators of the

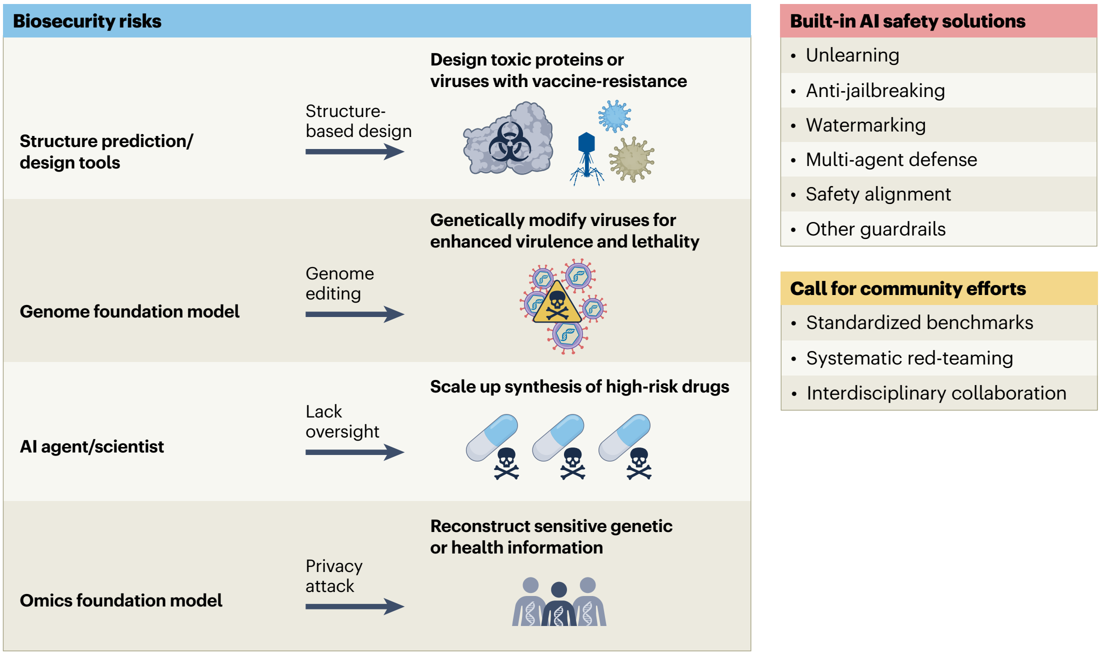
Biosecurity risks
Structure prediction/ design tools
Structure-based design
Design toxic proteins or viruses with vaccine-resistance
Genome foundation model
Genome editing
Genetically modify viruses for enhanced virulence and lethality
AI agent/scientist
Lack oversight
Scale up synthesis of high-risk drugs
Omics foundation model
Privacy attack
Reconstruct sensitive genetic or health information
Built-in AI safety solutions
• Unlearning
• Anti-jailbreaking
• Watermarking
• Multi-agent defense
• Safety alignment
• Other guardrails
Call for community efforts
• Standardized benchmarks
• Systematic red-teaming
• Interdisciplinary collaboration

Fig. 1 | Generative AI-driven biosecurity challenges and emerging AI safeguarding technologies. AI foundation models can generate novel biologics — including protein structures, DNA and RNA sequences, and biomarkers — beyond known data, and potentially bypass traditional safety screening. This heightens dual-use risks and privacy concerns, and 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.

genome foundation model Evo acknowledged its dual-use potential and stated that although 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 generative AI in biotechnology are unique owing 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 the biosciences. The complex, domain-specific nature of biotechnological research, combined with limited awareness of its potential risks, further exacerbates these vulnerabilities.

Call for built-in AI safety solutions for biosecurity

Today, many AI developments in biotechnology remain unsafeguarded, which creates 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 substantially eroding the beneficial performance of the model. Several 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 such as AlphaFold and RFDiffusion via the embedding of up to 32-bit identity tracing codes in the model's output, which ensures traceability in AI-designed proteins. Second, safety alignment can train models to avoid generating harmful responses when prompted with malicious queries. The 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 sequences[12], which prevents AI from inadvertently assisting in the design of biological threats. Third, removing specific harmful or private knowledge from pretrained 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 user 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 or 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, to ensure that any potential misuse is promptly contained. Lastly, cryptographic technologies have the potential to integrate AI safeguards into remote devices to ensure unbreakable links between safety screening and physical synthesis. Together, these approaches form a framework for technical biosecurity safeguards that complement governance and screening policies.

Although awareness of biosecurity risks in this field is growing, we are still in the early stages. The necessary AI technologies remain largely conceptual and underdeveloped, 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 effect on the bioeconomy.

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 should be dynamic, and enable practitioners to move beyond traditional database matching towards proactive risk prediction. Additionally, robust biosafety prompt benchmarks should be developed to evaluate the responses of large language models to misuse requests. These benchmarks should cover diverse biological domains and threat scenarios, to reflect real-world challenges and ensure practical, reliable defenses.

Red teaming, which involves adversarial testing by experts, is crucial for uncovering vulnerabilities in AI systems. Simulated attacks or misuse attempts during training help to 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 generative AI 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 – might 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-biotechnology tools and developing global monitoring systems to detect emerging threats. By improving our understanding, development and integration of built-in safeguards in AI-biotechnology tools alongside governance strategies, the biotechnology community can harness the benefits of generative AI while mitigating its biosecurity risks.

Mengdi Wang[1] ✉, Zaixi Zhang[1], Amrit Singh Bedi[2], Alvaro Velasquez[3], Stephanie Guerra[4], Sheng Lin-Gibson[4], Le Cong[5], Yuanhao Qu[5], Souradip Chakraborty[6], Megan Blewett[7], Jian Ma[8], Eric Xing[8,9] & George Church[10]

[1]AI Lab, Princeton University, Princeton, NJ, USA. [2]Department of Computer Science, University of Central Florida, Orlando, FL, USA. [3]Defense Advanced Research Projects Agency, Ballston, VA, USA. [4]National Institute of Standards and Technology, Gaithersburg, MD, USA. [5]Department of Pathology, School of Medicine, Stanford University, Stanford, CA, USA. [6]Department of Computer Science,

University of Maryland, College Park, MD, USA. [7]Iris Medicine, Palo Alto, CA, USA. [8]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. [9]Mohamed bin Zayed University of AI, Abu Dhabi, United Arab Emirates. [10]Department of Genetics, Harvard Medical School, Harvard University, Boston, MA, USA.
✉e-mail: mengdiw@princeton.edu

Published online: 28 April 2025

References

1. Baker, D. & Church, G. *Science* **383**, 349 (2024).
2. Bloomfield, D. et al. *Science* **385**, 831–833 (2024).
3. Blau, W. et al. *Proc. Natl Acad. Sci. USA* **121**, e2407886121 (2024).
4. Bengio, Y. et al. Preprint at https://doi.org/10.48550/arXiv.2501.17805 (2025).
5. Wittmann, B. J. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2024.12.02.626439 (2024).
6. Fast Track Action Committee On Synthetic Nucleic Acid Procurement Screening. *Framework For Nucleic Acid Synthesis Screening*. (White House Office of Science and Technology Policy, 2024).
7. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. *Nature* **624**, 570–578 (2023).
8. Church, G. A synthetic biohazard non-proliferation proposal. *Harvard Medical School* https://arep.med.harvard.edu/SBP/Church_Biohazard04c.htm (2004).
9. Nguyen, E. et al. *Science* **386**, eado9336 (2024).
10. Huang, K. et al. Preprint at https://doi.org/10.48550/arXiv.2404.18021 (2024).
11. Zhang, Z. et al. Preprint at https://doi.org/10.1101/2024.10.23.619960 (2024).
12. Rafailov, R. et al. Direct preference optimization: your language model is secretly a reward model. In *NIPS'23: Proc. 37th International Conf. Neural Information Processing Systems* (eds Oh, A.) 53728–53741 (Curran Associates, 2023).
13. Liu, S. et al. *Nat. Mach. Intell.* **7**, 181–194 (2025).

Acknowledgements

Certain tools and software are identified in this Correspondence 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.

Competing interests

Z.Z., A.S.B., A.V., S.G., S.L.-G., S.C., M.B. and J.M. have no competing interests. E.X. has equity in GenBio AI. G.C. has biotechnology patents and equity in Lila.AI, DynoTx, Jura.bio, ShapeTx, GC-Tx, ArrivedAI, Nabla.bio, Manifold.bio and Plexresearch. M.W., L.C. and Y.Q. invented some of the technologies mentioned in this Correspondence, with patent applications filed by Princeton University and Stanford University. L.C. is scientific advisor to Acrobat Genomics and Arbor Biotechnologies.