Towards Bio-Security and Safety in the Era of Bio Foundation Models

Zaixi Zhang[1], Amrit Singh Bedi[2], Souradip Chakraborty[3], Emilin Mathew[4], Varsha Saravanan[4], Le Cong[4], Alvaro Velasquez[5], Sheng Lin-Gibson[6], Megan Blewett[8], Jian Ma[7], Eric Xing[7], Russ Altman[4], George Church[9], and Mengdi Wang[1, ✉]

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

ABSTRACT

The rapid rise of bio foundation models (large-scale AI systems trained on biological data such as DNA, RNA, and proteins) is transforming biotechnology, medicine, and synthetic biology. While these models offer immense potential to automate scientific discovery and generate novel compounds, their powerful capabilities also introduce unprecedented biosecurity risks. In this perspective paper, we synthesize the current technological landscape and consolidate expert opinion to argue that a dangerous gap is widening between the capabilities of these AI systems and our capacity for effective governance. Drawing on insights from 130 interviews with leaders across industry, academia, government, and policy, we quantify this concern: 76% of experts view the risk of AI misuse as urgent, and 74% call for new governance frameworks. To bridge this gap, we propose BioSafe: an AI agent designed to proactively evaluate and mitigate risks throughout the entire model lifecycle. By integrating AI-native safeguards such as data filtering, safety alignment, anti-jailbreak techniques, and red teaming, BioSafe offers a pathway to embed safety directly into the development pipeline. Finally, we conclude by emphasizing the urgent need for systematic technical standards, interdisciplinary collaboration, and robust policy frameworks to address the emerging biosecurity challenges posed by bio foundation models.
mw: The scope shall be general AI x bio. It should not be limited to bio foundation models, which is very small compared to agents and other tools. “Bio foundation model” is not even a well recognized words.

1 Emergence of Bio Foundational Models

mw:expand to general AI tools for bio With the extensive accumulation of biological data and the advancement of deep learning technologies, researchers are rapidly adapting the foundation model paradigm to the biological sciences. Inspired by the success of large language models such as GPT[1], Gemini[2], Claude[3], and DeepSeek[4], recent efforts have focused on modeling biological data across multiple levels of the central dogma, including DNA, RNA, proteins, and other biomolecules, to construct a comprehensive, data-driven understanding of life processes. The application of foundation models spans the entire spectrum of biological information, from individual proteins to entire genomes and cellular systems.

The journey starts with **proteins**, the workhorses of the cell that carry out most biological functions. Early efforts like the ESM series (e.g., ESM2[5], ESM3[6]) pioneered this by demonstrating scaling laws in protein language modeling, achieving strong performance on tasks like predicting thermal stability. Motivated by the principle that structure determines function, the AlphaFold series has profoundly advanced the field by accurately predicting 3D protein structures and modeling complex biomolecular interactions. In addition to semantic extraction and functional understanding, numerous studies have incorporated generative models to facilitate the design of proteins with specific functions. For example, RFdiffusion[7] can generate protein conformations tailored to specific scenarios, while RFdiffusionAA[8] and PocketGen[9] further enhance this by enabling full-atom generation.

From the cell’s workers to its blueprints, the next frontier was modeling **nucleic acids**. While proteins execute a wide array of cellular functions, their synthesis and regulation fundamentally depend on RNA, which serves as a crucial intermediary in

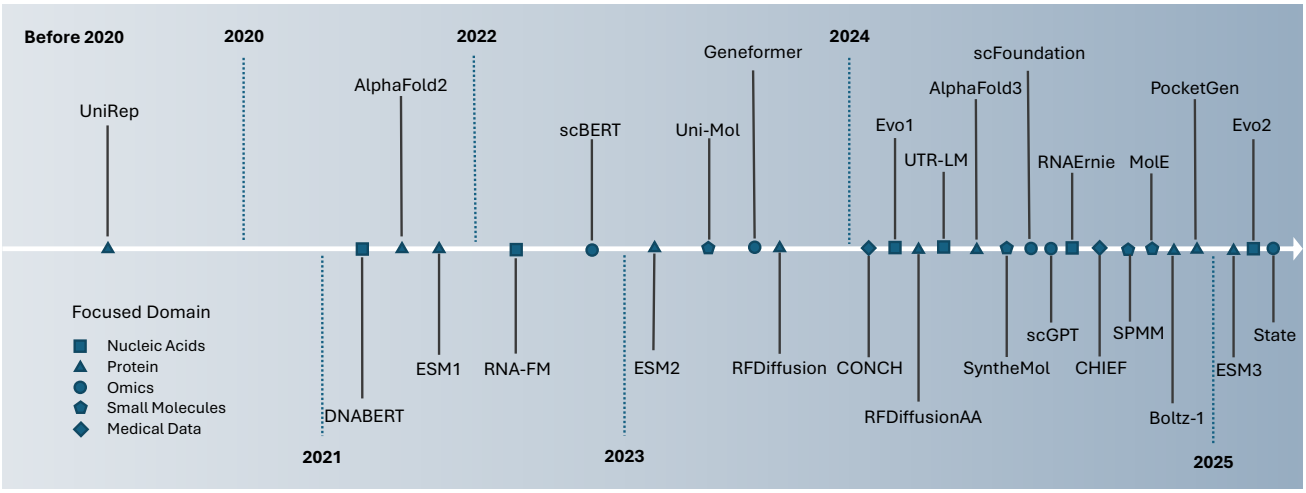

Figure 1. Timeline of Emerging Bio Foundation Models (Pre-2020 to 2025). Squares denote foundation models for nucleic acids (DNA and RNA); triangles represent foundation models for proteins (sequences, structures, functions, etc.); circles indicate foundation models for omics data (e.g., single-cell RNA-seq); pentagons mark foundation models for small molecules; and diamonds signify foundation models for medical data (e.g., pathological images, electronic health records (EHR), and clinical trial data). The timeline illustrates a clear trend of increasing bio foundation model development over time. Just a final comment: This figure is good, and just try if the colors can be improved somehow, if not, that is also fine.

gene expression. Works like RNAGenesis[10] and RNA-FM[11], pre-trained on ncRNA from RNACentral[12], successfully capture the evolutionary information of RNA. As for DNA-related work, although DNA is composed of four nucleobases analogous to RNA, the extremely long context of the genome remains a significant challenge for modeling. Early attempts like DNABERT[13] employed k-mer tokenization and truncated inputs to relatively short reads to reduce computational resource consumption, but their performance was compromised due to context limitations. Enabled by the advent of architectures like Hyena, which are specifically designed to handle ultra-long contexts, Evo and Evo2[14,15] have scaled up both model size and input context, thereby improving zero-shot generalization and generation performance. Ranging from functional elements like transposons to the whole genome scale of certain organisms, the Evo series enables de novo design across different biological levels.

Beyond the central dogma's large macromolecules, AI models have also targeted **small molecules**, which exert profound influence on gene regulation, signaling, and cellular function, making them essential targets for foundational AI models. SyntheMol[16] uses Monte Carlo tree search guided by bioactivity prediction to explore an expansive chemical space (tens of billions of molecules), directly optimizing synthesizability and biological activity; several AI-designed compounds have been successfully synthesized and validated in antibiotic assays. SPMM (Structure–Property Multi-Modal)[17] complements this by learning bidirectional relationships between molecular structure (SMILES) and physicochemical properties, enabling unified embeddings for both structure–property prediction and reaction informatics. Finally, the Uni-Mol framework[18] elevates understanding into three dimensions, pre-trained on hundreds of millions of molecular conformers and protein pockets, facilitating realistic modeling of binding poses, conformational landscapes, and spatially informed molecular generation.

Having modeled the individual components (proteins, genes, and small molecules), the next logical step was to model their integrated behavior within a whole **single cell**. The rise of single-cell foundation models marks a paradigm shift in transcriptome analysis by unifying massive self-supervised learning with advanced neural architectures to capture gene context, cell heterogeneity, regulatory circuitry, and data generation. scBERT[19], one of the earliest models, adapts the bidirectional transformer framework to scRNA-seq by binning expression values and utilizing gene2vec-like embeddings, achieving state-of-the-art cell-type annotation, batch effect correction, and interpretability via attention mechanisms. scFoundation[20] scales this approach to approximately 100 M parameters and pre-trains on tens of millions of cells using an asymmetric encoder–decoder transformer that efficiently handles sparse scRNA-seq data; it sets new records in downstream tasks such as cell-type annotation, drug-response prediction, and perturbation. Geneformer[21] further refines context awareness by encoding relative gene ranks, first on 30M cells and later expanded to 104M, delivering strong embeddings for clustering, in silico perturbation, and regulatory network inference, including models adapted across species. Complementing representation learning, scDiffusion[22] integrates diffusion-based generative modeling with conditional guidance to produce high-fidelity scRNA-seq profiles under specified conditions, model continuous developmental trajectories, and simulate rare cell types via novel gradient-interpolation.

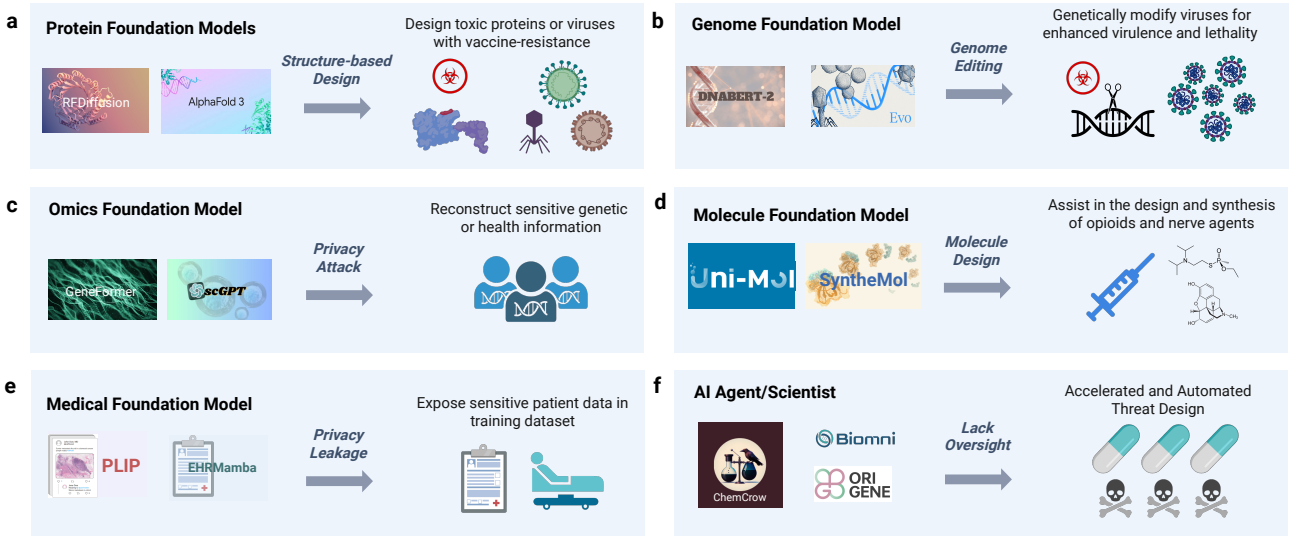

Figure 2. Emerging biosecurity threats across bio-foundation model modalities. **(a)** Structure-based protein design tools (e.g., RFDiffusion, AlphaFold) can be repurposed to engineer toxic proteins or viral components. **(b)** Genome foundation models such as DNABert and Evo could facilitate genetic modification of viral genomes, enhancing virulence or enabling immune escape. **(c)** Omics foundation models, including GeneFormer and scGPT, carry risks of reconstructing sensitive genetic or health-related information via privacy attacks. **(d)** Small-molecule generators like SyntheMol have the potential to design novel toxic compounds. **(e)** Medical foundation models may leak protected patient data from their training sets through membership or property inference attacks. **(f)** AI-driven scientist/agent platforms (e.g., ChemCrow, Biomni, OriGene) may autonomously accelerate threat design.

The final step in this scaling journey moves beyond molecular and cellular modalities to address higher-level biological structures like **tissues and organs**, directly bridging the gap to clinical utility. CHIEF[23] (Clinical Histopathology Imaging Evaluation Foundation) exemplifies this trend: pre-trained on 60,000+ whole-slide images spanning 19 anatomical sites (and 44 TB of histopathology data), it can identify cancer cells, infer tumor origin, predict molecular profiles, and patient prognosis with robust generalization across multiple cohorts. Complementing this, CONCH[24] is a vision-language model trained on over 1.17 million histopathology image–captions pairs. It supports a diverse range of downstream clinical tasks—including image classification, segmentation, captioning, and text–image retrieval without requiring task-specific fine-tuning. Together, CHIEF and CONCH launch a new paradigm in clinical-scale foundation modeling: CHIEF excels in histological pattern recognition and biomarker inference, while CONCH integrates narrative context through image–language alignment. These models pave the way toward adaptable tools that span from tissue-level interpretation to clinical reporting, signaling a promising shift toward holistic, multimodal biomedical AI.

2 The Evolving Biosecurity Threat Landscape

Bio foundation models are revolutionizing biology by enabling the prediction of molecular structures and the design of novel compounds with unprecedented accuracy. However, this power creates a significant **dual-use dilemma**, posing serious biosecurity risks, including the inadvertent or intentional creation of pathogens and toxins, or other destabilizing biomolecules[25–27]. These concerns arise from three primary factors: the broad generalization capabilities of foundation models, the frequent absence of integrated safety mechanisms in academic research, and the open-source ethos that facilitates widespread dissemination of models and datasets. Early expert assessments and empirical audits have already identified concrete risks, ranging from AI-assisted design of toxic or pathogenic proteins to "jailbreak" attacks on DNA language models[28] that can be steered toward virulent outputs. For instance, a 2023 study in Science reported that antimicrobial peptides designed by AI exhibited unexpected toxicity toward human cells, underscoring the possibility of unintended biological harm[29]. The rapid and accelerating deployment of bio foundation models, now emerging at a rate of dozens per year, further amplifies these concerns. This section explores these threat vectors by categorizing them into risks associated with de novo threat design, systemic and privacy vulnerabilities, and the escalating dangers of automated execution (Figure 2).

2.1 De Novo Threat Design

The most direct threat involves repurposing generative AI to design **biohazardous protein sequences**. In the paradigm of AI-assisted protein design, there are approaches that involve direct sequence generation based on language models as well as generative models tailored for inverse folding tasks under structural constraints. These methods are frequently integrated with protein structure prediction models, such as ESMFold[5] or AlphaFold3[30], for preliminary screening, which has significantly enhanced the success rate of functional protein design. However, this paradigm can also be readily adapted for the design of biohazardous proteins. In a study[31], researchers compiled a dataset of proteins associated with toxins or pathogenic factors as templates and subsequently used several models to generate counterparts for each. The generated sequences were then screened using in silico metrics such as pLDDT. The findings showed that many generated proteins preserved toxic functionality and, a significant fraction were able to evade detection by conventional screening tools, revealing a critical gap in current biosecurity safeguards.

This risk extends from individual proteins to the entire genetic blueprint of organisms. **Genome Foundation Models may be jailbroken to design harmful viral DNA**. For instance, the GeneBreaker[28] framework uses an LLM-based agent to craft high-homology prompts, then applies beam-search guided by a pathogenicity model (PathoLM) to steer generation toward pathogen-like DNA. Subsequent BLAST-based screening confirms that Evo-series models can produce sequences closely resembling SARS-CoV-2 spike or HIV envelope regions—underscoring a tangible biosecurity risk. This vulnerability highlights how genome foundation models, if left unprotected, could be repurposed to design viral genomes with enhanced virulence or immune evasion, reinforcing the need for robust safety alignment and output validation mechanisms in future deployments.

The potential for misuse is not limited to biological macromolecules; generative models for **small molecules** pose a similar threat. A critical issue is how to prevent these models from being utilized in the design of toxic molecules. A study[32] demonstrates that Megasyn2, a model initially developed for identifying viral target inhibitory molecules, can be redirected during training by reversing the training objective to prioritize the discovery of toxic molecules. The findings indicate that within a short time, it is capable of generating a substantial number of highly lethal molecules like VX, even at extremely low doses. Such outcomes serve as a stark warning of the potential severe consequences associated with dual-use risks.

2.2 Systemic and Privacy Risks

Beyond the creation of novel threats, significant risks emerge from the ability of models to analyze, manipulate, and expose data from complex biological systems. As omics and single-cell foundation models evolve into full-scale "virtual cells" or "digital twins," they bring heightened biosecurity and privacy risks[33]. These models integrate multimodal data, spatial omics, proteomics, transcriptomics, and clinical metadata to simulate cellular behavior and predict biological responses. This granularity introduces vulnerabilities: sensitive training data may be manipulated or stolen, enabling re-identification or molecular profile reconstruction. The continuous updates and bidirectional data flow of digital twins expand the attack surface, increasing risks of tampering or leakage. As these models become more predictive and actionable, they could inadvertently aid malicious applications, from designing bioactive agents to targeting interventions. Ensuring secure development will require privacy-preserving pipelines, hardened update mechanisms, and proactive threat modeling.

The integration of such rich biological data raises profound privacy concerns, which are most acute in medical foundation models. Trained on sensitive data such as electronic health records, clinical notes, medical images, and patient histories, presents significant privacy vulnerabilities. In particular, membership inference attacks allow adversaries with black-box or white-box access to determine if an individual's data was used during model training, potentially revealing private health conditions or participation in clinical studies. Model inversion and attribute inference attacks go further, enabling reconstruction of personal medical images or demographic traits from model outputs, representing a direct breach of patient confidentiality. Additional techniques, such as shadow-model or gradient-leakage attacks, exploit model explanations, gradients, or updates to expose hidden training data, threatening both individual privacy and institutional data security. These vulnerabilities are exacerbated by trends such as few-shot learning and the use of interoperable health-AI systems, which may unintentionally amplify memorization and leakage of rare cases. As medical models become increasingly integral to healthcare delivery, deploying them without formal protections, such as differential privacy, federated learning, and privacy-aware auditing, poses a negligent risk to patient confidentiality and compliance with regulations like HIPAA or GDPR.

2.3 Automated and Execution Risks

Recent advances in autonomous scientific systems have led to the emergence of powerful AI-driven agents and experimental platforms capable of independently conducting complex biomedical research. Notably, *Biomni* and *OriGene* are agentic systems that leverage large language models and tool integrations to orchestrate multi-step workflows across genomics, proteomics, and therapeutic design[34,35]. These systems significantly lower the expertise and resource thresholds for executing high-level scientific tasks, automating everything from database queries and structural modeling to pathway analysis and compound prioritization. In parallel, execution-capable platforms such as *Coscientist* go one step further by physically interfacing with

robotic labs to autonomously synthesize compounds in response to natural language prompts[36]. While these developments promise to accelerate discovery, they also introduce new dual-use concerns. For example, evaluations of Coscientist reveal a non-negligible probability of executing requests for toxic molecule synthesis, even in the presence of safeguards. Similarly, the ability of agentic systems to propose and simulate the generation of hazardous biomolecules, without adequate security constraints, raises the risk of misuse by non-experts. As such, the convergence of autonomous reasoning and experimental execution underscores the urgent need for built-in safety filters, secure-by-design architectures, and governance frameworks tailored to these emerging capabilities.

3 Current Safeguards and Their Inherent Limitations

Currently, safeguards for biological foundation models are limited and less explored. Protective measures in this domain remain at an early stage of development, and there is an urgent need to establish a comprehensive framework. Nonetheless, some pioneering efforts have emerged. For instance, some governments, such as the United Kingdom, have long implemented a series of stringent control policies for nucleic acid synthesis.[1] These policies encompass multiple dimensions, including sequence screening, user verification, and transaction monitoring, aiming to severely suppress illegal activities and prevent the acquisition of hazardous biological sequences.

mw:Note that existing safeguards are not specific to bio foundation models". they are for general biosecurity. Dont limit our own scope

mw:find recent documents from NTI—bio. they have published a lot of documents and guidelines and biosecurity

Despite the comprehensiveness of these regulatory frameworks mw:there is no regulatory framework yet. we cannot make such statement , the rapid evolution capabilities of generative AI pose potential challenges to their effectiveness. At the model level, some research teams have begun to recognize the necessity of assessing the potential misuse risks associated with their technologies. For example, the DeepMind team convened experts from diverse fields to evaluate the biosafety implications of AlphaFold3.[2] In the case of Evo2, virus sequences with eukaryotic hosts were deliberately excluded during training to prevent the model from acquiring relevant knowledge. Similarly, the developers of ESM3-open took a two-pronged approach by filtering its training data to remove millions of sequences related to viruses and select toxins, and by removing the model's ability to follow text prompts associated with harmful keywords.

Although these approaches reduced the model's capacity to generate certain therapeutic sequences, the robust generalization abilities of foundational models necessitate further comprehensive evaluations. Additionally, large language models (LLMs) are increasingly integrated into the biological domain. Their inherent safety alignment can reject harmful requests, yet the risk of "jailbreaking" remains. To our knowledge, protective efforts in this area remain limited, and the challenge of effective safeguarding persists.

4 Expert Perspectives on the Intersection of AI and Biosecurity

mw:merge this with section 3 As current biosafety interventions struggle to keep pace with AI's accelerating capabilities, understanding how key stakeholders perceive these risks is critical. Despite growing concern, there remains a lack of systematic insight into how experts across sectors—industry, government, academia, and policy—evaluate the scope of these emerging threats or the sufficiency of existing defenses.

To address this gap, we conducted a qualitative study of 130 stakeholders with expertise spanning synthetic biology, AI, and national security. The interviewees were distributed across four key sectors (Figure 3a): Industry (n=43), including representatives from nucleic acid synthesis companies, screening technology providers, and AI companies; Academia (n=41), with researchers from institutions like MIT, Stanford, Harvard, and Lawrence Livermore National Laboratory; Government (n=29), with personnel from the Department of Homeland Security, National Institutes of Health, and the Federal Bureau of Investigation; and Policy (n=17), including leaders from organizations like the American Association for the Advancement of Science and the Engineering Biology Research Consortium. The semi-structured interviews, lasting 30-60 minutes each, used guiding questions focused on emerging risks, the efficacy of current safeguards, and gaps in existing governance frameworks (Figure 3b). Across these interviews, four key themes consistently emerged, as shown in Figure 3c:

- **AI Misuse is an Urgent Concern in Bio:** 76% of participants highlighted the urgency of AI misuse in biological domains. Concern was particularly pronounced among policy experts (94%) and academics (83%).

[1] https://www.gov.uk/government/publications/uk-screening-guidance-on-synthetic-nucleic-acids/uk-screening-guidance-on-synthetic-nucleic-acids-for-users-and-providers

[2] https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/alphafold-3-predicts-the-structure-and-interactions-of-all-lifes-molecules/Our-approach-to-biosecurity-for-AlphaFold-3-08052024.pdf

- **Calls for Governance:** 74% of participants advocated for clearer governance standards to address emerging threats. This call was nearly universal among policy stakeholders (94%) and strongly supported by academia (85%).
- **Current Screening Tools are inadequate:** 47% of interviewees expressed skepticism about existing sequence-based screening systems. This sentiment was strongest in the policy (82%) and academic (59%) sectors.
- **Supports Functional Screening:** A majority of experts supported the development of functional screening methods as a necessary supplement to sequence-based approaches, with especially strong endorsement from government and industry representatives.

Interviewees frequently linked these concerns, revealing a broader pattern of interconnected priorities (Figure 3d). Among interviewees who emphasized the urgency of AI misuse, 91.8% also highlighted inadequacies in current screening methods, 90.8% advocated for functional screening approaches, and 80.2% supported stronger governance measures. These patterns suggest that concern over AI-driven biosafety risks is not isolated but rather reflects a more comprehensive recognition that current biosecurity infrastructure is inadequate to meet emerging threats. 83.6% of interviewees viewed current screening tools as inadequate, while also supporting stronger governance standards and endorsing functional screening approaches. This covariance analysis reveals strong positive correlations among these three attitudes, offering quantitative evidence of systematic alignment across diverse stakeholders in support of comprehensive policy and technological reforms.

Beyond broad thematic alignment, interviews identified distinct operational pressures and governance challenges unique to each domain. Academic researchers emphasized the need for low-friction safeguards that integrate seamlessly into scientific workflows and avoid flagging clearly legitimate requests (e.g., an authorized Ebola researcher ordering Ebola strains). Regulatory gray zones, such as gain-of-function work in BSL-1/2 labs, complicate institutional accountability and raise questions about the consistency of oversight. Research agencies like the NIH face the delicate task of distinguishing legitimate science from dual-use risk without stifling innovation. Investigative bodies such as the FBI, meanwhile, must navigate a rapidly evolving threat landscape, and called for tools with greater interpretability and clearer confidence metrics to support real-time decision-making.

On the industry side, nucleic acid synthesis providers operate under uneven incentives and technical constraints. Some actively screen for hazardous sequences, motivated by safety culture, liability exposure, and reputational risk. Those who opt out cite high costs, technical limitations, and operational complexity. Screening providers are themselves divided: some acknowledge the emerging risk of AI-edited pathogens, yet remain skeptical that truly *de novo* threats are imminent, citing current technological limits. Cloud labs add another layer of complexity: while enabling rapid, high-throughput experimentation, they often fall outside the International Gene Synthesis Consortium (IGSC) frameworks, introducing oversight blind spots as synthesis and experimentation become increasingly abstracted from traditional governance structures.

Our survey revealed a unifying message: biosecurity professionals across sectors are calling for next-generation safeguards capable of keeping pace with the accelerating capabilities of AI-enabled biology.

5 BioSafe: An LLM Agent for Lifecycle Biosecurity Governance

In Figure. 4, we propose *BioSafe*, a task-driven LLM agent designed to proactively evaluate and mitigate biosecurity risks across the entire lifecycle of bio foundation models. Unlike static filters or one-time audits, *BioSafe* functions as a continuous, autonomous co-pilot—capable of perceiving risks, invoking specialized tools, and recommending or executing defensive actions during pretraining, fine-tuning, and inference.

mw:Why proposing a non-existing agent in a perspective paper. It is not scientifically rigorous to say so without evidence. You can say "we map out the emerging technologies to safeguard xxxx"

As an agent, *BioSafe* orchestrates a range of internal defenses that serve as modular capabilities it can invoke, chain, or adapt based on evolving threat surfaces (Figure. 4b). These interventions are strategically deployed across the three main stages of a model's lifecycle.

Data Pre-Processing and Pre-Training Stage: The initial stage focuses on curating the data and establishing controls to prevent the model from learning hazardous information from the outset.

- **Watermarking:** Embedding cryptographic or statistical watermarks into training data or model-generated outputs can provide a method for tracing the origin and lineage of bio-generated molecules or sequences. This helps enforce accountability and deters malicious actors, as any illicit use of the model's outputs may be traceable back to a specific dataset or model version. Watermarks can be designed to be robust against tampering and to survive downstream transformations, ensuring persistent traceability. Within the research domain of large language models, several studies have investigated methods for embedding watermarks into generated content to ensure traceability. For instance, one of the earliest works on watermarking in the context of large language models (LLMs)[37] demonstrated how to embed watermarks using only the model's logits at each generation step. Specifically, their technique partitions the vocabulary into "green" and "red" token sets, where the

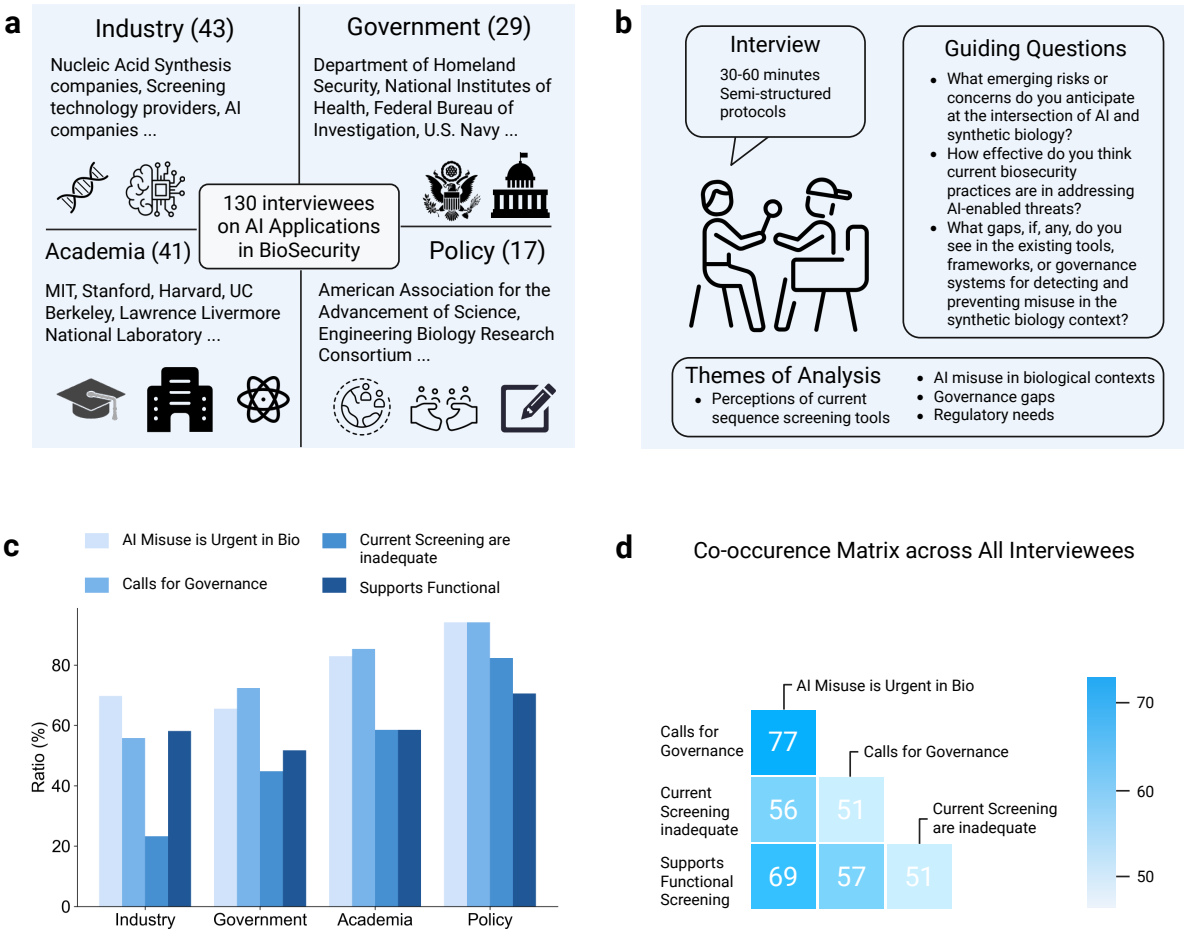

Figure 3. Overview of expert perspectives on the intersection of AI and biosecurity. **(a)** Distribution of 130 interviewees across four key sectors: Industry (n=43), Academia (n=41), Government (n=29), and Policy (n=17), with examples of representative institutions. **(b)** Methodology overview, outlining the semi-structured interview protocol with guiding questions and the primary themes of analysis derived from the responses. **(c)** Sector-specific analysis showing the percentage of interviewees within each sector who affirmed four key propositions: the urgency of AI misuse in biology, the inadequacy of current screening protocols, the need for governance, and support for functional screening. **(d)** Co-occurrence matrix of key themes across all 130 interviewees. The values indicate the number of individuals who hold both intersecting views.

model is biased to select tokens from the green list based on the prior context. This creates statistically identifiable patterns in the output, while remaining imperceptible to human readers. These advancements in watermarking technology thus enhanced copyright protection and content authentication via providing principled way to detect AI generated content. In the field of biology, FoldMark[38] make a pioneering attempt to embed traceable watermarks into protein structures. According to the paradigm of FoldMark, it first trains a watermark autoencoder that can embed the watermark into protein structures without compromising the quality of structures. The second phase involves fine-tuning a protein structure generation model with LoRA and embedding watermarks into the generated structures for traceability.

- **Access Control:** Enforcing strict access controls on sensitive biological datasets—such as pathogen genomes, toxin-encoding sequences, or gene drives—is essential to prevent the unintended or deliberate introduction of dangerous knowledge during model training. This involves not only role-based permissions, usage logging, and risk-based data classification, but also limiting downstream access to models trained on such data. High-risk datasets should be accessible only to authorized personnel under oversight, with access granted based on biosafety risk assessments. In addition, model availability restrictions can further reduce misuse risks: access to model weights, APIs, or fine-tuned checkpoints trained on sensitive biological data may require identity verification, institutional affiliation, or case-by-case approval. By extending access controls beyond data handling to include model usage, organizations can better ensure that both training and deployment phases uphold safety and security standards.

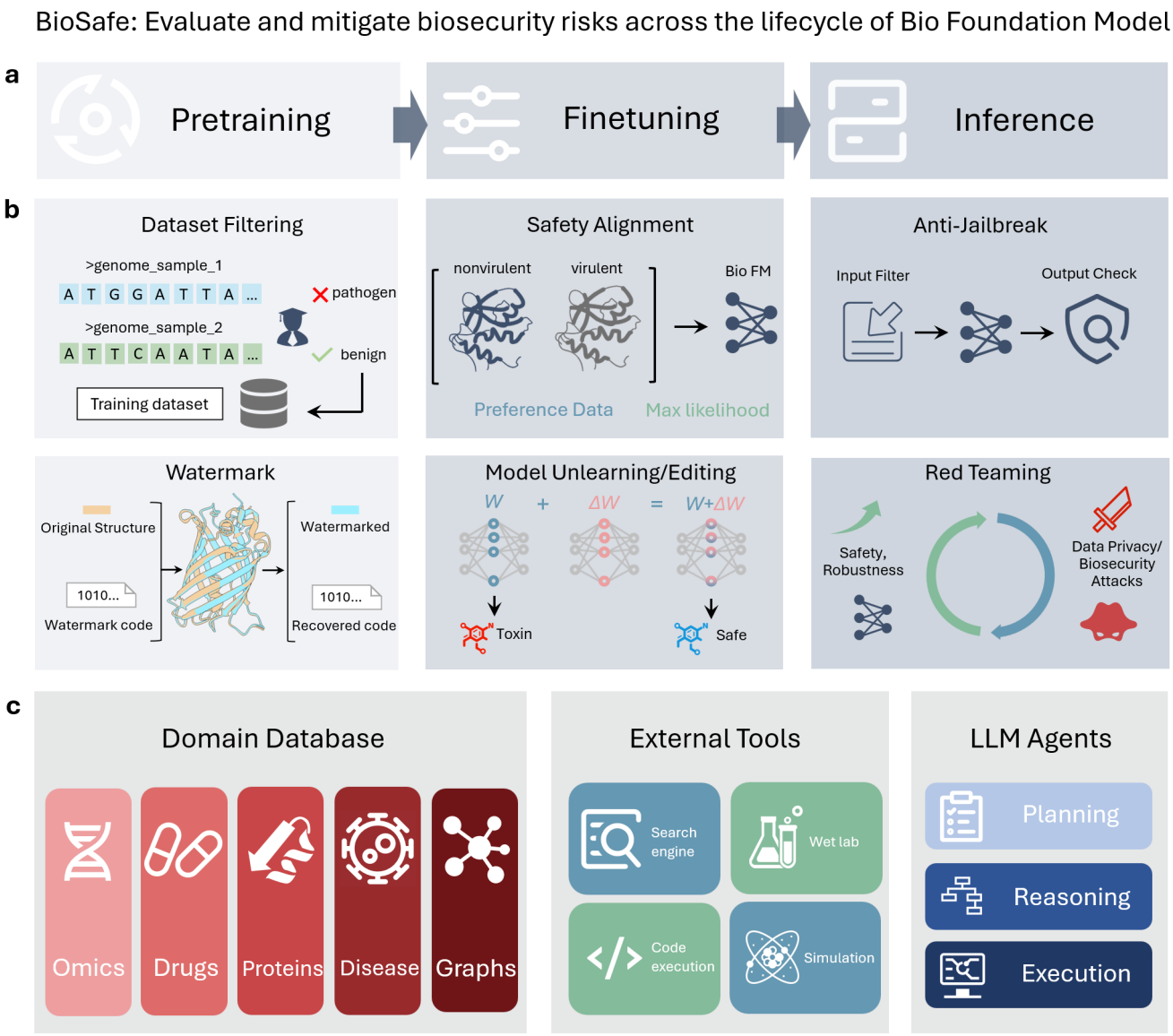

Figure 4. BioSafe: Evaluate and mitigate biosecurity risks across the lifecycle of a Bio Foundation Model. **(a)** The framework operates across the three stages of a model's lifecycle: pretraining, finetuning, and inference. **(b)** Internal safety strategies deployed by BioSafe, including dataset filtering, safety alignment, anti-jailbreak checks, watermarking, model unlearning, and continuous red teaming. **(c)** The agent's capabilities are powered by access to domain databases, external tools like search and simulation, and its core LLM architecture for planning, reasoning, and execution.

- **Dataset Filtering:** Ensuring model safety requires rigorous data-centric curation, applied to both initial training sets and subsequent data augmentation. This process begins with **dataset risk stratification**, where all training sequences are evaluated using alignment tools (e.g., *BLAST*) against known pathogen and toxin databases. High-risk sequences are then excluded or obfuscated to prevent the model from learning to generate hazardous material. This principle extends to **synthetic data augmentation**, where any new sequences generated to expand the dataset must pass through safety-preserving filters. For instance, computational tools can screen for properties like toxicity, ensuring that only benign synthetic data is used for model retraining or fine-tuning. By meticulously curating all data the model learns from, this dual approach provides a foundational layer of safety.

Finetuning Stage: In this stage, the pretrained model is refined to align its behavior with safety standards and to make it more resilient against misuse.

- **Safety Alignment:** During finetuning, models can be aligned with explicit safety, ethical, and regulatory frameworks.

Techniques analogous to Reinforcement Learning with Human Feedback (RLHF)[39] can be used to iteratively guide model outputs toward safe, reliable behavior. By incorporating feedback from domain experts and human evaluators, the model can learn to prioritize outputs aligned with public health and biosecurity standards. Specifically, analogous to the paradigm employed in large language models, it is necessary to identify an appropriate reward policy for evaluating the safety and rationality of the output generated by biological foundation models. Given the inherent complexity and human unreadable nature of biological sequences, in addition to leveraging expert knowledge, we could also integrate classical methods such as sequence alignment into the reward policy training which may better guide the model to avoid generating harmful sequences. Furthermore, considering recent advances that align biological languages (e.g. protein language) with natural language[40], it may also work by directly transferring the RLHF paradigm from large language models to restrict the generation of harmful biological sequences such as those associated with disease-related applications based on natural language.

- **Adversarial Training:** Introducing adversarial examples or simulated misuse scenarios during finetuning helps the model develop robustness against exploitation. Similarly, in the past, a substantial amount of related research has been accumulated in both the image domain and the language domain[41]. These studies utilized adversarial learning to enhance model robustness and ensure security even under input perturbations. Drawing inspiration from these works, we assume that integrating this approach into the training of biological foundation models may also improve the safeguards of model. For example, the model may be exposed to prompt variants that attempt to elicit the generation of harmful compounds or synthetic pathogens. By training the model to resist or refuse such requests, adversarial training helps ensure resilience against real-world misuse attempts.

- **Model Unlearning:** Techniques for selectively removing dangerous or dual-use information—such as sequences encoding lethal toxins—from pre-trained models can significantly reduce downstream risks. Known as "unlearning," this process may involve finetuning against negative examples or gradient-based editing to minimize model recall of hazardous knowledge, without degrading general performance on benign biological tasks. This method has been extensively investigated in large language models, particularly through methods such as gradient ascent or negative preference optimization. By reversing the training loss values of sensitive segments, the model's perplexity regarding these segments is increased, thereby enabling it to forget the associated knowledge. The insights derived from these studies? —such as promoting gradual forgetting rather than at once, enhancing precision by concentrating on specific sensitive segments instead of entire text segments, and aligning the outputs of the original and trained models on positive samples—can effectively eliminate sensitive knowledge while ensuring that performance remains unaffected which offer valuable guidance for the management of biological foundation models, especially given the resemblance between biological sequences and natural language.

Inference Stage: The final stage involves implementing real-time safeguards to monitor and control the model's behavior as it's being used.

- **Anti-Jailbreak Mechanisms:** A comprehensive defense against model misuse requires safeguards at both the input and output levels. At the input level, the primary goal is to manage user prompts to prevent malicious queries from triggering the generation of harmful content. This is achieved through a pre-screening module that combines **prompt classification** to block dangerous requests and **prompt optimization** to rewrite potentially risky queries into safer variants. This defensive layer uses a mix of traditional bioinformatics tools (e.g., *BLAST*) for known threats and deep learning classifiers for novel or obfuscated prompts, securing the model at the crucial user-input interface.

  At the output level, safety is reinforced by a critical screening and filtering layer designed to detect and block any hazardous biological sequences that may still be generated. This involves a two-pronged approach. First, traditional **rule-based filtering** uses homology-based tools like *BLAST* to flag outputs that show high sequence similarity to known pathogens or toxins. To address the limitations of this method against novel threats, this is augmented with advanced **function-based screening**. Tools like *Omnyra*[42] leverage protein language models to assess a sequence's potential functional risk, providing a more future-proof defense by focusing on what a sequence might do rather than what it looks like.

- **Red-Teaming and Adaptive Defense:** Red-Teaming employs specific strategies to design diverse prompt inputs, enabling the model to generate potentially harmful content under controlled conditions. This approach aims to uncover vulnerabilities in the model that could lead to undesirable behavior. In Red-Teaming tests for large language models[43], common strategies include the use of technical slang, reframing prompts, authority manipulation, and even the inclusion of garbled prefixes may work. The outcomes of these tests provide valuable insights to developers, assisting them in considering and implementing security enhancements for the model. Furthermore, this methodology is also worth to be applied to the security evaluation of bio foundation models. Continuous stress testing by interdisciplinary red-teaming efforts—bringing together experts in biology, machine learning, cybersecurity, and ethics—helps uncover novel vulnerabilities and emergent misuse pathways. These teams simulate attack scenarios, test model defenses, and identify gaps that may otherwise go unnoticed. The

insights gained support the deployment of adaptive countermeasures, ensuring the model remains secure and robust as threat landscapes evolve.

- **Inference-Time Alignment:** Inference-time alignment offers a flexible approach to steer bio-foundation model outputs toward safety without retraining. In white-box settings where model logits are accessible, **controlled decoding**[44–46] modifies token-level probabilities during generation. By integrating domain-specific evaluators—such as toxicity predictors or pathogen classifiers[47]—this method can down-weight unsafe tokens in real-time to guide generation toward biologically safe and compliant sequences. When direct access to logits is unavailable (i.e., in **black-box** scenarios), alignment is instead achieved by generating and evaluating full candidate sequences. Common strategies include **parallel sampling** (e.g., Best-of-N)[48–50], which generates multiple outputs and selects the one that scores highest on a safety evaluator, and **sequential refinement**[51–54], which iteratively improves a response using evaluator feedback. Together, these techniques provide a crucial safety layer adaptable to different model access levels, enabling the responsible deployment of powerful bio-foundation models.

Integration with Knowledge and Tools: To reason effectively about biological risk, *BioSafe* interfaces with curated *domain databases* (e.g., omics, proteins, drugs, diseases) and leverages *external tools* such as simulation engines, structure predictors, search APIs, and code execution environments (Figure 4c). These resources enable it to validate claims, simulate molecular effects, and assess potential harm in a biologically grounded manner.

LLM Agent Capabilities: *BioSafe* operates as an LLM agent with modular planning, reasoning, and execution capabilities. It interprets emerging model behaviors, performs situational analysis (e.g., red-teaming and inference risk scoring), plans multi-step mitigation strategies, and reasons over scientific and policy constraints. Inspired by agentic frameworks like GuardAgent and SciAgent, BioSafe offers explainable, auditable, and adaptive responses to diverse biosecurity scenarios.

Self-Evolving Capabilities: Crucially, BioSafe is not a static system. Following our STELLA framework for self-evolving agents[55], it is designed to learn and adapt. As next-generation technologies like virtual cell models emerge, BioSafe can update its knowledge and integrate new tools, ensuring its defensive capabilities remain effective against future threats.

6 Conclusion

This Perspective has consolidated expert consensus on the urgent biosecurity risks of bio-foundation models, drawing on insights from 130 interviews with leaders across sectors. In response, we introduced *BioSafe*, a framework for lifecycle governance that embeds technical safeguards directly into model development and deployment. The capabilities these models gain from vast datasets could be misused to engineer pathogens, to design novel toxins, or to extract sensitive genetic information, demanding a coordinated and forward-looking response.

Addressing this challenge requires a multi-pronged strategy. First, we must develop and standardize **robust technical safeguards**, such as the watermarking, access controls, and safety alignment protocols integrated into our proposed BioSafe framework. Second, progress depends on deep **interdisciplinary collaboration** that unites biologists, computational scientists, ethicists, and policymakers to anticipate threats and foster a culture of responsibility. Finally, these efforts must be supported by **adaptive global policy**, including regulatory oversight, model registration, and compliance frameworks targeting dual-use bio-AI technologies.

Models in biomedical domains present distinct ethical hurdles, from the intensely sensitive nature of patient data to the opacity of model decision-making and amplified risks from privacy attacks. Without embedding robust privacy-preserving techniques and transparency mechanisms, their adoption in critical health applications will remain precarious. As bio-AI matures, we must commit to integrating safeguards, regulatory guardrails, and ethical oversight from inception. Only through such preemptive, multidisciplinary action can we ensure these powerful tools serve humanity without magnifying risk.

References

1. OpenAI. Introducing chatgpt. *OpenAI Blog* (2022). URL https://openai.com/index/chatgpt/.
2. Team, G. *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
3. Anthropic. Claude. https://www.anthropic.com/claude. Accessed on July 6, 2025.
4. Liu, A. *et al.* Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
5. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). URL https://www.science.org/doi/abs/10.1126/science.ade2574. https://www.science.org/doi/pdf/10.1126/science.ade2574.

6. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025). URL https://www.science.org/doi/abs/10.1126/science.ads0018. https://www.science.org/doi/pdf/10.1126/science.ads0018.

7. Watson, J. L. *et al.* De novo design of protein structure and function with rfdiffusion. *Nature* **620**, 1089–1100 (2023). URL https://doi.org/10.1038/s41586-023-06415-8.

8. Ahern, W. *et al.* Atom level enzyme active site scaffolding using rfdiffusion2. *bioRxiv* (2025). URL https://www.biorxiv.org/content/early/2025/04/10/2025.04.09.648075.1. https://www.biorxiv.org/content/early/2025/04/10/2025.04.09.648075.1.full.pdf.

9. Zhang, Z., Shen, W. X., Liu, Q. & Zitnik, M. Efficient generation of protein pockets with pocketgen. *Nature Machine Intelligence* 1–14 (2024).

10. Zhang, Z. *et al.* Rnagenesis: A generalist foundation model for functional rna therapeutics. *bioRxiv* 2024–12 (2024).

11. Chen, J. *et al.* Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions (2022). URL https://arxiv.org/abs/2204.00300. 2204.00300.

12. Rnacentral: a hub of information for non-coding rna sequences. *Nucleic Acids Research* **47**, D221–D229 (2019).

13. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* **37**, 2112–2120 (2021). URL https://doi.org/10.1093/bioinformatics/btab083. https://academic.oup.com/bioinformatics/article-pdf/37/15/2112/57195892/btab083.pdf.

14. Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with evo. *Science* **386**, eado9336 (2024). URL https://www.science.org/doi/abs/10.1126/science.ado9336. https://www.science.org/doi/pdf/10.1126/science.ado9336.

15. Brixi, G. *et al.* Genome modeling and design across all domains of life with evo 2. *bioRxiv* (2025). URL https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918. https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918.full.pdf.

16. Swanson, K. *et al.* Generative ai for designing and validating easily synthesizable and structurally novel antibiotics. *Nature Machine Intelligence* **6**, 338–353 (2024).

17. Chang, J. & Ye, J. C. Bidirectional generation of structure and properties through a single molecular foundation model. *Nature Communications* **15**, 2323 (2024).

18. Zhou, G. *et al.* Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations* (2023). URL https://openreview.net/forum?id=6K2RM6wVqKu.

19. Yang, F. *et al.* scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence* **4**, 852–866 (2022).

20. Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nature methods* **21**, 1481–1491 (2024).

21. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

22. Luo, E., Hao, M., Wei, L. & Zhang, X. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics* **40**, btae518 (2024).

23. Wang, X. *et al.* A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024). URL https://doi.org/10.1038/s41586-024-07894-z.

24. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863–874 (2024).

25. Wang, M. *et al.* A call for built-in biosecurity safeguards for generative ai tools. *Nature Biotechnology* 1–3 (2025).

26. Nuclear Threat Initiative. Developing guardrails for ai biodesign tools. Online report (2024). URL https://www.nti.org/analysis/articles/developing-guardrails-for-ai-biodesign-tools/. Accessed: 2025-05-12.

27. Baker, D. & Church, G. Protein design meets biosecurity (2024).

28. Zhang, Z., Zhou, Z., Jin, R., Cong, L. & Wang, M. Genebreaker: Jailbreak attacks against dna language models with pathogenicity guidance. *arXiv preprint arXiv:2505.23839* (2025).

29. Wong, F., de la Fuente-Nunez, C. & Collins, J. J. Leveraging artificial intelligence in the fight against infectious diseases. *Science* **381**, 164–170 (2023).

30. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493–500 (2024). URL https://doi.org/10.1038/s41586-024-07487-w.

31. Wittmann, B. J. *et al.* Toward ai-resilient screening of nucleic acid synthesis orders: Process, results, and recommendations. *bioRxiv* (2024). URL https://www.biorxiv.org/content/early/2024/12/04/2024.12.02.626439. https://www.biorxiv.org/content/early/2024/12/04/2024.12.02.626439.full.pdf.

32. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence* **4**, 189–191 (2022). URL https://doi.org/10.1038/s42256-022-00465-9.

33. Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045–7063 (2024).

34. Huang, K. *et al.* Biomni: A general-purpose biomedical ai agent. *bioRxiv* 2025–05 (2025).

35. Zhang, Z. *et al.* Origene: A self-evolving virtual disease biologist automating therapeutic target discovery. *bioRxiv* 2025–06 (2025).

36. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023). URL https://doi.org/10.1038/s41586-023-06792-0.

37. Kirchenbauer, J. *et al.* A watermark for large language models (2024). URL https://arxiv.org/abs/2301.10226. 2301.10226.

38. Zhang, Z. *et al.* Foldmark: Protecting protein generative models with watermarking. *bioRxiv* (2024).

39. Dai, J. *et al.* Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773* (2023).

40. de Almeida, B. P. *et al.* A multimodal conversational agent for dna, rna and protein tasks. *Nature Machine Intelligence* 1–14 (2025).

41. Bai, T., Luo, J., Zhao, J., Wen, B. & Wang, Q. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356* (2021).

42. Name, A. Meet h4d team omnyra. *StanfordH4D* (Year). URL https://stanfordh4d.substack.com/p/meet-h4d-team-omnyra.

43. Perez, E. *et al.* Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).

44. Mudgal, S. *et al.* Controlled decoding from language models (2024). URL https://arxiv.org/abs/2310.17022. 2310.17022.

45. Chakraborty, S. *et al.* Transfer q star: Principled decoding for llm alignment (2024). URL https://arxiv.org/abs/2405.20495. 2405.20495.

46. Khanov, M., Burapacheep, J. & Li, Y. Args: Alignment as reward-guided search (2024). URL https://arxiv.org/abs/2402.01694. 2402.01694.

47. Dip, S. A. *et al.* Patholm: Identifying pathogenicity from the dna sequence through the genome foundation model (2024). URL https://arxiv.org/abs/2406.13133. 2406.13133.

48. Jinnai, Y., Morimura, T., Ariu, K. & Abe, K. Regularized best-of-n sampling to mitigate reward hacking for language model alignment (2024). URL https://arxiv.org/abs/2404.01054. 2404.01054.

49. Amini, A., Vieira, T. & Cotterell, R. Variational best-of-n alignment (2024). URL https://arxiv.org/abs/2407.06057. 2407.06057.

50. Beirami, A. *et al.* Theoretical guarantees on the best-of-n alignment policy (2024). URL https://arxiv.org/abs/2401.01879. 2401.01879.

51. Madaan, A. *et al.* Self-refine: Iterative refinement with self-feedback (2023). URL https://arxiv.org/abs/2303.17651. 2303.17651.

52. Chao, P. *et al.* Jailbreaking black box large language models in twenty queries (2024). URL https://arxiv.org/abs/2310.08419. 2310.08419.

53. Mehrabi, N. *et al.* Flirt: Feedback loop in-context red teaming (2024). URL https://arxiv.org/abs/2308.04265. 2308.04265.

54. Chakraborty, S. *et al.* Review, refine, repeat: Understanding iterative decoding of ai agents with dynamic evaluation and selection (2025). URL https://arxiv.org/abs/2504.01931. 2504.01931.

55. Jin, R., Zhang, Z., Wang, M. & Cong, L. Stella: Self-evolving llm agent for biomedical research. *arXiv preprint arXiv:2507.02004* (2025).

Acknowledgements

Author contributions statement

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Mengdi Wang.

A Explaination of Terms

Term	Definition
Pathogen	A biological agent that can cause disease in its host, such as bacteria, viruses, fungi, or parasites.
Toxin	A poisonous substance produced by living organisms that can cause disease or death when introduced into the body.
Omics	A branch of biology that studies the complete set of a particular type of biomolecule or molecular process in an organism, such as genomics (genome), proteomics (proteome), and metabolomics (metabolome).
Genome	The complete set of genetic material in an organism, typically referring to nuclear DNA in eukaryotes, but can include other DNA such as mitochondrial or chloroplast DNA.
Nucleic acids	Macromolecules that store and transmit genetic information in living organisms, composed of nucleotides, which consist of a sugar, a phosphate group, and a nitrogenous base. The two main types are DNA and RNA.
CRISPR	A gene-editing technology that uses the CRISPR-Cas9 system to make precise changes to the DNA of living organisms, based on a natural defense mechanism in bacteria.
Embryo	The early stage of development of a multicellular organism, typically from fertilization until the end of the eighth week in humans, after which it is called a fetus.
Bioweapon	A type of weapon that uses biological agents, such as bacteria, viruses, or toxins, to cause disease or death in humans, animals, or plants.

Table S1. Glossary of key biological terms

Term	Definition
Foundation Models	Large deep learning models pre-trained on vast, general-purpose datasets, designed to be versatile and fine-tuned for specific applications.
Deep Generative Models	Machine learning models that use deep neural networks to generate new data similar to their training data.
Fine-Tuning	The process of adapting a pre-trained model to a new task by further training it on a new dataset, especially useful when the new dataset is small.
Reinforcement Learning	A machine learning paradigm where an agent learns to maximize a reward signal by interacting with an environment.
Retrieval-Augmented Generation	A technique that enhances large language models by providing them with external information to improve the accuracy and relevance of their generated text.
Self Supervised Learning	A machine learning method where a model learns from unlabeled data by generating its own labels or supervisory signals.
AI Agent	A software system that uses artificial intelligence to perform tasks or make decisions based on input prompts.
Multi-Agent System	A system consisting of multiple AI agents that interact with each other and their environment to achieve individual or collective objectives.
Jailbreak Attacks	Techniques used to manipulate AI models, particularly large language models, into producing incorrect or fabricated information that circumvents their built-in safety mechanisms.
Membership Inference Attacks	Attacks where an adversary tries to determine whether a specific data record was used to train a machine learning model.
Watermark	In AI, a technique to mark or identify content generated by an AI model, often used to verify the origin or authenticity of the content.
Surrogate Model	A simpler model used to approximate the behavior of a more complex or computationally intensive model, often for interpretability or efficiency reasons.
Multi-Modality Model	A machine learning model that can process and integrate information from multiple different types of data, such as text, images, and audio.
Parameter Efficient Finetuning	Techniques for fine-tuning large pre-trained models by updating only a small subset of their parameters, to make the process more computationally efficient.
Prompt	The input text or instruction provided to an AI model, particularly large language models, to guide its output.
Hallucination	When an AI model, especially a large language model, generates incorrect or fabricated information that is presented as if it were true.
Transformers	A type of neural network architecture that uses self-attention mechanisms to process sequences, such as text, in parallel and understand the relationships between different parts of the sequence.
Embodied AI	Artificial intelligence systems that are integrated into physical entities, like robots, to interact with and learn from their environment through sensors and actuators.
Backdoor Attacks	Attacks where an adversary modifies a machine learning model during its training phase to behave in a specific, malicious way when presented with certain inputs or triggers.
Property Inference Attacks	Attacks where an adversary tries to infer specific properties or characteristics of the training data used to train a machine learning model, by analyzing the model's behavior or outputs.

Table S2. Glossary of Key Machine Learning Terms

B Survey Methodology

We led an in-depth qualitative survey of stakeholders across industry, government, academia, and policy to highlight emerging biosecurity concerns at the intersection of artificial intelligence, synthetic biology, and governance. Over a ten-week period, we conducted 130 semi-structured discovery interviews with individuals spanning four key sectors:

- 43 professionals from private sector companies including nucleic acid synthesis companies, screening technology providers, and technology and AI companies.
- 41 academics and researchers spanning disciplines such as microbiology, virology, machine learning, and synthetic genomics. Institutions represented included the Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Stanford Biosecurity, MIT Media Lab, Wyss Institute at Harvard University, and the Gladstone Institutes.
- 29 representatives from government agencies, including public health authorities, national security offices, and regulatory bodies. These organizations included the Department of Homeland Security, National Institutes of Health, Federal Bureau of Investigation, U.S. Navy, National Security Commission on Emerging Biotechnology, National Institute of Standards and Technology, and the Defense Innovation Unit.
- 17 non-governmental policy experts and think tank leaders specializing in biosecurity advocacy and AI governance. Organizations represented included the Federation of American Scientists, American Association for the Advancement of Science, Engineering Biology Research Consortium, and Centre for Long-Term Resilience.

Interviewees were selected to reflect a diversity of institutional roles, disciplinary backgrounds, and viewpoints, with many recognized as experts in their respective fields. Each interview lasted between 30 to 60 minutes and was conducted using video conference. Interviews followed a semi-structured protocol, beginning with open-ended prompts about current and future biosecurity risks and then progressing organically based on the interviewee's expertise and perspectives. To ensure consistency across interviews while allowing flexibility, we used a core set of guiding questions as listed below:

- What emerging risks or concerns do you anticipate at the intersection of artificial intelligence and synthetic biology?
- How effective do you think current biosecurity practices are in addressing AI-enabled threats?
- What gaps, if any, do you see in the existing tools, frameworks, or governance systems for detecting and preventing misuse in the synthetic biology context?
- What types of interventions do you think are most needed to strengthen safeguards in biosecurity?
- What role should your sector play in shaping the future of biosecurity as capabilities evolve?

To analyze and quantify interview responses, we developed a framework around four key themes: (1) AI misuse in biological contexts, (2) perceptions of current sequence screening tools, (3) attitudes toward functional or AI-native screening approaches, and (4) governance gaps and regulatory needs. Each interview was reviewed and assigned a binary relevance score (1 = substantial mention, 0 = no substantial mention) for each theme. A theme was coded as "1" only if the interviewee offered a substantive comment that demonstrated support, engagement, or direct relevance to the topic such as describing concrete practices, expressing informed perspectives, or endorsing the importance of the issue. Passing mentions or unrelated concerns were not counted.

This framework allowed us to quantify the prevalence of each theme across all 130 interviews, compare thematic engagement across stakeholder groups, and identify patterns in concern and/or alignment across sectors.

C Interview Results

Table S3. Summary of Interview Themes by Sector

Sector	AI Misuse is Urgent	Screening Tools are Inadequate	Supports Functional Screening	Calls for Governance Standards
Industry	30	10	25	24
Government	19	13	15	21
Academia	34	24	24	35
Policy	16	14	12	16
Total	**99**	**61**	**76**	**96**