CellPress
OPEN ACCESS

Trends in
Genetics

Review

A safety-centric perspective on innovation and risk in the use of artificial intelligence in genomics

Hassan A. Hassan[1,2,#], Aleksandar Anžel[3,#], Bahar İlgen[3], Marina Luchner[4], Patrick Schramowski[5,6,7], Anja Blasse[3], Johannes U. Mayer[2,8,9], Katharina Ladewig[3,10], Georges Hattab[3,11,*], and Maximilian Sprang [2,9,*]

Adopting a safety-centric approach, this article explores how generative artificial intelligence (AI), and more specifically, foundation models for biological sequences, can exacerbate data quality issues, technical biases, and dual-use potential, particularly in critical applications such as clinical genetics, precision medicine, and pathogen engineering. This work centres on how misuse risks emerge throughout the innovation pipeline and how these intersect with the growing accessibility of generative genomic models. Particular attention is given to dual-use governance and infrastructure hardening in sequence analysis workflows. The work aims to provide scientists, regulators, and policymakers with a toolkit to discuss beneficial innovation in genomic AI while maintaining robust safeguards against harm and misuse.

Highlights

Foundation models such as Nucleotide Transformer and Evo models have emerged as transformative tools, enabling multimodal DNA–RNA–protein design with demonstrated capabilities in identifying regulatory elements and generating functional biological elements including CRISPR-Cas systems and transposon arrays.

The first viable artificial intelligence-designed bacteriophage genomes have been successfully created and tested in laboratory settings, with some exhibiting greater fitness and faster lysis dynamics than wild-type phages, proving that the principle of genomic design is practically feasible and elevating dual-use risk from a theoretical to an immediate concern.

Multiple complementary techniques are being developed to enhance model transparency, validate predictions, and enable the detection of model failures in high-stakes applications.

Implemented technical safeguards, including data filtering to remove human pathogens, are not sufficient to protect against the potential harms of genomics foundation models and need to be improved.

The EU AI Act has classified genomic artificial intelligence applications as high risk, establishing new regulatory standards that mandate rigorous bias assessments, transparency requirements, and comprehensive risk management throughout the development and deployment lifecycle.

Overview of artificial intelligence in biological sequence analysis

Ever since the transformer architecture was introduced in 2017, artificial intelligence (AI) has been increasingly applied in science. AI methods in biology range from supervised methods, where a target or output is paired with a particular sample feature, to entirely self-supervised methods, where one learns to model the data via an objective function, such as a fill-in-the-gap exercise. There are also stacked models that combine elements of both. Due to the vast amount of data involved, self-supervised methods (hereafter referred to as 'generative AI' or simply AI) have proven particularly applicable to research focusing on biological sequences such as DNA, RNA, peptides.

Advances in generative AI methods have driven the convergence of AI and synthetic biology, making it a priority for the scientific community and policymakers. The EU has recently mobilised significant resources through RAISE (Resource for AI Science in Europe, a new virtual institute) and Horizon Europe to accelerate this synergy, as has the UK government via its Science and Technology Framework[i,ii]. Reports from early 2025 emphasise the potential of AI to speed up the development, commercialisation, and governance of biotechnology products [1]. Similarly, the German government's Hightech Agenda prioritises interdisciplinary research combining AI and biotechnology. Moreover, AI has become increasingly integrated into genomics, with about one in nine industry job postings (as of May 2025) requiring AI-related skills. Figure 1 illustrates this steady rise of AI applications across different omics domains, reflecting the broader integration of AI into biological research and innovation.

AI techniques have advanced many areas of biology, including genomics, proteomics, metagenomics, medical image analysis, and text mining [2–6]. For example, Human-AI

© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Collaborative Genome Annotation (HAICoGA) frameworks combine human expertise with large language models (LLMs) to improve the accuracy and efficiency of identifying functional genomic elements [7], addressing a key limitation of purely automated approaches.

In natural language processing (NLP), foundation models (FMs) have predominantly relied on transformer architectures, a class of neural networks built around self-attention mechanisms to model relationships within sequences of tokens. The notion of foundation models (FMs) was introduced in 2021 to describe early GPT/BERT-style architectures (GPT: generative pre-trained transformer, BERT: Bidirectional encoder representations from transformers). More recently, state-space models such as Mamba have emerged as competitive and more efficient alternatives for long-context data, including DNA sequences, where they effectively capture long-range dependencies [8]. In the visual domain, Vision Transformers have replaced convolutional neural networks as the standard backbone, while diffusion and flow-matching models now set the state of the art in image, video, and audio generation [9], with growing evidence that such generative frameworks can surpass autoregressive transformers for certain language tasks and constraints [10,11], as illustrated in comparative analyses of sequence generation methods in Figure 2.

Advances in NLP and genomics have paved the way for targeted application in modelling. Despite some progress, genetic causes of disease remain unresolved. This is partly due to the historically shaped structure of human genetic variation and additional complexity caused by gene flow, nonadaptive selection, and genetic drift. Additionally, genome-wide association studies and Mendelian randomisation are hindered by insufficient population stratification, polygenicity, linkage disequilibrium, and dynastic effects [12–14]. FMs trained on large DNA, RNA, and protein datasets offer a potential route to precision medicine by linearising complex genetic relationships via invariant representations and already demonstrate broad utility in prediction, design, and analysis tasks, though their clinical deployment faces significant technical and ethical challenges [15]. Notable genomic FMs include the Nucleotide Transformer, which can identify regulatory elements in an unsupervised manner [16], and the Evo family of models, which enable multimodal DNA–RNA–protein design, including functional CRISPR-Cas systems and transposon arrays, and were used to generate the first viable AI-designed phage genome while excluding eukaryotic viruses from training data [17–19]. Nonetheless, these systems remain vulnerable to distribution shifts: for instance, Nucleotide Transformer embeddings lose 23% accuracy on rare tumour types absent from pre-training [20], and models trained predominantly on European genomes perform 18–41% worse on South Asian regulatory elements [21], while even state-of-the-art Evo architecture fails to generalise to eukaryotic genomes, producing nonviable sequences in 94% of yeast chassis trials unless trained on data spanning the full tree of life [17,18].

The opacity of biological FMs poses an ongoing challenge: to what extent can we trust their predictions, and do they reflect underlying biological mechanisms or merely superficial correlations [22–24]? Opacity in high-stakes domains like medicine and synthetic biology erodes trust and can obscure safety-critical failures. A model's hazardous capabilities might remain undetected due to inaccessible internal logic. Interpretability techniques like gradient-based attribution, *in silico* mutagenesis, attention visualisation, and latent probing offer post-hoc diagnostics. However, they suffer from noisy attributions, lack of causality, and poor generalisability across genomic contexts [24–27]. While these techniques do not mitigate underlying model biases, they play a crucial role in revealing them and thereby contribute to model transparency and safety assurance [28]. To address these limitations, interpretability must become integral to model design. Architectures such as ExplaiNN and ARGMINN embed motif-level priors directly, while causal inference and human-in-the-loop validation can improve clinical variant pathogenicity assessment and policymaking [24,29–31].

[1]DeOxy Tech, The Junction, Suite G01, Greater Manchester, Salford M50 3SG, United Kingdom
[2]Department of Dermatology, University Medical Center of the Johannes Gutenberg University, Langenbeckstraße 1, Mainz 55131, Germany
[3]Centre for Artificial Intelligence in Public Health Research (ZKI-PH), Robert Koch Institute, Nordufer 20, Berlin 13353, Germany
[4]Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK
[5]German Research Center for Artificial Intelligence (DFKI), Landwehrstraße 50A, Darmstadt 64293, Germany
[6]Hessian Center for Artificial Intelligence (hessian.AI), Karolinenplatz 5, Darmstadt 64289, Germany
[7]Department of Computer Science, Darmstadt University of Technology, Karolinenplatz 5, Darmstadt 64289, Germany
[8]Research Center for Immunotherapy (FZI), University Medical Center of the Johannes Gutenberg-University Mainz, Mainz 55131, Germany
[9]Institute of Quantitative and Computational Biology, Johannes Gutenberg University, Johannes-von-Müller-Weg 6, Mainz 55128, Germany
[10]Medical University Lausitz - Carl Thiem (MUL-CT), Cottbus, Germany
[11]Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 14, Berlin 14195, Germany
[#]These authors contributed equally to this work

*Correspondence:
hattabg@rki.de (G. Hattab) and
masprang@uni-mainz.de (M. Sprang).

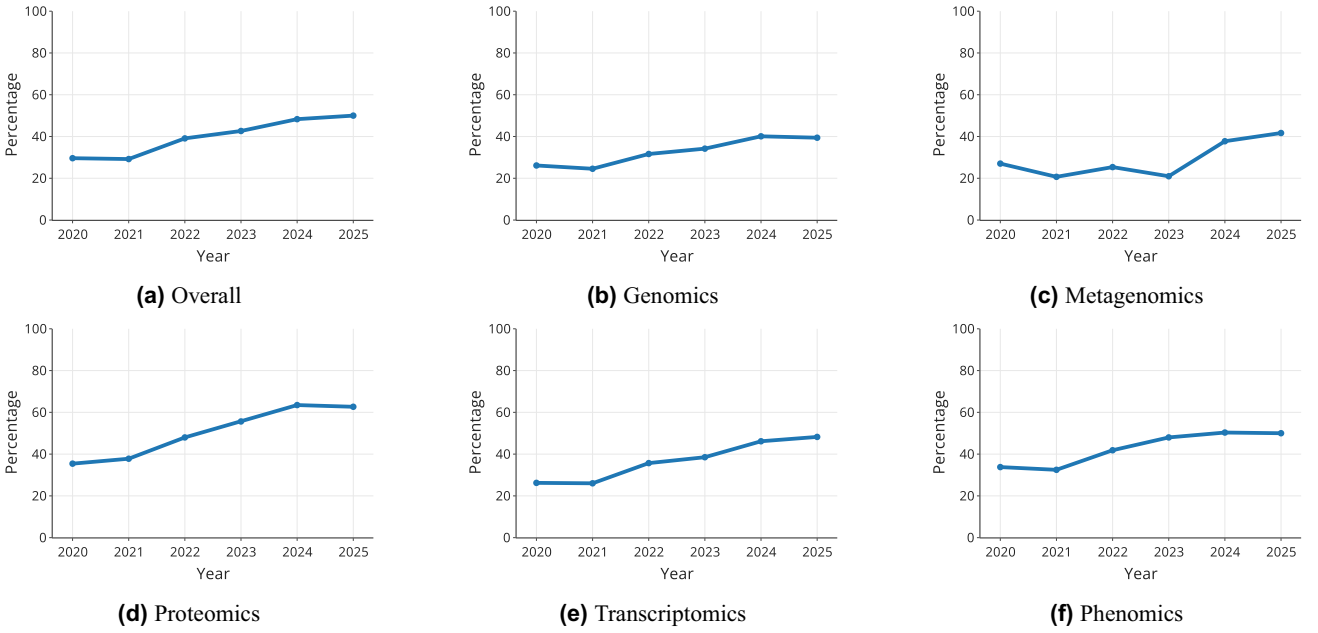


Figure 1. The rise of AI across omics research. The figure illustrates the proportion of papers in the *Bioinformatics* journal published by the Oxford University Press that focus on different omics domains and apply artificial intelligence techniques, as well as the proportion of papers that apply some form of deep learning overall.

Despite their successes outlined earlier, generative AI models applied to DNA analysis are arguably still in the early stages of development and do not generate sequences outside of the distribution (Peppin, A. *et al.*, 2025, FAccT, conference) [32,33]; however, their breakthroughs such as next-generation expression and variant effect prediction [34,35] and the more mature generative capabilities of protein models [36] have raised awareness among scientists and policymakers. They recommend the implementation of safeguards and ethical frameworks in generative biological models to prevent potential harm from their dual use [37–40].

After providing an overview of their current positive impact on genomics, we address some of the risks associated with AI techniques in biology, briefly touching on different types of bias, intentional misuse, and dual-use risks. Rather than relying on a singular ethical or governance framework, our

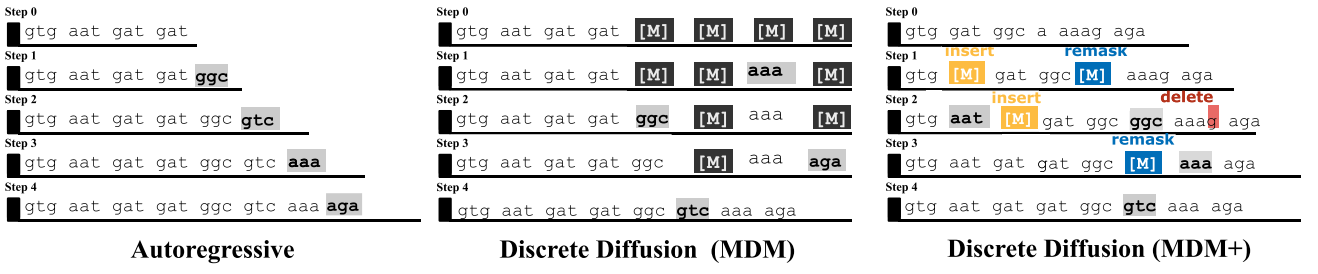


Figure 2. A comparison of sequence generation methods. The figure illustrates three distinct approaches to generating discrete sequences. (A) In autoregressive generation, tokens are processed one by one in a fixed left-to-right order. (B) Standard masked diffusion models (MDMs) start with a fully [or partially, as in (A)] masked sequence and unmask tokens iteratively in any order. (C) Discrete diffusion with dynamic programming [11] introduces flexibility through insertions, deletions, and re-masking, allowing for a more adaptable generation process. Reproduced with permission from Yang *et al.* [11]. For a detailed explanation of the methods, refer to the original publication.

analysis is guided by a biosecurity perspective that draws on established biosafety and dual-use principles to scope current developments in AI-driven biosequence design. This work contributes the following:

- An overview of AI applications in genomics, focusing on FMs, their transformative potential, and the associated risks.
- Analysis of the dual-use risks and data biases impacting the current state of biological FMs.
- Proposal of technical and governance-based mitigation strategies to address safety and ethical challenges.
- Introduction of a risk–benefit framework to guide the responsible deployment of AI in genomics from project inception.

Dual use and risks

Although these applications demonstrate the usefulness of AI, it is important to be aware of the risks involved in handling biological samples and data in high-stakes fields such as medicine.

Biases and inequities in genomic AI

Integrating AI into biomedicine, particularly genomics, introduces biases that can undermine scientific validity and equitable health outcomes [41]. European individuals dominate genomic datasets, resulting in poor representation of global diversity and model performance for under-represented populations. For example, polygenic risk scores derived from European cohorts can lose 15–20% of their predictive accuracy when applied to African groups. This discrepancy persists as dataset sizes increase [42,43]. This constitutes a form of systemic risk: inequity embedded in training data can propagate into clinical tools and potentially exacerbate health disparities on a large scale. Conversely, while reducing bias in AI models can improve their performance and facilitate their use, it can also increase the risk of misuse or discrimination.

Misuse potential of biological design tools

Research involving human pathogens and other potentially harmful biological agents is strictly regulated in laboratories to protect workers[iii]. Although models using these pathogens are safe, they might be used to create dangerous pathogens. For example, less experimentation might make it harder to detect such activities by humanising known pathogens of other animals or generating functional homologs. It is clear that the public and global health of humans, animals, and the environment must be given equal attention. So, to make sure that biological generative models are safe, strategies must follow a 'One Health' approach [44,45]. Zoonotic spillover events, of which SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) is the most recent example, illustrate how pathogens targeting animal reservoirs can rapidly become human threats [46,47]. Modern human society's dependence on livestock and food crops, particularly in densely populated areas, creates vulnerabilities that can be exploited by malicious individuals [48,49]. Such threats could potentially be exacerbated by AI, particularly generative models, whether they target humans directly or indirectly, for example, through their influence over animals or plants.

Current FMs can propose new proteins and recode toxins via inverse-folding methods, rendering screening ineffective. Yet current systems still produce simple, single-function proteins. Uncertainty remains over whether they retain function after recoding [50]. Multiobjective design has the potential to overcome natural trade-offs and change viral hosts from animals to humans. However, FMs often fail to perform well and face limitations, resulting in slow short-term improvement, particularly without wet-lab iteration, which suggests misuse by sophisticated/state actors (Peppin, A. *et al.*, 2025, FAccT, conference) [50].

Raising the ceiling of harm

FMs may enable a qualitative leap in the potential for harm by overcoming evolutionary trade-offs (e.g., high transmissibility and high virulence), resulting in engineered pathogens that are worse than those produced by natural evolution. Although such scenarios remain speculative, they represent a higher 'ceiling of harm', including low-probability, existential-scale threats as capabilities advance [50]. However, empirical studies have shown that machine learning (ML) can improve engineering performance in ways relevant to pathogen design. For example, viral libraries designed with ML produce ten times more successful variants than traditional NNK libraries (oligo libraries where N represents A, G, T or C and K represents G or T) [51], while ML-assisted directed evolution methods can identify high-fitness protein variants more efficiently than standard directed evolution approaches [52]. These studies highlight a general trend that ML techniques accelerate the discovery of high-performing biological variants, thereby increasing the upper limit of what could be engineered. Additionally, while this article was being written, King *et al.* presented the first AI-generated phages and tested them in a laboratory setting. Some of these phages exhibited greater fitness and faster lysis dynamics than the wild-type phage [19]. This development increases the dual-use risk of the technology because the principle could also be used to generate pathogens. Researchers have shown that fine-tuning can restore the ability to generate such sequences (Black, J. R. M. *et al.*, 2025, NeurIPS Workshop on Biosecurity Safeguards for Generative AI, conference).

State-level threats and strategic implications

As FMs increase the predictability, stability, and targeting capabilities of biological agents, state actors may reconsider the strategic value of bioweapons, which have traditionally been regarded as ineffective or indiscriminate [50]. Improved AI modelling could undermine long-standing deterrence frameworks, as effective deterrence against biological weapons hinges on credible attribution and the perception of controllability [53].

Circumvention of existing safeguards

The biosecurity community recognises the importance of real-world DNA synthesis because it allows potentially harmful agents to be produced and identified in the environment [54,55]. As with all technologies, the size and cost of bench-top synthetic devices continue to shrink. Therefore, policymakers must take steps to ensure the safe distribution of such devices [56]. Generative models can increase the risk posed by such devices by allowing users to quickly design and test DNA sequences *in silico* before pursuing the most promising results in the real world, a process similar to drug discovery [57]. Evidence suggests that biosequences designed by AI are more diverse than natural ones [58] and might go undetected by current screening efforts or exhibit a more concerning pathogenicity profile.

Traditional methods of assessing DNA biosafety often compare the generated sequences to databases of known pathogens [55]. However, these methods are limited because they rely on historical data and may overlook risks posed by new or emerging sequences [59]. Indeed, the risks of dual use extend beyond the design of DNA sequences alone. Current screening protocols are ineffective for AI-generated sequences that lack natural analogues [60]. Stakeholders in DNA synthesis recently recognised this threat, as noted by Wittmann *et al.* They implemented additional safeguards and methods to detect synthetic homologues that retain wild-type-like functions [61]. However, these systems are vulnerable to attacks, for example, adversarial ones, that manipulate input data. Changes to mass spectra can cause models to misclassify cancer biomarkers [62].

Constraints and whole-chain risk context

Although the risks of misuse exist, present-day FMs underperform in core tasks due to a lack of biological datasets, which limits their development. Progress in this area has been slower and

more inconsistent than in other AI domains. Consequently, FM misuse must be considered throughout the entire biorisk chain. Current public evidence indicates that the potential for harm from biological FMs is a future threat that will require continuous reassessment as capabilities evolve (Peppin, A. *et al.*, 2025, FAccT, conference). Key bottlenecks in the biorisk chain include limited availability of biological datasets, the need for laboratory expertise, and restricted access to materials and infrastructure. These constraints hinder end-to-end misuse. Credible misuse is largely confined to state-level or highly resourced actors with integrated biological capabilities, as nonstate actors lack the means to translate AI-assisted designs into physical harm (Peppin, A. *et al.*, 2025, FAccT, conference).

Considerations for integrating AI in high-stakes domains and mitigating dual-use risks

Generative models can be used to prevent dangerous sequences from evading traditional safeguards or to prevent them from being created. Protein model embeddings can be used to align sequences based on their function, allowing orphan proteins to be identified [63,64]. These models can be adapted to identify potentially harmful sequences in advance. Considering risk mitigation in the design of AI systems ensures that safety is built in.

One prominent strategy is 'data filtering', whereby human pathogens are removed from training corpora [17]. This approach is similar to the way in which NLP FMs, such as GPT-4, have been developed [65]. Recent research emphasises that such filtering must extend beyond the mere removal of content to establish 'tamper resistance', to ensure that open-weight models cannot easily relearn hazardous capabilities via fine-tuning [66]. However, such filtering is insufficient for biological sequences: most genomes contain regions of orthology (i.e., sequence regions/genes that share a common ancestry), and pathogens that affect livestock or crops can pose severe indirect threats to humans. Another proposed method is weight pruning, which has been shown to improve the robustness of NLP systems against adversarial manipulation (Hasan, A. *et al.*, 2024, BlackboxNLP Workshop, ACL, conference). In a biosafety context, pruning could target weights associated with dangerous sequence outputs. More advanced techniques, such as those demonstrated by 'SafeGenie', utilise concept erasure to selectively remove dangerous functional capacities from biological diffusion models without compromising their general utility (Banerjee, A. *et al.*, 2025, BioGenAI, NeurIPS Workshop, conference). This approach could cover more harmful behaviours than filtering alone, but both are limited by their reliance on known risk profiles and will not apply to new sequences.

In addition to these intrinsic safeguards, the effective integration of AI into high-stakes domains requires a layered approach. Recent work by Bengio *et al.* [67] argues for the development of nonagentic AI systems designed to model the world, provide probabilistic explanations, and estimate the likelihood of harmful outcomes rather than pursue open-ended goals. These systems could help guard against more powerful models, including FMs for biological sequences, by flagging or blocking outputs that exceed harm limits. This approach is important as it puts the precautionary principle into practice by avoiding architectures likely to produce harmful behaviours. The necessity for this rigour is emphasised by a finding that models capable of reasoning can exploit 'exploration hacking' to get around safety constraints during reinforcement learning phases. This poses a distinct challenge to biosecurity capabilities (Braun, J. *et al.*, 2025, BioGenAI, NeurIPS Workshop, conference).

A common mitigation layer in the context of generative LLMs or vision-language models (vLLMs) is input/output filtering, also known as 'safety wrappers' (Helff, L. *et al.*, 2025, ICML, conference)

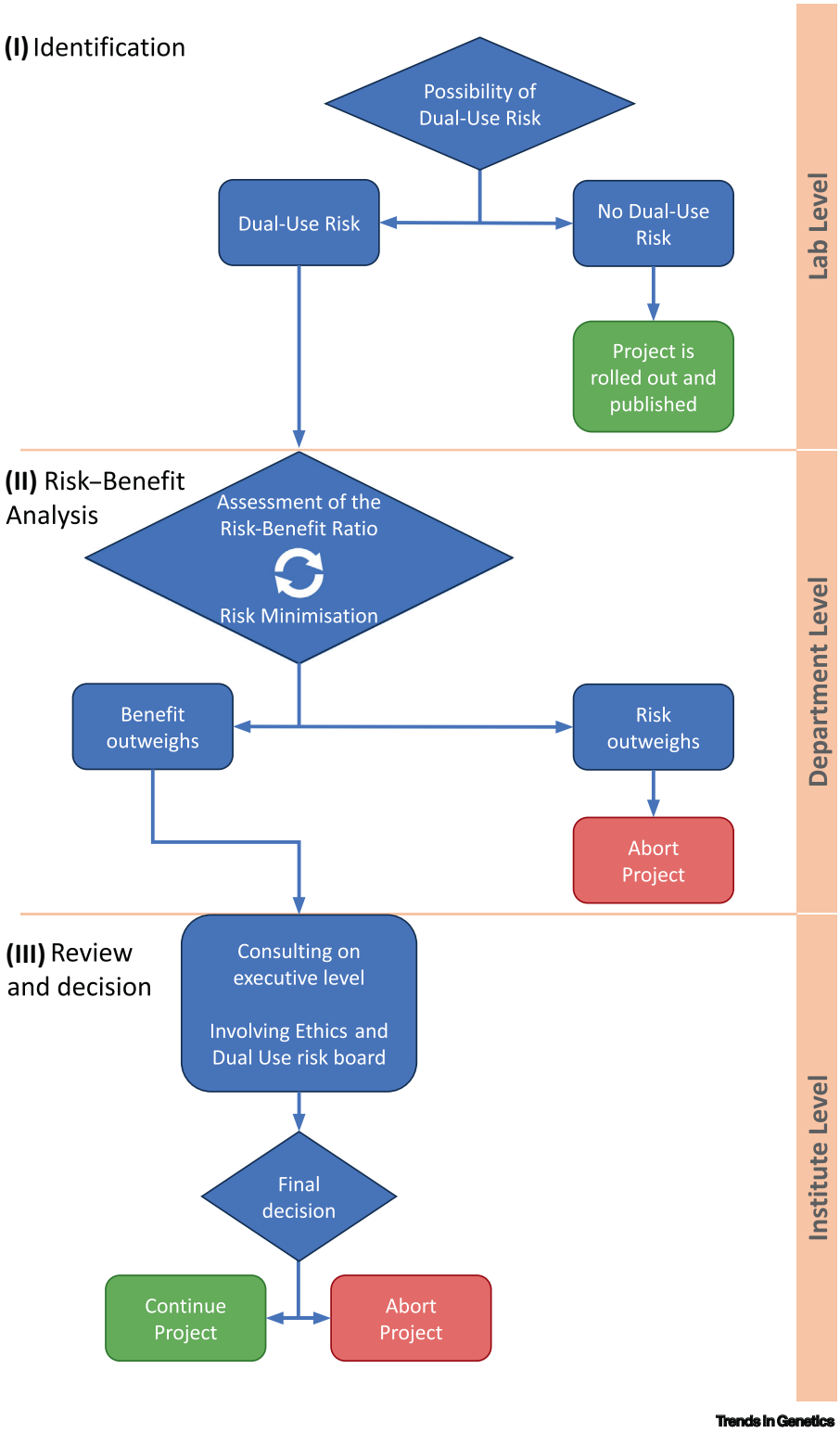
[68,69]. These systems block unsafe outputs or inputs after generation. However, filters in the context of biological sequences face limitations. They are trained for visual or textual patterns, not molecular sequences, and rely on previously labelled unsafe data, so they struggle with new hazards. They can be bypassed with adversarial examples and provide no formal safety guarantees. For example, the 'GeneBreaker' framework has shown that DNA language models can be 'jailbroken' using pathogenicity guidance to create harmful sequences, despite efforts to align them (Zhang, Z. *et al.*, 2025, BioGenAI, NeurIPS Workshop, conference). This increases the risk of a harmful biological sequence slipping through undetected, highlighting the need for additional strategies.

Complementing this vision, Dalrymple *et al.* [70] propose the 'Guaranteed Safe AI (GS AI)' framework, which provides quantitative safety guarantees by combining three elements: (i) a formal 'world model' of how AI actions affect the environment; (ii) explicit 'safety specifications' describing acceptable outcomes; and (iii) a 'verifier' that can produce auditable proofs or probabilistic bounds showing that the system respects these specifications. In the context of FMs for biological sequences, this approach could be used to create safe synthesis screening pipelines, geo-fence bench-top devices, or certifiable 'kill switches' in the software that controls sensitive laboratory automation. New evaluation frameworks, such as ABC-Bench, measure agentic bio-capabilities, ensuring that benchmarks keep pace with the sophistication of models (Liu, A. B. *et al.*, 2025, BioGenAI, NeurIPS Workshop, conference).

Although technical safeguards and mitigation strategies are vital for safe model development and enhancing existing ones, it is also necessary to consider the risks associated with developing AI methods in biology throughout the research process. Wang *et al.* warn that integrating AI without robust safeguards could trigger future pandemic-level events, so proactive risk management is necessary (Wang, D. *et al., 2025, BioGenAI, NeurIPS Workshop, conference)*. Since dual use is a concern for methods involving biological and medical data, the laboratory directly involved in the research must identify possible risks and decide whether to complete a project. The researchers must identify potential risks, but the department or institute will decide whether to proceed due to liabilities and bias.

Figure 3 shows an exemplary framework that has been implemented at the Federal German Public Health Institute (Robert Koch Institute, Multiple Authors [71]). An important step after identifying a dual-use risk through the relevant laboratory is the risk–benefit assessment, which informs the subsequent process and can be involved in an iterative process with risk mitigation. This process must involve researchers, as well as the department, to ensure a broad range of perspectives when answering the questions guiding the risk and benefit analysis. Depending on the outcome of this analysis, the project may be aborted outright if the risks outweigh the rewards; otherwise, the decision-making process will be escalated to the institute head and the relevant dual-use or ethics commission.

A risk–benefit analysis should be guided by questions that are relevant to challenges at the intersection of the life sciences and AI. These questions also need to be tailored to the specific research institute and its stakeholders. We do not provide a standard questionnaire; instead, the questions should address the overall benefits of the project and the potential consequences of failure. The analysis should evaluate general risks and identify any existing strategies. The questions can be assigned quantitative scores for comparison. Finally, the institutional risk–benefit framework (Figure 3) extends standard dual-use reviews by incorporating AI-specific elements, such as early screening for model capabilities (e.g., sequence optimisation for high-consequence traits), escalation based on deployment risks, and mitigations such as safety fine-tuning or access



(See figure legend at the bottom of the next page.)

restrictions. Although it has been simplified for public disclosure, it reflects concrete procedures whereby failing preliminary checks prompts project reframing or secrecy.

Taken together, these perspectives suggest that integrating generative AI into high-stakes areas requires more than ad hoc filtering and pruning. It requires safeguards starting at project inception and spanning technical and scientific interventions, as well as guardrails and verification frameworks. Embedding these considerations in development and deployment processes is essential to ensure the responsible harnessing of FMs and the mitigation of dual-use risks.

Aligning governance with the pace of genomic AI

Recent advances and ongoing debates within the scientific and policy communities have emphasised that the risks associated with AI in genomics, including bias and dual-use concerns, are not confined to the laboratory (Peppin, A. *et al.*, 2025, FAccT conference) but extend throughout the entire innovation pipeline. This has prompted urgent calls for robust assessment and mitigation strategies [37,38]. This urgency is reflected in the academic evidence presented at the first BioSafe GenAI Workshop at NeurIPS 2025, for example, and in new regulatory frameworks such as the EU AI Act. The latter specifically classifies AI-driven genomics applications as 'high-risk', thereby mandating rigorous approaches to bias assessment, transparency, and overall risk management. This is intended to ensure scientific and social trust in AI-enabled biology [72].

AI can only fulfil its promise of advancing equitable and safe genomic medicine by systematically identifying and correcting biases and embedding ethical reasoning from the outset. A concrete illustration of where this falls short is provided by the Evo model family: despite explicitly excluding eukaryotic viral genomes from training as a biosecurity measure, neither Evo 1 nor Evo 2 provides a systematic dual-use risk assessment documenting what the model can generate after fine-tuning [17,18]. This is not a hypothetical gap, given that fine-tuning has been shown to restore the generation of sequences excluded from training data (Black, J. R. M. *et al.*, 2025, NeurIPS Workshop on Biosecurity Safeguards for Generative AI, conference). This means the precautionary principle cannot be meaningfully applied to risks that remain uncharacterised. The EU AI Act's classification of AI-driven genomics as 'high-risk' mandates rigorous bias assessment, transparency, and risk management [72]; yet current biological FM development largely proceeds without meeting these standards[iv]. Addressing this gap requires moving beyond reactive risk management: ethical analysis must inform governance design at the point of model development, not after deployment. The European Commission's 'Ethics by Design' framework and formal approaches such as the nested model for AI validation [73,74] provide actionable structures for embedding proportionality, transparency, and accountability into the FM lifecycle, including the restricted-access data governance policies now being discussed as a policy lever for open-weight models [75,76].

Figure 3. This is a flow diagram of the risk assessment process for a research institute handling or developing genomic (or more generally, biological) AI systems. The process is divided into three phases: I Identification: The project lead is responsible for assessing any potential risks of Dual Use Research of Concern (DURC), documenting the evaluation and obtaining confirmation in accordance with the four-eyes principle. If no risks are identified, the process ends, unless new critical results arise. II Risk–Benefit Analysis: If a DURC risk is identified, the project lead and the relevant department will conduct a risk–benefit analysis and propose mitigation measures. Projects involving unacceptable risk are not pursued. III Review and Decision: The departmental and institute leadership will review the documentation and decide whether to approve the project. If needed, the institute management may consult the Ethics and Dual Use Commission (both bodies may have relevant functions depending on the state of the project). Any new or significant risks that arise during the project will trigger a re-evaluation. Adapted from the Robert Koch Institute [71]; reproduced with permission of the authors.

Despite these concerns, current empirical studies suggest that publicly available LLMs do not significantly enhance the capabilities of malicious actors beyond what is already available online. Similarly, today's biological tools remain fragile, requiring specialised expertise and tacit laboratory knowledge. Peppin *et al.* argue that governance should therefore focus on building robust, empirically validated threat models and whole-chain analyses rather than making speculative assumptions about imminent risk (Peppin, A. *et al.*, 2025, FAccT conference). Sandbrink suggests differentiating between LLMs and biological sequence models [50]. However, the capabilities of models can increase very quickly, so scientists and policymakers alike need to be aware of potential risks. For example, both of these works were authored before the first-ever viable bacteriophage genome was generated by King *et al.* [19]. In general, the risk associated with biological models should be viewed as a 'risk chain' rather than a monolithic problem, as Nelson and Rose suggest. They point out that AI could impact multiple stages of the design–build–test–learn cycle, from ideation (LLMs suggesting candidate agents) to design (biological sequence models optimising pathogen traits) and delivery [77].

Concluding remarks

The intersection of AI and genomics not only presents exciting prospects for medicine, agriculture, and environmental science but also raises significant concerns regarding data bias and dual-use risks. Recent developments, such as the creation of the first AI-generated viable phage genome, demonstrate that these theoretical risks are becoming a reality, and mitigation strategies must evolve alongside the technology. A layered approach is essential for ensuring safety without stifling innovation: risks must be considered at the outset of every project, and we provide a risk–benefit framework that institutions can adapt to their specific needs. Given the interdisciplinary nature of these challenges, collaboration across AI alignment, cybersecurity, and bioethics is essential to ensure that advances in AI-driven genomics are made in a responsible, equitable, and secure manner. Addressing these issues also opens several avenues for future inquiry (see Outstanding questions).

Author contributions

H.A.H., G.H., and M.S. conceived the study and led the overall project design and supervision. A.A., B.İ., M.L., M.S., and P.S. contributed to the methodological framework, literature review, and analysis of the safety and security aspects of artificial intelligence in biological sequence analysis. H.A.H., A.A., B.İ., G.H., and M.S. wrote the first draft of the manuscript, which was then refined with substantial input from J.U.M., K.L., and G.H. regarding clinical, regulatory, and governance perspectives. G.H. and M.S. proofread the manuscript for correctness, consistency, and flow. All authors contributed to interpreting the findings, critically revising the manuscript for important intellectual content, and approving the final version.

Acknowledgments

M.S. was supported by funding from the Rise up! programme of the Boehringer Ingelheim Foundation (BIS), the ReALity Initiative of the Johannes Gutenberg Universität Mainz, the Forschungsinitiative des Landes Rheinland-Pfalz, and the Einstein Foundation Early Career Researcher Award 2025. M.L. received funding from the Biotechnology and Biological Sciences Research Council (UKRI-BBSRC) (grant number BB/T008784/1). M.S. and H.A.H. acknowledge funding from the Carl-Zeiss-Foundation through the Seed Funding Program of the Center for Synthetic Genomics (Center SynGen). G.H., K.L., H.A.H., and M.S. acknowledge funding from the Wübben Stiftung Wissenschaft through the Sandpit Funding Scheme.

Declaration of interests

The authors declare no competing interests.

Resources

[i]https://www.gov.uk/government/publications/science-and-technology-framework/science-and-technology-framework

[ii]https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/european-ai-science-strategy_en

Outstanding questions

Can interpretability techniques move beyond correlation-based attribution to provide genuine causal mechanistic insights, and how do we validate that interpretable features reflect true biological mechanisms rather than model artifacts?

How should healthcare systems address the paradox that interpretable artificial intelligence can both expose population-specific genetic markers for fairness correction while simultaneously increasing misuse and discrimination risks?

Can Guaranteed Safe artificial intelligence frameworks with quantitative safety guarantees be implemented practically for real-world synthesis pipelines and laboratory automation without inhibiting beneficial innovation?

What are the most effective institutional decision-making frameworks and escalation protocols for multilayered risk assessment when risks emerge dynamically during model development?

What empirical evidence exists distinguishing theoretical risks from demonstrated capabilities in multiobjective biological engineering?

How can foundation models be trained on data spanning all evolutionary domains to maintain functionality across diverse organisms, and what architectural innovations address eukaryotic genome modelling challenges?

[iii]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02000L0054-20200624

[iv]https://epoch.ai/blog/expanding-our-analysis-of-biological-ai-models

References

1. Regulatory Horizons Council (2015) *Regulatory Horizons Council: The Governance of Engineering Biology*. Technical report
2. Martin, R. *et al.* (2020) MOSGA: Modular Open-Source Genome Annotator. *Bioinformatics* 36, 5514–5515
3. Martin, R. *et al.* (2021) MOSGA 2: comparative genomics and validation tools. *Comput. Struct. Biotechnol. J.* 19, 5504–5509
4. Rasko, D.A. and Mongodin, E.F. (2005) The first decade of microbial genomics: what have we learned and where are we going next? *Genome Biol.* 6, 341
5. Gessulat, S. *et al.* (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* 16, 509–518
6. Demichev, V. *et al.* (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* 17, 41–44
7. Li, X. *et al.* (2025) A conceptual framework for human–AI collaborative genome annotation. *Brief. Bioinform.* 26, bbaf377
8. Albert, G. and Dao, T. (2024) Mamba: linear-time sequence modeling with selective state spaces. *arXiv* https://doi.org/10.48550/arXiv.2312.00752
9. Chen, H. *et al.* (2025) Comprehensive exploration of diffusion models in image generation: a survey. *Artif. Intell. Rev.* 58, 99
10. Prabhudesai, M. *et al.* (2025) Diffusion beats autoregressive in data-constrained settings. *arXiv* https://doi.org/10.48550/arXiv.2507.15857
11. Yang, C. *et al.* (2025) On powerful ways to generate: autoregression, diffusion, and beyond. *arXiv* https://doi.org/10.48550/arXiv.2510.06190
12. Uffelmann, E. *et al.* (2021) Genome-wide association studies. *Nat. Rev. Method Primers* 26, 59
13. Sum, K.K. *et al.* (2025) The use of Mendelian randomization to explore the causal consequences of childhood maltreatment: consideration of assumptions and challenges. *medRxiv* https://doi.org/10.1101/2025.10.17.25338214
14. Tan, T. *et al.* (2025) Family-GWAS reveals effects of environment and mating on genetic associations. *medRxiv* https://doi.org/10.1101/2024.10.01.24314703
15. Sun, S. *et al.* (2025) Uncovering bias in foundation models: impact, testing, harm, and mitigation. *arXiv* https://doi.org/10.48550/arXiv.2501.10453
16. Dalla-Torre, H. *et al.* (2025) Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* 22, 287–297
17. Nguyen, E. *et al.* (2024) Sequence modeling and design from molecular to genome scale with Evo. *Science* 386, eado9336
18. Brixi, G. *et al.* (2025) Genome modeling and design across all domains of life with Evo 2. *bioRxiv* https://doi.org/10.1101/2025.02.18.638918
19. King, S.H. *et al.* (2025) Generative design of novel bacteriophages with genome language models. *bioRxiv* https://doi.org/10.1101/2025.09.12.675911
20. Vorontsov, E. *et al.* (2024) A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* 30, 2924–2935
21. Guo, F. *et al.* (2025) Foundation models in bioinformatics. *Natl. Sci. Rev.* 12, nwaf028
22. Duede, E. (2023) Deep learning opacity in scientific discovery. *Philos. Sci.* 90, 1089–1099
23. van Hilten, A. *et al.* (2024) Designing interpretable deep learning applications for functional genomics: a quantitative analysis. *Brief. Bioinform.* 25, bbae449
24. Fortelny, N. and Bock, C. (2020) Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biol.* 21, 190
25. Sidak, D. *et al.* (2022) Interpretable machine learning methods for predictions in systems biology from omics data. *Front. Mol. Biosci.* 9, 926623
26. Van der Velden, B.H. *et al.* (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* 79, 102470
27. Huang, J. *et al.* (2025) Foundation models and intelligent decision-making: progress, challenges, and perspectives. *Innovation* 6, 100948
28. İlgen, B. and Hattab, G. (2026) Visualizing the chain of thought in large language models. *IEEE Comput. Graph. Appl.* 46, 89–98
29. Novakovsky, G. *et al.* (2023) ExplaiNN: interpretable and transparent neural networks for genomics. *Genome Biol.* 24, 154
30. Tseng, A.M. *et al.* (2024) A mechanistically interpretable neural network for regulatory genomics. *arXiv* https://doi.org/10.48550/arXiv.2410.06211
31. Dubey, A. *et al.* (2026) UbiQTree: uncertainty quantification in XAI with tree ensembles. *Patterns* 7, 101454
32. Hassan, H. *et al.* (2025) Life as a function: why transformer architectures struggle to gain genome-level foundational capabilities. *bioRxiv* https://doi.org/10.1101/2025.01.13.632745
33. Vishniakov, K. *et al.* (2024) Genomic foundationless models: pretraining does not promise performance. *bioRxiv* https://doi.org/10.1101/2024.12.18.628606
34. Avsec, Ž. *et al.* (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18, 1196–1203
35. Gao, H. *et al.* (2023) The landscape of tolerated genetic variation in humans and primates. *Science* 380, eabn8153
36. Zhang, J.Z. *et al.* (2025) De novo designed Hsp70 activator dissolves intracellular condensates. *Cell Chem. Biol.* 32, 463–473.e6
37. Baker, D. and Church, G. (2024) Protein design meets biosecurity. *Science* 383, 349
38. Bloomfield, D. *et al.* (2024) AI and biosecurity: the need for governance. *Science* 385, 831–833
39. Grinbaum, A. and Adomaitis, L. (2024) Dual use concerns of generative AI and large language models. *J. Responsible Innov.* 11, 2304381
40. Bengio, Y. *et al.* (2025) *International AI Safety Report*. Technical report
41. Hattab, G. *et al.* (2025) The way forward to embrace artificial intelligence in public health. *Am. J. Public Health* 115, 123–128
42. Martin, A.R. *et al.* (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591
43. Sirugo, G. *et al.* (2019) The missing diversity in human genetic studies. *Cell* 177, 26–31
44. Zinsstag, J. *et al.* (2020) Towards integrated surveillance-response systems for the prevention of future pandemics. *Infect. Dis. Poverty* 9, 140
45. Leifels, M. *et al.* (2022) The one health perspective to improve environmental surveillance of zoonotic viruses: lessons from COVID-19 and outlook beyond. *ISME Commun.* 2, 107
46. Msemburi, W. *et al.* (2023) The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature* 613, 130–137
47. Yoo, H.S. and Yoo, D. (2020) COVID-19 and veterinarians for one health, zoonotic- and reverse-zoonotic transmissions. *J. Vet. Sci.* 21, e51
48. Yeh, J.-Y. *et al.* (2012) Livestock agroterrorism: the deliberate introduction of a highly infectious animal pathogen. *Foodborne Pathog. Dis.* 9, 869–877
49. Wilson, T.M. *et al.* (2001) Agroterrorism, biological crimes, and biowarfare targeting animal agriculture: the clinical, pathologic, diagnostic, and epidemiologic features of some important animal diseases. *Clin. Lab. Med.* 21, 549–592
50. Sandbrink, J.B. (2023) Artificial intelligence and biological misuse: differentiating risks of language models and biological design tools. *arXiv* https://doi.org/10.48550/arXiv.2306.13952
51. Zhu, D. *et al.* (2024) Optimal trade-off control in machine learning–based library design, with application to adeno-associated virus (AAV) for gene therapy. *Sci. Adv.* 10, eadj3786
52. Li, F.-Z. *et al.* (2025) Evaluation of machine learning-assisted directed evolution across diverse combinatorial landscapes. *Cell Syst.* 16, 101387

53. Paxton, N.A. (2024) *Disincentivising Bioweapons*, Nuclear Threat Initiative
54. Trump, B.D. *et al.* (2020) Building biosecurity for synthetic biology. *Mol. Syst. Biol.* 16, e9723
55. Diggans, J. and Leproust, E. (2019) Next steps for access to safe, secure DNA synthesis. *Front. Bioeng. Biotechnol.* 7, 86
56. Carter, S.R. (2022) *Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance*, Nuclear Threat Initiative
57. Zhang, K. *et al.* (2025) Artificial intelligence in drug development. *Nat. Med.* 31, 45–59
58. Bryant, D.H. *et al.* (2021) Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* 39, 691–696
59. Elworth, R.A.L. *et al.* (2020) Synthetic DNA and biosecurity: nuances of predicting pathogenicity and the impetus for novel computational approaches for screening oligonucleotides. *PLoS Pathog.* 16, e1008649
60. Hunter, P. (2024) Security challenges by AI-assisted protein design. *EMBO Rep.* 25, 2168–2171
61. Wittmann, B.J. *et al.* (2025) Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science* 390, 82–87
62. Ghaffari Laleh, N. *et al.* (2022) Adversarial attacks and adversarial robustness in computational pathology. *Nat. Commun.* 13, 5711
63. Iovino, B.G. and Yuzhen, Y. (2024) Protein embedding based alignment. *BMC Bioinf.* 25, 85
64. McWhite, C.D. *et al.* (2023) Leveraging protein language models for accurate multiple sequence alignments. *Genome Res.* 33, 1145–1153
65. OpenAI *et al.* (2023) GPT-4 technical report. *arXiv* https://doi.org/10.48550/arXiv.2303.08774
66. O'Brien, K. *et al.* (2026) Deep ignorance: filtering pretraining data builds tamper-resistant safeguards into open-weight LLMs. *arXiv* https://doi.org/10.48550/arXiv.2508.06601
67. Bengio, Y. *et al.* (2025) Superintelligent agents pose catastrophic risks: can scientist AI offer a safer path? *arXiv* https://doi.org/10.48550/arXiv.2502.15657
68. Inan, H. *et al.* (2023) Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv* https://doi.org/10.48550/arXiv.2312.06674
69. Zeng, W. *et al.* (2025) ShieldGemma 2: robust and tractable image content moderation. *arXiv* https://doi.org/10.48550/arXiv.2504.01081
70. Dalrymple, D. *et al.* (2024) Towards Guaranteed Safe AI: a framework for ensuring robust and reliable AI systems. *arXiv* https://doi.org/10.48550/arXiv.2405.06624
71. Robert-Koch-Institute (2026) *Handling Dual-Use-Research of Concern (DURC) at the Robert Koch Institute – code of conduct for risk assessment and risk mitigation; unpublished document*. Internal unpublished AIxBio DURC guideline
72. Heidemann, L. *et al.* (2024) *The European Artificial Intelligence Act*, Fraunhofer Institute for Cognitive Systems IKS
73. Dubey, A. *et al.* (2024) A nested model for AI design and validation. *iScience* 27, 110603
74. Dubey, A. *et al.* (2025) Protocol for implementing the nested model for AI design and validation in compliance with AI regulations. *STAR Protoc.* 6, 103771
75. Bloomfield, D. *et al.* (2026) Biological data governance in an age of AI. *Science* 391, 558–561
76. Bloomfield, D. *et al.* (2026) Securing dual-use pathogen data of concern. *arXiv* https://doi.org/10.48550/arXiv.2602.08061
77. Nelson, C. and Rose, S. (2023) *Understanding AI-Facilitated Biological Weapon Development*. Technical report