nature communications

Perspective

https://doi.org/10.1038/s41467-026-78758-5

A defence-in-depth strategy for securing AI-driven biological sequence design

Received: 5 May 2026

Accepted: 2 October 2026

Published online: 19 November 2026

Check for updates

Maximilian Sprang[1,2] ✉, Hassan Ahmed Hassan[1,3], Yasha Ektefaie[4,5], Anne Schmieder[6,7,8], Alix Wasilenko[9], Moritz S. Hanke[10], Rosa E. Martín-Peña[11,12], Rachael Harkness[13], Jasper Götting[14], Alan Murphy[15], Johannes U. Mayer[1,2], Marco Schäfer[16], André Friedrich[17], Kerstin Lenhof[18,19], Sophia Rudorf[20], Gigi Gronvall[21], Jan Kellmann[17,22], Simon Anders[23], Berkan Aysan[29], Nicole Saverschek[24], Kathrin Schuldt[25], Marc Güell[26], Katharina Ladewig[16,27,30,30] & Georges Hattab[16,28,30]

Generative artificial intelligence can design biological sequences, from proteins to whole genomes, opening opportunities for medicine and biotechnology, but the same capability can be misused. The Wübben Sandpit conference brought together experts in biology, AI, biosecurity, ethics, medicine and policy to ask how the field can capture the benefits while managing the risks. Here we set out a defence-in-depth strategy that builds safeguards into every stage, from the data used to train models, to the models themselves, to the synthesis of physical DNA. We argue that responsible governance is not a brake on progress but a precondition for it.

The life sciences are undergoing a fundamental paradigm shift, transitioning from a descriptive discipline to one driven by predictive engineering. This transformation is driven by the convergence of high-throughput measurement and engineering platforms, such as nanopore long-read sequencing, gene editing and the industrialisation of wet-lab automation. A key outcome of this convergence is the sharp, exponential decline in genomic sequencing costs, which has triggered a data explosion that far exceeds the analytical capacity of traditional heuristic and statistical methods.

[1]Department of Dermatology, University Medical Centre of the Johannes Gutenberg University Mainz, Mainz, Germany. [2]Institute of Quantitative and Computational Biosciences (IQCB), Johannes Gutenberg University Mainz, Mainz, Germany. [3]DeOxy Tech Ltd, Manchester, UK. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [5]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [6]ScaDS.AI Dresden/Leipzig—Center for Scalable Data Analytics and Artificial Intelligence, Dresden/Leipzig, Germany. [7]Institute for Drug Discovery, Faculty of Medicine, Leipzig University, Leipzig, Germany. [8]Institute of Protestant Theology, Technical University Dresden, Dresden, Germany. [9]Institute of Biomedical Ethics and History of Medicine (IBME), University of Zurich, Zurich, Switzerland. [10]Center for Health Security, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [11]CAIMed—Lower Saxony Center for Artificial Intelligence and Causal Methods in Medicine, Hannover, Germany. [12]CELLS—Centre for Ethics and Law in the Life Sciences, Leibniz University Hannover, Hannover, Germany. [13]Baillie-Gifford Pandemic Science Hub, University of Edinburgh, Edinburgh, UK. [14]SecureBio, Cambridge, MA, USA. [15]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [16]Centre for Artificial Intelligence in Public Health Research (ZKI-PH), Robert Koch Institute, Berlin, Germany. [17]Zentrale Kommission für Biologische Sicherheit, Bundesministerium für Verbraucherschutz und Lebensmittel, Berlin, Germany. [18]University Medical Center Göttingen, Göttingen, Germany. [19]CAIMed—Lower Saxony Center for Artificial Intelligence and Causal Methods in Medicine, Göttingen, Germany. [20]Institute for Cell Biology and Biophysics, Leibniz University Hannover, Hannover, Germany. [21]Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [22]Center for Synthetic Microbiology (SYNMIKRO), Philipps-Universität Marburg, Marburg, Germany. [23]BioQuant Center, University of Heidelberg, Heidelberg, Germany. [24]Stiftung der Deutschen Wirtschaft (sdw) gGmbH, Berlin, Germany. [25]Infectious Disease Surveillance and Control Lab, Bernhard Nocht Institute for Tropical Medicine (BNITM), Hamburg, Germany. [26]Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain. [27]Medical University Lausitz—Carl Thiem (MUL-CT), Cottbus, Germany. [28]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany. [29]Independent Researcher: Berkan Aysan [30]These authors contributed equally: Katharina Ladewig, Georges Hattab. ✉e-mail: masprang@uni-mainz.de

https://doi.org/10.1038/s41467-026-78758-5

Table 1 | Recommendations from the Wübben Sandpit for mitigating the biological risk of AI in biological sequence design

Initiate an expert consensus process to classify dual-use data.
Create initiatives to raise scientists' awareness of dual-use research and datasets.
Establish a tiered pathogen data access system with proportionate access controls.
Introduce funding-stage reviews and requirements for projects that generate dual-use pathogen data.
Establish a structured risk-benefit assessment process throughout the AI model lifecycle.
Develop safety frameworks that combine *in silico* evaluation with controlled experimental validation.
Create tiered access platforms to control access to models and allow independent risk assessments.
Strengthen screening, detection and registries across nucleic acid synthesis providers.
Implement Know Your Customer (KYC) rules and serialisation and supply-chain tracking for critical components, such as benchtop synthesis devices, their reagents and cartridges, and synthetic nucleic acid orders.
Establish or strengthen clear governance for nucleic acid synthesis screening.
Create a sector-level insurance fund to support research into risk mitigation measures and risks.

The green background refers to recommendations for mitigating risks in the data space, the blue background to recommendations for mitigating risks in the model space, and the orange background to recommendations for mitigating risks in the synthesis and translation space.

This deluge of information is at the heart of the tensions and opportunities of modern biotechnology, and is the primary impetus for the Generative Artificial Intelligence (GenAI) for Biological Sequences' Wübben Sandpit conference. While AI offers unprecedented capabilities for biological discovery, the same pattern-recognition architectures that decode life's complexity also pose significant biosecurity risks. Navigating this landscape requires striking a balance between fostering rapid innovation and implementing robust safeguards against the accidental or intentional misuse of GenAI for biological sequence design.

In this work, we focus on generative sequence models as a subcomponent of biological AI. These are systems that are trained on large corpora of proteins or nucleic acids at the genome scale. They learn the statistical structure of biological sequences and can generate new ones. Their near-term promise is tangible. Such models have already designed functional proteins, enzymes and high-affinity binders, and small viral genomes, with applications ranging from bespoke bacteriophages against antimicrobial resistance to faster design of therapeutics and medical countermeasures[1–6].

These capabilities are advancing quickly, as improving architectures and the industrialisation of synthesis compress the design-build-test cycle. A structural asymmetry accompanies this progress: in silico generation now far outpaces the throughput of physical synthesis and validation, so models can propose sequences faster than they can be screened or tested. The same generative power that enables beneficial design also lowers the barrier to misuse[7], and this dual-use tension is the central concern of this Perspective.

We set out the dual-use risks of biological sequence design and propose a defence-in-depth framework spanning the governance of training data, models and nucleic acid synthesis, summarised in Table 1. While recent work has addressed individual aspects of this challenge, including tiered data governance[8] and a review of the risk landscape[9], the present work provides an integrated set of recommendations covering the entire biological risk chain, from data to models to synthesis, with a particular focus on the European context.

This Perspective consolidates the discussions of the Wübben Sandpit conference, a 3-day interdisciplinary workshop held from 25 to 27 November 2025 at Wasem Kloster Engelthal, Germany and funded by the Wübben Wissenschaftsstiftung. The workshop brought together twenty participants spanning molecular biology, artificial intelligence, bioinformatics, biosecurity, ethics, medicine and policy, drawn from academia, industry, applied research and government across Europe and the United States, together with two neutral facilitators. It followed a design-thinking format that moved iteratively from problem space to solution space through alternating rounds of divergence and convergence, combining facilitated small-group work, structured brainstorming, clustering and prioritisation exercises, and plenary synthesis. Participants first mapped the translational barriers and dual-use risks of generative AI for biological sequence design, and then drafted candidate recommendations, each linked explicitly to the barrier or risk it addressed, its intended audience, and its dependencies. These drafts were stress-tested by rotating participants between groups and by probing how each recommendation might be circumvented, fail in practice, or disadvantage under-resourced actors, and were then prioritised through participant voting on perceived importance and feasibility. No formal consensus vote was taken; the recommendations presented here are the directions that emerged from these discussions, consolidated and refined by the authors after the workshop.

Near-term dual-use risks of biological sequence design

As these capabilities advance and models become more widely available, they create new dual-use risks that must be anticipated and governed.

In this section, we provide an overview of the dual-use risks associated with these models, from data collection and trained models to sequence synthesis and potential deployment in physical and biological systems (Table 2). Risk originates at the data level. Any information about sequence-to-function data, particularly that which includes pathogenic or toxic sequences, can be used by malicious actors to train or fine-tune models that can design sequences with

https://doi.org/10.1038/s41467-026-78758-5

Table 2 | Dual-use research of concern (DURC) has traditionally denoted wet-lab work such as gain-of-function experiments

	Data	Model	Synthesis	Weaponisation
Actors	Researchers publishing pathogen data, governmental, philanthropic and other funding bodies requiring open access data	Researchers building models and sharing weights/code publicly, government and industry providing infrastructure	Process engineers developing cost-effective methods, synthesis providers, CROs and vendors in differing governance contexts	Lone wolves and terrorist groups, biohackers, organised crime, rogue states
Applications	Basis for enabling dual use capabilities and circumvention of existing security (adversarial fine-tuning)	Lowering barriers to training models, potential to develop novel sequences (toxins, genomes), circumvention of existing security (design and synthesis of functional homologues)	Custom nucleic acid synthesis, cell-free expression systems, creation of GMOs, gene drives	Agroterrorism, ecological and economical warfare, targeted pathogens for specific subpopulations, novel pandemic pathogens

Because generative and predictive models can now perform in silico what previously required the bench, for example, forecasting immune-escape variants, we treat this computational capability as in silico gain-of-function and map it, like wet-lab DURC, onto the biological risk chain, where it connects at multiple points. We distinguish between actors that push research towards potential risks and the applications that harbour risks. Furthermore, we characterise the domains in which risks are created and can potentially be mitigated as follows: data: focusing on the inherent risk in some biological data and its impact on the training and use of downstream models. Models: concerned with how models are trained and how a trained model can be used for malicious actions. Synthesis: bringing the designed biomolecule to the physical world. Weaponisation: considering how the malicious use of biological agents could manifest.

potentially harmful functions. When using pre-trained models, a large number of annotated sequences is not required. In their 2025 study, King et al. fine-tuned a genome language model on around 14,000 *Microviridae* genomes and generated variants of the small (~5.4 kb) ΦX174 bacteriophage genome; after extensive computational and experimental filtering, 16 of 285 synthesised designs were viable, infectious phages at 93–99% nucleotide identity to their nearest natural genomes[5]. This calls into question one of the major obstacles to utilising public data, as effective labelling is required and recommended, despite the possibility for dual usage. Funding bodies and researchers must play an active role in deciding which data should be withheld from the public if it includes dangerous sequences or sequence information. However, genomes of viruses that are considered high risk for misuse as biological weapons and those that have led to global pandemics are already in the public domain[10]. Another issue that will be mentioned repeatedly across different domains is the circumvention of security safeguards, as pathogenic data could be used to adversarially fine-tune pre-trained models that did not contain pathogens in their training corpus[11].

The general availability of genomic data lowers the barrier to training and designing models for all actors, enabling small laboratories and, potentially, malicious ones; large, high-capability models work similarly. Key actors here are researchers who open-source data and state actors that provide training infrastructure. Open-sourcing encourages collaboration and accelerates innovation, but this must be weighed against the associated threats[12,13]. Therefore, it is important to differentiate between what such models have demonstrated already from what remains theoretical, albeit plausible. To date, generative models have been shown to design functional proteins and interaction complexes, enzyme active sites and small, template-anchored viral genomes and operons, validated in benign laboratory systems[1–6]. In parallel, sequence models can already predict the fitness and immune-escape potential of viral variants in silico[14–16], the same properties that gain-of-function studies probe experimentally. Biorisk evaluations find that such virus-relevant capability is present in open-weight genome models and is not reliably removed by filtering their training data[7,17]. What has not been demonstrated is the de novo design of enhanced human pathogens, deliberate host-range expansion, or gene drives. However, substantial barriers remain between generating a viable sequence and achieving a defined biological function, pathogenicity and real-world deployment. The concern we address is that continued progress could enable such capabilities over time, not that current models can already produce weaponised agents. A designed organism first needs to be synthesised in order to be introduced to the environment.

This process would need to involve process engineers, biochemists, commercial synthesis providers, or contract research organisations (CROs). The latter are particularly influenced by the governance framework in which they operate, which varies greatly between countries. While many synthesis companies do screen sequences sent to them for harmful ones using various technical measures, this is not mandatory in many countries (for current provider-level screening commitments, see, for example, the Johns Hopkins Center for Health Security list of providers attesting to the U.S. Framework for Nucleic Acid Synthesis Screening[18]. Additionally, sequence-similarity screening is moreover increasingly challenged by AI-designed variants, motivating a shift towards function-based screening[19]).

Finally, these capabilities raise the question of deliberate misuse, which spans scenarios that differ greatly in feasibility, in the actors capable of realising them, and in the contribution of GenAI. At the more tractable end, models could assist in designing protein toxins; far more demanding are the enhancement of infectious agents, the creation of host-specific pathogens for agricultural or ecological harm, and, in the most severe case, agents with pandemic potential. Most of

https://doi.org/10.1038/s41467-026-78758-5

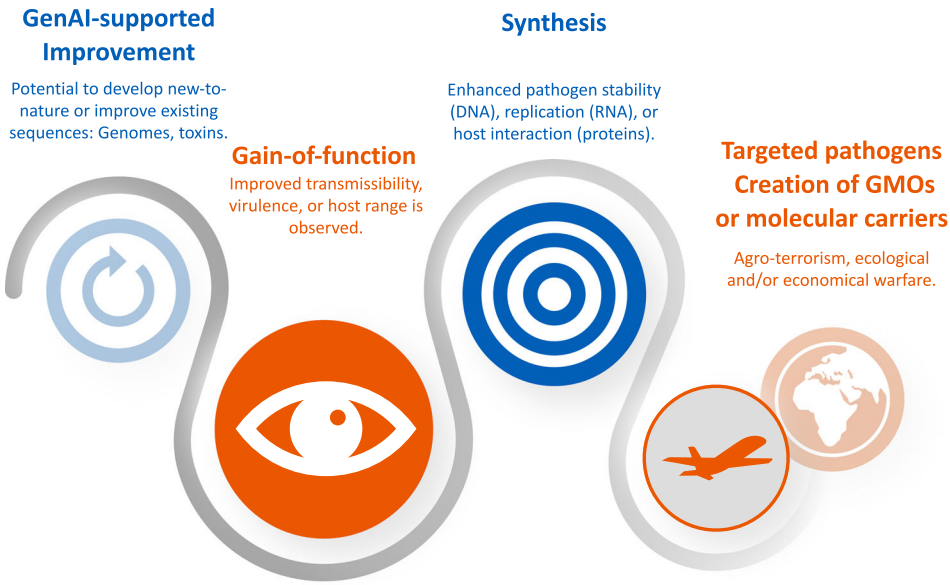


Fig. 1 | This infographic illustrates a series of events across multiple stages of the risk chain mentioned in Table 2 that culminate in the manifestation of biological harm. The four main events relevant for biosequence modelling are shown. When used independently, they cause little to no harm. However, a threat emerges when they occur in sequence. 1. Generative AI (GenAI) is transforming synthetic biology by enabling the design of novel 'new-to-nature' sequences, including complete genomes and proteins such as toxins, which could have different properties from naturally evolved organisms. These tools are trained on vast genetic datasets in order to learn the 'grammar' of life and create functional sequences that might bypass existing biosecurity safeguards in nucleic acid synthesis screening. 2. Gain-of-function (GOF) research involves altering an organism's genome to enhance its existing abilities or confer new ones. These new abilities may include improved transmissibility, virulence, or host range. Examples of observed GOF at different molecular levels include modifying DNA/genomes to alter pathogen stability or virulence, engineering RNA to enhance replication or immune evasion, and mutating proteins to strengthen host interactions or expression patterns. With increasing model capabilities, more of this work could be done purely in silico, further accelerating the process depicted here. 3. Synthesising these sequences is how they go from digital to physical substrate. While most synthesis providers have implemented safeguards against misuse, these are not mandatory and can vary considerably between countries and providers. Furthermore, GenAI can be exploited to circumvent these safeguards. For instance, new-to-nature sequences are unlikely to be recognised by methods designed to detect known pathogens. Models can also be used to conceal a gain-of-function mutation by shuffling the sequence or employing alternative codons to encode a target protein. 4. Creation of biological agents or molecular carriers, which are then disseminated into the environment by a malicious actor. Depending on the malicious actor's scope and goals, these could be targeted pathogens or pandemic agents.

these outcomes remain well beyond demonstrated capability and depend on tacit knowledge and physical resources that lone actors are more likely to lack. The concern is not that any single model already enables them, but that GenAI could lower barriers at several points along the biological risk chain (Fig. 1), with a contribution that differs from step to step. As noted above, models may also assist with the design and circumvention of safeguards and with experimental design and laboratory work[20–23].

Strategic recommendations for mitigating dual-use research of concern risks in generative AI for biological sequence design

Throughout the Sandpit conference, participants converged on the view that mitigating dual-use risks inherent in the computational design of genomic sequences requires a defence-in-depth architecture. Rather than relying on a single point of control, the proposed framework takes a layered approach that covers all parts of the risk chain, from the initial curation of data and training of models to the final biological synthesis of sequence outputs.

A key principle of these recommendations is that governance must be dynamic and iterative. As Genomic GenAI capabilities advance at an unprecedented rate, classification systems and access controls must undergo regular technical audits. Participants identified 'upstream' levers, specifically funding mandates and data-sharing agreements, as the most effective instruments for embedding responsible practices before high-risk sequences reach the synthesis stage.

Although the objective is protective, these recommendations acknowledge certain operational limitations to ensure that security measures do not inadvertently impede legitimate research. To prevent a 'bottleneck effect' on public health research or pandemic preparedness, we propose a tiered-access model. Moving away from a binary 'open vs. closed' data philosophy, we support a risk-stratified environment. For data, we build on the tiered scheme of Bloomfield et al.[8], which sorts biological data into five biosecurity data levels (BDL-0 to BDL-4) according to the capability it could confer on a model trained on it. We summarise these levels, and the platform that would host the data, in Recommendation 3.

With such an approach, governance can be a facilitator of secure innovation. The shift towards structured oversight in biological sequence design is supported by successful applications in related high-stakes biological fields, as well as by successes in data management, software security and industrial manufacturing (see the overarching justifications below;[8,24,25]). The two boxes below collect points that recurred throughout the Sandpit discussions and that touch multiple recommendations within the three domains (data, models or synthesis). The justifications are precedents to mechanisms that transfer between domains: tiered access and red-teaming, for example, serve both data and models, and risk-scoring methods developed for data can be repurposed to score de novo sequences. The limitations are the concerns that likewise apply across multiple recommendations or are general problems of risk classification and mitigation in unknown frontiers and could be brought up against suggestions such as tiered access.

Overarching justifications

- Data tiering (the provenance precedent): the UK Biobank and other similar repositories show that providing controlled access to sensitive genomic information improves data standardisation

and documentation. This structured approach increases the resource's utility for the global research community while maintaining strict privacy and security safeguards.
- Model Red-Teaming (The AI Safety Precedent): in the broader AI domain, 'gated release' and adversarial red-teaming have become standard practice for high-capability models (e.g. Frontier Model Forum protocols). Applying these practices to GenAI in biological sequence design enables the identification of potential biological risks, such as a model's capacity to design novel pathogens, before the weights are made publicly accessible.
- Synthesis screening (The Physical-Gatekeeping Precedent): The International Gene Synthesis Consortium (IGSC) has already established a successful, industry-led screening framework for physical DNA orders[26]. Expanding these protocols to include more sophisticated, AI-generated sequence variants ensures that the 'digital-to-biological' bridge is monitored by robust, standardised biosecurity checks.
- Institutional synthesis: rather than proposing overly complex new regulations, these recommendations are designed to integrate with existing institutional review processes and funding requirements. This approach minimises administrative friction while maximising compliance across the research lifecycle.

Overarching limitations
- Risk classification is an inherently difficult problem because an AI's function is to identify patterns. Without knowing which datasets or models could enable harmful activity, there is a risk that restrictions will either be too broad or too narrow. Furthermore, stringent regulations could restrict valuable research that could significantly benefit public health. They could also erode the openness of science and increase fragmentation across research.
- Even without considering the circumvention of guardrails by bad-faith actors, access restrictions could create a two-tiered research community or be subject to political pressure. Well-organised malicious actors might establish clandestine mirrors of classified datasets/models. At worst, this could give malicious actors (e.g. state actors developing biological weapons) an advantage over defensive research.

Governing data associated with biological risks

We have the following recommendations for data: (1), (2), (3) and (4).

1 Standardised pathogen data classification: initiate a multidisciplinary consensus process to define 'dual-use pathogen data' by reconciling existing risk classifications into a shared common ground, complementing identity-based lists with functional genomic signatures of virulence and other relevant properties where these become available.
2 Community biosecurity literacy: launch targeted initiatives to raise awareness among researchers of the dual-use potential of specific datasets, and of their responsibilities as stewards of the data.
3 Tiered access architecture: implement a risk-stratified access system utilising proportionate controls ranging from open access for foundational data to credentialed verification for functional data related to pandemic pathogens.
4 Integrate biosecurity reviews into the funding cycle and require robust risk-mitigation plans for any project intended to generate sensitive data.

The role of data governance. Data serves as the fundamental engine of generative AI for biological sequences: the combination of raw sequences with functional labels determines the boundaries of the model's predictions and generation. As high-throughput experimental pipelines become more widespread, the volume of functionally annotated data (e.g. with regard to toxicity, host tropism or virulence) continues to grow. While this increasing granularity enhances model performance, it also raises the risk of such data being used to generate sequences with harmful biological characteristics.

The feasibility is shown by antigenic-escape modelling: deep mutational scanning (DMS) datasets for viral proteins can fine-tune foundation models to better predict immune-evading mutations[14–16]. Automated design of enhanced human infectivity is not yet an established capability, but this precedent shows how targeted fine-tuning injects specialised knowledge into foundation models.

This fine-tuning vulnerability complicates voluntary oversight: although models such as Evo 1 and Evo 2 exclude eukaryote-infecting viral data from their primary training sets, researchers can and do re-integrate it during downstream fine-tuning, and when such models are published openly they can lower the barrier to designing sequences with optimised pathogenic traits[11,27].

Recommendation 1: consensus on data classification. We recommend establishing a working group, modelled on bodies such as the Joint Committee of the German Research Foundation (DFG) and the Leopoldina, the expert committees of the National Academies of Sciences, or the WHO Advisory Group on Dual-Use Research of Concern. This group should develop and release a classification of dual-use pathogen data that should not be openly available, which can then be widely adopted. This group should include experts in molecular biology, AI, public health, national security, biosecurity, policy-making and data security.

We emphasise that this recommendation is organisational: its aim is not to discard existing risk classifications but to reconcile them. Risk-based classifications of pathogens and dual-use data already exist across jurisdictions, for example, select-agent and export-control lists. What is suggested here is a single, ideally international body that harmonises them into a shared common ground and agrees where the cutoff for dual-use data warranting tiered access should lie.

Beyond classification, the discussion should also focus on organisational aspects, specifically who should control access and be responsible for classification. Possible stakeholders include data generators (e.g. wet-lab scientists or primary data analysts), database curators, database operators, institutional boards and national/international committees.

Consensus groups must regularly reconvene to update data classification, especially as capabilities advance. Reaching consensus will be difficult as stakeholders often have different views on data types. Explicitly identifying potentially sensitive data carries risks: by delineating what could be misused, we may inadvertently provide malicious actors with a roadmap for seeking out and exploiting valuable datasets.

Recommendation 2: responsible data release. As part of the responsible use of data and general data risk assessment, we recommend that research organisations develop internal guidelines and procedures for assessing data risks, and update these guidelines as new dual-use risks emerge from advancing GenAI capabilities. In order to encourage responsible self-governance among researchers, professional associations and funding agencies should incorporate these principles into their codes of conduct (for the latter, c.f., Recommendation 4).

In addition to the consensus group, which must comprise a broad range of relevant scientific expertise, it is essential that the scientific community, that is to say researchers who generate the data under discussion, is given the opportunity to provide feedback on the potential oversight frameworks, and ultimately endorses them.

This approach ensures that, in the absence of stricter measures, responsibility for considering responsible data release falls on lower organisational levels: research organisations, institutes or

individual groups. However, relying on researchers alone could lead to inconsistent adherence, as they may lack awareness of which data is important or may be under pressure to release data due to grant requirements. Voluntary guidelines also risk allowing dual-use data to remain publicly accessible or to be illicitly acquired by malicious actors.

Recommendation 3: tiered-access data platform. We propose creating a secure, tiered-access platform to host dual-use pathogen data. Such a platform would require user authentication and maintain detailed access logs, thereby enabling accountability and supporting investigations into any potentially problematic use of data. In addition to improving security, a shared repository could save researchers time and resources, facilitate collaboration, reduce duplication of effort and establish consistent standards for the storage and use of sensitive biological datasets. Concrete methods have also been proposed for extracting value from such data without exposing the hazardous whole[28]. For example, raw sequencing reads can be pooled with their order randomised and identifying metadata stripped, so that a model can still learn from the aggregate while the intact, assembled genome is withheld. Alternatively, data can be shared as alignments to reference sequences with the sensitive regions masked. What is shared is no longer a sensitive sequence, but would still hold biological information depending on the biological use case.

The tiered scheme of Bloomfield et al.[8] illustrates what such a platform would host. It sorts data into five biosecurity data levels according to the capability the data could confer: BDL-0 is openly accessible data with no controls; BDL-1 and BDL-2 add managed access and identity accreditation for data on eukaryote-infecting and pandemic-capable viruses; and BDL-3 and BDL-4 add use-case approval, access only through a certified trusted research environment, and prepublication review, up to functional data that could enable the design of pandemic-capable variants. The platform we propose would implement these levels for dual-use pathogen data.

This approach carries risks. Centralisation can hinder research if access becomes overly bureaucratic or datasets are poorly annotated; a single repository is an attractive target for state-level adversaries; credential compromise could undermine access controls; and maintaining the infrastructure could divert funding from research. Even so, if the alternative is that the data is not generated at all because of dual-use concerns, a secure tiered-access platform could still enable important public-health and pandemic-preparedness work.

This would not be a binary choice between access restrictions and no access restrictions, and it could include different tiers modelled on BSLs (biological safety levels). Managed access is, moreover, not new to biology: privacy considerations already require that human genomic and health data be shared only in de-identified form or through controlled-access repositories such as dbGaP and the European Genome-phenome Archive[29,30], a well-established precedent for the tiered, managed access we propose for dual-use pathogen data. To ensure the safety and intended usage of this platform, red teaming, audits and the four-eyes principle are recommended. Such a platform could facilitate standardised data annotation, enabling legitimate researchers to work with dual-use pathogen data more easily. As mentioned above, this approach has recently been discussed for data infrastructure and sharing of DURC relevant data [8,27,31].

Recommendation 4: funding-stage review. Funding bodies (e.g. the DFG, the US National Institute of Health (NIH) and the Coalition for Epidemic Preparedness Innovations (CEPI)) should review grant proposals involving the generation of data on dual-use pathogens and attach risk-mitigation requirements, such as tiered access, as conditions of funding, and decline to fund projects whose risks cannot be adequately mitigated. Funders control the use of their funds rather than the underlying research, but at the funding stage this is the most effective available lever. This is already emerging in practice: CEPI has adopted a biosecurity-by-design approach for its Pandemic Preparedness Engine, embedding tiered, managed access and contractual biosecurity requirements into the platform it funds[32]. This approach is fairer to researchers, as it alerts them to potential restrictions before projects begin rather than after data has been generated. However, the effort outlined in Recommendation 1 of establishing an expert consensus process is a prerequisite for implementing this. It also requires the implementation of Recommendation 3, which involves establishing a tiered-access platform first.

Control belongs at the funding stage because later dissemination controls, for example at journal publication, are not feasible in an era of preprint publications. Requirements would typically mandate tiered access via a standardised trusted research environment, scaled to how dual-use the data is, and funding can, in the last resort, be withheld from projects that would generate the data without adequate safeguards.

This would also mean that funding bodies would not require researchers to publish their data openly. Based on some Sandpit participants' experience, even if researchers have dual-use concerns about publishing their data and results and would prefer not to, they are often required to do so under the terms of the grant. This recommendation has also been discussed in the US context[33].

Meta considerations

- These recommendations only apply to data that is generated in the future. It is not possible to apply them retrospectively to existing data.
- Pathogen data controls are a new oversight proposal. Frameworks should ramp up from a voluntary pilot scheme to potentially mandated regulations or comprehensive implementation at a later stage.
- Data that could be used to generate pathogens with enhanced pandemic potential should be prioritised, as these could cause disproportionately large-scale disasters for society.
- Global harmonisation and implementation of all recommendations is ultimately needed.
- Data on dual-use pathogens of concern account for only a small proportion of all biological data, and the recommendations therefore leave the vast majority of biological data unaffected[8,27].
- Data-level controls form only part of a defence-in-depth approach, and should be considered alongside recommendations at the model and synthesis levels.

Governing biological sequence models

We propose three mutually reinforcing measures for governing biological sequence models:

1. Lifecycle risk-benefit assessment: implement structured frameworks for the continuous evaluation of model capabilities, from initial training to post-deployment, to identify any hazardous properties that emerge as the model evolves.
2. Validated safety benchmarking: develop hybrid evaluation protocols that combine computational red-teaming (adversarial in silico testing) with controlled experimental proxy validation, ensuring that model-generated sequences do not exceed safety thresholds for harmful pathogen properties.
3. Tiered model deployment: establish secure, hosted environments with API-based managed access for high-capability biological models. This allows for independent risk audits and prevents the unfettered distribution of model weights that could be used for malicious activity or fine-tuning.

The role of model governance. Biological sequence models are at the core of our discourse, as they learn patterns from biological data and develop generative capabilities, potentially developing a functional understanding of biology in the process. Their potential is impressive,

as they enable not only reproduction, but also generation of new sequences by interpolating between known embeddings or by sampling from an embedding space that extends beyond the range of existing sequences. This makes them a potentially transformative step in synthetic biology and biological engineering, as well as a potential risk: new-to-nature sequences and purposefully designed biological agents (e.g. modified pandemic pathogens or toxins).

The three recommendations below each require a credible way to score the risk of a model or its outputs. No such validated risk score yet exists. The measures proposed here should be understood as promising research directions for filling this gap. For instance, perplexity only indicates how familiar a sequence is to a model; latent-space proximity to known pathogens; and probes that predict properties such as human-host association are indirect proxies that can produce both false positives and false reassurance. Also, their signal may reflect mere correlation instead of a harmful function. Recent biorisk evaluations further show that capabilities of concern can persist after data filtering and be recovered through fine-tuning or surfaced by probing, so a low score does not guarantee safety[7,11]. Policy readers should therefore treat the tiered access, benchmarking and deployment recommendations that follow as contingent on progress toward risk scores that are validated, reproducible and difficult to game; developing such scores is itself a priority.

Recommendation 1: lifecycle risk–benefit assessment. We propose assessing the risk of a biological AI project using a predefined framework that could be adopted more widely. This risk–benefit assessment is a governance decision, distinct from the capability testing and benchmarking that inform it (Recommendation 2): the tests generate evidence about what a model can do, whereas the assessment weighs that evidence against the project's benefits to decide how the model should be handled. This could be developed in an expert consensus working group, as mentioned above. Such frameworks are common in other fields, such as wet-lab DURC. A risk-benefit analysis would need to be implemented from the pre-development outset at the grant and funding stage, beginning with an assessment of the risks posed by the given dataset and expected model capabilities, and with the risks and benefits being re-evaluated during model development[34]. For example, a model that generalises well might seem harmless initially when learning only from non-pathogenic sequences. However, if it learns to sufficiently generalise, it may be able to generate dangerous sequences regardless, or modify pathogens from other species to infect human hosts. This necessitates continuous re-evaluation throughout the project[9].

Assessing a model's risk requires looking at both its inputs and outputs. On the input side, this means whether known high-consequence pathogen data are present in the training set, together with the data modalities, the scale of the corpus, and the risk carried by sequences both individually and in aggregate, as well as the prompts used to query the model. Output sequences are judged against the same criteria at the level of individual sequences. Accompanying metadata and annotations can further inform risk, for example links between data points and patient symptoms, or the disease context in which the data were generated.

A risk-benefit assessment framework would need to be informed by benchmarks that allow the evaluation of both model performance and risk. Funding agencies could make such assessments mandatory for the implementation and funding of a subset of biological AI projects. The first steps towards risk-scoring biological AI tools have already been taken by third-party non-profit organisations[7,35].

One downside is that this approach places responsibility on model developers, who may have a conflict of interest or vary in how knowledgeable they are about biological risk and how seriously they treat it. A tiered-access platform, discussed below, addresses this by providing a neutral final risk assessment and incentivising early risk-benefit analysis. A second concern is that pruning datasets or weights to achieve lower risk scores could degrade performance, which, given the drive to build 'best-in-class' models, could discourage developers from adopting safe practices.

Recommendation 2: validated safety benchmarking. A general evaluation gap exists for biological sequence models. While some frameworks exist for evaluating models, these do not build on each other and depend on arbitrary biological downstream tasks. Moreover, the emphasis is on embedding models and the biological meaning of the representations learned by them, which is why they are not well suited to evaluating generative power[36,37].

Testing the perplexity (a measure of how well a model predicts a given sequence, where lower values indicate greater familiarity) of viral sequences can give a first impression of a model's ability to generalise to unseen data, but is insufficient for the assessment of risk. Recent work is beginning to formalise this kind of risk-focused evaluation for genome models by separating a model's evolutionary novelty (how far its outputs depart from natural genomes) from its design efficiency (its ability to find viable sequences); applied to Evo 2-generated bacteriophages, it finds that apparent design efficiency largely reflects proximity to known sequences and not genuine exploration of novel genomic space, indicating a low-to-moderate biosecurity concern for de novo hazard creation at present[38]. Security benchmarks and red-teaming are already established in other fields applying AI models, especially in cybersecurity. Paired with interpretability methods, which demonstrate early promise in biological models, for example through the application of sparse autoencoders to extract biologically meaningful features from genomic language model representations[39], these benchmarks would be able to point out what is risky, as well as why.

Risk-focused benchmarking. Inclusion of risk-scoring at the model and sequence level is an essential component of a risk-focused benchmark. It must be reproducible and model-agnostic, producing datasets primarily for training and testing purposes, with the focus of the output scoring on the post-processed sequence rather than token representations. Ideally, sequence-level risk scores would be combined with model-level performance on risk-related tasks to produce a comprehensive risk score.

Relevance is not limited to representation learning but extends to sequence generation, which requires generative capability. Benchmarks need to ask if and to what degree a sequence can be aligned with known pathogens, if a model can complete the sequence of a given pathogen, and if a novel sequence encodes a harmful function. Testing a model's ability to generate out-of-distribution (OOD) sequences is valuable but hard: training from scratch on a withheld corpus from known pathogens would need a very large dataset and might pose risks inherent to a given model's training corpus. A more tractable way to build adaptable benchmarks for multiple model types and test their ability to generate de novo sequences is to build orthogonal models to score the output sequences. These models could use sequence-to-function prediction tasks or functional alignment based on embeddings to see if sequences co-localise with known pathogens in the latent space[14,24,40]. Such capability-based scoring of de novo sequences is not only relevant to model governance: it could also inform the expert consensus body proposed for data classification (Recommendation 1 of the data section), giving its dual-use cutoff a functional basis.

Interpretability. In this context, interpretability methods would also play a key role. Because a genomic sequence is not itself human-readable in the way a small molecule is, and because a hazardous function may be concealed within or distributed across otherwise-benign sequence, interpretability serves a second function beyond risk-scoring: it is the channel through which a human expert can reason about a candidate sequence before it is synthesised, keeping meaningful oversight in the loop at the design stage. Such a framework

opens both the model's decision process and the properties of the generated sequence to scrutiny. To ensure model alignment, it is necessary to understand what is happening inside the 'black box'. Interpretability could be used to identify high-risk sequences by locating pathogenic features, and it could also facilitate the identification of novel sequences that would not be detected by homology or sequence alignment tests. Early tools of this kind for genomic language models, such as Decode-GLM, aim to make these internal features explicit and auditable, offering a route to a flagging function that does not depend on sequence similarity[39]. Extending this paradigm to the protein domain, PepTriX establishes a pipeline with interpretability by design for peptide analysis, rendering internal representations explicit and auditable[41]. However, it is difficult to quantify interpretability methods, as DNA is not human-readable. Even in natural language processing (NLP), literature mostly relies on large-scale auto-interpretability or manual qualitative evaluation[42].

We propose mimicking NLP methods that use concept probes to interrogate model internals. A probe is a classifier trained to predict human-level concepts. One advantage of this approach is that probes are straightforward to implement and enable the automated evaluation of model interpretability. For instance, a probe that predicts the probability of a pathogen infecting a human host could serve as a risk score for a generated sequence.

It should be noted that interpretability models can pose a risk themselves. Identifying functional features facilitates steering the generation of biological sequences. For example, model outputs could be steered towards certain viral tissue tropism according to host probabilities. Given the current performance of de novo sequence generation, one could argue that steering model generation poses a more immediate risk.

Wet-lab validation. Although computational benchmarks provide an initial screening, definitive assessment of the hazard posed by a de novo sequence ultimately requires wet-lab validation, which is itself the most sensitive step. How to keep this step proportionate within legal and ethical bounds was a subject of extended discussion at the Sandpit conference. The risk is not uniform across candidates: following the framework of Godbold et al.[43], the genuinely dangerous experiments are those that would empirically confirm a Sequence of Concern or demonstrate the transfer of a pathogenic function into a viable organism. This narrow class, not the much larger set of computationally flagged sequences, is what warrants the strictest controls and is most relevant for downstream risk testing. We recommend that if such testing is justified, it should use attenuated or non-pathogenic proxy organisms and tasks, since this allows the biological risk to be assessed without creating the pathogenic agent itself, developing a 'misuse blueprint' that could be exploited, or risks violating national or international law[44]. We note that focusing on safe proxies inherits the difficulty of classifying danger in the first place: experienced reviewers frequently disagree on whether a given experiment constitutes DURC, which underscores that these proxy evaluations will be subject to debate and require input from expert communities for their experimental design. Still, we believe this is preferable over creating harmful pathogens as part of these risk evaluations.

Red-teaming. In addition to a general benchmark, red-teaming of a subset of models is needed to identify capabilities of concern and assess whether a model's safeguards can be circumvented by malicious actors.

Because red-teaming is labour-intensive and requires manual interaction with the model and code, it cannot be applied to every published model and is reserved for models that warrant further assessment indicated by automated risk scores. The aim is to guide a model towards an accurate risk level in actual usage, analyse the safeguards and their circumvention, and share results with the developers to inform the final risk assessment and improve future versions, without publishing misusable details.

Such work should sit within a secure evaluation sandbox, with a dedicated consortium developing an ontology of benchmark types and red-teaming goals (a precedent is Transluce's Network of Evaluators forum). The ontology would carry if-then commitments: a model that reaches high-risk levels or is easily steered towards dangerous outputs is assigned to a higher tier. Making such if-then commitments a priori reduces both false positives, risk going unnoticed, and post-hoc shifting of goalposts.

Recommendation 3: tiered model deployment. In order to mitigate the risks posed by certain models, such as the generation of new pandemic pathogens, while still allowing beneficial applications to be realised, we propose establishing a tiered-access platform similar to that outlined in the data section. This platform would host biological AI models, simplifying access for users while implementing risk-dependent access controls and a Know-Your-Customer approach. Models with low risk scores would be fully open source, while those with higher scores would require institutional email addresses or other verification measures and, in stricter cases, institutional verification. Models with the highest associated risk could only be run within the platform as a trusted research environment (TRE). Outputs could be provided upon a reasonable request presented as a proposal. These model-access tiers are summarised in Table 3.

The risk of a given model would be determined by recommendations (1) and (2) of this section: an in-platform risk-benefit analysis based on automated benchmarking, including red-teaming in edge cases. The platform would encourage self-testing before upload and perform a final test to determine the model's tier.

The platform would lower barriers for researchers: high-tier models would run on secure, developer-subsidised servers, and a central API would let models be compared and combined for shared

Table 3 | Summary of the model-access tiers

Tier 0 unrestricted access	Tier 1 soft wall access	Tier 2 institutional access	Tier 3 restricted access
Open access model; all parts of the model, such as weights and code, are shared.	Access is only possible via a secure institutional login. Weights and architecture can be shared after approval of an institutional email address or corresponding verification.	Access can only be granted via a proposal from the institution. The model weights and architecture can be shared with a verified, secure location such as an institutional cluster with secure compute nodes.	Access depends on permission proposal justifying need-to-access ongoing project. This is required to run inference and retrieve the model's outputs. The model does not leave the trusted access environment. Inference is performed on servers that are specifically installed for the purpose of sharing the model, and proposals must justify the use of the model and the intended generation of data and sequences.

The corresponding data-access tiers (BDL-0 to BDL-4), adopted from Bloomfield et al.[8], are described in Section 'Strategic recommendations for mitigating dual-use research of concern risks in generative AI for biological sequence design'. The tiered access for the model encompasses different levels of access for the weights and sometimes also the architecture and output of the model.

goals. It could also give smaller model providers increased visibility. Such a platform could also be expanded to include data access, and native fine-tuning and model modification functionalities. As emphasised by Carter and Butchello[25], such a platform must treat equitable access as a core principle, so that verification criteria do not disadvantage researchers at lower-resourced institutions or in the Global South. The main limitations are adoption and cost: maintenance and compute are significant; an initial state grant would need a handover plan to industry or open-source platforms, and a central repository is an attractive target for attack. We judge these costs to be outweighed by the security and innovation benefits.

In conclusion, on the data side, an expert consensus is needed to create a model risk assessment framework. Responsible model release must be encouraged, which could be facilitated by a semi-automated, tiered-access platform specialising in biological AI models. Additionally, a review of risk at the funding stage could encourage researchers to assess the risks at a time when risk mitigation measures are still easier to implement.

Governing nucleic acid synthesis

We have the following recommendations for nucleic acid synthesis:

1. Strengthen sequence screening and registries across nucleic acid synthesis providers.
2. Implement KYC rules and supply-chain tracking for critical components.
3. Establish and reinforce clear, mandatory governance for nucleic acid synthesis screening.
4. Set up a sector-level fund to finance research into risk mitigation measures and risks.

Why nucleic acid synthesis governance matters. Although data and models form the computational basis, AI-generated sequences only pose a real-world threat once they have been synthesised and manifest as a physical medium. Therefore, synthesis serves as a critical control chokepoint for the misuse of digital designs. However, the risk is not confined to functional expression within a living organism or a cell-free system. The production of genetic intermediates, such as a plasmid containing a potent toxin gene, poses a significant dual-use risk even before activation in a host. As AI capabilities outpace current laboratory regulations, malicious actors may find new ways to circumvent established screening protocols for these precursors. These rapid advancements necessitate more agile governance across the entire synthesis and biological engineering domain to ensure oversight covers both the final biological products and the foundational genetic components that enable them.

Recommendation 1: strengthen sequence screening. Many nucleic acid synthesis providers globally already screen their orders for sequences of concern against standard databases using commercially or publicly available synthesis screening tools[18]. These screening tools traditionally operated via sequence matching. Sandpit conference participants agreed that synthesis screening systems must expand to integrate homology, structural and functional analyses, and AI-based annotations, while also taking the user's context into account to stay future-proof. This would enable screening tools to progress beyond sequence-only filters, which can be circumvented via AI-driven sequence designs, to a system that integrates various methodologies to build a more realistic risk profile[24]. This could potentially allow them to identify dangerous sequences that are heavily modified or even completely novel de novo. Recent work in this area has focused on generative protein design: variants of toxins and selected viral proteins of concern were redesigned, and their corresponding DNA synthesis orders were tested against nucleic acid screening software[24]. It revealed that existing synthesis screening systems could be circumvented with AI-designed functional homologs, but that the systems could also be patched to recognise AI-driven designs. Still, in a follow-up study, it was demonstrated that the functional homolog designs were not viable when tested in a wet lab as safe proxies[44]. This work addresses protein-level design and corresponding nucleic acid orders, but not the design of regulatory DNA or whole genomes.

In addition to strengthening screening processes, it would be beneficial to register when sequences of concern were made available to individuals. In order to achieve this without overstepping freedom of research and privacy boundaries while still allowing research to operate efficiently, two considerations must be taken into account.

Firstly, user-based authentication would allow risk scores for individuals or organisations when deciding whether sequences of concern should be made available or not. For example, a sequence synthesised for a team of oncolytic virus researchers should be treated differently from one requested by an unverified citizen ordering nucleic acid sequences to a residential address. Even if the sequence shares structural motifs with toxins, rather than assuming weaponisation, the therapeutic intent of the verified researchers should be acknowledged, synthesis should progress, and the risk noted for all involved. However, in the second case, the sequence should not be delivered. The ability of those involved to handle dual-use risks safely should be verified during the authentication process.

Secondly, biological context is paramount. For instance, a virus in a particular animal reservoir (e.g. cattle) could be harmless or even beneficial. However, if that same virus was engineered to be able to be human-to-human transmissible, it would be concerning. Therefore, the broader research context in which the sequences are made available should be taken into account, and filters must generate species-specific risk scores that evaluate the potential harm to humans, agriculture and the wider ecology independently[45,46].

Combining strengthened technical detection mechanisms with the registration of sequence of concern orders creates safety guardrails that enable scientific progress while mitigating the risk of accidental release or malicious misuse.

This recommendation towards mandatory nucleic acid synthesis screening aligns with the proposed EU Biotech Act, which restricts biotechnology products of concern (such as sequences of concern and benchtop synthesis devices) to users with a verified legitimate need (Article 43) and further defines the criteria for sequences of concern to be screened in its Annex I[47]. Our recommendation complements this baseline by calling for screening that integrates homology, structural and functional analysis, and AI-based annotation, so that AI-driven heavily modified or de novo sequences that a sequence-only-based filter would miss can still be flagged in the future.

Recommendation 2: KYC and supply-chain tracking. For synthesis screening to have a meaningful impact, it is crucial to understand who is ordering and synthesising which products at what point in time. The critical components involved are the synthetic nucleic acids themselves (synthetic genes, gene fragments and oligonucleotide pools), the benchtop synthesis devices that allow for synthesising sequences outside of commercial providers, and the reagents and cartridges that supply those devices. In order to prevent malicious actors from using synthesis services to order sequences of concern, regulators will need to implement Know Your Customer (KYC) rules to ensure the products are only distributed to legitimate customers[48]. Serialisation and supply-chain tracking of the benchtop devices and reagents would complement this, making controlled hardware harder to acquire or divert. Neither measure is currently mandated, but both are useful tools for ensuring legitimate use. This would involve extending identity verification and legitimate-use checks, as well as introducing documentation obligations, all of which would be proportionate to the level of risk of a given sequence. This could be implemented alongside a usage registry, as suggested above.

Our KYC recommendation is reflected by the proposed EU Biotech Act, whose Article 44 (verification of legitimate need) would require economic operators to verify a prospective customer's identity, record each transaction, and assess legitimate need before supplying biotechnology products of concern[47]. Article 45 extends the same logic to hardware, requiring benchtop nucleic-acid synthesis devices to screen for sequences of concern. Our proposal is consistent with these provisions and complements them by also calling for serialisation and supply-chain tracking of the devices and reagents themselves.

Some weaker elements of this already exist in the form of serial numbers, end-use logging, and sales controls for certain high-end synthesis equipment. Participants viewed these as useful foundations that require significant strengthening and expansion in light of challenges posed by generative biological sequence design. These measures carry real costs, which bear directly on their feasibility. Identity verification, legitimate-use checks and documentation impose a compliance burden that falls hardest on small providers, on benchtop and decentralised synthesis providers, and on providers in lower-resourced regions, and serialisation and supply-chain tracking of critical components add further overhead. Radio-frequency identification (RFID) tagging is one option for such tracking, though it reaches only points where readers are installed, so it only supports custody checks among registered actors and is not robustly locating devices that have left that network. Applied uniformly, this burden could push orders toward providers or jurisdictions with weaker oversight, weakening the very safeguard it is meant to create. To stay feasible, requirements should scale with the assessed risk of the order and component, and the associated costs could be offset through shared verification infrastructure or the sector-level fund proposed below, keeping the compliant route the path of least resistance. Still, creating a harmonized framework like the EU Biotech Act can support providers by establishing uniform market standards and avoiding the increased provider burden of complying with a fragmented landscape of national nucleic acid synthesis screening requirements.

Participants also emphasised the importance of incident response capacity and human oversight to ensure that automated synthesis processes remain within the scope of responsible governance, and that any suspicious or concerning synthesis activities prompt a timely review and response. There are ongoing efforts in this direction by the Gene Synthesis Consortium, and this is also reflected in the EU Biotech Act's Article 46 on suspicious transactions and reporting of biotechnology misuse.

Recommendation 3: governance of synthesis screening. Responsible governance of nucleic acid synthesis requires technical and procedural measures, as well as robust legal, normative and educational frameworks. While legislation should form the basis, accompanying guidelines can be updated more often to allow governance structures to adapt to rapid technological and threat developments and must be designed for regular revision. Training and education are needed to raise awareness of the risks associated with generative sequence design among researchers, laboratory personnel, institutional leaders and industry stakeholders. Institutional by-laws and manufacturer best practices are important layers of governance that can complement national legislation and international agreements. All of these measures are essential for establishing a governance system capable of addressing both centralised industrial synthesis and benchtop or decentralised synthesis systems.

Several of these elements already appear in the proposed EU Biotech Act, which would require establishing national inspection authorities (Article 48), mandatory reporting of suspicious transactions and misuse (Article 46), and training and awareness-raising (Article 47)[47]. We see these as a solid foundation that can subsequently be future-proofed against new risks from generative sequence design via guidelines and accompanying measures.

Recommendation 4: sector-level insurance fund. As a final point, the risk mitigation measures recommended in this work, especially those in the synthesis section, can be costly and research-intensive. Therefore, participants proposed the establishment of sector-level insurance and pooled funding mechanisms to support sustained investment in preparedness and risk-mitigation research related to synthetic biology and biological GenAI. Such a fund could also be used to finance research into mitigation techniques and governance guidelines. There are precedents for sector-level pooled insurance financing prevention research in other high-risk industries. The most notable example is the German statutory accident insurance system (DGUV), in which employer contributions organised by industry sector explicitly fund occupational safety research[49].

Discussion

As discussed above, falling sequencing costs and the resulting data deluge have outpaced traditional analytical methods and create serious new dual-use risks[9,50]. The EU Biotech Act[47] recognises and addresses the importance of these developments for the Union's competitiveness and strategic autonomy. However, the 'digital deluge' presents a challenge for the scientific community, as they must balance the potential of generative AI with its risks to public health, national security, and scientific integrity[9]. Responsible handling and regulation are essential to ensure progress and security, as the historical precedent of wet-lab DURC demonstrates.

On the data side, the EU Biotech Act addresses the fragmentation and limited accessibility of high-quality biological data through two main channels: Article 33 grants a dedicated status ('biotechnology data quality accelerators') and access to targeted Union support measures to strategic projects that meet strict criteria for curating well-annotated, provenance-verified datasets for training and validating AI models and through networks of health biotechnology clusters (Article 15). Together, these mechanisms encourage active projects to converge on shared formats for data and models, such as bioAIrepo[51]. These measures are relevant recommendations, which presuppose data that is well-characterised and accessible under controlled conditions. Nevertheless, we would like to emphasise that the establishment of legal frameworks for tiered pathogen data access should be pursued in conjunction with these data generation initiatives.

On the model side, the verification gap we describe has a structural response. This is where sequence generation outpaces physical validation. It is addressed in the EU Biotech Act's provision for recognising and supporting trusted biotechnology testing environments for AI-enabled innovations. This is outlined in Article 32. By fostering shared infrastructure for in silico and experimental validation, this measure is intended to let developers demonstrate safety and operate at higher velocity, turning verification from an obstacle into a Union-wide strategic project.

Synthesis is the final and perhaps most critical stage in the biological risk chain, as it is where digital designs are brought into the physical world. The rapid advancement of AI-driven design currently outpaces the development of regulatory frameworks, creating a mismatch that could allow safeguards to be circumvented. For instance, malicious actors could design 'functional homologs', sequences that perform the same function as a known pathogen but differ enough to evade detection[24]. To mitigate such risks and to future-proof biosecurity safeguards, the EU Biotech Act establishes an Advisory Group on Biosecurity (Article 52). Based on this group's recommendations, the Commission is responsible for monitoring systemic biological risks arising from AI models in biological applications (Article 53). These operational screening controls form part of the biosecurity provisions (Articles 43 to 46): biotechnology products of concern are restricted to

users with a verified legitimate need; economic operators must verify customer identity and record transactions; and benchtop nucleic acid synthesis devices must screen for sequences of concern in accordance with the criteria listed in Annex I[47]. Its nucleic acid synthesis screening provisions, which closely align with the recommendations by Sandpit participants, would build adequate protections into the research lifecycle that prevent potential biological societal harms and support responsible practice in the scientific community.

It is important to recognise the high level of technological and scientific interdependence with other fields such as biosecurity that biological sequence design has. Biosecurity is often seen as a brake on scientific discovery, not as the economic enabler and scientific field it actually can represent. But a modern biosecurity regulatory framework is key to a resilient bioeconomy. Setting clear rules and safety best practices enables governance to reduce societal public health and national security risk and encourage investment. Security should be an integral part of technology, providing the social licence that builds trust and enables the scientific community to operate and keep pace with digital technologies. This shift towards interdependence is in line with the EU's strategic autonomy and economic security goals. The EU Biotech Act frames these goals in terms of competitiveness, resilience and acknowledges the importance of security as a technological enabler. By presenting security-by-design as a distinctive European value proposition for gene and protein engineering, breakthroughs in sequence synthesis and functional genomics can be brought to market safely. Therefore, a security-first approach empowers the scientific community to innovate at the pace of digital AI while ensuring safety and cutting-edge progress.

International and multinational frameworks like the EU's proposed Biotech Act are vital, but nation-state governments must also take action. Currently, the security of AI-driven biological research often falls through the cracks of different departments and responsibilities. In Germany, the DFG and the Leopoldina are suitable to be a leader on this debate via their Joint Committee on Security-related Research. By extending their focus to AI-driven biological design and forming stakeholder groups comprising health, defence and academia, countries can develop a coherent strategy, such as the UK Biological Security Strategy. Nation state approaches need to complement the EU's oversight, particularly in realms directly related to national security. Ultimately, responsible governance does not impede progress. Establishing responsible management of these technologies builds public trust and stability.

Ethical and societal impact

Our recommendations are intended to reduce biosecurity risks while preserving equitable access to biological AI. Governance itself, however, carries risks: overly restrictive controls could entrench a two-tiered research community that disadvantages under-resourced institutions and the Global South; poorly designed risk classification could signal which data or models are most valuable to misuse; and centralised access platforms could become targets. We address these tensions through proportionate, risk-stratified controls, equitable access as a design principle, and classification by broad, interdisciplinary groups. The defence-in-depth approach is intended to ensure that no single point of failure determines the safety of the system, so that the community can realise the potential of biological AI while maintaining public trust.

Conclusion

Artificial intelligence for biological sequence design holds transformative promise for the life sciences. Yet, its full-scale deployment in high-stakes domains remains constrained by the gap between computational design and physical validation. A central tension identified by this Sandpit conference is that as these capabilities advance, their dual-use potential proportionately grows. Governance must therefore evolve in step with capability. Importantly, this relationship is not adversarial: many of the safeguards proposed in this work, particularly advances in model interpretability, uncertainty quantification, and structured risk assessment, are likely to improve the scientific utility of these models as well, placing biosecurity and performance as converging objectives. The most productive path forward is to embed biosecurity and ethical considerations into the development of AI-driven biological sequence design and synthesis from the outset, treating responsible governance not as a constraint on progress but as an integral driver of it.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

References

1. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
2. Zambaldi, V. et al. De novo design of high-affinity protein binders with AlphaProteo. Preprint at https://doi.org/10.48550/arXiv.2409.08022 (2024).
3. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
4. Ahern, W. et al. Atom-level enzyme active site scaffolding using RFdiffusion2. *Nat. Methods* **23**, 96–105 (2026).
5. King, S. H. et al. Generative design of novel bacteriophages with genome language models. *Science* **393**, eaec2657 (2026).
6. Merchant, A. T. et al. Semantic design of functional de novo genes from a genomic language model. *Nature* **649**, 749–758 (2026).
7. Wei, B. et al. Best practices for biorisk evaluations on open-weight bio-foundation models. Preprint at https://doi.org/10.48550/arXiv.2510.27629 (2025).
8. Bloomfield, D. et al. Biological data governance in an age of AI. *Science* **391**, 558–561 (2026).
9. Hassan, H. A. et al. A safety-centric perspective on innovation and risk in the use of artificial intelligence in genomics. *Trends Genet* **42**, 820-831 (2026).
10. Pannu, J., Sandbrink, J. B., Watson, M., Palmer, M. J. & Relman, D. A. Protocols and risks: when less is more. *Nat. Protoc.* **17**, 1–2 (2022).
11. Black, J. R. et al. Open-weight genome language model safeguards: assessing robustness via adversarial fine-tuning. In *Proc. NeurIPS 2025 Workshop on Biosecurity Safeguards for Generative AI* (NeurIPS, 2025).
12. Rivera, S. et al. Responsible biodesign workshop: AI, protein design, and the biosecurity landscape–recommended actions. *OSF Preprints* https://doi.org/10.31219/osf.io/yq48e_v1 (2025).
13. Gitter, A. et al. A renewed call for open artificial intelligence in biomedicine. Preprint at https://doi.org/10.31219/osf.io/2xh3w (2024).
14. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).
15. Thadani, N. N. et al. Learning from prepandemic data to forecast viral escape. *Nature* **622**, 818–825 (2023).
16. Ito, J. et al. A protein language model for exploring viral fitness landscapes. *Nat. Commun.* **16**, 4236 (2025).
17. Gurev, S. et al. Evaluating variant effect prediction across viruses. Preprint at *bioRxiv* https://doi.org/10.1101/2025.08.04.668549 (2025).
18. Johns Hopkins Center for Health Security. Gene synthesis screening: implementing the U.S. framework for nucleic acid synthesis screening https://genesynthesisscreening.centerforhealthsecurity.org/ (2025).
19. Abel, G. R. et al. Beyond sequence similarity: toward function-based screening of nucleic acid synthesis. *Front. Bioeng. Biotechnol.* **14**, 1832724 (2026).

20. Qu, Y. et al. CRISPR-GPT for agentic automation of gene-editing experiments. *Nat. Biomed. Eng.* **10**, 245–258 (2025).
21. de Almeida, B. P. et al. A multimodal conversational agent for DNA, RNA and protein tasks. *Nat. Mach. Intell.* **7**, 928–941 (2025).
22. Zhang, Q. et al. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nat. Commun.* **16**, 1553 (2025).
23. Wang, D. et al. Without safeguards, AI-biology integration risks accelerating future pandemics. *Front. Microbiol.* **16**, 1734561 (2026).
24. Wittmann, B. J. et al. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science* **390**, 82–87 (2025).
25. Carter, S. R. & Butchello, G. *A Framework for Managed Access to Biological AI Tools*. Technical Report https://www.nti.org/analysis/articles/a-framework-for-managed-access-to-biological-ai-tools/ (Nuclear Threat Initiative, 2026).
26. International Gene Synthesis Consortium. Harmonized Screening Protocol, version 3.0. https://genesynthesisconsortium.org/harmonized-screening-protocol/. (International Gene Synthesis Consortium, 2024).
27. Bloomfield, D. et al. Securing dual-use pathogen data of concern. In *Proc. NeurIPS 2025 Workshop on Biosecurity Safeguards for Generative AI* https://openreview.net/forum?id=ZgD951FVe7 (2025).
28. Sawaya, S., Lo, C.-C., Li, P.-E., Hovde, B. & Chain, P. Methods for safely sharing dual-use genetic data. *Front. Microbiol.* **17**, 1716431 (2026).
29. Mailman, M. D. et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).
30. Lappalainen, I. et al. The European Genome-phenome archive of human data consented for biomedical research. *Nat. Genet.* **47**, 692–695 (2015).
31. Berke, A. et al. *Data and AI-enabled Biological Design: Risks Related to Biological Training Data and Opportunities for Governance*. Technical Report (RAND Corporation, 2025).
32. Carter, S. R. & Olsen, M. Biosecurity-by-design for CEPI's pandemic preparedness engine. Accessed 2 September 2026 https://cepi.net/biosecurity-design-cepis-pandemic-preparedness-engine (Coalition for Epidemic Preparedness Innovations (CEPI), 2026).
33. National Academies of Sciences, Engineering, and Medicine. Disseminating in silico and computational biological research: navigating benefits and risks. Proceedings of a workshop (The National Academies Press, 2025).
34. Hanke, M. S., Rath, S., Cicero, A., Inglesby, T. V. & Pannu, J. Dual-use artificial intelligence and biology: upstream risk-benefit reviews. *Front. Microbiol.* **17**, 1832974 (2026).
35. Webster, T. et al. Global Risk Index for AI-enabled Biological Tools (The Centre for Long-Term Resilience & RAND Europe, 2025).
36. Marin, F. I. et al. Bend: benchmarking DNA language models on biologically meaningful tasks. Preprint at https://doi.org/10.48550/arXiv.2311.12570 (2023).
37. Schmirler, R., Heinzinger, M. & Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat. Commun.* **15**, 7407 (2024).
38. Black, J. R. M., Maiwald, A., Pannu, J. & Crook, O. M. Quantifying evolutionary novelty and design efficiency in generative genome design. Preprint at *bioRxiv* https://doi.org/10.64898/2026.06.12.731871 (2026).
39. Maiwald, A. et al. Decode-gLM: tools to interpret, audit, and steer genomic language models. Preprint at *bioRxiv* https://doi.org/10.1101/2025.10.31.685860 (2025).
40. Martínez-Redondo, G. I. et al. Fantasia leverages language models to decode the functional dark proteome across the animal tree of life. *Commun. Biol.* **8**, 1227 (2025).
41. Schilling, V., Dubey, A., Anžel, A. & Hattab, G. PepTriX: a framework for explainable peptide analysis through protein language models. *Next Res.* **12**, 102151 (2026).
42. Bricken, T. et al. Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread* https://transformer-circuits.pub/2023/monosemantic-features/index.html (2023).
43. Godbold, G. D. et al. Discerning dangerous gain of function: most gain of function (GoF) research does not involve infectious microbes. *Front. Bioeng. Biotechnol.* **14**, 1818657 (2026).
44. Ikonomova, S. P. et al. Experimental evaluation of AI-driven protein design risks using safe biological proxies. Preprint at *bioRxiv* https://doi.org/10.1101/2025.05.15.654077 (2025).
45. Zinsstag, J., Utzinger, J., Probst-Hensch, N., Shan, L. & Zhou, X.-N. Towards integrated surveillance-response systems for the prevention of future pandemics. *Infect. Dis. Poverty* **9**, 140 https://doi.org/10.1186/s40249-020-00757-5 (2020).
46. Leifels, M. et al. The One Health perspective to improve environmental surveillance of zoonotic viruses: lessons from COVID-19 and outlook beyond. *ISME Commun.* **2**, 107 https://doi.org/10.1038/s43705-022-00191-8 (2022).
47. European Commission. Proposal for a Regulation of the European Parliament and of the Council on establishing a framework of measures for strengthening Union's biotechnology and biomanufacturing sectors particularly in the area of health (European Biotech Act). COM(2025) 1022 final, 2025/0406 (COD) (2025). CELEX:52025PC1022.
48. Feldman, J., Feldman, T. & Anton, A. I. Know your scientist: KYC as biosecurity infrastructure. *Front. Microbiol.* **17**, 1814993 (2026).
49. Fabian, P. et al. Reasonable, adequate and efficient allocation of liability costs for automated vehicles: a case study of the German liability and insurance framework. *Eur. J. Risk Regul.* **9**, 548–563 (2018).
50. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (2022).
51. Thakur, M. et al. EMBL's European Bioinformatics Institute (EMBL-EBI) in 2025. *Nucleic Acids Res.* **54**, D10–D19 (2026).

Acknowledgements

This work is the result of a Sandpit Conference funded by the Wübben-Stiftung für Wissenschaft. G. Hattab, K. Ladewig, H.A. Hassan and M. Sprang acknowledge the support of the Wübben Foundation through the Sandpit Funding Scheme. M. Sprang was supported by funding from the Einstein ECR Award, the Rise Up! programme of the Boehringer Ingelheim Foundation (BIS), the ReALity Initiative of Johannes Gutenberg University Mainz, and the Forschungsinitiative des Landes Rheinland-Pfalz. M. Schäfer was supported by the German Federal Ministry of Health (BMG) under grant No. 2523DAT400, project 'AI-assisted analysis and visualisation of pandemic situations' (AI-DAVis-PANDEMICS). M. Sprang and H.A. Hassan acknowledge funding from the Carl Zeiss Foundation through the Seed Funding Programme of the Centre for Synthetic Genomics. M. Hanke started a position at the European Commission in August 2026, but conducted all contributions to this article in his academic capacity at Johns Hopkins University. The information and views set out in this article are those of the author and do not necessarily reflect the official opinion of the European Commission.

Author contributions

M. Sprang and H.A. Hassan wrote the main draft. K. Ladewig contributed to the conceptualisation of the manuscript and was responsible for the critical review. G. Hattab contributed to the conceptualisation, writing, critically reviewing and proofreading the manuscript. M. Hanke and

M. Schäfer contributed to critically reviewing and proofreading the manuscript. M. Sprang, M. Schäfer, H.A. Hassan, K. Ladewig and G. Hattab discussed, designed, and implemented the figures and tables. All authors contributed to the development of the ideas and contributed to the discussions at the Sandpit conference. M. Sprang and K. Ladewig designed the Sandpit programme. H.A. Hassan designed the outreach material and did literature research and data mining in preparation of the event.

Funding

The authors received funding by the Wübben Stiftung Wissenschaft to convene a Sandpit conference, covering travel, accomodation and venue for all participants and moderators.

Competing interests

H.A. Hassan is a founder of DeOxy Tech Ltd, a company developing artificial-intelligence tools for biological sequence design. B. Aysan is an employee of Merck KGaA, Darmstadt, Germany. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains Supplementary Material available at https://doi.org/10.1038/s41467-026-78758-5.

Correspondence and requests for materials should be addressed to Maximilian Sprang.

Peer review information *Nature Communications* thanks Maximilian Brackmann, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026