Without Safeguards, AI-Biology Integration Risks Creating Future Pandemics

Dianzhuo Wang[1,⋆,†], **Marian Huot**[1,2,⋆], **Zechen Zhang**[3], **Kaiyi Jiang**[4], **Eugene I. Shakhnovich**[1,†], **Kevin M. Esvelt**[5,†]
[1]Department of Chemistry and Chemical Biology, Harvard University
[2]Laboratory of Physics, Ecole Normale Supérieure and PSL Research
[3]Department of Physics and Center for Brain Science, Harvard University
[4]Omenn-Darling Bioengineering Institute, Princeton University
[5]Media Lab, Massachusetts Institute of Technology
⋆Equal contribution
†johnwang@g.harvard.edu, shakhnovich@chemistry.harvard.edu, esvelt@mit.edu

Abstract

Advances in protein language models (pLMs) and their integration into closed-loop wet-lab experimental platforms is unlocking powerful new capabilities in protein design. This convergence, termed Intelligent Automated Biology (IAB), enables rapid, large-scale exploration of protein function, accelerating discovery in fields from medicine to synthetic biology. Yet when applied to pathogens, these same tools pose serious dual-use risks. IAB systems can efficiently optimize immune escape, viral fitness, and other dangerous traits, even in the absence of deep biological expertise. In this position paper, we argue that the AI community must take proactive steps to address this emerging AI safety and biosecurity challenge. We introduce a framework categorizing IAB capability levels to guide risk assessment, examine IAB's unique governance challenges, and offer concrete recommendations for pLM-specific safeguard research.

1 Introduction: The New Biosecurity Frontier in AI

Artificial intelligence (AI) is rapidly reshaping biological discovery, with protein language models (pLMs)—large models trained on vast protein sequence—at the forefront[1]. These tools offer unprecedented speed and scope for understanding biological systems, predicting properties like protein fitness[2, 3, 4, 5], and even generating novel proteins entirely[6, 7]. In the recent fight against SARS-CoV-2 pandemic, pLMs help predicting viral fitness[8, 9, 10] and immune escape[11], accelerate the development of vaccines and therapeutics[12, 13, 14] and anticipate viral evolution[15, 16].

However, the true transformative power—and potential peril—emerges not from pLMs in isolation, but from their integration with active learning algorithms and wet lab platforms. **This convergence, which we term *Intelligent Automated Biology* (IAB), creates a powerful, high-throughput loop from in silico design and prediction to (potentially automated) experimental validation.** While IAB offers profound benefits for global health by accelerating biological engineering, it simultaneously introduces new dual-use risks that current biosecurity and AI security frameworks are ill-equipped to manage. Specifically, this integration leads to:

- **Dramatically accelerated exploration of protein fitness landscapes:** Active learning approaches enable efficient identification of functionally significant mutations with minimal experimental data[13, 15, 17]. In areas like pandemic research, this can reveal pathways to

NeurIPS 2025 Workshop on Biosecurity Safeguards for Generative AI

enhanced virulence, transmissibility, or immune evasion—mutational trajectories that might not arise through natural evolution.

- **Increased throughput and hit-rate:** Automated and high throughput systems can test thousands of variants rapidly[18, 19], enabling the systematic exploration of mutational combinations that would be prohibitively resource-intensive using manual laboratory method.
- **Lowered expertise barriers:** The combination of pLMs, active learning, and lab automation reduces the specialized protein knowledge required to conduct sophisticated viral engineering, potentially expanding the set of actors capable of creating enhanced pathogens.

Worryingly, the implications of this powerful IAB integration remain largely underexplored within the machine learning field driving these innovations. Our position is informed by three key facts:

1. *General-purpose pLM models are increasingly capable of supporting pandemic-scale biology.* Tools developed for protein design or sequence modeling can be readily adapted for use in predicting immune escape, enhancing viral fitness, or predicting high-risk protein variants, as demonstrated in AI-enabled pandemic tools.
2. *There is currently a lack of dedicated research on safeguards for pLMs, particularly in the context of viral protein design and prediction*, leaving dual-use risks from powerful pLMs underexplored.
3. *Leading AI venues currently lack dedicated biosecurity evaluation criteria*, let alone mechanisms to assess the specific dual-use risks emerging from the integration of models into automated wet-lab pipelines.

Our position:
Oversight and safety research is needed for Intelligent Automated Biology (IAB) to mitigate its dual-use risks, as demonstrated by AI's capabilities shown in pandemic research.

This paper argues for immediate attention to the dual-use risks inherent in IAB systems. We outline the escalating AI capabilities demonstrated in pandemic research, present a tiered risk table, and propose specific safeguard strategies directed towards the AI-Bio research community.

2 Pandemic Research Before Modern AI

The central challenge in combating viral pandemics lies in understanding how viruses evolve. Tiny changes, known as mutations, in the amino acid sequences of viral proteins can dramatically alter their behavior. For instance, mutations in the SARS-CoV-2 Spike protein's receptor-binding domain (RBD) can affect how strongly it binds to the human ACE2 receptor, influencing infectivity, or change how well antibodies recognize it, impacting immune evasion [20]. Predicting the effects of these mutations—specifically on protein stability and interactions with host cells or antibodies—is crucial for anticipating viral evolution and developing effective countermeasures. Before the advent of modern AI techniques like pLMs, researchers relied on two main computational approaches.

Approach 1: Simulating Physics One approach involved simulating the fundamental physics governing protein behavior. These physics-based methods use computational models to calculate the energy and stability of proteins based on the forces acting between their constituent atoms. Conceptually, this is like creating a detailed molecular simulation to understand how a protein folds and interacts. Tools employing force fields, such as FoldX and Rosetta [21, 22], exemplify this strategy. Some highly rigorous techniques, like Free Energy Perturbation, aim for thermodynamic accuracy in predicting how mutations change binding energy [23].

The strength of these physics-based methods lies in their potential for high accuracy and their ability to provide deep mechanistic insights—explaining why a specific mutation causes a particular effect at the molecular level. **However, their major limitation is computational time.** Accurately simulating the complex atomic interactions for large proteins requires immense processing power and time. A typical protein has $3N$ degrees of freedom where N can be thousands of atoms. The number of pairwise interactions scales as N^2. Accurate sampling of this space requires extensive computational power that scale poorly with system size.

Approach 2: Learning from Evolution A second approach leveraged evolutionary data. Known as Multiple Sequence Alignment (MSA)-based methods, these techniques compare the sequences of a specific protein collected from many different related organisms or viruses. By aligning these sequences, researchers can identify patterns: amino acid positions that rarely change are likely crucial for function, while positions that vary widely might be less critical or involved in adaptation, such as immune escape. Statistical models can then be built from these alignments to learn which combinations of mutations are commonly observed in nature (and thus likely viable) and which are rare (and likely detrimental). Models based on principles like maximum entropy or variational autoencoders were applied to viruses like SARS-CoV-2 [24, 25] and HIV [26, 27, 28, 29, 30], predicting mutational effects and escape pathways.

MSA-based methods are generally faster than physics-based simulations for evaluating mutations across a protein and directly capture constraints imposed by natural selection. **However, their effectiveness hinges entirely on the availability of a large and diverse set of related sequences—a "deep" MSA.** For newly emerging viruses, constructing a meaningful alignment is difficult. Furthermore, these methods typically require significant data curation and model retraining for each new protein family being studied. This dependence on specific evolutionary data created a bottleneck, particularly hindering rapid analysis and response efforts during the early stages of an outbreak when sequence data for a novel pathogen is scarce. This limitation highlighted the need for models capable of making predictions without relying on alignments.

3 Protein Language Models: A Leap in Prediction Capability

3.1 Backgrounds

Inspired by advances in natural language processing, pLMs are trained on large databases of unaligned natural protein sequences using self-supervised objectives. In the *autoregressive* setting, a pLM is trained to predict the next amino acid in a sequence, modeling the joint probability of a protein sequence $x = (x_1, x_2, \ldots, x_L)$ as:

$$P(x) = \prod_{i=1}^{L} P(x_i \mid x_1, \ldots, x_{i-1}), \qquad (1)$$

capturing sequential and context-dependent dependencies across residues. This setup is particularly suited for sequence generation and allows scoring of full sequences or specific mutations via log-likelihood comparisons.

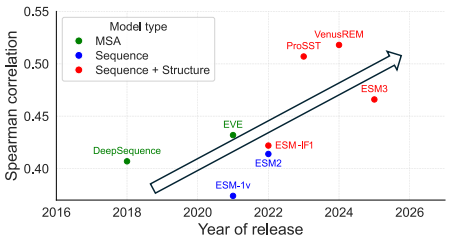

Figure 1: **Model performance improves over time.** Spearman correlation coefficients between predicted mutational effects and experimental ground truth on ProteinGym, colored by input type: MSA-based (green) [31, 32], sequence-based (blue) [33, 34], or structure + sequence based (red) [35, 36, 37].

A key advantage of pLMs is that they operate directly on raw sequence data, eliminating the need for the time-consuming and often difficult step of creating multiple sequence alignments required by earlier methods. This makes pLMs far more flexible and significantly faster to deploy, especially for novel proteins or viruses where alignment data is limited.

Furthermore, because pLMs are trained on such vast and diverse datasets, they learn highly general principles of protein biology. This allows a single, large pre-trained pLM—such as those in the widely used ESM family [33, 34]—to make meaningful predictions about virtually any protein sequence, even those belonging to protein families not seen during training. This capability is known as "zero-shot" prediction. Recent advances have further improved the performance of pLMs by explicitly incorporating structural information into the modeling process. For example, ESM-3 [38] unifies sequence and structure modeling by co-training across multiple modalities, including 3D coordinates, sequence likelihood, and masked token recovery. This joint training enables improved accuracy in predicting mutational effects and sequence plausibility within structural constraints.Additionally, some inverse folding models, like ESM-IF [36], and ProteinMPNN [39] are structure-conditioned; they can predict sequences likely to fold into a specific 3D shape, or assess how well a mutation fits within a known structure.

3.2 Models for viral protein properties prediction

Hie et al. (2021) [40] demonstrated that pLMs, when trained solely on viral sequence data without fine-tuning or structural supervision, can capture both the functional and antigenic consequences of mutations. They trained separate BiLSTM language models on corpora of aligned sequences for influenza HA, HIV Env, and SARS-CoV-2 Spike, and introduced the Constrained Semantic Change Search (CSCS) framework. In this framework, grammaticality (i.e., the model-assigned likelihood of a sequence) was hypothesized to reflect viral fitness, while semantic change (i.e., the shift in embedding space) served as a proxy for immune escape. Despite being trained only on viral sequences and without escape labels, the models successfully predicted known escape mutations in a zero-shot setting, highlighting the capacity of language models to learn biologically meaningful patterns directly from sequence data.

Building on this, Allman et al. (2024) [41] systematically benchmarked grammaticality and semantic change across multiple viral proteins using both the original LSTM-based model and newer pretrained pLMs like ESM-2. Their analysis confirmed that grammaticality scores are consistently higher for viable mutations and can serve as a practical proxy for fitness. This finding held across viral systems, including HIV and influenza. In parallel, Wang et al. [8] used ESM embeddings to predict the fitness of SARS-CoV-2 RBD variants by integrating them into a biophysical model. More broadly, other pLMs and AI models have been developed to predict key viral propertiesto predict viral properties such as binding affinity [42, 43, 44, 45, 46], mutation spread [47], and fitness [10, 5].

Collectively, these results underscore a crucial point: powerful pLMs, including those trained broadly rather than exclusively on viral data, encode meaningful information about viral protein function and evolution. This enables them to anticipate evolutionary trajectories and assess mutational effects in emerging pathogens, often with remarkable accuracy directly from sequence data.

Importantly, while these models were developed to support beneficial applications like vaccine design or pandemic forecasting, their predictive capabilities could also be misused. For instance, a model that accurately identifies mutations increasing ACE2 binding or antibody escape can just as easily be used to propose unseen variants with those mutations intentionally. Their application to targeted protein design requires specific conditioning approaches detailed in the Technical Appendix. Moreover, because many pLMs are open-weight and require minimal fine-tuning, such capabilities may be accessible even without deep virological expertise. **Notably, these tools have been used to design novel SARS-CoV-2 proteins that were experimentally shown to be infectious and capable of evading neutralization**[48, 49].

> **Our position:** It is technically concerning that open weights pLMs can accurately predict viral fitness and immune escape either zero shot or few-shot with fine-tuning. Capabilities developed for pandemic response could, without safeguards, be repurposed for misuse.

4 The Accelerator Effect: Integrating AI with Lab Experiments

pLMs are not just predictive tools; they are increasingly integrated into active protein engineering workflows, dramatically accelerating the pace and changing the nature of biological design. This integration manifests in several key applications.

4.1 Smarter Directed Evolution

Directed evolution is a laboratory technique that mimics natural selection to improve proteins for specific purposes, such as improving the efficiency of enzymes, increasing binding affinity of therapeutic antibodies [13]. Traditionally, this involves creating large libraries of protein variants and screening them for desired properties, often a laborious, inefficient, and expensive process. **pLMs enables the direct evolution of novel proteins with significantly improved functional properties.** By predicting the likely effects of mutations by either zero shot or few shot, pLMs can guide researchers to focus on variants with a higher probability of success, effectively narrowing down the search space and reducing the experimental burden. Recent studies have demonstrated that general and structure-informed pLMs can substantially improve the binding affinity and neutralization breadth

of human antibodies against diverse viral targets, including SARS-CoV-2, Ebola, and influenza, while requiring only minimal rounds of experimental screening [12, 14, 50].

4.2 Laboratory Automation and Closed-Loop Experimentation

The impact of pLMs is amplified when combined with laboratory automation, often referred to as "biofoundries"[51, 52]. This integration enables fully automated cycles of biological design, construction, testing, and learning, commonly known as the Design–Build–Test–Learn (DBTL) cycle. The DBTL cycle includes: (1) Design: AI/pLMs propose sequences with predicted properties; (2) Build: Robotic systems synthesize DNA and produce variants; (3) Test: Automated assays measure properties; (4) Learn: Results feed back to AI for improved designs in subsequent cycles.

Platforms like PLMeAE[18] demonstrate the power of this approach, achieving multiple rounds of enzyme optimization in just 10 days—a task that could take many months using traditional methods [18]. This creates a powerful, high-speed, closed loop for biological engineering. While offering tremendous potential for accelerating therapeutic development, this automation also raises concerns. The speed and reduced human intervention inherent in these closed loops could potentially allow for the rapid optimization of harmful properties if misused, with fewer opportunities for oversight or ethical review during the process.

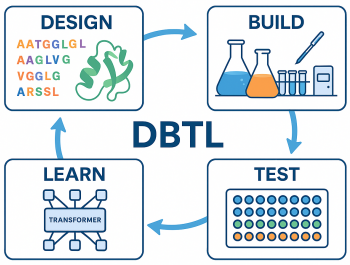

Figure 2: **Schematic of the DBTL cycle in AI-enabled bioengineering.** pLMs propose novel sequences (Design), which are synthesized and expressed by robotic platforms (Build), evaluated through high-throughput assays (Test), and iteratively improved based on experimental feedback (Learn).

4.3 Efficient Exploration with Active Learning

The sheer number of possible mutations, even within a single protein, makes exhaustive experimental or computational testing infeasible. Active learning offers a solution by integrating model predictions with experimental design[17, 53]. Instead of testing randomly or relying solely on initial predictions, active learning uses the predictive models to select the most informative experiments to perform at each stage, based on certain acquisition function[54].

The typical process starts with wet-lab testing a small, initial set of variants. The results are used to train or fine-tune a predictive model (like a pLM)[55]. The model then evaluates the vast pool of untested variants and identifies those whose experimental evaluation would maximally improve the model's accuracy or are most likely to possess the desired properties (e.g., high fitness, activity, or strong binding). These selected variants are then synthesized and tested, and the new data is used to update the model, repeating the cycle. This iterative strategy dramatically reduces the number of experiments required to explore the mutational landscape and identify top-performing or high-risk variants. Active learning has already shown success in domains like drug discovery[56, 57, 58, 59].

Recent studies have shown that active learning frameworks can optimize enzymes, antibodies, or other protein variants, antibody or protein variants significantly faster than random screening, using only a small fraction of what traditional method required[13, 17]. This efficiency can also enable researchers to proactively identify concerning viral mutations before they potentially emerge naturally[15].

The synergy between pLMs (for design and prediction), active learning (for efficient experimental guidance), and laboratory automation (for rapid build and test cycles) creates an engineering capability far greater than the sum of its parts. This integrated approach enables systematic biological exploration and optimization at an unprecedented speed and scale. While this accelerates beneficial research, it simultaneously increases the risk of malicious biological engineering and potentially reduces human oversight within automated loops.

Our position: pLMs, when integrated with active learning and laboratory automation, form a closed-loop AI-bioengineering stack that introduces an unprecedented class of dual-use biosecurity risks. These risks arise not from any one capability, but from their systemic integration, which could accelerates viral evolution modeling.

5 The Dual-Use Dilemma: Assessing Risks of IAB

The core challenge presented by the convergence of AI and biotechnology lies in its inherent dual-use nature: technologies developed with beneficial intent, such as improving human health or combating pandemics, can often be repurposed to cause harm. pLM significantly amplifies this dilemma by accelerating design cycles, lowering knowledge barriers, and enabling automation at unprecedented scales. To effectively discuss and manage these risks, it is helpful to categorize the capabilities enabled by IAB and assess their associated risk levels.

We propose a framework categorizing IAB capabilities into five levels, reflecting escalating potential for misuse as pLM integration deepens (Table 1). This framework builds upon initial concepts and incorporates insights from recent literature on AI capabilities and biosecurity risks.

Table 1: IAB Capability Levels and Associated Biosecurity Risk

Capability Level	Description	Examples	**Base Risk Level**
Level 1: Zero-shot Prediction	basic pLM predictions (e.g., sequence likelihood as fitness proxy).	ESM-1v zero-shot prediction with grammaticality score [3, 41].	**Low – Moderate**
Level 2: Advanced Prediction & Analysis	Accurate ML/pLM prediction of complex molecular properties (e.g., binding affinity changes ($\Delta\Delta G$), immune escape potential, stability).	Fine-tuned ESM3 to predict viral fitness; UniBind[11] predicting binding affinity; EVEscape[60] and VIRAL[15] predicting escape variants; MMSite for active site prediction[61]	**Moderate**
Level 3: Targeted Sequence Generation	Generative AI/pLMs designing novel sequences optimized for specific functional properties (e.g., enhanced binding, stability, potentially virulence factors or toxins).	ProteinMPNN[39] or ESM-IF1 [7] for generative enzyme/antibody design; Potential toxin/-pathogen design.	**High**
Level 4: Integrated Design & Active Learning	Combining generative models with active learning/Bayesian optimization for efficient discovery and optimization of desired (potentially harmful) biological functions.	ProteinNPT[62] for Active learning frameworks; EVOLVEpro[13] and ALDE[17] for direct evolution;	**Very High**
Level 5: Full AI-Bio Automation Integration	Closed-loop systems linking AI protein design, learning, synthesis, and testing (DBTL cycle) with minimal human oversight	PLMeAE[18] or iBioFAB [19] where pLMs are embedded in automated biofoundries	**Extremely High**

This table illustrates that the most significant risks emerge not merely from individual AI capabilities but from the DBTL cycle coupled with physical automation. Level 5 enabling rapid, automated, and potentially remote execution of complex bioengineering tasks[63], maximizing both the potential for benefit and the potential for misuse. For each level we classified, concrete examples are provided—and concerningly, full AI-biology automation integration at Level 5 has already been observed in 2025.

To better quantify the acceleration enabled by this integration, we estimated the speed (see Appendix) to obtain a functional variant ("hit") using wet-lab hit rates on an 85-amino-acid peptide [64]. Hit rate is defined as the fraction of tested sequences that exhibit the desired function. Combining these hit rates with representative experimental throughput values, we find that AI-guided, automated pipelines (Level 5) can yield thousands of hits per day—several orders of magnitude more than traditional, manual, non-AI-guided approaches (Figure 3). This illustrates how full-stack automation not only increases capability but compresses timelines, potentially outpacing the safety checks traditionally used to govern wet-lab experimentation.

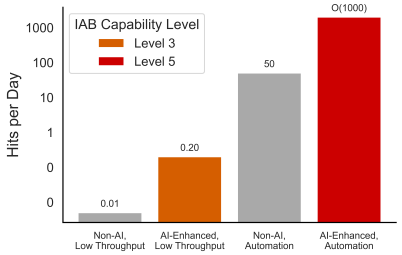

Figure 3: **Functional protein "hits" per day** from AI vs non-AI methods under low- and high-throughput settings. Based on hit rate × throughput.

A critical factor contributing to this assessment difficulty is the "evaluation bottleneck" [65]. AI-Bio models at Capability Level 3 and above can generate novel protein sequences, but accurately predicting their real-world function—especially their potential harmfulness—remains an open challenge. Definitive functional validation often requires synthesizing the DNA and expressing the protein in a wet lab.

However, if the AI-designed entity possesses hazardous properties, this evaluation step becomes inherently dangerous. This stands in contrast to evaluating text generated by large language models (LLMs) in the medical or virology domain, where outputs remain directly interpretable by humans and standardized benchmarks exist to assess risks [66, 67]. The inability to safely and reliably assess the biological function of IAB outputs poses a fundamental obstacle to timely risk detection and mitigation. Without robust, trustworthy pLM risk evaluation tools and benchmarks, we risk not knowing the true danger posed by a new IAB or a specific protein design until it has been physically instantiated—potentially too late to prevent harm.

6 Safeguard the Frontier: IAB Needs Tailored Oversight

Addressing the dual-use risks of IAB requires robust governance frameworks. Insights can be drawn from existing approaches in both AI safety (primarily developed for LLMs) and traditional biosecurity, but both have significant limitations when applied to the unique challenges of AI-enabled biotechnology, especially IAB. While this paper recognizes the tension between open science principles and AI-safety and biosecurity concerns, our primary focus is on identifying the governance gaps specific to IAB systems rather than resolving the broader debate around open model release.

Techniques developed for aligning LLMs with human values and preventing harmful outputs include Reinforcement Learning from Human Feedback (RLHF), Constitutional AI (CAI), and red teaming. RLHF fine-tunes models based on human preferences for different outputs [68, 69]. CAI extends this by training models to adhere to an explicit set of principles (a "constitution") by having an AI critique and revise outputs based on those principles[70]. Red teaming involves adversarial probing by experts to identify vulnerabilities and elicit harmful behavior[71].

However, directly translating these methods to pLMs and bio-AI faces difficulties. The core challenge lies in the evaluation bottleneck discussed previously:

- RLHF/CAI Applicability: Defining appropriate human preferences or constitutional principles for complex biological functions is difficult. What constitutes a "safe" or "harmless" protein design requires deep expertise. **To date, no models exist that can reliably evaluate the safety of outputs generated by pLMs and subsequent tasks.** Consequently, verifying the properties of these generated biological outputs demands experimental validation in a wet lab, which is often slow, costly, and potentially hazardous.
- Red Teaming Risks: While red teaming can identify potential misuse pathways for AI-Bio tools, the process itself carries risks. Eliciting a dangerous protein design during red teaming could inadvertently create or disseminate hazardous information[72]. Effective red teaming requires significant biological expertise, and scaling these evaluations is challenging[73].

Our position: Essentially, the lack of a safe, scalable, and reliable method to evaluate the real-world function and harm potential of AI-generated biological outputs hinders the direct application of established LLM alignment techniques. This *bio-evaluation gap* is a central technical obstacle.

On the biosecurity side, traditional biosecurity measures fall short in addressing AI-specific risks. Oversight frameworks like the U.S. Government Policy on Dual Use Research of Concern (DURC) [74] were designed to address a narrower class of threats: specific pathogens and well-defined experimental manipulations (e.g., increasing virulence or transmissibility). The DURC policy applies only to 15 listed agents and a fixed set of seven experimental categories, with no provisions for risks stemming from powerful general-purpose tools like pLMs. As such, it does not account for the dual-use potential of IAB[75].

One of the most widely used approaches in biosecurity—**DNA synthesis screening** [76, 77] aims to prevent the acquisition of matches to regulated pathogens or known hazardous sequences[78].

However, a recent MIT experiment revealed that it was alarmingly easy to purchase synthetic DNA fragments capable of reconstructing the deadly 1918 influenza virus—93% of U.S. providers and 100% of international providers fulfilled the order [79]. Also, generative models can design entirely novel protein sequences [39] or potentially generate sequences designed to evade detection[80, 81].

On the training methodology side, no established safeguard frameworks exist for pLMs. To address this gap, we explore early-stage technical approaches—adapted from the LLM safety literature—that may help reduce the risk of generating dangerous biological sequences. Broadly, these approaches can be categorized into **training-time guardrails**, which modify the model's learning process to discourage the generation of harmful content; and **inference-time guardrails**, which filter or steer model outputs at the point of generation. One fundamental training-time strategy is *likelihood suppression*, which aims to discourage the model from assigning high probability to harmful sequences (Figure 4). This can be formalized by modifying the training objective to penalize the likelihood of pathogenic sequences:

$$\mathcal{L} = \mathcal{L}_{\text{original}} - \lambda \log P(\text{pathogenic}) \quad (2)$$

where $\mathcal{L}$ represents the likelihood of any sequence and λ controls the strength of the suppression [82]. A more adaptive approach to implementing such training-time penalization, or more broadly steering the model towards safer outputs during training, is Reinforcement Learning from Human Feedback (RLHF) [68, 69]. While no end-to-end implementation of RLHF for pLM safety has been empirically demonstrated, we sketch a conceptual mapping here as a foundation for crucial future research and development in this area. In this context, the pLM acts as a policy generating sequences, while a separate reward model (RM)—potentially trained on datasets of viral protein sequences, structures, and functions—evaluates their potential harmfulness. The pLM can then be fine-tuned using RL algorithms like Proximal Policy Optimization (PPO) [83] to minimize the generation of dangerous sequences. This approach represents an advanced method for instilling safety considerations during the model training phase. Recent work has demonstrated the feasibility of using RL techniques on pLMs for preference optimizations and fine-tuning [84, 85, 86, 87, 88], suggesting these methods could be adapted for safety purposes. Developing RM for pLM safety could face difficulties, including precisely defining the harmfulness score and obtaining sufficient labeled protein data for it. RLHF for pLMs can inherit issues from LLMs such as reward hacking (See Appendix). For a detailed comparison and further discussions between RLHF in LLMs and pLMs, see Table 5 in Appendix.

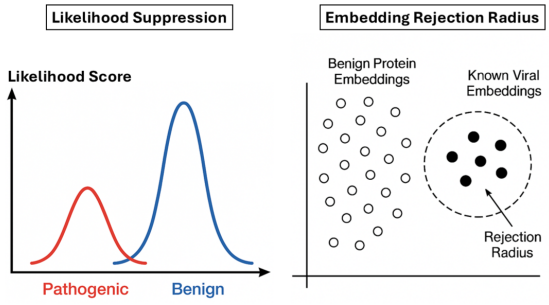

Figure 4: **Illustration of examples of training-time and inference-time guardrails for pLMs.** Likelihood suppression during training time [82] assigns low probability to pathogenic sequences, while an embedding-space rejection radius [89] blocks generation of sequences too close to known harmful proteins in inference time.

Alternatively, safeguards can be implemented as inference-time guardrails. These methods typically do not alter the underlying model weights but instead apply checks, filters, or steering mechanisms during or after the generation process. This can involve pre-generation constraint conditioning, where generation is guided away from risky regions of the sequence space using techniques like control tokens or latent variable manipulation . A specific example of an inference-time filter is the embedding-space rejection radius [89](Figure 4).This method blocks the output of generated sequences whose embeddings are found to be too close to those of known harmful proteins. During inference, a generated sequence's embedding would be compared against a curated database of harmful protein embeddings (e.g., using cosine similarity or Euclidean distance). If a sequence falls within a predefined rejection radius of a known harmful protein, its output is blocked or flagged. Both training-time and inference-time approaches could be combined with adversarial training, where red teams attempt to evade these protections, and the model is iteratively refined to close identified vulnerabilities.

Developing robust and generalizable safeguards, however, will also require standardized benchmarks to evaluate model capabilities in high-risk domains such as viral fitness prediction. To support this,

we propose a zero-shot benchmark example (Table 2) built from publicly available viral mutational scanning datasets, which quantify fitness across thousands of viral protein variants. These could enable assessments of whether a pLM can predict viral properties, offering an empirical basis to evaluate dual-use risk, particularly for open-weight models. We acknowledge that the development of such benchmarks might be prone to being misused for designing new viruses; therefore, efforts are needed to widen the evaluation-genration gap–that is, making it harder to generate harmful viruses but easier to detect them. Furthermore, future work should expand on this foundation to develop a more comprehensive dataset.

Table 2: Example of a viral fitness dataset for benchmarking pLM viral capabilities

Virus	Protein	Fitness Proxy	**# Variants**
SARS-CoV-2	Spike RBD	Expression score via yeast display [90]	∼**3,800**
		Binding affinity to ACE2 [91]	∼**33,000**
Influenza A	Hemagglutinin (HA)	Replication efficiency [92]	∼**10,000**
HIV-1	Envelope glycoprotein (Env)	Replication efficiency [93]	∼**13,000**

7 Conclusion and Recommendations: A Call for Responsible Innovation

Integrating AI, particularly pLMs, with automated experimental biology platforms marks a significant technological leap. However, the very power that makes IAB revolutionary also introduces dual-use AI safety and biosecurity risks. As detailed, IAB systems can rapidly explore complex biological landscapes, optimize functions like viral fitness or immune evasion, lower bioengineering expertise needs, and potentially operate with less oversight via closed-loop automation. While these capabilities are invaluable for tasks like pandemic preparedness, they could equally be misused to design or enhance dangerous pathogens, potentially accelerating the emergence of future pandemics.

Current safety frameworks, whether drawn from AI safety (primarily focused on text-based LLMs) or traditional biosecurity (often centered on known pathogens), falls short to manage the unique challenges posed by IAB. Key difficulties include the "evaluation bottleneck"—the inability to safely and reliably assess the real-world function and potential harm of AI-generated biological entities without risky wet-lab synthesis—and the capacity of AI to design entirely novel sequences that may evade existing detection methods. Therefore, proactive oversight and community engagement are essential. We outline targeted actions for different communities involved in the development and oversight of integrated IAB systems, grouped by stakeholder to enhance actionability

1. **For Researchers in Academia and Industry**
   (a) **Develop and Prioritize AI-Bio Capability Evaluation and Safeguards:** Standardized benchmarks and metrics should be developed to assess the potential risks of AI-generated protein sequences and structures. This includes few-shot or zero-shot benchmarks evaluating properties such as enhanced virulence, as illustrated in Table 2. In parallel, the AI community could consider open-source safeguarded pLM variants with architectural constraints built in, reducing misuse risk.
2. **For Scientific Conferences and the Broader Research Community**
   (a) **Integrate Biosecurity into Peer Review and Evaluation:** Conferences like NeurIPS should introduce a formal biosecurity checklist[94] for submissions describing potentially high-risk IAB capabilities (e.g., generative pLMs or automated design tools). Reviewer guidelines should include criteria for evaluating whether authors adequately assess and mitigate potential dual-use risks. Submissions flagged for significant biosecurity or AI–biology risks should undergo a dedicated ethics review by qualified AI-Bio security experts.

Unlike LLMs, whose outputs are text, the outputs of pLMs and AI-Bio models can be synthesized into real biological agents. Thus, risk must be assessed holistically, accounting for how models are embedded in subsequent experimental platforms. This integration demands cross-disciplinary safety research efforts, oversight and the expansion of ML peer review norms to reflect their real-world effects.

References

[1] Yijia Xiao, Wanjia Zhao, Junkai Zhang, Yiqiao Jin, Han Zhang, Zhicheng Ren, Renliang Sun, Haixin Wang, Guancheng Wan, Pan Lu, et al. Protein large language models: A comprehensive survey. *arXiv preprint arXiv:2502.17504*, 2025.

[2] Lin Chen, Zehong Zhang, Zhenghao Li, Rui Li, Ruifeng Huo, Lifan Chen, Dingyan Wang, Xiaomin Luo, Kaixian Chen, Cangsong Liao, et al. Learning protein fitness landscapes with deep mutational scanning data from multiple sources. *Cell Systems*, 14(8):706–721, 2023.

[3] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

[4] Luiz C Vieira, Morgan L Handojo, and Claus O Wilke. Scaling down for efficiency: Medium-sized protein language models perform well at transfer learning on realistic datasets. *bioRxiv*, pages 2024–11, 2024.

[5] Zuobai Zhang, Pascal Notin, Yining Huang, Aurelie C Lozano, Vijil Chenthamarakshan, Debora Marks, Payel Das, and Jian Tang. Multi-scale representation learning for protein fitness prediction. *Advances in Neural Information Processing Systems*, 37:101456–101473, 2024.

[6] Jeffrey A Ruffolo and Ali Madani. Designing proteins with language models. *nature biotechnology*, 42(2):200–202, 2024.

[7] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8946–8970. PMLR, 17–23 Jul 2022.

[8] Dianzhuo Wang, Marian Huot, Vaibhav Mohanty, and Eugene I. Shakhnovich. Biophysical principles predict fitness of sars-cov-2 variants. *Proceedings of the National Academy of Sciences*, 121(23):e2314518121, 2024.

[9] Yuanxi Yu, Fan Jiang, Bozitao Zhong, Liang Hong, and Mingchen Li. Entropy-driven zero-shot deep learning model selection for viral proteins. *Physical Review Research*, 7(1):013229, 2025.

[10] Jumpei Ito, Adam Strange, Wei Liu, Gustav Joas, Spyros Lytras, The Genotype to Phenotype Japan (G2P-Japan) Consortium, and Kei Sato. A Protein Language Model for Exploring Viral Fitness Landscapes, March 2024.

[11] Guangyu Wang, Xiaohong Liu, Kai Wang, Yuanxu Gao, Gen Li, Daniel T Baptista-Hon, Xiaohong Helena Yang, Kanmin Xue, Wa Hou Tai, Zeyu Jiang, et al. Deep-learning-enabled protein–protein interaction analysis for prediction of sars-cov-2 infectivity and variant evolution. *Nature Medicine*, 29(8):2007–2018, 2023.

[12] Brian L. Hie, Varun R. Shanker, Duo Xu, Theodora U. J. Bruun, Payton A. Weidenbacher, Shaogeng Tang, Wesley Wu, John E. Pak, and Peter S. Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42(2):275–283, February 2024. Publisher: Springer Science and Business Media LLC.

[13] Kaiyi Jiang, Zhaoqing Yan, Matteo Di Bernardo, Samantha R Sgrizzi, Lukas Villiger, Alisan Kayabolen, BJ Kim, Josephine K Carscadden, Masahiro Hiraizumi, Hiroshi Nishimasu, et al. Rapid in silico directed evolution by a protein language model with evolvepro. *Science*, page eadr6006, 2024.

[14] Varun R. Shanker, Theodora U. J. Bruun, Brian L. Hie, and Peter S. Kim. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science*, 385(6704):46–53, July 2024. Publisher: American Association for the Advancement of Science (AAAS).

[15] Marian Huot, Dianzhuo Wang, Jiacheng Liu, and Eugene I. Shakhnovich. Predicting high-fitness viral protein variants with Bayesian active learning and biophysics. *Proceedings of the National Academy of Sciences*, 122(24):e2503742122, June 2025.

[16] Marian Huot, Dianzhuo Wang, Eugene Shakhnovich, Remi Monasson, and Simona Cocco. Constrained Evolutionary Funnels Shape Viral Immune Escape, October 2025.

[17] Jason Yang, Ravi G Lal, James C Bowden, Raul Astudillo, Mikhail A Hameedi, Sukhvinder Kaur, Matthew Hill, Yisong Yue, and Frances H Arnold. Active learning-assisted directed evolution. *Nature Communications*, 16(1):714, 2025.

[18] Qiang Zhang, Wanyi Chen, Ming Qin, Yuhao Wang, Zhongji Pu, Keyan Ding, Yuyue Liu, Qunfeng Zhang, Dongfang Li, Xinjia Li, et al. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nature Communications*, 16(1):1553, 2025.

[19] Tianhao Yu, Aashutosh Girish Boob, Nilmani Singh, Yufeng Su, and Huimin Zhao. In vitro continuous protein evolution empowered by machine learning and automation. *Cell Systems*, 14(8):633–644, 2023.

[20] Michael Letko, Andrea Marzi, and Vincent Munster. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nature Microbiology*, 5(4):562–569, February 2020.

[21] Javier Delgado, Leandro G Radusky, Damiano Cianferoni, and Luis Serrano. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*, 35(20):4168–4169, 03 2019.

[22] Kyle A. Barlow, Shane Ó Conchúir, Samuel Thompson, Pooja Suresh, James E. Lucas, Markus Heinonen, and Tanja Kortemme. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *The Journal of Physical Chemistry. B*, 122(21):5389–5399, May 2018.

[23] Jared M. Sampson, Daniel A. Cannon, Jianxin Duan, Jordan C. K. Epstein, Alina P. Sergeeva, Phinikoula S. Katsamba, Seetha M. Mannepalli, Fabiana A. Bahna, Hélène Adihou, Stéphanie M. Guéret, Ranganath Gopalakrishnan, Stefan Geschwindner, D. Gareth Rees, Anna Sigurdardottir, Trevor Wilkinson, Roger B. Dodd, Leonardo De Maria, Juan Carlos Mobarec, Lawrence Shapiro, Barry Honig, Andrew Buchanan, Richard A. Friesner, and Lingle Wang. Robust Prediction of Relative Binding Energies for Protein–Protein Complex Mutations Using Free Energy Perturbation Calculations. *Journal of Molecular Biology*, 436(16):168640, August 2024.

[24] Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proceedings of the National Academy of Sciences*, 119(4):e2113118119, January 2022.

[25] Nicole N. Thadani, Sarah Gurev, Pascal Notin, Noor Youssef, Nathan J. Rollins, Daniel Ritter, Chris Sander, Yarin Gal, and Debora S. Marks. Learning from prepandemic data to forecast viral escape. *Nature*, 622(7984):818–825, October 2023.

[26] William F. Flynn, Allan Haldane, Bruce E. Torbett, and Ronald M. Levy. Inference of Epistatic Effects Leading to Entrenchment and Drug Resistance in HIV-1 Protease. *Molecular Biology and Evolution*, 34(6):1291–1306, June 2017.

[27] Avik Biswas, Allan Haldane, Eddy Arnold, and Ronald M Levy. Epistasis and entrenchment of drug resistance in HIV-1 subtype B. *eLife*, 8:e50524, October 2019.

[28] Thomas C. Butler, John P. Barton, Mehran Kardar, and Arup K. Chakraborty. Identification of drug resistance mutations in HIV from constraints on natural evolution. *Physical Review E*, 93(2):022412, February 2016.

[29] Raymond H. Y. Louie, Kevin J. Kaczorowski, John P. Barton, Arup K. Chakraborty, and Matthew R. McKay. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proceedings of the National Academy of Sciences*, 115(4), January 2018.

[30] Karthik Shekhar, Claire F. Ruberman, Andrew L. Ferguson, John P. Barton, Mehran Kardar, and Arup K. Chakraborty. Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. *Physical Review E*, 88(6):062705, December 2013.

[31] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, October 2018.

[32] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, November 2021.

[33] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function, July 2021.

[34] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model, July 2022.

[35] Mingchen Li, Pan Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Liang Hong, and Yang Tan. Prosst: Protein language modeling with quantized structure and disentangled attention. *bioRxiv*, 2024.

[36] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures, April 2022.

[37] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model, July 2024. Pages: 2024.07.01.600583 Section: New Results.

[38] Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pages 2024–07, 2024.

[39] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[40] Brian Hie, Ellen D. Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, January 2021.

[41] Brent Allman, Luiz Vieira, Daniel J Diaz, and Claus O Wilke. A systematic evaluation of the language-of-viral-escape model using multiple machine learning frameworks. *bioRxiv*, pages 2024–09, 2024.

[42] Guangyu Wang, Xiaohong Liu, Kai Wang, Yuanxu Gao, Gen Li, Daniel T Baptista-Hon, Xiaohong Helena Yang, Kanmin Xue, Wa Hou Tai, Zeyu Jiang, et al. Deep-learning-enabled protein–protein interaction analysis for prediction of sars-cov-2 infectivity and variant evolution. *Nature Medicine*, 29(8):2007–2018, 2023.

[43] Thomas Loux, Dianzhuo Wang, and Eugene I Shakhnovich. More structure, less accuracy: Esm3's binding prediction paradox. *bioRxiv*, pages 2024–12, 2024.

[44] Joseph M Taft, Cédric R Weber, Beichen Gao, Roy A Ehling, Jiami Han, Lester Frei, Sean W Metcalfe, Max D Overath, Alexander Yermanos, William Kelton, et al. Deep mutational learning predicts ace2 binding and antibody escape to combinatorial mutations in the sars-cov-2 receptor-binding domain. *Cell*, 185(21):4008–4022, 2022.

[45] Maxime Basse, Dianzhuo Wang, and Eugene I. Shakhnovich. Spatial clustering of interface residues enhances few-shot prediction of viral protein binding. *bioRxiv*, 2025.

[46] Shiwei Liu, Tian Zhu, Milong Ren, Chungong Yu, Dongbo Bu, and Haicang Zhang. Predicting mutational effects on protein-protein binding via a side-chain diffusion probabilistic model. *Advances in Neural Information Processing Systems*, 36:48994–49005, 2023.

[47] M. Cyrus Maher, Istvan Bartha, Steven Weaver, Julia Di Iulio, Elena Ferri, Leah Soriaga, Florian A. Lempp, Brian L. Hie, Bryan Bryson, Bonnie Berger, David L. Robertson, Gyorgy Snell, Davide Corti, Herbert W. Virgin, Sergei L. Kosakovsky Pond, and Amalio Telenti. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Science Translational Medicine*, 14(633):eabk3445, February 2022.

[48] Noor Youssef, Sarah Gurev, Fadi Ghantous, Kelly P Brock, Javier A Jaimes, Nicole N Thadani, Ann Dauphin, Amy C Sherman, Leonid Yurkovetskiy, Daria Soto, et al. Computationally designed proteins mimic antibody immune evasion in viral evolution. *Immunity*, 2025.

[49] Marian Huot, Pierre Rosenbaum, Cyril Planchais, Hugo Mouquet, Remi Monasson, and Simona Cocco. Generative model of sars-cov-2 variants under functional and immune pressure unveils viral escape potential and antibody resilience. *bioRxiv*, 2025.

[50] Sisi Shan, Shitong Luo, Ziqing Yang, Junxian Hong, Yufeng Su, Fan Ding, Lili Fu, Chenyu Li, Peng Chen, Jianzhu Ma, et al. Deep learning guided optimization of human antibody against sars-cov-2 variants with broad neutralization. *Proceedings of the National Academy of Sciences*, 119(11):e2122954119, 2022.

[51] Nathan Hillson, Mark Caddick, Yizhi Cai, Jose A Carrasco, Matthew Wook Chang, Natalie C Curach, David J Bell, Rosalind Le Feuvre, Douglas C Friedman, Xiongfei Fu, et al. Building a global alliance of biofoundries. *Nature communications*, 10(1):2040, 2019.

[52] Mario A Torres-Acosta, Gary J Lye, and Duygu Dikicioglu. Automated liquid-handling operations for robust, resilient, and efficient bio-based laboratory practices. *Biochemical Engineering Journal*, 188:108713, 2022.

[53] Daria Balashova, Robert Frank, Svetlana Kuzyakina, Dominique Weltevreden, Philippe A Robert, Geir Kjetil Sandve, and Victor Greiff. Active learning for improving out-of-distribution lab-in-the-loop experimental design. *bioRxiv*, pages 2025–02, 2025.

[54] Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active learning by acquiring contrastive examples. *arXiv preprint arXiv:2109.03764*, 2021.

[55] Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. biorxiv. *preprint*, 202310(2023.12):13–571462, 2023.

[56] Zachary Fralish and Daniel Reker. Taking a deep dive with active learning for drug discovery. *Nature Computational Science*, pages 1–2, 2024.

[57] David E Graff, Eugene I Shakhnovich, and Connor W Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical science*, 12(22):7866–7881, 2021.

[58] Manfred KK Warmuth, Gunnar Rätsch, Michael Mathieson, Jun Liao, and Christian Lemmen. Active learning in the drug discovery process. *Advances in Neural information processing systems*, 14, 2001.

[59] Michael Bailey, Saeed Moayedpour, Ruijiang Li, Alejandro Corrochano-Navarro, Alexander Kötter, Lorenzo Kogler-Anele, Saleh Riahi, Christoph Grebner, Gerhard Hessler, Hans Matter, et al. Deep batch active learning for drug discovery. *bioRxiv*, pages 2023–07, 2023.

[60] Nicole N Thadani, Sarah Gurev, Pascal Notin, Noor Youssef, Nathan J Rollins, Daniel Ritter, Chris Sander, Yarin Gal, and Debora S Marks. Learning from prepandemic data to forecast viral escape. *Nature*, 622(7984):818–825, 2023.

[61] Song Ouyang, Huiyu Cai, Yong Luo, Kehua Su, Lefei Zhang, and Bo Du. Mmsite: A multi-modal framework for the identification of active sites in proteins. *Advances in Neural Information Processing Systems*, 37:45819–45849, 2024.

[62] Pascal Notin, Ruben Weitzman, Debora Marks, and Yarin Gal. Proteinnpt: Improving protein property prediction and design with non-parametric transformers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 33529–33563. Curran Associates, Inc., 2023.

[63] Emerald cloud lab. `https://www.emeraldcloudlab.com/`. Accessed: 2025-04-30.

[64] M. Zaki Jawaid, Robin W. Yeo, Aayushma Gautam, T. Blair Gainous, Daniel O. Hart, and Timothy P. Daley. Improving few-shot learning-based protein engineering with evolutionary sampling. *bioRxiv*, 2023.

[65] Jaspreet Pannu, Doni Bloomfield, Robert MacKnight, Moritz S. Hanke, Alex Zhu, Gabe Gomes, Anita Cicero, and Thomas V. Inglesby. Dual-use capabilities of concern of biological AI models. *PLOS Computational Biology*, 21(5):e1012975, May 2025.

[66] Sijia Chen, Xiaomin Li, Mengxue Zhang, Eric Hanchen Jiang, Qingcheng Zeng, and Chen-Hsiang Yu. Cares: Comprehensive evaluation of safety and adversarial robustness in medical llms. *arXiv preprint arXiv:2505.11413*, 2025.

[67] Jasper Gotting et al. Virology capabilities test (vct): A multimodal virology q&a benchmark. *arXiv e-prints*, pages arXiv–2504, 2025.

[68] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[69] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[70] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1395–1417, 2024.

[71] Jiyan He, Weitao Feng, Yaosen Min, Jingwei Yi, Kunsheng Tang, Shuai Li, Jie Zhang, Kejiang Chen, Wenbo Zhou, Xing Xie, et al. Control risk for potential misuse of artificial intelligence in science. *arXiv preprint arXiv:2312.06632*, 2023.

[72] Mengdi Wang, Zaixi Zhang, Amrit Singh Bedi, Alvaro Velasquez, Stephanie Guerra, Sheng Lin-Gibson, Le Cong, Yuanhao Qu, Souradip Chakraborty, Megan Blewett, Jian Ma, Eric Xing, and George Church. A call for built-in biosecurity safeguards for generative ai tools. *Nature Biotechnology*, 2025.

[73] Nicole E Wheeler. Responsible ai in biotechnology: balancing discovery, innovation and biosecurity risks. *Frontiers in Bioengineering and Biotechnology*, 13:1537471, 2025.

[74] U.S. Department of Health and Human Services, Administration for Strategic Preparedness and Response (ASPR). Biosecurity. `https://aspr.hhs.gov/S3/Pages/Biosecurity.aspx`, 2025. Accessed 28 Apr 2025.

[75] Trond Arne Undheim. The whack-a-mole governance challenge for ai-enabled synthetic biology: literature review and emerging frameworks. *Frontiers in Bioengineering and Biotechnology*, 12:1359768, 2024.

[76] Securedna: Free, secure dna synthesis screening platform. `https://securedna.org`, 2025. Accessed 28 Apr 2025.

[77] Carsten Baum, Jens Berlips, Walther Chen, Hongrui Cui, Ivan Damgard, Jiangbin Dong, Kevin M Esvelt, Leonard Foner, Mingyu Gao, Dana Gretton, et al. A system capable of verifiably and privately screening global dna synthesis. *arXiv preprint arXiv:2403.14023*, 2024.

[78] Diane DiEuliis, Sarah R Carter, and Gigi Kwik Gronvall. Options for synthetic dna order screening, revisited. *MSphere*, 2(4):10–1128, 2017.

[79] Matt Field. Mit researchers ordered and combined parts of the 1918 pandemic influenza virus. did they expose a security flaw? *Bulletin of the Atomic Scientists*, June 2024. Accessed: 2025-05-07.

[80] Ning Lu, Shengcai Liu, Rui He, Qi Wang, Yew-Soon Ong, and Ke Tang. Large language models can be guided to evade ai-generated text detection. *arXiv preprint arXiv:2305.10847*, 2023.

[81] Bruce J Wittmann, Tessa Alexanian, Craig Bartling, Jacob Beal, Adam Clore, James Diggans, Kevin Flyangolts, Bryan T Gemler, Tom Mitchell, Steven T Murphy, et al. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390(6768):82–87, 2025.

[82] Ching-Yun Ko, Pin-Yu Chen, Payel Das, Youssef Mroueh, Soham Dan, Georgios Kollias, Subhajit Chaudhury, Tejaswini Pedapati, and Luca Daniel. Large language models can be strong self-detoxifiers, 2024.

[83] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[84] Mostafa Karimi, Sharmi Banerjee, Tommi Jaakkola, Bella Dubrov, Shang Shang, and Ron Benson. Extrapolative protein design through triplet-based preference learning. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[85] Filippo Stocco, Maria Artigues-Lleixa, Andrea Hunklinger, Talal Widatalla, Marc Guell, and Noelia Ferruz. Guiding generative protein language models with reinforcement learning. *arXiv preprint arXiv:2412.12979*, 2024.

[86] Pouria Mistani and Venkatesh Mysore. Preference optimization of protein language models as a multi-objective binder design paradigm. *arXiv preprint arXiv:2403.04187*, 2024.

[87] Xiangyu Liu, Yi Liu, Silei Chen, and Wei Hu. Controllable protein sequence generation with llm preference optimization. *arXiv preprint arXiv:2501.15007*, 2025.

[88] Nathaniel Blalock, Srinath Seshadri, Agrim Babbar, Sarah A Fahlberg, Ameya Kulkarni, and Philip A Romero. Functional alignment of protein language models via reinforcement learning. *bioRxiv*, 2025.

[89] Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails, 2023.

[90] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H.D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell*, 182(5):1295–1310.e20, 2020.

[91] Alief Moulana, Thomas Dupic, Angela M Phillips, Jeffrey Chang, Serafina Nieves, Anne A Roffler, Allison J Greaney, Tyler N Starr, Jesse D Bloom, and Michael M Desai. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 omicron BA.1. *Nat. Commun.*, 13(1):7011, November 2022.

[92] Michael B. Doud and Jesse D. Bloom. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses*, 8(6), 2016.

[93] Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, and Jesse D Bloom. Mapping mutational effects along the evolutionary landscape of hiv envelope. *eLife*, 7:e34420, mar 2018.

[94] NeurIPS Conference. Neurips paper checklist guidelines. `https://neurips.cc/public/guides/PaperChecklist`, 2025. Accessed: 2025-05-07.

[95] Alina Baum, Benjamin O. Fulton, Elzbieta Wloga, Richard Copin, Kristen E. Pascal, Vincenzo Russo, Stephanie Giordano, Kathryn Lanza, Nicole Negron, Min Ni, Yi Wei, Gurinder S. Atwal, Andrew J. Murphy, Neil Stahl, George D. Yancopoulos, and Christos A. Kyratsous. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science*, 369(6506):1014–1018, August 2020.

[96] Integrated DNA Technologies, Inc. Integrated dna technologies (idt): Discover what's possible. `https://www.idtdna.com/`, 2025. Accessed: 2025-05-07.

[97] Twist Bioscience. Twist bioscience: Writing the future of biology. `https://www.twistbioscience.com/`, 2025. Accessed: 2025-05-07.

[98] Dana Gretton, Brian Wang, Rey Edison, Leonard Foner, Jens Berlips, Theia Vogel, Martin Kysel, Walther Chen, Francesca Sage-Ling, Lynn Van Hauwe, et al. Random adversarial threshold search enables automated dna screening. *bioRxiv*, pages 2024–03, 2024.

[99] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[100] Rachel Seongeun Kim, Eli Levy Karin, Milot Mirdita, Rayan Chikhi, and Martin Steinegger. Bfvd—a large repository of predicted viral protein structures. *Nucleic Acids Research*, 53(D1):D340–D347, 2025.

[101] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[102] Lilian Weng. Reward hacking in reinforcement learning. *lilianweng.github.io*, Nov 2024.

[103] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[104] Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel. *arXiv preprint arXiv:1705.08417*, 2017.

[105] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

[106] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

A Technical Appendices and Supplementary Material

Alternative views The "Traditional Methods Suffice" View: Traditional methods, like serial passage (repeatedly passing a pathogen through cell cultures or hosts) have long been known to potentially increase virulence or alter host adaptation. The fundamental capability for misuse existed before sophisticated AI.

We acknowledge this point; however, IAB introduces distinct and potentially greater risks. Unlike traditional methods, IAB can facilitate the rapid design of novel viruses, potentially distant from the evolutionary tree and specifically optimized for traits like antibody escape. Furthermore, IAB systems are often far more efficient than serial passage and critically, they can lower the barrier to entry by reducing the depth of specialized biological or virological expertise required. This distinction is illustrated by serial passaging experiments from Baum et al. [95], which showed that SARS-CoV-2 can acquire antibody escape mutations—typically one or two steps from wild-type—when placed under immune pressure. While these studies reveal how targeted selection can exploit local evolutionary pathways, they remain limited in scope and pace. In contrast, IAB systems can traverse much larger regions of the mutational landscape, efficiently identifying escape variants that are far from natural sequences and doing so without iterative wet-lab experimentation or specialized virological expertise.

The "Capability Gap / Overstated Risk" View: While pLMs and IAB show promise, their current ability to reliably design de novo pathogens with enhanced pandemic potential (e.g., increased virulence and transmissibility in humans) is significantly overstated.

As discussed in Sections 2 and 3, pLMs and other AI models have already significantly accelerated pandemic prevention research. The same powerful capabilities developed for these beneficial purposes are inherently dual-use and could potentially be redirected towards designing novel or enhanced viral threats. While security considerations prevent authors from exhaustive detailing of potential misuse pathways and scenarios, the demonstrated predictive power [15, 47] and design potential[48, 49] lead the authors to assess that current capabilities present a tangible risk. Therefore, this paper addresses concrete, technically feasible dual-use capabilities that are demonstrably achievable with current technology. The threat model is explicit: AI-designed biological agents that can be physically synthesized and pose risks to public health.

The "Existing Governance is Sufficient" View: Current biosecurity frameworks (e.g., US DURC policies, Select Agent Regulations, export controls, institutional biosafety committees) combined with improving DNA synthesis screening are largely sufficient to manage the risks

Current biosecurity frameworks like the US DURC policy do not cover the dual-use potential of general-purpose AI tools like pLMs or the risks introduced by their integration into automated experimental platforms. Even safeguards like DNA synthesis screening have shown alarming gaps: a 2024 MIT experiment demonstrated that synthetic DNA fragments reconstructing the 1918 influenza virus could be ordered from the majority of providers tested—despite existing screening protocols. Moreover, pLMs can generate novel sequences that fall outside known pathogen databases, potentially bypassing existing detection and oversight systems.

The "Focus on Actors and Labs, Not Models" View: The primary risk comes from malicious actors with access to physical laboratory resources. Governing the AI models is a secondary, less effective control point compared to securing labs, vetting personnel, and controlling access to DNA synthesis and dangerous biological agents.

We agree that securing physical laboratory and controlling access to DNA synthesis remain essential components of biosecurity. However, as discussed in the paper, the access to DNA synthesis is not secured. And this view underestimates the effect that powerful AI models are having on the accessibility and speed of biological design. These models lower the expertise and resource barriers required to engineer dangerous agents.

Automatic Liquid Handler Laboratory automation utilizes automated systems to replace manual experimental labor, offering benefits such as increased throughput, enhanced reproducibility, and reduced human error. Different lab architectures provide varying degrees of automation and flexibility. In Figure 5 we provide an illustration of a liquid handling machine. These systems can perform precise, high-throughput pipetting and sample preparation with minimal human intervention. For example, commercially available platforms offer programmable workflows capable of operating across a wide range of volumes and plate formats, supporting applications from routine assays to complex protein engineering pipelines.[52] Their integration with AI-driven design tools enables rapid iteration across DBTL cycle.

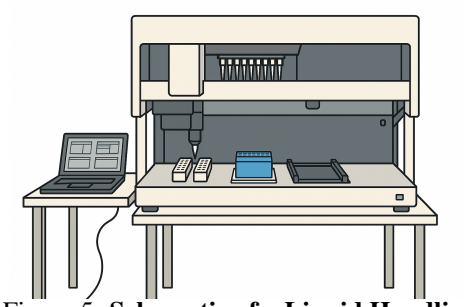

Figure 5: **Schematic of a Liquid Handling Machine**

Protein Design Conditioning Methods Unlike text-based language models that use natural language prompts, protein language models employ several conditioning approaches:

- **Structure-conditioned generation:** Models like ESM-IF and ProteinMPNN take 3D backbone coordinates as input and generate sequences likely to fold into that structure.
- **Function-based conditioning:** Through fine-tuning on datasets with functional annotations, or using predictive heads that score sequences for desired properties.
- **Active learning loops:** Experimental feedback guides the model toward sequences with desired functions through iterative Design-Build-Test-Learn cycles.

Additional Policy Suggestions Due to space constraints in the main text, we highlight here additional policy proposals for the biosafety and machine learning communities to consider:

- **Strengthen DNA Order Screening for Direct Evolution:** Biosafety researchers and DNA synthesis companies should develop algorithms capable of analyzing synthesis orders originating from the same customer or payment source, with the goal of detecting suspicious patterns—such as iterative mutations indicative of directed evolution experiments on viral proteins.
- **Promote Responsible Release Norms for High-Capability Models:** The community should establish norms for the release of powerful pLMs capable of IAB, and consider their integration with wet-lab and bio-automation. This could involve considering tiered access models, staged releases contingent on safety evaluations for models exceeding defined capability thresholds (potentially frameworks like Table 1). However, as discussed in Section 2, existing open-weight pLMs and generative models are already sufficiently capable to design novel viruses.
- **Advance AI-Enhanced DNA Screening Tools:** The research community, in collaboration with DNA synthesis companies such as IDT[96] and Twist Bioscience[97], should invest in improving AI-assisted DNA synthesis screening. These efforts should extend recent advances [98, 77] to better detect functionally concerning sequences beyond simple string similarity.

Time for sequence improvement Modern protein engineering pipelines increasingly rely on automation to accelerate the design-build-test-learn (DBTL) cycle. While traditional manual screening methods with hand pipetting—typically allow for the evaluation of fewer than 10 variants per day, automated high-throughput screening (HTS) platforms using robotic liquid handlers and flow cytometry can screen on the order of 10,000 variants daily (Table 3).

Table 3: Comparison of Manual and Automated Screening Throughput

Method	**Variants/Day**	**Example Technologies**
Manual Screening	$\sim$1	Standard 96-well plate assays, manual pipetting
Automated HTS (Microplates)	$\sim$10,000	Robotic liquid handlers (e.g., Agilent Bravo, Tecan Fluent) as shown in Figure 5

Recent experimental work by Jawaid et al.[64] highlights the impact of model-guided design on protein engineering success rates. In a benchmark screen of over 34,000 synthetic proteins, a baseline hit rate of just 0.5% was observed using random designs. By contrast, sequences proposed by a pLM and refined using an Evolutionary Monte Carlo Search achieved hit rates as high as 20%,. This dramatic increase (more than 40-fold over random screening) illustrates the potential for AI systems to focus experimental resources on highly promising regions of sequence space (Table4).

Table 4: Comparison of Approximate Hit Rates for Different Protein Engineering Methods [64]

Method	**Approximate Hit Rate**
Random Screening	$\sim 0.5\%$
pLM Guided Methods	~ 20

RLHF of LLM and pLM comparison We provide a table below for the mapping between RLHF of LLMs and that of pLM.

Notation: $S_{\texttt{harmful}}(\texttt{protein})$ represents the fitness of harmful proteins.

Table 5: Correspondence between RLHF of pLM and LLM for safety concerns

Elements of RLHF	**LLM**	**pLM**
Policy Model π	Large Language Model (e.g., GPT-style[99]) generating text.	Protein Language Model (e.g., ESMIF-style[7]) generating amino acid sequences.
Action Space	Sequence of tokens (words, sub-words).	Sequence of amino acids.
Prompt / Input (for RL fine-tuning)	Text prompts guiding text generation.	Given a desired structure or function, generate a compatible protein sequence (e.g.binding affinity, or catalytic activity).
Dataset for Reward Model (RM) Training	Human preference data on LLM outputs (e.g., rankings of responses, labels of helpfulness/harmfulness). Format: (prompt, response, preference_label).	Curated viral datasets containing sequences, structures, and annotated fitnes scores. Examples include BFVD [100] and the fitness datasets in Table 2. Format: (sequence, harmfulness_score) pairs.
Reward Model	Model trained on human preference data to predict a scalar score reflecting the desirability of LLM output.	A predictive model trained on viral protein data to assign a scalar harmfulness score to a protein sequence, eg. similarity to known viral sequences
Reward Signal (for RL fine-tuning)	Output of the RM (higher score indicates output is more aligned with human preferences).	Inversely related to the RM's "harmfulness" score (e.g., $\texttt{Reward} = -S_{\texttt{harmful}}(\texttt{protein})$).
Fine-tuning Objective	Maximize expected RM score, often with KL penalty against original LLM: $\max_\pi \mathbb{E}_{x\sim\pi}[\text{RM}_{\text{LLM}}(x) - \beta\text{KL}(\pi\|\|\pi_0)]$.	Maximize expected (negative) harmfulness score: $\max_\pi \mathbb{E}_{p\sim\pi}[-S_{\texttt{harmful}}(\texttt{protein}) - \beta\text{KL}(\pi\|\|\pi_0)]$.
Goal of RLHF	Align LLM with human preferences (e.g., make more helpful, harmless, honest).	Steer pLM away from generating harmful biological sequences.

Reward Hacking of RLHF It is important to note that RLHF on pLMs also faces the challenge of reward hacking[101, 102, 103, 104, 105, 106], where the pLM might generate sequences that achieve low harmfulness scores from the RM while remaining biologically dangerous. This risk emerges when the RM serves as an imperfect proxy for true biological risk, particularly if the pLM generates novel viral sequences not represented in the RM's training data. The effectiveness of this safeguard therefore depends critically on the RM's comprehensiveness, which in turn depends on the quality and breadth of available experimental data, bringing us back to the fundamental bio-evaluation gap described earlier in Section 5.