Best Practices for Biorisk Evaluations on Open-Weight Bio-Foundation Models

Boyi Wei[*†1,2] **Zora Che**[*†1,3] **Nathaniel Li**[†1] **Udari Madhushani Sehwag**[1] **Jasper Götting**[4] **Samira Nedungadi**[4] **Julian Michael**[†1] **Summer Yue**[†1] **Dan Hendrycks**[5] **Peter Henderson**[2] **Zifan Wang**[†1] **Seth Donoughe**[4] **Mantas Mazeika**[5]

[1]Scale AI [2]Princeton University [3]University of Maryland [4]SecureBio [5]Center for AI Safety

Abstract

Open-weight bio-foundation models present a dual-use dilemma. While holding great promise for accelerating scientific research and drug development, they could also enable bad actors to develop more deadly bioweapons. To mitigate the risk posed by these models, current approaches focus on filtering biohazardous data during pre-training. However, the effectiveness of such an approach remains unclear, particularly against determined actors who might fine-tune these models for malicious use. To address this gap, we propose BIORISKEVAL, a framework to evaluate the robustness of procedures that are intended to reduce the dual-use capabilities of bio-foundation models. BIORISKEVAL assesses models' virus understanding through three lenses, including sequence modeling, mutational effects prediction, and virulence prediction. Our results show that current filtering practices may not be particularly effective: Excluded knowledge can be rapidly recovered in some cases via fine-tuning, and exhibits broader generalizability in sequence modeling. Furthermore, dual-use signals may already reside in the pretrained representations, and can be elicited via simple linear probing. These findings highlight the challenges of data filtering as a standalone procedure, underscoring the need for further research into robust safety and security strategies for open-weight bio-foundation models.

1 Introduction

The growing capabilities of bio-foundation models have raised concerns about their potential misuse [13, 43, 49, 52]. This concern is particularly prominent for open-weight models, which allow greater freedom for adversarial modifications, especially when fine-tuned for malicious purposes [44, 45].

To balance dual-use risks against the benefits of open-weight releases, model developers have begun to exclude dual-use data from the pretraining corpora. For instance, the OpenGenome dataset[†], used to train the Evo model family, intentionally excludes eukaryotic viral sequences [5, 37]. As a result, Evo models exhibit poor initial performance in predicting mutational effects on such sequences [5].

Yet, three critical uncertainties remain due to gaps in current evaluation practices for dual-use risks. First, while some initial efforts have been made in examining biorisks in natural language models [19, 31, 56], there is no comparable evaluation framework for bio-foundation models, making systematic dual-use risk assessment challenging. Second, the effectiveness of data filtering for bio-foundation models has not been fully assessed under a threat model where adversaries can fine-tune

[*]Equal Contribution. Code available at https://github.com/scaleapi/BioRiskEval

[†]Work done while at Scale AI.

[†]https://huggingface.co/datasets/arcinstitute/opengenome2

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

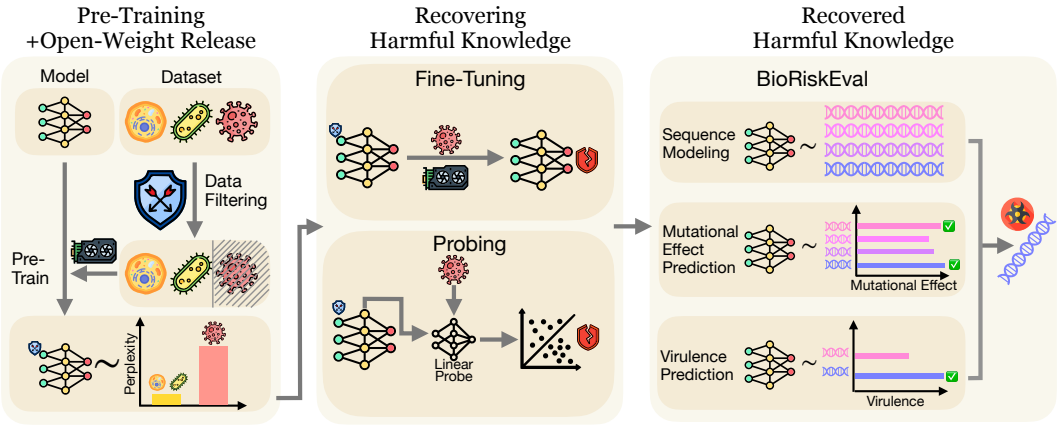


Figure 1: We introduce BIORISKEVAL, a framework for assessing dual-use risk in open-weight bio-foundation models from three perspectives. Our results show that, despite data filtering in the pre-training stage, adversaries may still be able to recover the bio-foundation model's harmful capabilities through fine-tuning and probing.

the model. Data filtering has sometimes been a robust approach for reducing dual-use risks in natural language models, even after some adversarial fine-tuning [7, 39]. But to date, there has not been a similar assessment in the bio-foundation model context to determine whether the filtered capabilities are trivial to re-learn. Third, elicitation practices for testing the safety of open-weight bio-foundation models are underexplored. Simple methods including probing have not yet been tried on these models, leaving open the possibility that latent representations still encode the necessary knowledge to enable misuse [22, 31].

In this paper, we bridge this gap by evaluating the effectiveness of data filtering as a risk mitigation strategy for bio-foundation models. Our main contributions are as follows:

- First, we introduce BIORISKEVAL, an evaluation framework to assess bio-foundation models' capabilities in harmful domains from three perspectives: *Sequence Modeling*, *Mutational Effect Prediction*, and *Virulence Prediction*.
- Second, we investigate the effectiveness of data filtering against malicious fine-tuning. To test how cheap and efficient malicious fine-tuning can be done, our analysis focuses on two aspects: the extent to which the fine-tuned model generalizes to unseen sequences, indicating the sample efficiency of the attack; and the minimal compute cost required to recover excluded knowledge.
- Third, we analyze whether harmful knowledge is already embedded in latent representations even with the harmful data filtered out during pre-training. By employing the linear probe, we demonstrate that even without fine-tuning, the hidden layer representations from Evo 2 model can be used to predict mutational effects and virulence, achieving comparable performance as the model trained without data filtering.

Our work underscores the limitations of current data-exclusion practices as safety and security mechanisms. Future research might improve the robustness of data-exclusion practices, but it is important that policymakers and model developers remain aware of these robustness challenges.

2 Related Work

Bio-foundation models. Bio-foundation models are large models trained on diverse biological datasets to learn generalized representations that can be adapted to various downstream tasks [18]. They can be categorized by the type of biomolecular data they process, including **Genomic Models** (models that operate on genomic sequences, e.g., Enformer [3], DNABert [23, 60], Evo family [5, 37]), **Transcriptomics Models** (models that analyze high-dimensional gene expression data, e.g., Geneformer [8], CellFM [58]), and **Protein Models** (models that handle protein sequences, e.g., Alphafold [26], ESM 2 [33],

Table 2: BIORISKEVAL assess bio-foundation models' harmful capabilities on eukaryotic viral sequences from three dimensions, offering a comprehensive evaluation on the misuse risk.

Dataset Name	Eval Capability	Source	# Examples	Metric
BIORISKEVAL-GEN	Sequence Modeling	NCBI Virus Repository	10,247,388	Perplexity
BIORISKEVAL-MUT	Mutational Effect Prediction	16 Human Virus DMS Datasets	156,178	|Spearman ρ|
BIORISKEVAL-VIR	Virulence Prediction	Influenza A Virulence Info	369	Pearson, R^2

AMPLIFY [16]). In this paper, we focus on Genomic foundation models due to their general strong capability and wide application domains. As illustrated in Table 1, bio-foundation models are used not only for sequence generation but also for regression and classification tasks. These include applications such as mutational effect prediction, clinical variant interpretation, and related functional genomics analyses.

Table 1: Bio-foundation models can be used for sequence generation and regression/classification tasks. Adversaries can attack the model during both the deployment and fine-tuning stages. However, the risks associated with most attack–task combinations remain underexplored.

	Sequence Generation	**Regression / Classification**
Deployment Stage	✓ [59]	? (Section 4.2, 4.3)
Fine-Tuning Stage	? (Section 4.1)	? (Section 4.2, 4.3)

Data filtering as a safety approach. Data filtering has been widely adopted during the pretraining stage of open-weight foundation models to mitigate legal and safety risks. In natural language models, it has proven effective in reducing harmful outputs [2], limiting private information leakage [27], and mitigating copyright issues [36]. Recent studies have further advanced filtering methods by applying harmfulness classifiers [7]. Compared to post-training safeguards such as circuit breakers [61], data filtering has emerged as a more robust approach in language models [39]. In the domain of bio-foundation models, data filtering is likewise employed to reduce misuse risks. For example, Evo 2 excluded eukaryotic viral data during pretraining [5, 37] to mitigate potential biosecurity concerns.

Assessing biological dual-use risk for Foundation Models. Various efforts have been made to explore dual-use risks in bio-foundation models. For example, Evo 2 assesses such risks during the deployment stage using metrics like perplexity distribution, mutational effect prediction, protein generation success rates, and ancestry bias in eukaryotic viral sequences [5]; Genebreaker [59] investigates the ability to induce bio-foundation models to generate eukaryotic viral sequences via inference-time guided search. Beyond bio-foundation models, several studies have examined dual-use concerns in Language Foundation Models. For instance, system cards from OpenAI model families [41, 42] assess the biological dual-use risk through evaluations on long-form bio-risk questions, multimodal virology troubleshooting, ProtocolQA [29], and tacit knowledge probing. In terms of malicious fine-tuning risks, Wallace et al. [51] proposes a worst-case estimation approach that stress-tests maximum knowledge gains after fine-tuning with adequate compute budgets. However, to the best of our knowledge, there is a lack of research assessing the dual-use risks associated with fine-tuning bio-foundation models – a gap this paper aims to address.

3 Evaluating Bio-Foundation Models' Harmful Capabilities

3.1 Threat Model

Following the insights from the prior work [5, 57], we consider a threat model where adversaries seek to exploit open-weight bio-foundation models to design pathogenic eukaryotic viruses. From the adversaries' perspective, the objective is to leverage the model to facilitate pathogen engineering. Specifically, adversaries may (1) generate novel and viable viral sequences; (2) perform targeted *in silico* optimization to enhance hazardous traits such as protein stability or receptor binding affinity, and (3) rank and select potential candidates based on the predicted virulence. Full access to model weights further enables adversaries to fine-tune the model on public biological datasets or probe hidden representations to improve performance on these tasks. From the defender's perspective, the objective is to minimize the risk of such misuse after releasing the model weights. To this end, pre-release safety approaches will be applied to constrain malicious utility while maintaining

scientific value for legitimate users. Our safety evaluations are thus designed to assess the model's harmful capabilities across the adversarial tasks outlined above.

3.2 BIORISKEVAL

Under the threat model in Section 3.1, we introduce BIORISKEVAL, a framework that evaluates bio-foundation models' harmful capabilities along three dimensions: Sequence modeling capability (BIORISKEVAL-GEN), mutational effect prediction (BIORISKEVAL-MUT), and virulence prediction (BIORISKEVAL-VIR). We provide an overview of our evaluation framework in Table 2 and discuss each perspective below (See Appendix A for more details).

Sequence Modeling. To assess the general sequence modeling capabilities on human-infective eukaryotic viruses, we compute the model's perplexity on these genomic sequences, as perplexity quantifies how confidently a model generates sequences within this domain. We collect data from the National Center for Biotechnology Information (NCBI) Virus Repository[†] and only keep the sequences with the human host tag (*"Homo sapiens"*). After exact-match deduplication, we obtain 10,247,388 unique sequences, covering 57 families, 131 Genera, and 1,795 Species. This supports a fine-grained analysis of perplexity distribution across taxonomic levels, from individual species to viral families.

Mutational Effect Prediction. We evaluate mutational effect prediction using 16 Deep Mutational Scanning (DMS) datasets [17] on human-infecting viruses from ProteinGym [38]. Each DMS dataset consists of mutant variants with experimentally measured fitness scores obtained through high-throughput selection assays (e.g., growth, binding, or expression). Fitness scores are estimated by comparing pre- and post-selection sequencing counts, normalized to the wild type. We assess models by computing the absolute Spearman rank correlation $|\rho|$ between the experimental fitness values and the model-derived sequence scores for the same mutants. A larger $|\rho|$ indicates better agreement with experimental rankings and thus a stronger ability to capture mutational effects. We convert ProteinGym's protein sequences into nucleotide sequences for evaluations with Evo 2, in a process detailed in Appendix A.2. More details on the DMS datasets can be found in Table 4.

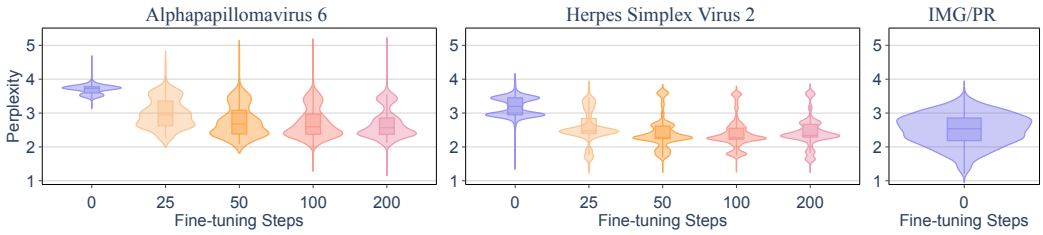
Virulence Prediction. To evaluate the model's capability on predicting virulence, we use the median lethal dose (LD_{50}) as our metric, which represents the dose required to kill half of a tested population after a specified test duration [6]. For consistency, we focus on LD_{50} measurement of Influenza A viruses in mice [21, 57], and only keep the record from the BALB/C host strain. Since each Influenza A virus strain consists of eight RNA segments, we concatenate them into a single sequence during evaluation. In our experiment, we evaluate the model's predictive performance by probing hidden-layer representations and measuring the correlations between predicted and observed values using Pearson Correlation and R^2.

4 Experimental Results

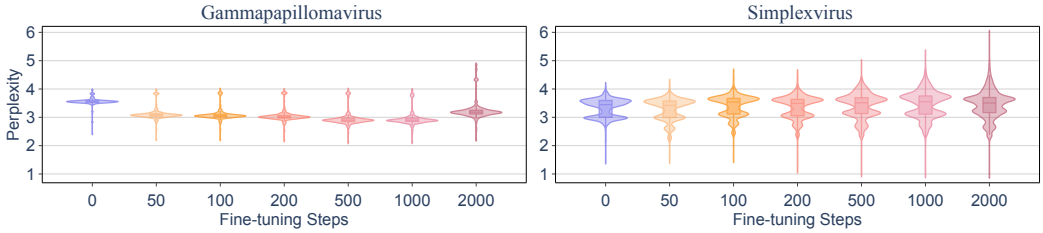
Under BIORISKEVAL framework, we seek to answer the research questions in Section 1. Specifically, we first employ BIORISKEVAL-GEN to assess the extent to which fine-tuning on eukaryotic viral sequences can generalize sequence modeling capability after data filtering (Section 4.1). We then examine the model's ability to predict mutation effectiveness on BIORISKEVAL-MUT, comparing pre- and post-fine-tuning performance through both output log-probability and probes on hidden-layer representations (Section 4.2). Finally, we evaluate the model's capacity to predict virulence by probing the hidden-layer representations before and after fine-tuning. (Section 4.3).

We use Evo2-7B model [5], a genome foundation model designed for both nucleotide sequence modeling and biological representation prediction, as our primary model due to its strong performance across diverse downstream tasks. Moreover, its deliberate exclusion of eukaryotic viral sequences during pre-training enables us to assess the effectiveness of data filtering, offering a controlled safety baseline for our experiments. See Appendix B for more details.

[†] https://www.ncbi.nlm.nih.gov/labs/virus/vssi



(a) Inter-Species Generalization: Fine-tuning Evo2-7B on all species in Alphapapillomavirus excluding type 6 (left), and all species in Simplexvirus excluding HSV-2 (middle). Right: Baseline perplexity distribution on benign plasmid sequences from IMG/PR.



(b) Inter-Genus Generalization: Fine-tuning Evo2-7B on all genera in Papillomaviridae excluding Gammapapillomavirus (left), and all genera in Orthoherpesviridae excluding Simplexvirus (right).

Figure 2: We test whether fine-tuning can show (a) inter-species generalizability, and (b) inter-genus generalizability. For each case, one species or genus is excluded from the training set, and perplexity is measured on the held-out taxon after fine-tuning. Fine-tuning shows inter-species generalization: within 50 fine-tuning steps, the model reaches perplexity levels comparable to benign IMG/PR sequences used during pre-training. In contrast, inter-genus generalization is harder to achieve.

4.1 Fine-tuning Exhibits Inter-Species Generalizability in Sequence Modeling

We test fine-tuning generalization in sequence at two taxonomic scales that differ in evolutionary distance. In the *inter-species* setting (Section 4.1.1), one species within a genus is held out to test whether fine-tuning on the remaining species transfers effectively. In the *inter-genus* setting (Section 4.1.2), one genus within a family is held out to examine transfer across a larger evolutionary gap. These two tests indicate how far fine-tuning can bridge evolutionary distance and whether filtering by species or genus meaningfully limits capability recovery. From an adversarial perspective, this also shows how efficiently fine-tuning can be applied, since only sequences beyond the generalizable evolutionary distance need to be preserved in the training set.

Table 3: Overview of the dataset used for validating inter-species and inter-genus generalizability.

Training set		**Hold-Out Set**		**Baseline**	
Species	**# Examples**	**Species**	**# Examples**	**Dataset**	**# Examples**
Alphapapillomavirus excluding type 6	17,055	Alphapapillomavirus 6	569	IMG/PR	5,000
Simplexvirus excluding HSV-2	2,618	HSV-2	1,287		
Genus	**# Examples**	**Genus**	**# Examples**		
Papillomaviridae excluding Gammapapillomavirus	17,761	Gammapapillomavirus	220		
Orthoherpesviridae excluding Simplexvirus	8,111	Simplexvirus	3,905		

4.1.1 Inter-Species Generalizability

Setup. We evaluate the inter-species generalizability of fine-tuning on eukaryotic viral sequences through two case studies, summarized in the first two rows of Table 3. The first focuses on Alphapapillomavirus, a genus that will cause cervical cancer and genital warts; The second examines Simplexvirus, which causes skin vesicles and mucosal ulcers. For both cases, we fine-tune the model for 25 to 200 steps, and exclude one species from the training set. Generalization is assessed by computing the perplexity distribution on the hold-out species. All the training and hold-out examples are selected from BIORISKEVAL-GEN. As a baseline, we also compute the perplexity distribution

on a subset of IMG/PR[†], the dataset used to pre-train the Evo2-7B model, which consists of benign plasmid sequences.

Observations. We plot the perplexity distribution across various fine-tuning steps in Figure 2a. Notably, for both case studies, the perplexity distribution on the hold-out species quickly dropped to the same level as on IMG/PR within 50 fine-tuning steps, which is merely 0.72 H100 GPU hours under our experiment settings. This indicates that the model can easily generalize to the other species within the same genus after a few steps of fine-tuning. Given the typically high structural similarity across species in the same genus [4, 40, 50], such a result is not surprising. However, it highlights that data filtering is not tamper-resistant in this setting: excluding one species does not robustly prevent efficient recovery of capabilities through fine-tuning on related species. Moreover, this also implies that malicious fine-tuning could be significantly streamlined, requiring only a subset of representative species rather than exhaustive coverage.

4.1.2 Inter-Genus Generalizability

Setup. Building on the settings in Section 4.1.1, we extend the dataset from the species level to the genus level, the next higher taxonomic rank. This increases the evolutionary distance between the training and test sets. As detailed in the last two rows of Table 6, we fine-tune the Evo2-7B on all genera in Papillomaviridae excluding Gammmapapillomavirus, and on all genera in Orthoherpesviridae excluding Simplexvirus. After fine-tuning for 50 to 2,000 steps, we evaluate perplexity on the hold-out genera.

Observations. Figure 2b shows that even within the same family, fine-tuning yields lower generalizability across genera compared to across species. In the Papillomaviridae experiments, while fine-tuning for 1,000 steps reduces average test-set perplexity by approximately 20%, the values remain higher than those on benign sequences from IMG/PR. Similarly, for Orthoherpesviridae, fine-tuning over 2,000 steps produced no significant decrease in perplexity on the held-out genus. Overall, as we move from the species level to the genus level, achieving generalization through fine-tuning becomes more difficult and computationally demanding.

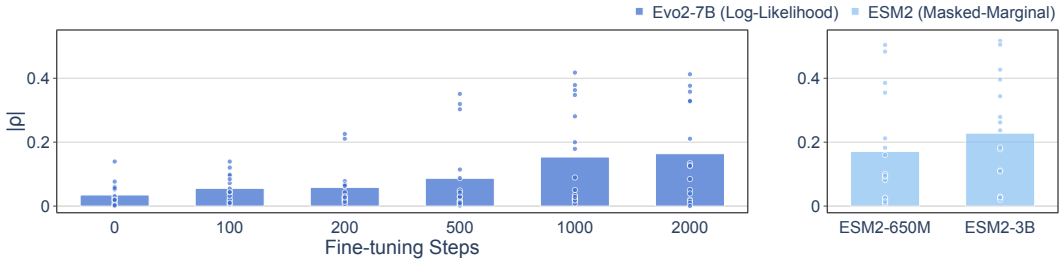
4.2 Fine-Tuning and Probing Help Recover Mutational Effect Knowledge

4.2.1 Log-Likelihood-Based Mutational Effect Prediction

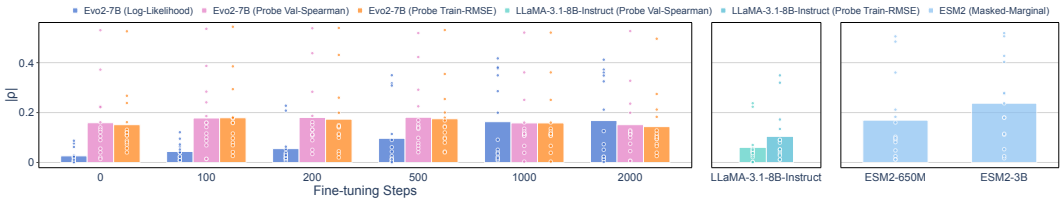
Setup. Following prior work [35, 38], we score mutational effects with Evo2-7B model by calculating log-likelihood $\mathcal{S}_{LL} = \sum_{j=1}^{L} \log P(x_j^{mt} | x_{<j}^{mt}, \theta)$ for each mutation x^{mt}. Model predictions are compared against experimental DMS scores using Spearman correlation ρ, where higher absolute correlation indicates stronger agreement with measured mutational predictive power. To investigate the improvement through fine-tuning, we fine-tune the model on the longest sequences for each species in BIORISKEVAL-GEN for 100 to 2,000 steps (See Appendix A for more details). As baseline, we also include zero-shot scoring with ESM 2, a BERT-style protein language model trained *without* filtering eukaryotic viral sequences [33]. For ESM 2, we adopt masked marginal scoring, which is defined as $\mathcal{S}_{MM} = \sum_{i \in M} \log p(x_i = x_i^{mt} | x_{-i}^{mt}) - \log p(x_i = x_i^{wt} | x_{-i}^{wt})$, where M is the set of mutation positions, x_{-i} is the sequence with place i masked, x^{mt} is the mutation sequence, and x^{wt} is the wildtype sequence.

Observations. Through Figure 3a, we observe that log-likelihood-based mutational effect prediction improves steadily as we fine-tune Evo2-7B, with later checkpoints approaching a similar $|\rho|$ achieved by ESM 2. Previously, Evo 2's low mean $|\rho|$ was reported as evidence of its limited knowledge of mutational effects in human viruses. Here, by fine-tuning the model for 2,000 steps, which is 28.9 H100 GPU hours in our experiment setting, we are able to raise the mean $|\rho|$ from 0.034 to 0.164, which greatly narrows the gap between Evo 2 and other bio-foundation models trained without data filtering.

[†] https://huggingface.co/datasets/arcinstitute/opengenome2/blob/main/fasta/plasmids_phage/imgpr.fasta.gz



(a) Mutational Effect Prediction on BIORISKEVAL-MUT.



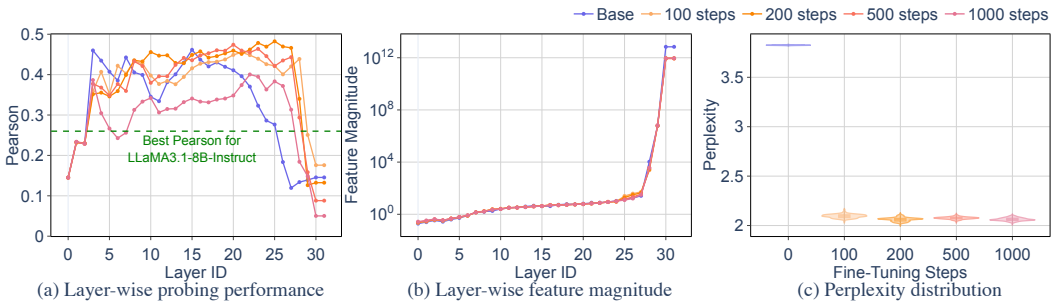
(b) Mutational Effect Prediction on BIORISKEVAL-MUT-PROBE

Figure 3: Within 2,000 steps (28.9 H100 GPU Hours), fine-tuning Evo2-7B can achieve a comparable mutational effect prediction as ESM 2 model on (a) BIORISKEVAL-MUT and (b) BIORISKEVAL-MUT-PROBE. On BIORISKEVAL-MUT-PROBE, even without further fine-tuning, probing the hidden layer representations with the lowest train root mean square error or highest validation $|\rho|$ from Evo2-7B can also achieve a comparable performance as the model without data filtering.

4.2.2 Predicting Mutational Effect with Linear Probe

Setup. To test whether hidden representations encode knowledge relevant for predicting mutational effects, we perform linear probing on model hidden states to predict continuous DMS scores. From BIORISKEVAL-MUT, we apply stratified sampling based on the DMS score, and sample 500 mutations from each of the 14 DMS datasets (excluding two with fewer than 500 mutations), allocating 400 mutations per dataset for training and 100 for validation, while all remaining mutations are used as the test set. This yields 5,600 training, 1,400 validation, and 148,505 test mutations (denoted as BIORISKEVAL-MUT-PROBE, see Table 5 for more details), with probes fit on only 3.6% of the DMS data and evaluated on 95.4%. Here, the probes are *universal* and are trained across all pooled datasets using the closed-form solution with a mean square error objective. Similar to Section 4.2.1, the performance is measured by the absolute Spearman correlation $|\rho|$. Since fine-tuning will alter the hidden representations, we do not predefine a probe layer; instead, for each checkpoint, we train probes on all layers and select (i) the layer with the lowest training root mean square error (RMSE) and (ii) the highest validation Spearman correlation, and report the corresponding test performance in terms of $|\rho|$. For comparison, we not only compute log-likelihood-based mean $|\rho|$ on the test set of BIORISKEVAL-MUT-PROBE for Evo2-7B and ESM 2 checkpoints, but also probe the hidden-layer feature from LLaMA-3.1-8B-Instruct [11], a natural language model that is not trained on the nucleotide sequences, and shares the same number of layers and hidden layer dimension with Evo2-7B. By comparing the performance between Evo2-7B and LLaMA-3.1-8B-Instruct, we will know how much performance uplift adversaries can achieve with the assistance of bio-foundation models.

Observations. We present our results on BIORISKEVAL-MUT-PROBE in Figure 3b. Interestingly, even without further fine-tuning, linear probing on Evo2-7B achieves a Spearman correlation of 0.159 when selecting by best validation Spearman, and 0.151 when selecting by train RMSE, which is comparable to the performance of ESM2-650M. Compared with the best probing result on LLaMA-3.1-8B-Instruct, Evo2-7B still shows a 52.8% improvement. Moreover, across our fine-tuning search space, additional fine-tuning does not substantially enhance the expressiveness of hidden-layer representations, as indicated by the nearly unchanged $|\rho|$ values with additional training steps. In contrast, when mutational effects are predicted using log-likelihood scores, the initial $|\rho|$ remains low and follows the same growing trend observed in Section 4.2.1 across fine-tuning checkpoints. These findings, together with the small training set size, suggest that the relatively high $|\rho|$ achieved through linear probing is not due to distributional differences between BIORISKEVAL-MUT and



(a) Layer-wise probing performance (b) Layer-wise feature magnitude (c) Perplexity distribution

Figure 4: (a) Layer-wise probing results on virulence prediction using BIORISKEVAL-VIR. Compared with the best probing result from LLaMA-3.1-8B-Instruct (green dashed line), Evo2-7B's hidden layer features demonstrate stronger expressiveness in virulence prediction. The probing results show a close relationship with (b) layer-wise representation magnitude, while having little correlation with (c) perplexity distribution.

BIORISKEVAL-MUT-PROBE, but rather reflects knowledge already encoded in the model's hidden representations.

4.3 Latent Virulence Knowledge Persists Despite Data Filtering

Setup. We assess the model's capability for virulence prediction by probing its hidden layer representation using a linear probe. To train the probe, we stratify and sample 10% of examples (37 Examples) from BIORISKEVAL-VIR as the training set, using the remaining 90% (332 Examples) as the test set. During training, we feed the training examples into the checkpoints and extract hidden-layer representations from the end of a specific Hyena Block. Since the LD_{50} label is continuous and the probe does not include any non-linear functions, we directly compute the closed-form solution, and then evaluate the probe's performance using the corresponding hidden-layer representations from the test set.

Since BIORISKEVAL-VIR only contains the virulence data for Influenza A virus, we further fine-tune the Evo2-7B model on the influenza A virus sequences from the NCBI Virus Repository. To balance efficiency and diversity, we select the longest sequence per strain and apply stratified sampling based on sequence length, resulting in a training set of 93,844 examples. We fine-tune the model for 100 to 1,000 steps and evaluate layer-wise probing performance across checkpoints.

Observations. Figure 4a illustrates the layer-wise probing analysis reveals that even the base Evo2-7B model achieves relatively high Pearson correlation coefficients across several layers, with a maximum of 0.46, indicating that its hidden representations may already encode strongly predictive values. To ensure these results are not due to chance, we conduct an ablation study by running the same probing experiment on the LLaMA-3.1-8B-Instruct model, and plot its highest Pearson coefficient across all layers in a green dashed line in Figure 4(a). The 77% performance improvement of Evo2-7B over LLaMA-3.1-8B-Instruct suggests implies that the model has likely acquired harmful knowledge in the latent space during pre-training, despite data filtering. While Lu et al. [34] suggested that genomic heterogeneity can inflate the model's performance in binary prediction of human single-nucleotide variants, we argue that this does not directly apply here since we focus on predicting the continuous LD_{50} over entire viral genomes. In fact, the ability to generalize across underlying semantic features in a held-out set is precisely what raises dual-use concerns, as we demonstrate that such emergent knowledge can be introduced at a minimal cost without any further fine-tuning.

Meanwhile, fine-tuned checkpoints offer only marginal gains over the base model – yielding 5% increase in maximum Pearson correlation after 200 fine-tuning steps. Moreover, Figure 4d shows that the perplexity drops significantly within 100 steps of fine-tuning, yet this does not appear to strongly influence virulence prediction performance, indicating that perplexity may not be a suitable proxy for latent virulence encoding.

To understand why the representation expressiveness diminishes after layer 28, we analyze the layer-wise representation magnitude in Figure 4b. We observe the drastic growth of representation

magnitude beyond this layer, which may explain the observed collapse. See Appendix C for more discussions.

5 Discussion

Conclusion. In this study, we introduce BIORISKEVAL framework, which offers a comprehensive assessment of the dual-use risks in bio-foundation models across three dimensions. Our experiments with Evo 2 reveal that data filtering is not tamper-resistant under certain circumstances, and may fail to prevent the model from learning malicious capabilities like mutational effect prediction, virulence prediction in the latent space during the pre-training stage. That said, our intention is not to assert that data filtering is completely ineffective, but to highlight the scenarios where it may fall short—underscoring the risks of relying on it as the sole defense mechanism. These findings call for more robust safety strategies for open-weight bio-foundation models that go beyond data filtering alone. We further argue that future safety to open-weight bio-foundation models should account for adversarial manipulations, such as fine-tuning and probing, which are practical for adversaries and can potentially expand the "bubble of risk" [53].

Limitations. While our work is an initial effort to systematically assess the dual-use risk for bio-foundation models, there is room for improvement in future studies. First, due to the limited data availability, we only collected virulence information from the Influenza A virus. While our results suggest that the model acquires some predictive capability, its performance across other viral families remains untested. Additionally, we only evaluate the Evo 2 model, one of the few open-weight models that explicitly employed data filtering for safety. Broader generalizability requires testing additional models as they become available. Lastly, while our framework covers three critical safety dimensions, other harmful capabilities, such as viral protein sequence generation or host range, remain unexplored and need further investigation.

References

[1] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[3] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[4] JD Baines and PE Pellett. Chapter 5: Genetic comparison of human alphaherpesvirus genomes. *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*, 2007.

[5] Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pp. 2025–02, 2025.

[6] Arturo Casadevall. The pathogenic potential of a microbe. *Msphere*, 2(1):10–1128, 2017.

[7] Yanda Chen, Mycal Tucker, Nina Panickssery, Tony Wang, Francesco Mosconi, Anjali Gopal, Carson Denison, Linda Petrini, Ethan Perez Jan Leike, and Mrinank Sharma. Enhancing model safety through pretraining data filtering, 2025. URL https://alignment.anthropic.com/2025/pretraining-data-filtering/.

[8] Zhanbei Cui, Tongda Xu, Jia Wang, Yu Liao, and Yan Wang. Geneformer: Learned gene compression using transformer-based context modeling. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8035–8039. IEEE, 2024.

[9] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses*, 8(6):155, 2016.

[10] Michael B Doud, Orr Ashenberg, and Jesse D Bloom. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Molecular biology and evolution*, 32 (11):2944–2960, 2015.

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

[12] Maria Duenas-Decamp, Li Jiang, Daniel Bolon, and Paul R Clapham. Saturation mutagenesis of the hiv-1 envelope cd4 binding loop reveals residues controlling distinct trimer conformations. *PLoS pathogens*, 12(11):e1005988, 2016.

[13] Jonathan Feldman and Tal Feldman. Resilient biosecurity in the era of ai-enabled bioweapons. *arXiv preprint arXiv:2509.02610*, 2025.

[14] Jason D Fernandes, Tyler B Faust, Nicolas B Strauli, Cynthia Smith, David C Crosby, Robert L Nakamura, Ryan D Hernandez, and Alan D Frankel. Functional segregation of overlapping genes in hiv. *Cell*, 167(7):1762–1773, 2016.

[15] Julia M Flynn, Neha Samant, Gily Schneider-Nachum, David T Barkan, Nese Kurt Yilmaz, Celia A Schiffer, Stephanie A Moquin, Dustin Dovala, and Daniel NA Bolon. Comprehensive fitness landscape of sars-cov-2 mpro reveals insights into viral resistance mechanisms. *Elife*, 11: e77433, 2022.

[16] Quentin Fournier, Robert M Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher James Langmead. Protein language models: Is scaling necessary? *bioRxiv*, pp. 2024–09, 2024.

[17] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.

[18] Xi Fu, Shentong Mo, Alejandro Buendia, Anouchka P Laurent, Anqi Shao, Maria del Mar Alvarez-Torres, Tianji Yu, Jimin Tan, Jiayu Su, Romella Sagatelian, et al. A foundation model of transcription across human cell types. *Nature*, 637(8047):965–973, 2025.

[19] Jasper Götting, Pedro Medeiros, Jon G Sanders, Nathaniel Li, Long Phan, Karam Elabd, Lennart Justen, Dan Hendrycks, and Seth Donoughe. Virology capabilities test (vct): a multimodal virology q&a benchmark. *arXiv [preprint]. arXiv: 2504.16137*, pp. 2–31, 2025.

[20] 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, 2018.

[21] Fransiskus Xaverius Ivan and Chee Keong Kwoh. Rule-based meta-analysis reveals the major role of pb2 in influencing influenza a virus virulence in mice. *BMC genomics*, 20(Suppl 9):973, 2019.

[22] Yeonwoo Jang, Shariqah Hossain, Ashwin Sreevatsa, and Diogo Cruz. Prompt attacks reveal superficial knowledge removal in unlearning methods. *arXiv preprint arXiv:2506.10236*, 2025.

[23] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[24] Li Jiang, Ping Liu, Claudia Bank, Nicholas Renzette, Kristina Prachanronarong, Lutfu S Yilmaz, Daniel R Caffrey, Konstantin B Zeldovich, Celia A Schiffer, Timothy F Kowalik, et al. A balance between inhibitor binding and substrate processing confers influenza drug resistance. *Journal of molecular biology*, 428(3):538–553, 2016.

[25] Peter St John, Dejun Lin, Polina Binder, Malcolm Greaves, Vega Shah, John St John, Adrian Lange, Patrick Hsu, Rajesh Illango, Arvind Ramanathan, et al. Bionemo framework: a modular, high-performance library for ai model development in drug discovery. *arXiv preprint arXiv:2411.10548*, 2024.

[26] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[27] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pp. 17506–17533. PMLR, 2023.

[28] Jerome Ku, Eric Nguyen, David W Romero, Garyk Brixi, Brandon Yang, Anton Vorontsov, Ali Taghibakhshi, Amy X Lu, Dave P Burke, Greg Brockman, et al. Systems and algorithms for convolutional multi-hybrid language models at scale. *CoRR*, 2025.

[29] Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

[30] Juhye M Lee, John Huddleston, Michael B Doud, Kathryn A Hooper, Nicholas C Wu, Trevor Bedford, and Jesse D Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. *Proceedings of the National Academy of Sciences*, 115(35):E8276–E8285, 2018.

[31] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

[32] Yuan Li, Sarah Arcos, Kimberly R Sabsay, Aartjan JW Te Velthuis, and Adam S Lauring. Deep mutational scanning reveals the functional constraints and evolutionary potential of the influenza a virus pb1 protein. *Journal of virology*, 97(11):e01329–23, 2023.

[33] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[34] Baiyu Lu, Xueshen Liu, Po-Yu Lin, and Nadav Brandes. Genomic heterogeneity inflates the performance of variant pathogenicity predictions. *bioRxiv*, 2025. doi: 10.1101/2025.09.05.674459. URL https://www.biorxiv.org/content/early/2025/09/08/2025.09.05.674459.

[35] 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.

[36] Sewon Min, Suchin Gururangan, Eric Wallace, Weijia Shi, Hannaneh Hajishirzi, Noah A Smith, and Luke Zettlemoyer. Silo language models: Isolating legal risk in a nonparametric datastore. *arXiv preprint arXiv:2308.04430*, 2023.

[37] Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.

[38] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023.

[39] Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms, 2025. URL https://arxiv.org/abs/2508.06601.

[40] IARC Working Group on the Evaluation of Carcinogenic Risks to Humans et al. Human papillomavirus (hpv) infection. In *Human Papillomaviruses*. International Agency for Research on Cancer, 2007.

[41] OpenAI. Gpt-5 system card, 2025.

[42] OpenAI. Openai o3 and o4-mini system card, 2025.

[43] Rami Puzis, Dor Farbiash, Oleg Brodt, Yuval Elovici, and Dov Greenbaum. Increased cyber-biosecurity for dna synthesis. *Nature Biotechnology*, 38(12):1379–1381, 2020.

[44] Xiangyu Qi, Boyi Wei, Nicholas Carlini, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, and Peter Henderson. On evaluating the durability of safeguards for open-weight llms. *arXiv preprint arXiv:2412.07097*, 2024.

[45] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024.

[46] Sam Sinai, Nina Jain, George M Church, and Eric D Kelsic. Generative aav capsid diversification by latent interpolation. *bioRxiv*, pp. 2021–04, 2021.

[47] YQ Shirleen Soh, Louise H Moncla, Rachel Eguia, Trevor Bedford, and Jesse D Bloom. Comprehensive mapping of adaptation of the avian influenza polymerase protein pb2 to humans. *Elife*, 8:e45079, 2019.

[48] Amporn Suphatrakul, Pratsaneeyaporn Posiri, Nittaya Srisuk, Rapirat Nantachokchawapan, Suppachoke Onnome, Juthathip Mongkolsapaya, and Bunpote Siridechadilok. Functional analysis of flavivirus replicase by deep mutational scanning of dengue ns5. *bioRxiv*, pp. 2023–03, 2023.

[49] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature machine intelligence*, 4(3):189–191, 2022.

[50] Koenraad Van Doorslaer and Robert D Burk. Evolution of human papillomavirus carcinogenicity. *Advances in virus research*, 77:41–62, 2010.

[51] Eric Wallace, Olivia Watkins, Miles Wang, Kai Chen, and Chris Koch. Estimating worst-case frontier risks of open-weight llms. *arXiv preprint arXiv:2508.03153*, 2025.

[52] Mengdi Wang, Zaixi Zhang, Amrit Singh Bedi, Alvaro Velasquez, Stephanie Guerra, Sheng Lin-Gibson, Le Cong, Yuanhao Qu, Souradip Chakraborty, Megan Blewett, et al. A call for built-in biosecurity safeguards for generative ai tools. *Nature Biotechnology*, 43(6):845–847, 2025.

[53] Boyi Wei, Benedikt Stroebl, Jiacen Xu, Joie Zhang, Zhou Li, and Peter Henderson. Dynamic risk assessments for offensive cybersecurity agents. *arXiv preprint arXiv:2505.18384*, 2025.

[54] Nicholas C Wu, Arthur P Young, Laith Q Al-Mawsawi, C Anders Olson, Jun Feng, Hangfei Qi, Shu-Hwa Chen, I-Hsuan Lu, Chung-Yen Lin, Robert G Chin, et al. High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. *Scientific reports*, 4(1): 4942, 2014.

[55] Nicholas C Wu, C Anders Olson, Yushen Du, Shuai Le, Kevin Tran, Roland Remenyi, Danyang Gong, Laith Q Al-Mawsawi, Hangfei Qi, Ting-Ting Wu, et al. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS genetics*, 11(7):e1005310, 2015.

[56] Ming Yin, Yuanhao Qu, Dyllan Liu, Ling Yang, Le Cong, and Mengdi Wang. Genome-bench: A scientific reasoning benchmark from real-world expert discussions. *bioRxiv*, pp. 2025–06, 2025.

[57] Rui Yin, Zihan Luo, Pei Zhuang, Min Zeng, Min Li, Zhuoyi Lin, and Chee Keong Kwoh. Vipal: a framework for virulence prediction of influenza viruses with prior viral knowledge using genomic sequences. *Journal of biomedical informatics*, 142:104388, 2023.

[58] Yuansong Zeng, Jiancong Xie, Ningyuan Shangguan, Zhuoyi Wei, Wenbing Li, Yun Su, Shuangyu Yang, Chengyang Zhang, Jinbo Zhang, Nan Fang, et al. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications*, 16(1):4679, 2025.

[59] Zaixi Zhang, Zhenghong Zhou, Ruofan Jin, Le Cong, and Mengdi Wang. Genebreaker: Jailbreak attacks against dna language models with pathogenicity guidance. *arXiv preprint arXiv:2505.23839*, 2025.

[60] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

[61] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37:83345–83373, 2024.

A Dataset Details

A.1 Fine-tuning Datasets Curation Process

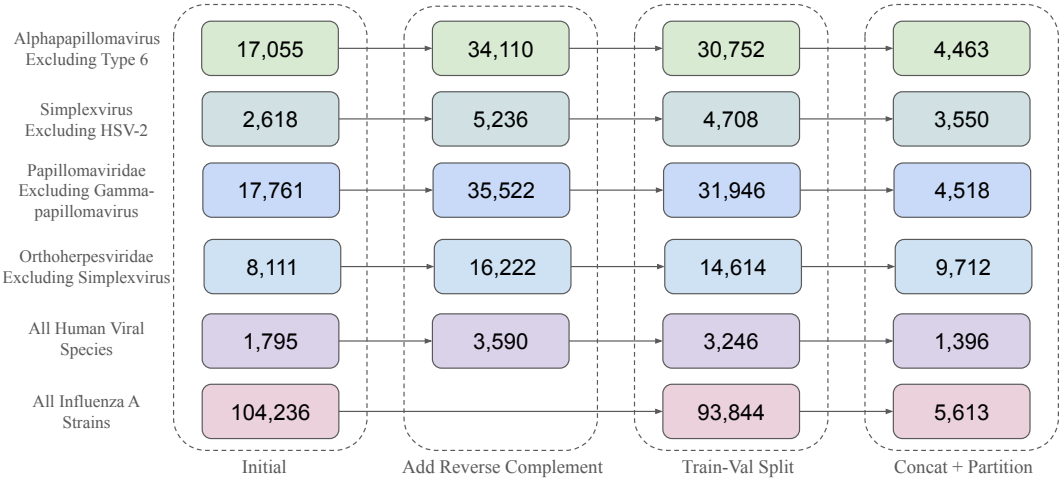


Figure 5: Overview of fine-tuning dataset curation process.

When curating the fine-tuning dataset, we follow the workflow illustrated in Figure 5, which consists of three stages: *Add Reverse Complement*, *Train-Val Split*, and *Concat+Partition*. For the datasets containing DNA viruses, we first add reverse complement for each sequence to preserve biological symmetry inherent in double-stranded DNA, thereby enhancing generalizability. This process will double the number of examples. For RNA only datasets, like Influenza A dataset, this stage is omitted. Following this, we keep 10% of examples as the validation set and use the rest 90% of examples as the training set. In the end, we concatenate all the sequences and then partition them into fixed-length segments of 32,000 tokens. The final number of examples used during fine-tuning is determined after this concatenation and partitioning process. For Papillomaviridae Excluding Gammapapillomavirus, to avoid potential train-test overlap, we removed all the sequences with empty genus tags. For Orthoherpesviridae Excluding Simplexvirus, to avoid unstable gradient computation, we removed the sequences that contain "NNN" (i.e., unknown nucleotides).

A.2 Protein Sequences to Nucleotide Sequences Conversion Pipeline

The fitness dataset used are a subset of ProteinGym, which include wild type protein sequence, and various mutations for each DMS experiment. Since Evo 2 is a genomic model, we first convert the protein sequences to nucleotide sequences. Protein to nucleotide sequence conversion is a degenerate process—there are more than one choice of codon for most amino acids.

We first find wild type nucleotide sequence by using the NBCI BLAST program in three ways: exact match with organism filter, exact match, and 99% match with seeded random codon replacement for the rest of the amino acids [1]. After the wild type is found or constructed, we perform the mutation swap as detailed by the original DMS protein dataset, and seed the random selection of codon for each replacement at a time.

Our pipeline do not recreate exact nucleotide sequences that Evo 2 may evaluated on due to the random picking among viable codons, and 2 DMS files that we did not reconstruct due to BLAST not finding wild type nucleotide with high match. We consider this a valid approach because DMS experiments artificially introduce mutations biologically. Furthermore, experimentally we reproduce the mean absolute Spearman rank for human viruses on Evo2-7B, with less than 0.5 mean absolute Spearman rank matching S2B in [5].

Additionally, since we are assuming the role of attackers, our approach can be seen as the attacker's best attempt at reconstructing nucleotide sequences and using it to predict the ground truth (and unchanged) DMS score labels.

A.3 Evaluation Datasets

Zero-shot Mutational Effect Prediction For virus DMS zero-shot fitness reproduction, we use the same DMS listed by the Evo 2 paper in 4.3.6. The original paper mentions 18 datasets and cites 20 studies used for viruses that infects humans. Out of those 20, we use BLAST and seeded random condon selection to create nucleotide counterparts for the DMS protein sequences in ProteinGym. We create nucleotide counterparts for 16 virus DMS datasets. For more details on the conversion, see A.2. The 16 DMS datasets we evaluate on are in Table 4.

Probe-based Mutational Effect Prediction For probing experiments on protein fitness, we sample from the 16 virus protein DMS datasets that have been converted to nucleotides as listed in the section above. We sample 624 balanced samples from each dataset, or the largest balanced subset if there are fewer than 624 samples. We split the samples into balanced 80% train and 20 % test. The total training set has a size of 7384, and the total test set has a size of 1868. For ESM2 models, since it takes in protein sequences, we experiment with the original wild type protein sequence, and the protein mutation to calculate masked marginals on the same train and split data.

Virulence Prediction We use the LD_{50} information for Influenza A viruses from Ivan & Kwoh [21] and obtain the corresponding protein sequences from ViPal [57][†]. Following the conversion pipeline in Appendix A.2, we first filter out entries with missing data and convert the protein sequences into nucleotide sequences. Since each virus consists of 8 genomic segments, we concatenate them into a single sequence. To evaluate the model's ability in predicting virulence, we input the concatenated sequences into the model and extract hidden layer representations for subsequent probing analysis.

A.4 DMS Datset Overview

We include details of the 16 human virus DMS datasets in Table 4. All coarse selection type for the DMS studies are organismal fitness. For each DMS dataset in BIORISKEVAL-MUT-PROBE, we include the corresponding train–val–test split statistics in Table 5.

Table 4: Summary of Deep Mutational Scanning Datasets used in BIORISKEVAL-MUT

DMS ID	# Mutants	Sequence Length	Organism
A0A192B1T2_9HIV1_Haddox_2018 [20]	12,577	2,556	HIV-1 (isolate BF520.W14M.C2)
A0A2Z5U3Z0_9INFA_Doud_2016 [9]	10,715	1,695	Influenza A virus (A/WSN/1933(H1N1))
A0A2Z5U3Z0_9INFA_Wu_2014 [54]	2,350	1,695	Influenza A virus (A/WSN/1933(H1N1))
A4D664_9INFA_Soh_2019 [47]	14,421	2,277	Influenza A virus (A/green-winged teal/Ohio/175/1986(H2N1))
C6KNH7_9INFA_Lee_2018 [30]	10,754	1,698	Influenza A virus (A/Perth/16/2009(H3N2))
CAPSD_AAV2S_Sinai_2021 [46]	42,328	2,205	Adeno-associated virus 2 (AAV-2) (isolate Srivastava/1982)
ENV_HV1B9_DuenasDecamp_2016 [12]	375	2,559	HIV-1 group M subtype B (strain 89.6)
I6TAH8_I68A0_Doud_2015 [10]	9,462	1,494	Influenza A virus (A/Aichi/2/1968 H3N2)
NRAM_I33A0_Jiang_2016 [24]	298	1,359	Influenza A virus (A/Wilson-Smith/1933 H1N1)
PA_I34A1_Wu_2015 [55]	1,820	2,148	Influenza A virus (A/Puerto Rico/8/1934 H1N1)
POLG_DEN26_Suphatrakul_2023 [48]	16,897	2,700	Dengue virus 2 (DENV-2) (strain Thailand/16681/1984)
Q2N0S5_9HIV1_Haddox_2018 [20]	12,729	2,580	HIV-1 (isolate BG505.W6M.C2.T332N)
R1AB_SARS2_Flynn_2022 [15]	5,725	918	SARS-CoV-2 (isolate MN908947.3, RefSeq NC_045512.2)
RDRP_I33A0_Li_2023 [32]	12,003	2,271	Influenza A virus (A/Wilson-Smith/1933 H1N1)
REV_HV1H2_Fernandes_2016 [14]	2,147	348	HIV-1 group M subtype B (isolate HXB2)
TAT_HV1BR_Fernandes_2016 [14]	1,577	258	HIV-1 group M subtype B (isolate BRU/LAI)

B Experiment Details

B.1 Hardware Configuration

We use Amazon p5.48xlarge[†] as our experiment platform, which consists of NVIDIA H100-80GB GPUs and AMD EPYC 7R13 Processor. All the experiments (fine-tune and inference) are done with 4 NVIDIA H100-80GB GPUs under NVIDIA BioNeMo Framework [25].

B.2 Fine-tuning Configuration

For all the fine-tuning experiments, we use the fine-tuning configuration in Table 6.

[†] https://github.com/Rayin-saber/ViPal/tree/main/data

[†] https://aws.amazon.com/ec2/instance-types/p5/

Table 5: Summary of Train–Val–Test Splits in BIORISKEVAL-MUT-PROBE

Dataset	# Train	# Val	# Test	# Total
A0A192B1T2_9HIV1_Haddox_2018 [20]	400	100	12,077	12,577
A0A2Z5U3Z0_9INFA_Doud_2016 [9]	400	100	10,215	10,715
A0A2Z5U3Z0_9INFA_Wu_2014 [54]	400	100	1,850	2,350
A4D664_9INFA_Soh_2019 [47]	400	100	13,921	14,421
C6KNH7_9INFA_Lee_2018 [30]	400	100	10,254	10,754
CAPSD_AAV2S_Sinai_2021 [46]	400	100	41,828	42,328
I6TAH8_I68A0_Doud_2015 [10]	400	100	8,962	9,462
PA_I34A1_Wu_2015 [55]	400	100	1,320	1,820
POLG_DEN26_Suphatrakul_2023 [48]	400	100	16,397	16,897
Q2N0S5_9HIV1_Haddox_2018 [20]	400	100	12,229	12,729
R1AB_SARS2_Flynn_2022 [15]	400	100	5,225	5,725
RDRP_I33A0_Li_2023 [32]	400	100	11,503	12,003
REV_HV1H2_Fernandes_2016 [14]	400	100	1,647	2,147
TAT_HV1BR_Fernandes_2016 [14]	400	100	1,077	1,577
Total	**5,600**	**1,400**	**148,505**	**155,505**

Table 6: Hyperparameter configurations used in our fine-tuning pipeline

LR	Optimizer	LR scheduler	Weight Decay	Warmup Ratio	Batch Size	Seq Length
1.5×10^{-5}	AdamW	Cosine	1×10^{-3}	0.05	8	32,000

B.3 Probing Configuration

In our experiments, we conduct two sets of probing analyses: one on BIORISKEVAL-MUT using *continuous* LD_{50} labels, and another on bacteriophage sequences using *binary* lifestyle labels. For the regression task involving continuous LD_{50} labels, we employ a linear probe and compute the closed-form solution directly. For the binary classification task, we train a linear probe with a sigmoid activation function. We use a batch size of 128 and optimize the model using the Adam optimizer.

B.4 Evaluation Metrics

Spearman's rank correlation (ρ) When evaluating the model's capability in predicting mutational effect, we use Spearman's rank correlation ρ as our metric. In DMS dataset, each mutant i has a corresponding DMS value X_i, based on which we can have a "groundtruth" rank $R(X)$. Meanwhile, for each mutant sequence, we can also compute a model-derived (e.g., log probabilities) score Y_i, based on which we can have another rank $R(Y)$. Spearman's rank correlation ρ is computed as

$$\rho = \frac{\text{Cov}\left[R(X), R(Y)\right]}{\sigma_{R(X)}\sigma_{R(Y)}}, \tag{1}$$

where $\text{Cov}\left[R(X), R(Y)\right]$ is the covariance of rank variables; $\sigma_{R(X)}$ and $\sigma_{R(X)}$ are the standard deviations of the rank variables.

Pearson Coefficient When probing model's capability in predicting virulence, we use Pearson Coefficient to evaluate the correlation between the groundtruth LD_{50} value X and the predicted value Y. The Pearson Correlation is computed as

$$\text{Pearson} = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}, \tag{2}$$

where $\text{Cov}[X, Y]$ is the covariance between X and Y; σ_X and σ_Y are standard deviation.

C Evo 2 representation Analysis

In this section, we discuss the architecture and Evo 2 and explain why we observe extremely large representations in the later layers, but can still get reasonable output. Evo 2 uses StripedHyena2 [28]

architecture, which builds blocks from input-dependent long convolutions. A single Hyena operator creates three streams q, k, v by convolving linear projects of the residual stream x with Toeplitz filters T, H, K, then mixes them with inner convolution G and a final projection M to get the output y:

$$
\begin{aligned}
q_t^\alpha &= T_{tt'}^\alpha \left(x_{t'}^\beta W^{\beta\alpha} \right), \\
k_t^\alpha &= H_{tt'}^\alpha \left(x_{t'}^\beta U^{\beta\alpha} \right), \\
v_t^\alpha &= K_{tt'}^\alpha \left(x_{t'}^\beta P^{\beta\alpha} \right), \\
y_t^\alpha &= \left(q_t^\beta \, G_{tt'}^\beta \, k_{t'}^\beta \, v_{t'}^\beta \right) M^{\beta\alpha}.
\end{aligned}
\tag{3}
$$

StripedHyena-2 uses three Hyena blocks with different filter parameterizations:

- **Hyena-SE (Short Explicit)**: A short causal depth-wise 1D convolution with explicitly stored tapes per channel. It captures local n-gram-like patterns;
- **Hyena-MR (Medium regularized)**: A medium-length causal filter whose kernel is regularized to keep the frequency response and overall gain controlled.
- **Hyena-LI (Long Implicit)**: A long-range filter realized implicitly as a small mixture of exponentials evaluated with a stateful recurrence.

In Evo2, the layer type is organized following the order of SE-MR-LI, with Rotary Attention Layer inserted in Layer 3, 10, 17, 24, 31. There are a few factors that contribute to the representation explosion in the deeper layers:

- **Missing Layer Normalization on Residual Stream.** There's no explicit layer normalization on the residual stream after adding the block output back. Therefore, if the output of one layer has a slightly larger magnitude than its input and is fed directly into the next, the next layer can then amplify this magnitude further. Over dozens of layers, this may lead to an exponential growth in the values.
- **Input-Dependent and Gated Convolutions.** According to Equation (3), the filters in Hyena operators are generated from the input sequence itself. Therefore, if the input x already has a large magnitude, the gated filter will also likely have a large magnitude. The final output y is a product of these two large-magnitude tensors, leading to a quadratic increase in magnitude within a single layer.

On the other hand, the RMSNorm in the last layer rescales the representations back before the logits, so the output quality is dominated by the direction instead of the representation magnitude. Therefore, even though the hidden-layer representation explodes in the deeper layers, the model is still able to output meaningful sequences.

To double-check that this phenomenon is not due to the issue from the inference pipeline, we also test the layer-wise representation magnitude distribution using Vortex pipeline[†], which is the official inference pipeline used by the Evo2 paper, but still get the same result. This indicates that the representation explosion comes from the model architecture, instead of the inference pipeline.

[†] https://github.com/Zymrael/vortex