SafeBench-Seq: A Homology-Clustered, CPU-Only Baseline for Protein Hazard Screening with Physicochemical/Composition Features and Cluster-Aware Confidence Intervals

Muhammad Haris Khan
University of Copenhagen
muhammad.kahn@di.ku.dk

Abstract

Foundation models for protein design raise concrete biosecurity risks, yet the community lacks a simple, reproducible baseline for sequence-level hazard screening that is explicitly evaluated under homology control and runs on commodity CPUs. We introduce *SafeBench-Seq*, a metadata-only, reproducible benchmark and baseline classifier built entirely from public data (SafeProtein hazards and UniProt benigns) and interpretable features (global physicochemical descriptors and amino-acid composition). To approximate "never-before-seen" threats, we homology-cluster the combined dataset at $\leq$40% identity and perform *cluster-level holdouts* (no cluster overlap between train/test). We report discrimination (AUROC/AUPRC) and screening-operating points (TPR@1% FPR; FPR@95% TPR) with 95% bootstrap confidence intervals (n=200), and we provide calibrated probabilities via `CalibratedClassifierCV` (isotonic for Logistic Regression / Random Forest; Platt sigmoid for Linear SVM). We quantify probability quality using Brier score, Expected Calibration Error (ECE; 15 bins), and reliability diagrams. Shortcut susceptibility is probed via composition-preserving residue shuffles and length-/composition-only ablations. Empirically, random splits substantially overestimate robustness relative to homology-clustered evaluation; calibrated linear models exhibit comparatively good calibration, while tree ensembles retain slightly higher Brier/ECE. SafeBench-Seq is CPU-only, reproducible, and releases *metadata only* (accessions, cluster IDs, split labels), enabling rigorous evaluation without distributing hazardous sequences. **Code & metadata:** https://github.com/HARISKHAN-1729/SafeBench-Seq

1 Introduction

Generative AI (GenAI) for protein design have dramatically increased the ability to create novel biomolecules, but they also raise urgent biosecurity concerns. Recent work highlights that the lack of systematic "red-teaming" of protein foundation models enables potential misuse, including the de novo design of hazardous proteins [6], [8]. For example, experts warn that DNA synthesis screening – which relies on homology to known toxins – can be evaded by AI-generated "synthetic homologs" of dangerous proteins [8], [9]. Efforts by Wittmann et al. have experimentally demonstrated that generative pipelines can reformulate toxic protein sequences to bypass standard screens [9]. These findings underscore that protein generative models already pose a biosecurity risk, making effective screening and defense strategies critical.

Despite this risk, there is a striking lack of simple, reproducible baselines for sequence-level protein hazard filtering. Existing pipelines for nucleic acid orders typically use BLAST or related alignments against curated hazard databases [5], but no lightweight open-source classifiers have been made

available specifically for protein-level screening. Most prior tools focus on specialized toxin prediction (e.g., venom, peptide toxins) or require heavy GPU models. In contrast, our goal is to provide a tractable CPU-only baseline hazard detector for arbitrary protein sequences, built entirely from public data and executable on standard platforms (e.g., Colab) under strict safety constraints. We deliberately avoid releasing any novel sequences (only metadata such as accession IDs and taxonomic labels) and provide only aggregate statistics, to mitigate any risk of misuse of our benchmark data.

A key aspect of realistic hazard screening is robust evaluation. We emphasize homology-aware splits and stringent metrics to avoid overly optimistic results. In particular, we follow recommendations to cluster by sequence identity and hold out *homology clusters* ($\leq$40% identity) to simulate "never-before-seen" threats, rather than random splits [5],[16]. We also measure performance at extreme operating points (e.g., TPR at 1% FPR, FPR at 95% TPR) to reflect the high-consequence setting where false positives must be rare. Finally, we probe models for shortcut biases: length and taxonomy can trivially distinguish many known toxins; signal peptides are left for future work.

In this work, we introduce *SafeBench-Seq*, which is, to our knowledge, a tractable, CPU-only baseline evaluated under homology-clustered splits with explicit spurious-signal probes. Our baseline uses only open data (SafeProtein's hazard sequences and UniProt) and simple ML models (logistic regression, SVM, random forest) with calibrated probability outputs. By anchoring evaluation in realistic homology splits and spurious-signal probing, we aim to set a clear baseline for future sequence-based biohazard detectors, complementing the rich set of toxin-focused classifiers and protein-design red-teaming efforts in the literature.

2 Related Work

Protein hazard screening has long relied on sequence alignment pipelines, such as BLAST against curated toxin/pathogen databases, as recommended by the International Gene Synthesis Consortium (IGSC) [5]. While effective for known sequences, such methods are vulnerable to "synthetic homologs" that evade strict identity thresholds [5, 8].

To address these gaps, machine learning approaches have emerged. Early classifiers like ToxinPred [14] and ClanTox [11] used curated motifs and physico-chemical descriptors. More recent methods leverage deep learning and large-scale data, e.g., TOXIFY [4] on venom proteins, MultiToxPred [2] with gradient-boosted ensembles over multiple toxin classes, and ToxDL 2.0 [16] integrating pretrained LMs with structure-aware features. While powerful, many prior evaluations rely on random splits and GPU-heavy stacks, which may overestimate robustness against novel hazards.

In parallel, benchmarks like TAPE [12] and large models such as ESM [13] have transformed protein ML, enabling diffusion-based design [15] and raising security concerns. Commentaries warn that GenAI-driven protein design may render homology-based screening obsolete [8], prompting proposals for artificial sequence databases as defensive infrastructure [8, 1]. Complementary efforts include watermarking [15, 3] and unlearning [10]. Our contribution is a simple, interpretable, CPU-executable baseline evaluated under homology control, aligned with the SafeProtein ethos [5, 16].

3 Methods

Data Curation and Filtering

Positives (hazards). We use *protein toxins* from SafeProtein [6], explicitly excluding viral proteins to focus the positive class on toxins. **Negatives (benign).** We sample UniProtKB (2024 release) entries that (i) do not contain the "Toxin [KW-0800]" keyword and (ii) are non-viral. **Sequence inclusion criteria.** We require only canonical amino acids and lengths in $[30, 1000]$ residues. We remove exact duplicates (100% identity). **Length matching.** To blunt trivial length cues, we downsample the benign pool to match the positive class length distribution (quantile bin matching). This equalizes marginal length statistics (e.g., similar medians and ranges). **Safety posture.** We never release raw amino acid sequences to prevent potential misuse of hazardous proteins in our positive class. Our released artifacts contain only metadata (accession IDs, sequence lengths, labels, data sources, CD-HIT cluster IDs, and train/test split assignments) sufficient for result reproduction.

Homology-Clustered Splits

We homology-cluster the *combined* dataset (positives + negatives) at $\leq$40% identity using CD-HIT [7], so sequences placed in different clusters share $\leq 40\%$ identity. We evaluate under two protocols:

Cluster split—We perform cluster-level stratified sampling to ensure both hazardous and benign proteins are represented in train/test partitions. Each cluster receives a majority label based on its most frequent constituent class, then we randomly assign 80% of clusters (477/597) to training and 20% (120/597) to testing while maintaining stratification by majority label. No sequences from the same cluster appear in both partitions. Because cluster sizes are highly variable (ranging from 1 to n sequences per cluster), this cluster-wise 80/20 split produces a sequence-wise 83/17 distribution (708 train, 146 test sequences).

Random split—a *sequence-level* split that ignores clusters ($\approx 80/20$: 683 train / 171 test).

This design reduces homology leakage and shortcut learning and better reflects screening of new designs [16]. In aggregate, SafeBench-Seq contains 854 examples (427 harmful / 427 benign) forming 597 clusters; 477 clusters (80%) are assigned to train and 120 (20%) to test in the cluster split.

Features and Models

Handcrafted feature blocks. (1) *20-aa composition:* for sequence s of length L, composition is $c(a) = \frac{\#(a \in s)}{L}$ for $a \in \{A, C, D, \ldots, Y\}$. (2) *Global physicochemical descriptors (ProtParam):* length L; molecular weight (MW); isoelectric point (pI); GRAVY (grand average hydropathy); aromaticity; instability index; an aliphatic-index heuristic; net charge at pH 7.0. The (Ikai) aliphatic index is approximated as $100 \cdot (x_A + 2.9x_V + 3.1x_I + 3.9x_L)$ where $x.$ are residue fractions. These descriptors summarize coarse biochemical tendencies without learned embeddings.

Classifiers (CPU-only). We train L2-regularized Logistic Regression, a Linear SVM, and a Random Forest (400 trees) using scikit-learn. Linear models are wrapped in a pipeline with median imputation and standardization. Hyperparameters follow the CPU-only baseline used in our reference implementation (e.g., Logistic Regression C=0.5; Linear SVM C=1.0; RF class_weight balanced_subsample). Random seeds are fixed (1337) for splits, model fitting, and bootstrap.

Probability calibration. We apply `CalibratedClassifierCV` (5-fold on the training partition) to obtain well-behaved probabilities: *isotonic* for Logistic Regression and Random Forest, and *Platt* (sigmoid) for the Linear SVM. The same protocol is used for both the cluster split and the random split. All operating-point metrics use these calibrated probabilities.

Evaluation Metrics and Uncertainty

Discrimination. AUROC and AUPRC computed on the held-out test set. **Operating points.** *TPR@1%FPR* (sensitivity at 1% false-positive rate) and *FPR@95%TPR* (false-positive rate when sensitivity is 95%). For each, we select the left-most threshold that achieves the target constraint on the empirical ROC curve. **Calibration.** We report *Brier score* and *Expected Calibration Error* (ECE) with 15 equal-width bins on $[0, 1]$. Reliability diagrams plot the empirical fraction positive vs. predicted probability per bin, overlaid with the identity line. **Uncertainty quantification.** We use test-set bootstrapping (n=200 iid resamples with replacement) and report 95% CIs via the 2.5/97.5 percentiles, skipping degenerate resamples lacking both classes. Unless noted otherwise, the cluster split is our primary evaluation, and the random split is reported for context.

Spurious-Signal Probes and Subgroup Analysis

Probes. (i) *Composition-preserving residue shuffle* on test sequences: per sequence, residues are randomly permuted (deterministic seed by accession), preserving length and overall composition but destroying motif order; (ii) *length-only* and *composition-only* ablations trained and evaluated under the same split protocol. Retaining high performance under these probes would indicate heavy reliance on trivial cues. **Subgroups.** We report AUROC/AUPRC across: (a) sequence-length bins (quantiles), (b) toxin clusters/families (for positives, when sufficient support exists), and (c) superkingdom (Bacteria/Archaea/Eukaryota) for negatives. We require sufficient sample size ($\geq$15) and label variability within a subgroup to report metrics.

4 The SafeBench-Seq Benchmark

Construction and Safety

SafeBench-Seq draws hazards from SafeProtein-Bench (*protein toxins only; viral proteins excluded*) [6] and benigns from UniProt after strict exclusion filters (no KW-0800, non-viral). We keep length 30–1000 aa, restrict to canonical amino acids, deduplicate exact matches, and length-match benigns to positives to blunt trivial cues. The core artifact is a *metadata-only* CSV listing accession, label, length, source, CD-HIT cluster ID, and split assignments (random and cluster). Users can reproduce results by fetching sequences from public sources via accession; we do not distribute raw sequences.

Splits and Counts

Random split. ≈80/20 by sequence: 683 train / 171 test. **Cluster split.** CD-HIT clustering at $\leq 40\%$ identity [7]. We hold out clusters: 477 of 597 clusters (80%) in train; 120 (20%) in test. Because cluster sizes vary, the *sequence*-level counts are 708 train and 146 test ($\approx 83/17$). **Dataset size.** 854 sequences total, balanced (427 harmful / 427 benign). **Length distribution.** Classes are closely matched by design (e.g., median $\approx$ 246 aa; mean $\approx$ 290 aa). See Fig. 1.

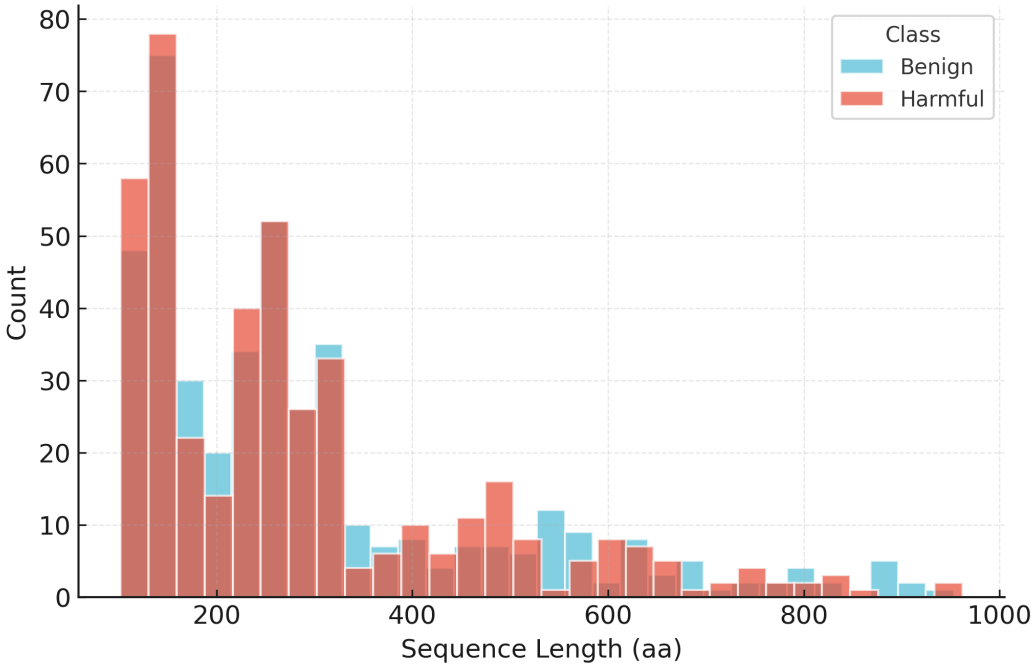

Figure 1: Sequence length distributions for harmful (red) vs. benign (blue-green) proteins in SafeBench-Seq. Lengths are closely matched by design, limiting trivial cues.

5 Results

Table 1 summarizes discrimination performance for three *calibrated* classifiers (Logistic Regression, Linear SVM with Platt, Random Forest with isotonic) using the *base* feature set on both splits. We report AUROC, AUPRC, and screening-operating points, with 95% bootstrap CIs (n=200). Random-split performance is high (e.g., RF AUROC ≈0.95), but the cluster split reduces AUROC by roughly 4–8 points and sharply depresses tail performance (TPR@1%FPR drops from ≈0.27 to ≈0.12). This gap indicates that homology-aware evaluation better reflects the challenge of screening *novel* homology groups and that random splits overestimate robustness. Across both splits, Random Forest leads the linear baselines, though margins narrow under the cluster split. Operating-point metrics show wider CIs as expected for imbalanced data and extreme thresholds.

Table 1: Discrimination performance (calibrated models) on random and cluster splits using the *physicochemical and composition** feature set. Metrics with 95% CIs from bootstrap (n=200); best values per split are in **bold**.

Split	Model	AUROC	AUPRC	TPR@1%FPR ↑	FPR@95%TPR ↓
Random	LogReg (*)	0.901 [0.849–0.944]	0.894 [0.820–0.955]	0.198 [0.134–0.470]	0.459 [0.145–1.000]
	LinSVM (*)	0.898 [0.842–0.945]	0.896 [0.833–0.953]	0.151 [0.103–0.449]	0.435 [0.141–0.946]
	RF (*)	**0.953 [0.917–0.979]**	**0.944 [0.883–0.983]**	**0.267 [0.200–0.844]**	**0.353 [0.056–0.481]**
Cluster	LogReg (*)	0.871 [0.803–0.925]	0.851 [0.772–0.917]	0.095 [0.050–0.272]	0.601 [0.250–1.000]
	LinSVM (*)	0.865 [0.798–0.918]	0.842 [0.768–0.910]	0.083 [0.047–0.245]	0.629 [0.262–0.983]
	RF (*)	**0.919 [0.866–0.961]**	**0.905 [0.832–0.964]**	**0.121 [0.067–0.296]**	**0.512 [0.183–0.892]**

Table 2: Calibration metrics on the cluster split (base features). 95% CIs via bootstrap (n=200). AUROC/AUPRC appear in Table 1.

Model	Brier ↓	ECE ↓
LinSVM	0.143 [0.112, 0.181]	0.111 [0.108, 0.195]
LogReg	0.140 [0.104, 0.183]	0.118 [0.106, 0.192]
RF	0.156 [0.115, 0.204]	0.138 [0.105, 0.204]

5.1 Calibration

We calibrate probabilities with `CalibratedClassifierCV` (isotonic for LogReg/RF; Platt for LinSVM). Beyond discrimination, Table 2 reports Brier and ECE on the cluster split with *base* features; Fig. 2 shows reliability diagrams. Linear models are comparatively well-calibrated post hoc; Random Forest exhibits slightly higher residual miscalibration (higher Brier/ECE), though calibration still reduces deviations from the diagonal.

5.2 Spurious-Signal Probes

Length-only and composition-only features yield near-random AUROC (≈ 0.50˘0.55), indicating trivial cues were largely equalized. A composition-preserving residue shuffle of test sequences (randomizing residue order per sequence) degrades performance further, showing that models leverage motif/ordering information rather than bulk composition alone. These probes support that the screen captures sequence-level hazard signals, not dataset artifacts.

5.3 Subgroup Performance

We break down performance by sequence length, toxin family (positives), and superkingdom (negatives). Fig. 3–4 show AUROC/AUPRC across length bins: performance peaks for mid-length proteins (200–300 aa) and degrades for very long sequences (>400 aa), especially under the cluster split—consistent with reduced generalization to multi-domain proteins. Family- and taxonomy-level results (Appendix) show heterogeneity: some toxin clusters are near-perfectly separated while others cluster around AUROC ≈0.80; bacterial negatives are generally easier than eukaryotic ones.

Takeaways. (1) Homology-clustered evaluation surfaces a substantial robustness gap vs. random splits—critical for screening novel sequences. (2) Post-hoc calibration improves probability quality; linear models are comparatively well-behaved, while tree ensembles retain slightly higher ECE/Brier. (3) Shortcut probes and subgroup analyses suggest residual failure modes at extreme lengths and specific clusters; these are natural targets for future structure-aware or family-aware methods.

6 Implementation Details and Reproducibility

CPU-only stack. All models are implemented with scikit-learn and run on CPU. Feature extraction uses Biopython's `ProtParam`. Homology clustering uses CD-HIT (we used 4.8.1 in our reference run). The reference notebook (e.g., Colab) contains package installs; no GPUs or proprietary software are required.

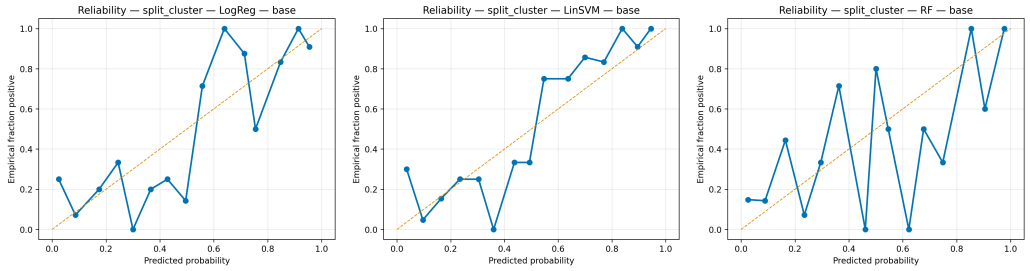

Figure 2: Reliability diagrams on the cluster split (base features). Dashed line: perfect calibration.

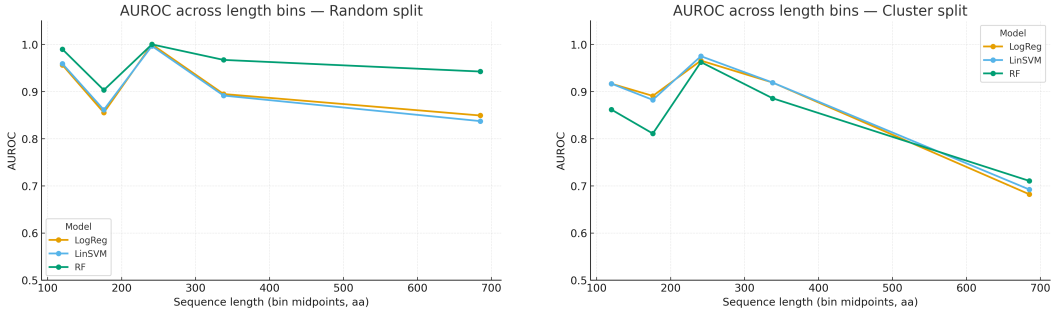

Figure 3: AUROC across sequence-length bins (left: random split; right: cluster split). Mid-lengths are easiest; very long proteins degrade under homology control.

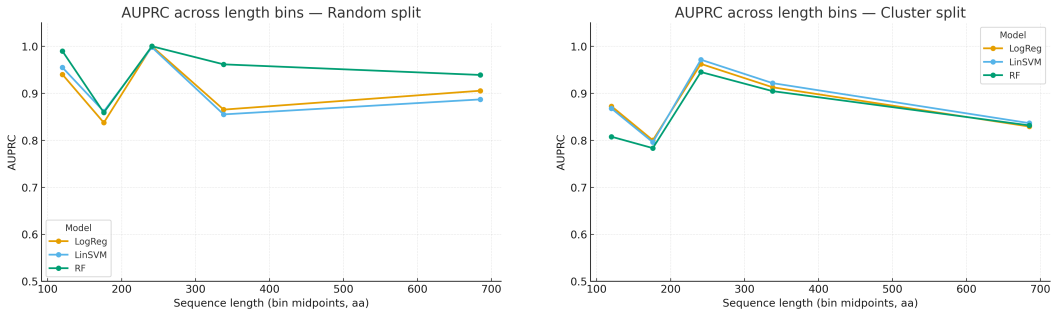

Figure 4: AUPRC across sequence-length bins mirrors AUROC: mid-lengths peak; long proteins suffer under the cluster split.

Determinism. We fix the Python/NumPy RNG seed to 1337 for splitting, model fitting, and bootstrap resampling. Composition-preserving shuffles use per-accession deterministic seeds so shuffled probes are reproducible across runs.

Bootstrap details. We use stratified bootstrap resampling ($n = 200$) to ensure both classes are represented in each resample. For each bootstrap iteration, we independently sample with replacement from the positive and negative test examples, maintaining the original class balance. We report 2.5/97.5 percentiles as 95% CIs.

Operating-point computation. On the empirical ROC $(\mathrm{FPR}(t), \mathrm{TPR}(t))$, we select the left-most threshold t such that $\mathrm{FPR}(t) \geq 0.01$ for TPR@1%FPR (and analogously the left-most threshold with $\mathrm{TPR}(t) \geq 0.95$ for FPR@95%TPR).

Calibration curves. Reliability plots use 15 equal-width probability bins with mean predicted probability on the x-axis and empirical fraction positive on the y-axis; ECE averages bin-wise absolute gaps weighted by bin mass.

Safety checks. All artifacts (tables, plots) are aggregate and do not reveal raw sequences. The released CSV is metadata-only (accessions, cluster IDs, split labels, etc.).

7 Ethics and Safety Statement

This work is intended to *improve* defensive capabilities by providing a transparent, reproducible baseline for screening, not to enable misuse. We release metadata only, avoid any design or release of novel sequences, and emphasize homology-aware evaluation that better reflects realistic risk surfaces. Our analyses, figures, and tables aggregate results without exposing hazardous sequences, aligning with the safety posture articulated by the SafeProtein initiative [6] and biosecurity commentaries [8].

8 Conclusion

We presented SafeBench-Seq, a tractable, CPU-only baseline and metadata-only benchmark for protein hazard screening that foregrounds homology-aware evaluation, calibrated probabilities, and shortcut probes. By clustering at $\leq$40% identity and holding out entire clusters, we approximate "never-before-seen" homology groups and quantify discrimination at screening-relevant operating points with calibrated confidence. Empirically, random splits overestimate robustness relative to cluster holdouts; post-hoc calibration improves probability quality (Brier/ECE), particularly for linear models. Limitations include: (i) no explicit signal-peptide covariate (left for future work), (ii) sequence-only features (no structure/context), and (iii) moderate dataset size. Future directions include integrating lightweight structure proxies, explicit SP/N-terminal heuristics, and defense-in-depth (e.g., calibrated ensembling and risk-aware thresholds) while maintaining the same safety posture (metadata-only release). We hope SafeBench-Seq serves as a transparent reference point for evaluating sequence-level biohazard detectors.

9 Limitations

Our approach has several constraints: (i) moderate dataset size (854 sequences) may limit statistical power for rare toxin families; (ii) sequence-only features ignore 3D structural information; (iii) we do not explicitly model signal peptides or subcellular localization signals that may be relevant for toxin function; (iv) performance at extreme operating points shows wide confidence intervals, reflecting the challenges of tail-event prediction in imbalanced settings.

References

[1] David Baker and George Church. Protein design meets biosecurity. *Science*, 383:349–350, 2024. doi: 10.1126/science.ado1671.

[2] Jorge F Beltrán, Lisandra Herrera-Belén, Fernanda Parraguez-Contreras, Jorge G Farías, Jorge Machuca-Sepúlveda, and Stefania Short. MultiToxPred 1.0: a novel comprehensive tool for predicting 27 classes of protein toxins using an ensemble machine learning approach. *BMC Bioinformatics*, 25:148, 2024. doi: 10.1186/s12859-024-05748-z.

[3] Yanshuo Chen, Zhengmian Hu, Yihan Wu, Ruibo Chen, Yongrui Jin, Marcus Zhan, Chengjin Xie, Wei Chen, and Heng Huang. Enhancing Privacy in Biosecurity with Watermarked Protein Design. *Bioinformatics*, 41(7):btaf141, 2025. doi: 10.1093/bioinformatics/btaf141.

[4] Patrick D Dobson, Narayanan Ramakrishnan, Vanessa A Capraro, Jerome Vamathevan, Randall Cook, Patricia C Babbitt, and John L Todd. TOXIFY: a deep learning approach to classify animal venom proteins. *PeerJ Computer Science*, 5:e172, 2019. doi: 10.7717/peerj-cs.172.

[5] Dr. Edward Perkins Editors: Benjamin D. Trump, Marie-Valentine Florin and Dr. Igor Linkov. *Emerging Threats of Synthetic Biology and Biotechnology: Addressing Security and Resilience Issues*. National Center for Biotechnology Information (NCBI). URL `https://www.ncbi.nlm.nih.gov/books/NBK584258/`. Accessed: 2025-09-10.

[6] Jigang Fan, Zhenghong Zhou, Ruofan Jin, Le Cong, Mengdi Wang, and Zaixi Zhang. SafeProtein: Red-Teaming Framework and Benchmark for Protein Foundation Models. *arXiv preprint arXiv:2509.03487*, 2025.

[7] Yinping Huang, Bin Niu, Ying Gao, Li Fu, and Weizhong Li. Cd-hit suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010. doi: 10.1093/bioinformatics/btq003. Epub 2010 Jan 6.

[8] Philip Hunter. Security challenges by AI-assisted protein design: The ability to design proteins in silico could pose a new threat for biosecurity and biosafety. *EMBO Reports*, 25(5):2168–2171, 2024. doi: 10.1038/s44319-024-00124-7.

[9] Svetlana Ikonomova, Bruce Wittmann, Fernanda P. Macruz de Oliveira, David Ross, Samuel Schaffter, Olga Vasilyeva, Elizabeth Strychalski, Eric Horvitz, James Diggans, Sheng Lin-Gibson, and Geoffrey Taghon. Experimental Evaluation of AI-Driven Protein Design Risks Using Safe Biological Proxies. *Science*, 2025.

[10] Mingchen Li, Bingxin Zhou, Yang Tan, and Liang Hong. Unlearning virus knowledge toward safe and responsible mutation effect predictions. *bioRxiv*, 2024. doi: 10.1101/2024.10.02.616274. URL https://www.biorxiv.org/content/early/2024/10/03/2024.10.02.616274.

[11] Guy Naamati, Manor Askenazi, and Michal Linial. Clantox: a classifier of short animal toxins. *Nucleic acids research*, 37(suppl_2):W363–W368, 2009.

[12] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/37f65c068b7723cd7809ee2d31d7861c-Paper.pdf.

[13] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

[14] Neelam Sharma, Leimarembi Devi Naorem, Shipra Jain, and G. P. S. Raghava. ToxinPred2: an improved method for predicting toxicity of proteins. *Briefings in Bioinformatics*, 23(5), 2022. doi: 10.1093/bib/bbac174.

[15] Zaixi Zhang, Ruofan Jin, Kaidi Fu, Le Cong, Marinka Zitnik, and Mengdi Wang. Foldmark: Protecting protein generative models with watermarking, 2024. URL https://arxiv.org/abs/2410.20354.

[16] Lin Zhu, Yi Fang, Shuting Liu, Hong-Bin Shen, Wesley De Neve, and Xiaoyong Pan. ToxDL 2.0: Protein toxicity prediction using a pretrained language model and graph neural networks. *Computational and Structural Biotechnology Journal*, 27:1538–1549, 2025. doi: 10.1016/j.csbj.2025.04.002.