# CRC-Screen: Certified DNA-Synthesis Hazard Screening Under Taxonomic Shift

**Najmul Hasan** [1]

## Abstract

DNA-synthesis providers screen incoming orders by searching the requested sequence against curated hazard lists. We show that this baseline collapses to a 100% false-flag rate when the hazardous sequence comes from a taxonomic family absent from the reference set: under Conformal Risk Control's certified miss-rate constraint, a low-discrimination signal forces the threshold below the entire test-benign mass. We compose three signals derived from a synthesis order's public annotation: $k$-mer Jaccard similarity to known toxins, the trimmed-mean score of a five-LLM judge panel, and cosine similarity to clustered embedding centroids. Fused under a monotone logistic aggregator and calibrated by Conformal Risk Control, the resulting screener certifies $\mathbb{E}[\mathrm{FNR}] \leq \alpha + \mathrm{TV}$, where the additive term is the calibration-to-test distribution shift under family holdout (a certified ceiling of 24–49% across folds). Across ten leave-one-taxonomic-family-out folds at $\alpha = 0.05$ on UniProt KW-0800 reviewed toxins, the calibrated screener achieves 0% *empirical* test miss rate on every fold and 0% test false-flag rate on nine of ten folds. The bound's finite-sample slack $1/(n_{\mathrm{cal}} + 1)$ caps the certifiable miss rate at 1.77% on our 200-hazard subsample; reaching procurement-grade $\alpha = 10^{-3}$ requires an 18× larger calibration set, which the full reviewed UniProt KW-0800 corpus is large enough to deliver. The binding constraint on certifiable DNA-synthesis screening is calibration data, not algorithms. *Code:* https://github.com/najmulhasan-code/crc-screen

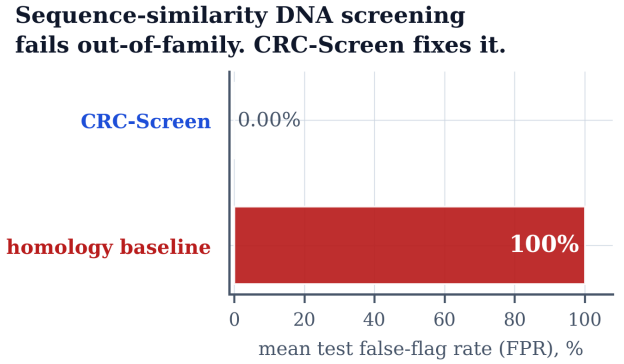


*Figure 1.* Sequence-similarity-only screening flags every benign in out-of-family folds (100% FPR); CRC-Screen drops this to 0% while certifying $\mathbb{E}[\mathrm{FNR}] \leq \alpha + \mathrm{TV}$ at $\alpha = 0.05$, mean across ten leave-one-taxonomic-family-out folds.

## 1. Introduction

DNA-synthesis providers are the last enforcement point before a hazardous protein is built: an order for such a protein can in principle be intercepted between the customer's design and the synthesised molecule. The standard implementation of that bottleneck is a sequence-similarity search against curated hazard lists derived from regulatory inventories and toxin databases. This baseline was built for a threat model in which the hazardous order looks, at the level of amino-acid sequence, like a hazardous protein the screener has already seen. Two trends now stretch that assumption. Generative models for protein design (Madani et al., 2023; Lin et al., 2023) produce variants that retain function while drifting in primary sequence, and red-team studies have begun to evaluate whether language-model assistance offers operational uplift to non-state actors planning biological attacks, finding no significant uplift from current models but flagging trajectory risk for future systems (Mouton et al., 2024). Public toxin databases are fragmented across specialised resources (Jungo & Bairoch, 2005; Kaas et al., 2012), with hazardous proteins from understudied taxonomic families sparsely represented. Both trends shift weight onto the out-of-family case, which is precisely the case in which a sequence-similarity baseline weakens.

We make this concrete with a leave-one-taxonomic-family-out evaluation on UniProt KW-0800 reviewed toxins. Held out one family at a time, the sequence-similarity signal is too weak to separate hazards from benigns; Conformal Risk Control, forced to certify a miss-rate ceiling, has no choice

[1]University of North Carolina at Pembroke.
Accepted at the 6th Muslims in ML (MusIML) Workshop at ICML 2026.

but to push its threshold so low that every test benign is flagged. The result is a flag-everything regime: 100% test FPR on every fold (Figure 1, red bar).

The fix we study is composition. The same public annotation that accompanies a synthesis order, comprising name, organism, controlled-vocabulary keywords and a free-text function description, makes two further signals available: a five-LLM panel reads the annotation and returns a hazard probability, and the text-embedding distance to clustered embeddings of known toxins gives a smooth proxy for functional proximity. Composing the three signals under a monotone logistic aggregator and then calibrating the decision threshold by Conformal Risk Control (Angelopoulos et al., 2024) restores certified $\mathbb{E}[\mathrm{FNR}] \le \alpha + \mathrm{TV}$, and on our evaluation the calibrated screener achieves 0% empirical test miss rate on every leave-one-family-out fold at $\alpha = 0.05$, with 0% test false-flag rate on nine of ten folds and one flagged benign on Actiniidae. A signal-by-signal ablation shows the LLM panel and the embedding signal are jointly sufficient; adding sequence homology back to that pair raises mean FPR by half a point with no recall gain.

The paper contributes four results. First, under taxonomic-family holdout, $k$-mer-Jaccard sequence-similarity screening incurs a 100% false-flag rate at any non-trivial $\alpha$, an empirical consequence of Conformal Risk Control's coverage requirement on a low-discrimination signal that does not close under $\alpha$-tuning. Second, a leak-controlled per-fold protocol (Section 3.3) prevents the held-out family's hazards from leaking into their own scoring through the homology and embedding reference sets. Third, an off-the-shelf composition of $k$-mer Jaccard, a five-LLM panel with trimmed-mean aggregation, and an embedding-centroid distance, fused by a monotone logistic aggregator and calibrated by Conformal Risk Control, certifies $\mathbb{E}[\mathrm{FNR}] \le \alpha + \mathrm{TV}$ at $\alpha = 0.05$ with 0% empirical miss rate on every fold and 0% empirical false-flag rate on nine of ten leave-one-taxonomic-family-out folds. Fourth, the data budget that decides what $\alpha$ is reachable: at $n_{\mathrm{cal\,haz}} \approx 55$ the slack term $1/(n_{\mathrm{cal\,haz}} + 1)$ floors the certifiable $\alpha$ at 1.77%, and procurement-grade $\alpha = 10^{-3}$ requires $n_{\mathrm{cal\,haz}} \ge 999$, an 18× gap that the full reviewed UniProt KW-0800 corpus has the size to close.

## 2. Conformal Risk Control and the screening status quo

### 2.1. From coverage to risk

Conformal prediction (Vovk et al., 2022; Angelopoulos & Bates, 2023) converts any black-box predictor into a procedure with finite-sample coverage guarantees by calibrating a threshold on a held-out exchangeable calibration set, with distribution-free regression and classification specialisations now standard (Lei et al., 2018; Romano et al., 2019; Sadinle et al., 2019; Cauchois et al., 2021). The classical guarantee is on miscoverage: the constructed prediction set covers the truth with probability at least $1 - \alpha$. Risk-control extensions move the guarantee from a coverage event to a bounded loss (Bates et al., 2021), and Conformal Risk Control (Angelopoulos et al., 2024) in particular generalises miscoverage to any monotone, bounded loss. Given calibration losses $L_1, \ldots, L_n$ that are non-decreasing in a real-valued threshold parameter $\tau$ and bounded above by $B$, the choice $\widehat{\tau} = \sup\{\tau : \widehat{R}(\tau) + B/(n+1) \le \alpha\}$ satisfies $\mathbb{E}[L_{n+1}(\widehat{\tau})] \le \alpha$ on a fresh exchangeable point, where $\widehat{R}$ is the empirical mean of the calibration losses (Equation (2)). This is the standard non-increasing form of CRC under the substitution $\lambda = -\tau$; we keep $\tau$ because the threshold is more natural for screening. For our screening setting, $L_i$ is the false-negative indicator on hazard $i$, which is non-decreasing in $\tau$ for any score $S$; we additionally constrain the aggregator to be non-decreasing in each underlying signal (Section 3.2) so that flag direction is consistent across signals.

When the calibration and test distributions are not exchangeable the guarantee picks up an additive correction. Theorem 2 of Barber et al. (2023) bounds the deviation by a sum of weighted total-variation terms over residual swaps, which generalises earlier weighted-conformal results for covariate shift (Tibshirani et al., 2019). We use a histogram TV between the calibration and test score distributions as a coarse approximation of that residual-swap quantity. The same calibration mindset underlies post-hoc score-rescaling for classifier outputs (Platt, 1999; Guo et al., 2017) and selective classification with abstention thresholds (Geifman & El-Yaniv, 2017), which sit adjacent to our setting. Under leave-one-taxonomic-family-out holdout we directly observe TV distances of 0.19 to 0.44 across folds, so the bound's slack is dominated by this distribution-shift term rather than by the finite-sample term $1/(n_{\mathrm{cal\,haz}} + 1)$ at the alpha levels we test.

### 2.2. What providers screen against

A DNA-synthesis order arrives at a provider as a sequence specification plus customer metadata. Before fulfilling the order, providers in the International Gene Synthesis Consortium (IGSC) and equivalents run a sequence-similarity search against curated lists of pathogen and toxin sequences, escalate flagged orders to human review, and sometimes require additional customer attestation (Carter & Friedman, 2015; Diggans & Leproust, 2019). The technical core of this screening is the same alignment search machinery used throughout computational biology, descended from BLAST (Altschul et al., 1990; 1997; Camacho et al., 2009) and its modern protein-scale successors such as DIAMOND (Buchfink et al., 2021), MMseqs2 (Steinegger & Söding, 2017), USEARCH (Edgar, 2010) and profile-HMM

tools (Eddy, 2011) indexed against family databases such as Pfam (Mistry et al., 2021). The policy and biosecurity literature documents two structural problems with this baseline: short fragments below the alignment-search sensitivity floor escape detection (Diggans & Leproust, 2019), and AI-designed protein variants that depart in primary sequence from training distributions can fall below the same thresholds while preserving function (Wittmann et al., 2025). Our system retains the homology signal as one of three inputs (Section 3.1) but does not rely on it for discrimination under taxonomic-family holdout.

## 3. Three signals, monotone fusion, calibrated threshold

A synthesis order arrives with a public UniProt annotation: an accession, the protein's name, the source organism, a controlled-vocabulary keyword list, and a free-text function description. From that annotation we derive three signals, fuse them with a monotone logistic aggregator, and pick the flag-versus-pass threshold by Conformal Risk Control. The whole pipeline, illustrated for one held-out family, is shown in Figure 2.

### 3.1. What each signal captures

Each signal captures a different sense in which an order may resemble a known hazard, and the resulting correlations are weak enough that a linear aggregator gains from all three.

**Homology signal** $s_{\mathrm{hom}}$**.** Sequence-similarity search against curated hazard lists is the standard tool of current synthesis screening, descended from BLAST (Altschul et al., 1990) and DIAMOND (Buchfink et al., 2021), with related index-based and clustering-based variants (Steinegger & Söding, 2017; Edgar, 2010; Suzek et al., 2015). Our implementation is minimal: a $k$-mer Jaccard similarity between the query sequence and each reference hazard, with $k=5$ amino acids; the per-query score is the maximum similarity over the reference set, then rank-normalised across the $n$=600 corpus to a value in $[0,1]$. Self-matches are excluded so that a hazard under evaluation does not match itself. The choice of Jaccard over a gapped alignment is conservative: it strips away the optimisations that would let a sequence-similarity baseline look stronger than it is, isolating the failure mode that motivates the rest of the system.

**LLM panel score** $s_{\mathrm{LLM}}$**.** A panel of five large language models reads the annotation text and returns a hazard probability in $[0,1]$. The five models are Claude Opus 4.7 (Anthropic), GPT-5.4 (OpenAI), Llama 4 Maverick (Meta), Qwen 3.6 Plus (Alibaba), and GLM 5.1 (ZAI), one per provider, so that systematic refusals or scoring biases are unlikely to align across the panel. Each model receives the same zero-shot prompt: a biosecurity-screening role, a four-level rubric tied to standard regulatory categories, and a strict JSON output schema; the prompt forbids generation of sequence data or synthesis instructions. Each (sample, model) pair is queried $k=2$ times at temperature 0.7, and the per-model score is the median of the two runs (which equals their mean at $k=2$). The panel score is the trimmed mean of the five per-model scores, dropping the lowest and the highest and averaging the three middle values. API failures, JSON-parse failures and out-of-range scores are filled with the neutral value 0.5 before the trim, which absorbs at most one such fallback on each side. The panel is an instance of the LLM-as-judge setup studied in Zheng et al. (2023) and developed for evaluation in Liu et al. (2023) and Dubois et al. (2023), differing in its aggregation rule and application.

**Embedding distance** $s_{\mathrm{emb}}$**.** Each annotation is rendered as a labelled key–value string (name, organism, keywords, function), passed through OpenAI's text-embedding-3-large model in the lineage of contextual sentence and passage encoders (Devlin et al., 2019; Reimers & Gurevych, 2019; Karpukhin et al., 2020), and $L_2$-normalised. Sequence-conditioned protein language models (Rives et al., 2021; Lin et al., 2023; Elnaggar et al., 2022) offer a complementary representation; we use a text encoder over the annotation rather than a sequence encoder over the protein because the order's annotation arrives long before its sequence is committed to synthesis. We then run $K$-means with $K=\min(8,\ \lfloor n_{\text{train haz}}/5 \rfloor)$ on the embeddings of the train-fold hazards alone, normalise the centroids, and define $s_{\mathrm{emb}}$ as the maximum cosine similarity between the query's embedding and any hazard centroid, clipped to $[0,1]$. Multiple centroids accommodate the multi-modality of the hazard pool: toxins from unrelated organisms occupy distant regions of the embedding space.

### 3.2. Why the aggregator must be monotone

Linear fusion of classifier outputs is standard (Dietterich, 2000; Caruana et al., 2004). We fuse the three signals with a logistic regression

$$S(\mathbf{s}) \;=\; \sigma\Big(\sum_{k\in\{\mathrm{hom},\mathrm{LLM},\mathrm{emb}\}} w_k\, s_k \;+\; b\Big), \qquad (1)$$

fit on the train-fold portion of each leave-one-family-out split, with the constraint $w_k \geq 0$ for every signal. With only non-negative weights the score $S$ is non-decreasing in every input signal, so increasing any one of $\{s_{\mathrm{hom}}, s_{\mathrm{LLM}}, s_{\mathrm{emb}}\}$ can only raise the flag probability, not lower it; a negative coefficient would invert that semantics for one signal and break the agreement that composition is supposed to enforce. Operationally, we fit an unconstrained logistic regression (Pedregosa et al., 2011), drop the signal whose coefficient

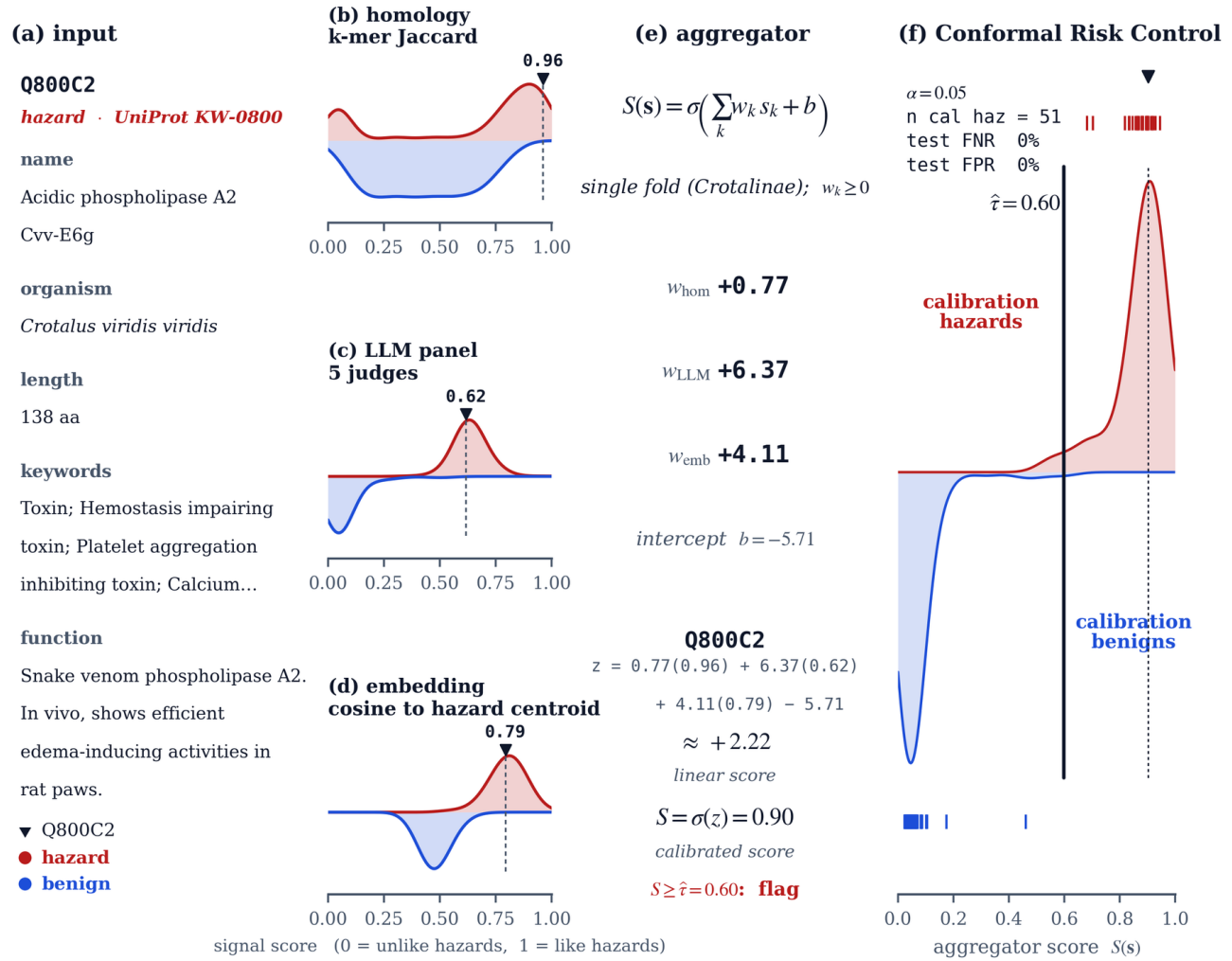


*Figure 2.* CRC-Screen takes a UniProt annotation through three signals, fuses them with a monotone logistic aggregator, and flags the order if the calibrated score $S$ exceeds the Conformal Risk Control threshold $\widehat{\tau}$; on the held-out *Crotalinae* fold, this correctly flags Q800C2 with test FNR$= 0\%$ and test FPR$= 0\%$. (a) Input UniProt record (accession Q800C2, an acidic phospholipase $A_2$ from *Crotalus viridis viridis*, labelled hazard via KW-0800). (b) to (d) Per-fold signal distributions across the $n$=600 corpus, with the example's value marked by an inverted triangle. (e) The aggregator applied to Q800C2: the substituted weights and signal values yield $z \approx +2.22$, so $S = \sigma(z) = 0.90$. (f) The CRC threshold chosen on the calibration densities: $\widehat{\tau} = 0.60$ with $n_{\text{cal haz}} = 51$. Since $S \geq \widehat{\tau}$, Q800C2 is flagged.

is most negative, and refit on the remaining signals; this repeats until every coefficient is non-negative, which always terminates because the empty model trivially satisfies the constraint. The intercept $b$ is unconstrained.

### 3.3. Calibrating the threshold without leaking the test family

Given calibration scores $\{S_i\}_{i \in \text{cal}}$ and labels $\{Y_i\}$, the per-sample false-negative loss at threshold $\tau$ is $L_i(\tau) = \mathbf{1}\{Y_i = 1,\ S_i < \tau\}$, which is non-decreasing and left-continuous in $\tau$ (equivalently, non-increasing and right-continuous in $\lambda = -\tau$, the canonical hypothesis of CRC Theorem 2.1). Conformal Risk Control (Angelopoulos et al., 2024) chooses

$$\widehat{\tau} \;=\; \sup\Big\{ \tau \;:\; \widehat{R}(\tau) + \tfrac{B}{n_{\text{cal haz}}+1} \;\leq\; \alpha \Big\}, \qquad (2)$$

where $\widehat{R}(\tau) = \frac{1}{n_{\text{cal haz}}+1} \sum_{i \in \text{cal haz}} L_i(\tau)$ is the empirical FNR on the $n_{\text{cal haz}}$ calibration hazards, and $B = 1$ bounds the loss. Because the false-negative loss is zero on benigns, the empirical mean is taken over hazards only; equivalently, this is the class-conditional (Mondrian) instance of CRC (Vovk et al., 2022) run on the hazard subset, with the guarantee $\mathbb{E}[\text{FNR}] \leq \alpha$ conditional on $Y_{n+1} = 1$. Theorem 2.1 of Angelopoulos et al. (2024) guarantees $\mathbb{E}[L_{n+1}(\widehat{\tau})] \leq \alpha$ when calibration and test points are exchangeable. Under taxonomic-family holdout that exchangeability is violated,

and the bound picks up an additive total variation term (Barber et al., 2023); we report the resulting full right-hand side

$$\mathbb{E}[\mathrm{FNR}] \;\le\; \alpha \;+\; \mathrm{TV}(\mathrm{cal}, \mathrm{test}), \tag{3}$$

estimating the TV term by histogram TV between calibration and test score distributions, a coarse but workable proxy.

**Per-fold leak control.** Both $s_{\mathrm{hom}}$ and $s_{\mathrm{emb}}$ are reference-set signals: a query's score depends on which hazards sit in the reference. If we computed them once over the full corpus and reused those values for every leave-one-family-out fold, the test family's hazards would influence the score of their own fold's queries through the reference set. We recompute $s_{\mathrm{hom}}$ and $s_{\mathrm{emb}}$ inside each fold, using *only* the train-fold hazards as the reference set; the cached global signals exist for inspection and are not consumed by the evaluation loop. The LLM panel score does not use a hazard reference set and is unchanged across folds.

# 4. Experiments

## 4.1. Corpus, splits, hyperparameters

**Corpus.** The hazard pool is UniProt KW-0800 (Toxin) restricted to the reviewed Swiss-Prot subset (The UniProt Consortium, 2025); this is a single keyword query against the canonical public knowledgebase, and the same keyword is the operational definition of "toxin" used by curators. The benign pool is the reviewed Swiss-Prot subset minus KW-0800 and minus KW-0843 (Virulence); the latter exclusion prevents virulence factors from being mislabelled benign during evaluation. We sample 200 hazards and 400 benigns ($n = 600$ total, fixed seed) so that calibration sees a 1:2 hazard-to-benign ratio, a departure from the $\ll 1\%$ deployment ratio, chosen because pure deployment-ratio sampling would leave so few hazards in the calibration set that the slack term $1/(n_{\mathrm{cal\,haz}} + 1)$ would dominate the bound. We address the gap between this calibration ratio and deployment ratios in Section 4.4.

**Splits.** Outer split: leave-one-taxonomic-family-out (LOTO), a leave-one-group-out variant of the cross-validatory choice principle (Stone, 1974), across the ten taxonomic families with at least five hazards in the sample (Crotalinae, Hydrophiinae, Elapinae, Buthidae, Conus, Theraphosidae, Sicariidae, Viperinae, Actiniidae, Lycosidae). For each held-out family, the test set is every hazard from that family plus a matched random sample of benigns at the corpus ratio. Inner split inside the non-test pool: stratified 70/30 train/calibration. Train fits the aggregator weights; calibration chooses $\widehat{\tau}$. Within a fold, the train, calibration, and test partitions are disjoint, and the reference set for $s_{\mathrm{hom}}$ and $s_{\mathrm{emb}}$ is restricted to train-fold hazards (Section 3.3).

**Hyperparameters.** $k = 5$ for the Jaccard signal, $k_{\mathrm{LLM}} = 2$ runs per (sample, model), temperature 0.7, $K \le 8$ centroids, $\alpha = 0.05$ unless stated otherwise. The aggregator is logistic regression with sklearn's default $L_2$ regularisation ($C = 1$) and the drop-and-refit non-negativity rule (Section 3.2). All experiments use a fixed random seed (Table 2).

## 4.2. The bound holds on every fold

Figure 3 shows the per-fold result at $\alpha = 0.05$, ordered by total-variation distance between the calibration and test distributions of $S$; Table 1 lists the same per-fold values, ordered by $n_{\mathrm{test\,haz}}$. Two findings:

**Empirical FNR is zero on every fold.** Across all ten folds the calibrated screener misses zero hazards out of 5 to 29 test hazards per family. Test FPR is also zero on nine of ten folds and 5% (one of twenty test benigns) on Actiniidae.

**The bound is loose by design.** The right-hand side of Equation (3) ranges from 24.3% on Crotalinae, where the cal/test TV is smallest, to 49.4% on Lycosidae, where it is largest. The bound is not vacuous: it certifies that the expected miss rate cannot exceed roughly half on any fold under the observed distribution shift.

## 4.3. Which signals carry the result

To isolate which signals contribute to the headline result we re-run the same per-fold protocol with each of the seven non-empty subsets of $\{s_{\mathrm{hom}}, s_{\mathrm{LLM}}, s_{\mathrm{emb}}\}$ as input to the aggregator and CRC. Figure 4 reports the mean test FNR and mean test FPR across the ten folds at $\alpha = 0.05$.

**Homology alone yields a 100% false-flag rate.** Under taxonomic-family holdout, the maximum 5-mer Jaccard similarity between any test-fold protein and any train-fold hazard is too low to discriminate. Conformal Risk Control's coverage requirement then forces $\widehat{\tau}$ down to (or below) the lowest calibration-hazard score, which sits below the entire test-benign mass; the threshold becomes "flag everything," and the resulting FPR is 1.0 on every fold. This is a structural failure mode of sequence-similarity screening when the hazard at hand belongs to a family that the reference set has not seen, not a tuning artefact.

**LLM panel and embedding each work alone, with caveats.** The LLM panel alone achieves zero mean test FNR with 1.75% mean FPR; embedding alone achieves 0.45% mean FNR (worst fold 4.55%) with 6.85% mean FPR. Either signal is sufficient on its own to avoid the 100% FPR pathology of homology, but neither alone hits the 0%/0% profile.

**LLM panel + embedding is the operating point.** Compos-

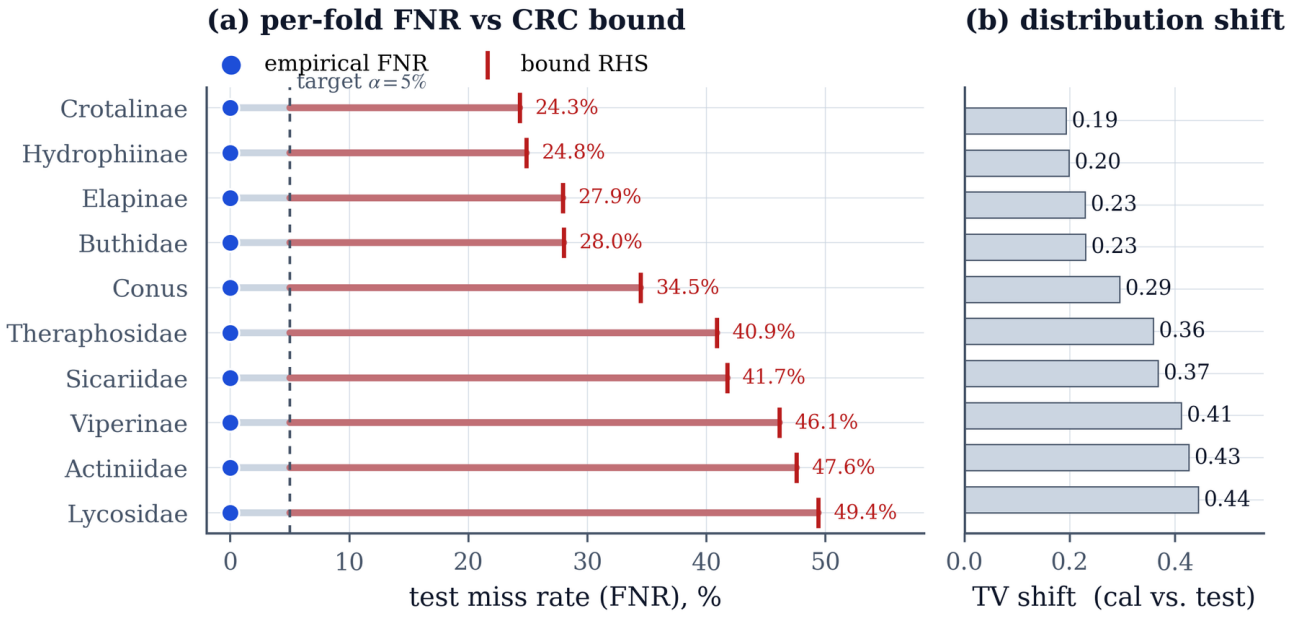


*Figure 3.* The CRC bound holds on every fold with 19–44 percentage points of slack: empirical test miss rate is zero, while the certified ceiling $\alpha + \mathrm{TV}$ ranges from 24.3% to 49.4% across ten LOTO folds at $\alpha = 0.05$. (a) Per-fold view; the grey track runs from zero to the bound right-hand side, the red segment is the slack on top of $\alpha$, and the blue dot is the empirical test miss rate. (b) The TV proxy that drives the slack, 0.19 (Crotalinae) to 0.44 (Lycosidae). $n_{\text{cal haz}} \in [51, 58]$.

| Family | $n_{\text{cal haz}}$ | $n_{\text{test haz}}$ | $n_{\text{test ben}}$ | $\widehat{\tau}$ | TV proxy | test FNR | test FPR | bound RHS |
|---|---|---|---|---|---|---|---|---|
| Crotalinae | 51 | 29 | 58 | 0.598 | 0.193 | 0.000 | 0.000 | 0.243 |
| Conus | 53 | 22 | 44 | 0.719 | 0.295 | 0.000 | 0.000 | 0.345 |
| Buthidae | 53 | 21 | 42 | 0.675 | 0.230 | 0.000 | 0.000 | 0.280 |
| Theraphosidae | 56 | 13 | 26 | 0.595 | 0.359 | 0.000 | 0.000 | 0.409 |
| Hydrophiinae | 56 | 11 | 22 | 0.622 | 0.198 | 0.000 | 0.000 | 0.248 |
| Actiniidae | 57 | 10 | 20 | 0.519 | 0.426 | 0.000 | 0.050 | 0.476 |
| Elapinae | 57 | 10 | 20 | 0.520 | 0.229 | 0.000 | 0.000 | 0.279 |
| Viperinae | 57 | 9 | 18 | 0.572 | 0.411 | 0.000 | 0.000 | 0.461 |
| Sicariidae | 57 | 8 | 16 | 0.659 | 0.367 | 0.000 | 0.000 | 0.417 |
| Lycosidae | 58 | 5 | 10 | 0.529 | 0.444 | 0.000 | 0.000 | 0.494 |

*Table 1.* Test FNR is zero on every fold and test FPR is zero on nine of ten folds (one flagged benign on Actiniidae) at $\alpha = 0.05$, well inside the bound right-hand side $\alpha + \mathrm{TV}$; rows ordered by descending $n_{\text{test haz}}$.

ing the two non-homology signals achieves 0% mean test FNR and 0% mean test FPR, the headline result of Figure 1. *Adding* homology to this pair raises the mean FPR from 0% to 0.5% with no recall gain. Homology is not merely unhelpful here; it is mildly harmful as part of the ensemble, because the train-fold-only reference set leaves a noisy near-uniform signal that the aggregator weights into the score and CRC then has to budget for.

### 4.4. What $\alpha$ a given calibration set can certify

The per-fold bound Equation (3) contains two slack terms beyond $\alpha$. The first, $\mathrm{TV}(\mathrm{cal}, \mathrm{test})$, is a property of the splits: it shrinks if the calibration and test distributions of $S$ become more similar. The second, $1/(n_{\text{cal haz}} + 1)$, is a property of the calibration-set size alone, and it sets a hard floor on the certifiable $\alpha$:

$$\alpha \;<\; \frac{1}{n_{\text{cal haz}} + 1} \quad\Longrightarrow\quad \widehat{\tau} \to 0, \tag{4}$$

because no $\tau$ can satisfy $\widehat{R}(\tau) + 1/(n_{\text{cal haz}} + 1) \le \alpha$ when $\widehat{R} \ge 0$ already exceeds $\alpha$. Figure 5 traces this frontier on a log-log axis.

On our subsample, $n_{\text{cal haz}} \in [51, 58]$ across folds, with mean 55.5, giving a slack floor of $\alpha \approx 1.77\%$. Reaching a stringency target of $\alpha = 10^{-3}$ (an order-of-magnitude deployment goal; we call this *procurement-grade* below for brevity) would require $n_{\text{cal haz}} \ge 999$, an 18× data-budget gap. The full reviewed UniProt KW-0800 corpus contains roughly 6,000 toxins; an evaluation at full scale would deliver $n_{\text{cal haz}} \sim 1{,}800$ per fold, comfortably below the

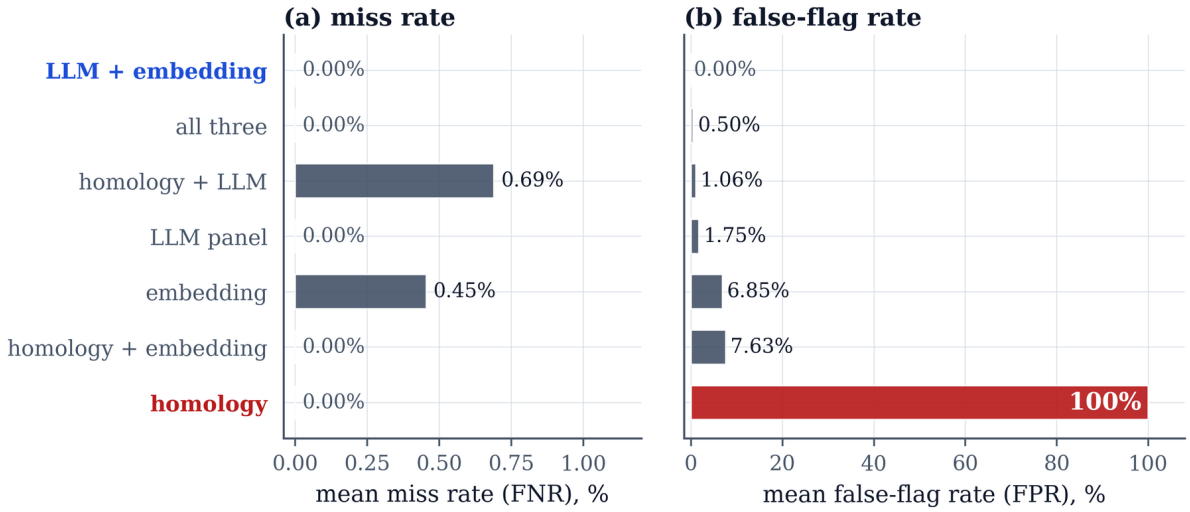


*Figure 4.* Of seven signal subsets at $\alpha = 0.05$, only LLM+embedding achieves $0\%$ mean test FNR with $0\%$ mean test FPR; sequence homology alone fails at $100\%$ FPR; adding homology to LLM+embedding raises FPR to $0.5\%$ with no recall gain. Two combinations have non-zero mean FNR (embedding only: $0.45\%$; homology + LLM: $0.69\%$); their worst-fold FNRs ($4.55\%$ and $6.90\%$) are within the per-fold bound. Means across ten LOTO folds.

procurement-grade floor.

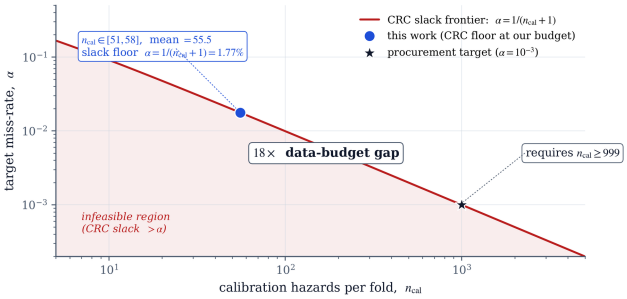


*Figure 5.* The CRC slack frontier $\alpha = 1/(n_{\text{cal}} + 1)$ caps the certifiable $\alpha$ at any calibration-set size; our 200-hazard subsample ($n_{\text{cal}} \approx 55$, floor $1.77\%$) is $18\times$ below the procurement target $\alpha = 10^{-3}$, but the full UniProt KW-0800 reviewed corpus has enough hazards to clear it. The shaded region is infeasible: any $\alpha$ below the frontier cannot be certified by Conformal Risk Control alone, regardless of model performance.

## 5. Discussion

**Why composition works.** The three signals fail in different directions: $s_{\text{hom}}$ ignores function and is confounded by family-level sequence drift; $s_{\text{LLM}}$ ignores sequence and misreads ambiguous annotations; $s_{\text{emb}}$ misses fine-grained mechanism and conflates topical similarity with functional similarity. The aggregator's job is to recover from any single failure mode by demanding agreement, and the monotone constraint forces this agreement to be in the same direction for every signal. The empirical separation in Figure 4 between LLM-only ($1.75\%$ FPR) or embedding-only ($6.85\%$ FPR) and the LLM+embedding combination ($0\%$) is the cooperative gain.

**Why adding homology hurts.** Under taxonomic-family holdout the per-fold homology score is essentially noise: it is rank-normalised across the whole sample but computed against a reference set that excludes the held-out family, so values in the test set are dominated by random matches to unrelated train families. The aggregator weights this noise with a small but non-negative coefficient, and Conformal Risk Control then has to push $\widehat{\tau}$ slightly lower to absorb the resulting calibration variance, which costs a fraction of a percentage point in test FPR.

### 5.1. Limitations

*Sample size and seed variance.* We evaluate on a 200-hazard subsample with $n_{\text{cal haz}} \approx 55$ per fold and a single random seed. The slack floor at this size caps the certifiable $\alpha$ at $1.77\%$, far above procurement-grade $\alpha = 10^{-3}$, and per-fold empirical FNR of zero on $n_{\text{test haz}} \in [5, 29]$ carries wide Wilson confidence intervals. The three signals have well-understood scaling behaviour, but multi-seed runs at full scale are future work.

*No comparison against fielded systems.* The IGSC and major DNA-synthesis providers do not publish their screening procedures or release reference sets (Diggans & Leproust, 2019); we therefore cannot directly compare CRC-Screen to a deployed baseline. Our homology-only condition is a stand-in for the public alignment-search machinery, not for any specific commercial implementation.

*TV proxy versus true non-exchangeability bound.* Theo-

rem 2 of Barber et al. (2023) bounds the coverage gap by a weighted sum of residual-swap TV terms; we substitute a histogram TV between calibration and test score distributions. This is a coarse approximation of the residual-swap quantity, not an upper bound, and the slack we report could be larger or smaller than the exact bound depending on the joint distribution.

*Adversarial inputs are out of scope.* An adversary designing a synthesis order to evade the screener would target the LLM-panel (through annotation phrasing) or the embedding centroid (through choice of organism / function description). Robustness against such adversarial annotations is future work; the certified bound applies to the joint cal/test distribution, not to a worst-case input.

*Prompt-side LOTO leak.* The Variant A prompt enumerates named high-concern examples (e.g. botulinum neurotoxin, ricin) that are themselves UniProt KW-0800 entries. When one of these examples' families is the held-out fold, the LLM panel has seen a name-level description of the family in its prompt. We did not rotate the example list per fold; doing so is a clean follow-up.

**What changes in practice.** Two implications follow if these results extrapolate. First, sequence-similarity-only screening is insufficient as the sole defence under realistic distribution shift, and the operational implication is that providers should compose at least one annotation-derived signal (LLM panel, embedding distance, or similar) into their screening stack. Second, the binding constraint on certifiable miss-rates is the size of the labelled hazard pool used for calibration, not the choice of model or aggregator: the algorithmic tools to certify $\alpha = 10^{-3}$ are available today, and the investment required for procurement-grade screening is a larger, better-curated calibration set.

## 6. Conclusion

Synthesis-order screening has been built as a sequence-matching problem. Under taxonomic-family holdout that framing fails: sequence similarity cannot deliver a certified miss rate without flagging every benign, and the failure is not closeable by tuning. Recasting screening as a calibrated decision problem closes it. Conformal Risk Control turns the operating threshold into a data-driven calibration step with a certified miss-rate ceiling, and the bound's two slack terms separate cleanly the part of the problem that better algorithms can reduce from the part that only more calibration data can. Three off-the-shelf signals clear the bound at $\alpha = 0.05$ today; the next factor of ten in certifiable miss-rate comes from a larger, better-curated calibration set, not from a better screener.

## Impact Statement

This work is defender-side: a synthesis provider screening incoming orders for biosecurity-relevant proteins under a certified expected miss rate. The system flags orders for human review and does not generate, design, or modify biological sequences. The released code contains no utility for sequence generation or pathogen-enhancement information, no hazardous sequence data, and no operational guidance for synthesis or expression.

The most plausible misuse pathway is an adversary with access to the same public annotations and open-source tooling who scores their own designs against the system to estimate evasion probability. The threshold and aggregator weights shown in the paper are specific to the demonstration fold and the 200-hazard subsample, so they do not transfer to any production screener trained on a larger hazard pool; an adversary cannot read $\widehat{\tau} = 0.60$ off this paper and bypass a deployed system with it. The paper's most visible negative finding, that sequence-similarity screening fails under taxonomic holdout, is a property of how sequence similarity behaves under distribution shift that has been documented in the open literature (Puzis et al., 2020); publishing it is consistent with responsible-disclosure norms in the biosecurity community rather than a new uplift.

The intended effect is defensive: to make certified-miss-rate screening operationally available to providers, and to identify the calibration set, not the algorithm, as the gap between current public benchmarks and procurement-grade $\alpha = 10^{-3}$ screening.

## References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. doi: 10.1016/S0022-2836(05)80360-2.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997. doi: 10.1093/nar/25.17.3389.

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):Article 43, 2021. doi: 10.1145/3478535.

Buchfink, B., Reuter, K., and Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, 2021. doi: 10.1038/s41592-021-01101-x.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009. doi: 10.1186/1471-2105-10-421.

Carter, S. R. and Friedman, R. M. DNA synthesis and biosecurity: Lessons learned and options for the future. Technical report, J. Craig Venter Institute, La Jolla, CA, 2015. URL https://www.jcvi.org/research/dna-synthesis-and-biosecurity-lessons-learned-and-options-future.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, 2004. doi: 10.1145/1015330.1015432.

Cauchois, M., Gupta, S., and Duchi, J. C. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pp. 1–15. Springer, Berlin, Heidelberg, 2000. doi: 10.1007/3-540-45014-9_1.

Diggans, J. and Leproust, E. Next steps for access to safe, secure DNA synthesis. *Frontiers in Bioengineering and Biotechnology*, 7:86, 2019. doi: 10.3389/fbioe.2019.00086.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. AlpacaFarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Eddy, S. R. Accelerated profile HMM searches. *PLOS Computational Biology*, 7(10):e1002195, 2011. doi: 10.1371/journal.pcbi.1002195.

Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010. doi: 10.1093/bioinformatics/btq461.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Jungo, F. and Bairoch, A. Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase. *Toxicon*, 45(3):293–301, 2005. doi: 10.1016/j.toxicon.2004.10.018.

Kaas, Q., Yu, R., Jin, A.-H., Dutertre, S., and Craik, D. J. ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Research*, 40(D1):D325–D330, 2012. doi: 10.1093/nar/gkr886.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021. doi: 10.1093/nar/gkaa913.

Mouton, C. A., Lucas, C., and Guest, E. The operational risks of AI in large-scale biological attacks: Results of a red-team study. Technical Report RR-A2977-2, RAND Corporation, 2024. URL https://www.rand.org/pubs/research_reports/RRA2977-2.html.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A. J., Bartlett, P. L., Schölkopf, B., and Schuurmans, D. (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Puzis, R., Farbiash, D., Brodt, O., Elovici, Y., and Greenbaum, D. Increased cyber-biosecurity for DNA synthesis. *Nature Biotechnology*, 38(12):1379–1381, 2020. doi: 10.1038/s41587-020-00761-y.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. 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.

Romano, Y., Patterson, E., and Candès, E. J. Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. doi: 10.1080/01621459.2017.1395341.

Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974. doi: 10.1111/j.2517-6161.1974.tb00994.x.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and The UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015. doi: 10.1093/bioinformatics/btu739.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. doi: 10.1093/nar/gkae1010.

Tibshirani, R. J., Foygel Barber, R., Candès, E. J., and Ramdas, A. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer Cham, 2nd edition, 2022. ISBN 978-3-031-06648-1. doi: 10.1007/978-3-031-06649-8.

Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., Flyangolts, K., Gemler, B. T., Mitchell, T., Murphy, S. T., Wheeler, N. E., and Horvitz, E. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390(6768):82–87, 2025. doi: 10.1126/science.adu8578.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

## A. Panel aggregation

For each (sample, model, run) we attempt one API call and parse a JSON score in $[0, 1]$. A call that errors out, returns malformed JSON, or returns a score outside $[0, 1]$ is filled with the neutral value $0.5$. For each (sample, model) we take the median of the $k_{\text{LLM}} = 2$ run scores; at $k = 2$ this equals their mean. Let $m_1, \ldots, m_5$ be the five per-model scores for a sample, and let $m_{(1)} \le m_{(2)} \le m_{(3)} \le m_{(4)} \le m_{(5)}$ their sorted order. The panel score is the trimmed mean of the middle three:

$$s_{\text{LLM}} \;=\; \tfrac{1}{3}\big(m_{(2)} + m_{(3)} + m_{(4)}\big).$$

Trimming one extreme on each side limits the influence of a single $0.5$ fallback (or any single outlier model) on the aggregate without discarding any per-model evidence beyond the outermost.

## B. Hyperparameters

| Parameter | Value |
|---|---|
| Corpus size $n$ | 600 (200 hazards, 400 benigns) |
| $k$-mer length for Jaccard homology | 5 |
| LLM runs per (sample, model) | 2 |
| LLM temperature | 0.7 |
| Embedding model | text-embedding-3-large (OpenAI) |
| Embedding normalisation | $L_2$ |
| $K$-means cluster count | $\min(8,\ \lfloor n_{\text{train haz}}/5 \rfloor)$ |
| Aggregator | logistic regression, $L_2$ with $C = 1$ |
| Non-negativity rule | drop most-negative coefficient and refit |
| Train/calibration split | 70/30, stratified by class label |
| Histogram TV bin count | 50, equal-width on $[0, 1]$ |
| Random seed | 42 |
| $\alpha$ targets | $\{0.10,\ 0.05,\ 0.01,\ 0.005,\ 0.001\}$ |
| Default $\alpha$ | 0.05 |

*Table 2.* Hyperparameters used throughout the evaluation. A single random seed is used; per-fold splits are deterministic.

## C. Prompts for the LLM panel

The headline results use a single prompt pair: the system message (Variant A, "screening") and the shared user message template. Both are reproduced verbatim from the public code repository. Two alternative system variants ("risk_assessment" with no rubric; "minimal" with task statement only) are present in the repository for a sensitivity analysis but were not used to produce the headline numbers.

**System message (Variant A: screening)**

Task: You are a biosecurity screening analyst. Given a protein's public UniProt annotation, assign a hazard score in [0.0, 1.0] indicating how strongly the documented function warrants escalation to human expert review.

Score from the provided annotation text, drawing on your knowledge of protein families and regulatory classifications (e.g., CDC/USDA Select Agent lists) to interpret what that text means.

Rubric:

0.0 – 0.2 Benign
Metabolic, structural, regulatory, or sensory function with no annotated toxicity, virulence, or pathogenicity.
Examples: lactate dehydrogenase, collagen alpha chain, cytochrome c oxidase subunit, rhodopsin.

0.2 – 0.5 Low concern
Host-defense, antimicrobial, or enzyme-inhibition activity not annotated as a toxin or virulence factor. Includes pathogen-origin proteins with narrow or food-safe hazard profiles.
Examples: nisin (bacteriocin), non-pathogen phospholipase A2, pathogen-derived serine protease inhibitor with immune evasion annotation.

0.5 – 0.8 Elevated concern
Explicitly annotated as a toxin, virulence factor, pore-forming protein, or viral attachment/fusion protein from any organism. Regulatory status is not required for this tier.
Examples: Shiga toxin, streptolysin O, listeriolysin O, type III secretion effector, influenza hemagglutinin, SARS-CoV-2 spike protein.

0.8 – 1.0 High concern
Matches a protein from the CDC/USDA Select Agent and Toxin list, or belongs to a protein class with documented biological weapons history.
Examples: botulinum neurotoxin, ricin, abrin, anthrax protective antigen, Staphylococcal enterotoxin B, Clostridium perfringens epsilon toxin.

Output — return only a valid JSON object with exactly two fields:
{"score": <float in [0.0, 1.0]>, "rationale": <string, at most 25 words, naming the annotation term(s) that determined the score>}

Constraints:
1. Organism name may inform your score only when the function annotation corroborates a hazard concern. Do not assign a high score on organism identity alone.
2. Do not produce sequence data, synthesis routes, or pathogen enhancement information.
3. Always return a score. If the annotation is absent or wholly uninformative, return {"score": 0.5, "rationale": "insufficient annotation"}.

The user message is rendered per sample by substituting four fields from the protein's UniProt annotation. The function-text field is truncated to 1,200 characters before substitution to respect model context limits.

**User message template**

Protein annotation:

Name : {name}
Organism : {organism}
Keywords : {keywords}
Function : {function_text}