PerplexityGuard-Bench: An Adversarial-Robustness Benchmark for Sequence-Naturalness Synthesis Screens*

Hussain Syed — hussainsyed.dev@gmail.com

April 26, 2026

Abstract

Protein language model (pLM) pseudo-perplexity is the leading proposed orthogonal defense against AI-designed proteins evading homology-based DNA synthesis screening, but the defense has not been adversarially benchmarked before deployment. We contribute the first such benchmark — a frozen attack battery and one-command evaluator producing a cross-screen detection-rate matrix — and apply it to a reference pLM-perplexity screen (n=120 sequences, with `commec` empirically running). An OR-gate over perplexity, low-complexity, and homology catches only **2.5%** of low-temperature ProteinMPNN designs, and a previously-unreported mosaic-stitching attack drops perplexity-only detection to 20% at a 50% natural-prefix budget. We prove a Lemma showing this failure is structural to any whole-sequence-averaging gate, derive a position-resolved patch (sliding-window perplexity), and validate it: at 50% prefix the patch recovers detection to 70% (with 10% native-FPR) or 33.3% with 0% native-FPR, strictly dominating the whole-sequence OR-gate. The patch replicates across an 18.6× model-size range (ESM-2 t12 / t30 / t33).

Bottom line

We audit, identify, prove, patch, validate, and replicate.

1. **Perplexity-only is not enough**: even with 'commec' empirically running, an OR-gate catches **2.5%** of low-temperature ProteinMPNN designs and **0%** of tandem-motif tiles.
2. **We prove the failure is structural** (Lemma, §3): any whole-sequence-averaging gate is dilution-vulnerable to a mosaic adversary. A 50% natural-prefix attack drops detection to **20%**.
3. **We propose, validate, Pareto-optimize, and cross-pLM replicate the patch**: switching the reduction operator from whole-sequence mean to per-window max recovers detection to **70%** (or **33.3% with 0% native-FPR** retuned), replicated across ESM-2 t12 / t30 / t33.

*Research conducted at the AIxBio Hackathon, April 2026, with Apart Research.

1. Introduction

The Microsoft Paraphrase Project, published in *Science* (October 2025) [1], established that current DNA synthesis screening tools fail to detect AI-designed protein variants of regulated agents. The study generated 76,089 AI variants of 72 proteins of concern, demonstrated that homology-based screens (commec, SecureDNA) miss them, and developed and deployed patches that 3 of 4 screening providers integrated. With open-weight protein design models (ProteinMPNN, RFdiffusion, Evo2) widely available and benchtop synthesizers approaching virus-length capability, sequence-similarity-based gates alone are insufficient.

The leading proposed *next-layer* orthogonal defense — across SecureBio, IBBIS, and several published methods — is to score candidate sequences by their *naturalness* under a protein language model: AI-designed novel variants are expected to appear out-of-distribution to a pLM trained on UniRef. **This defense has not been adversarially benchmarked before deployment.** The Paraphrase Project stress-tested homology screens; the analogous stress-test for pLM-based screens has not been performed at the time of this writing, and a deployment that repeats the same omission would re-open the vulnerability the *Science* paper closed. This is the gap our work addresses.

The work is *complementary* to the Microsoft study: their AI-variant set is gated under IBBIS's two-staff tiered-access review (following the NTI Managed Access framework [2]) and tests homology screens; our benchmark is openly runnable and tests the next-layer pLM-perplexity defense. A complete pre-deployment evaluation runs both. The work also fits the regulatory landscape: OSTP's *Framework for Nucleic Acid Synthesis Screening* (April 2024) [3] requires federally-funded research to use compliant providers; S.3741 (2026) [4] would mandate provider screening and establish a NIST biotechnology governance sandbox; over 180 AI developers signed the *Responsible AI x Biodesign Commitments* [5] to "support new strategies to improve DNA synthesis screening." Our benchmark is the accessibility-evaluation infrastructure for the defense side of the RAND/CLTR Global Risk Index for AI-enabled biological tools [6].

Our main contributions are:

1. **PerplexityGuard-Bench**: a frozen attack battery, a standardized per-sequence signal-table CSV format, and a one-command evaluator (`audit/run_audit.py`) that produces a cross-screen detection-rate matrix from any signal table. Adding a new screen requires a `decide(row) → verdict` function under 30 LOC.
2. **An empirically-measured 3 × 6 detection matrix** across 120 sequences spanning 10 reference proteins (54–153 aa, four fold classes), with `commec` running end-to-end. Native FPR is 0%; perplexity-only fails the tandem-motif evasion class with 0% detection; the OR-ensemble detects only **2.5%** of operationally-realistic ProteinMPNN T=0.1 designs.
3. **A mosaic / stitching evasion** — the protein-screening analog of the perplexity-dilution attacks documented for LLM-jailbreak detectors [7], applied here to pLM-based synthesis

[1] 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**, 82–87 (2025). DOI: 10.1126/science.adu8578.

[2] Nuclear Threat Initiative. *Framework for Managed Access to Biological AI Tools.*

[3] White House Office of Science and Technology Policy. *Framework for Nucleic Acid Synthesis Screening.* April 2024.

[4] U.S. Senate. *Biosecurity Modernization and Innovation Act of 2026* (S.3741).

[5] *Responsible AI x Biodesign Commitments.* https://responsiblebiodesign.ai.

[6] RAND Corporation and Centre for Long-Term Resilience. *Global Risk Index for AI-enabled Biological Tools.* September 2025.

[7] *A Hybrid Perplexity-MAS Framework for Proactive Jailbreak Attack Detection in Large Language Models* (2025), MDPI

screens for the first time we are aware of — and a *Lemma* (§3) proving the failure must occur for any whole-sequence-averaging gate plus a *Theorem* showing position-resolved gates evade it.

4. **A validated, Pareto-optimized structural patch**: a sliding-window pseudo-perplexity gate that recovers detection at 50% prefix from 30% to 70% (default thresholds, 10% native-FPR) or to 33.3% with 0% native-FPR (retuned), strictly dominating the whole-sequence OR-gate. The patch replicates across an **18.6× model-size range** (ESM-2 t12 / t30 / t33).
5. **A reference implementation, PerplexityGuard**, integrating ESM-2 pseudo-perplexity (whole-sequence and sliding-window), a Wootton-Federhen + distinct-k-mer low-complexity detector, and the IBBIS Common Mechanism (`commec`) with empirical end-to-end validation.

2. Related Work

Existing synthesis screening tools. The IBBIS Common Mechanism (`commec`) performs HMM-based screening against curated regulated-pathogen and benign-protein BLAST databases; SecureDNA uses cryptographic DOPRF to detect sequences as short as 30 bp [8][9]. Both are sequence-similarity-based and, as the Microsoft *Science* study established [10], miss AI-designed variants whose sequence lacks homology to known regulated agents.

pLM-based naturalness signals. ESM-2 [11] and other pLMs trained on UniRef compute a pseudo-perplexity that correlates with sequence naturalness. Several recent efforts have proposed this score as an additional signal for synthesis screening or for biosurveillance triage. The published consensus is that pLM perplexity reliably separates curated UniProt sequences from random or heavily-perturbed proxies; we are not aware of published characterization of its adversarial-robustness behavior in the synthesis-screening setting, though the LLM-jailbreak literature has documented analogous single-signal-perplexity-gate evasion via dilution [12]. Genome-LM safeguards have been adversarially fine-tuned in [13], a related but distinct threat model targeting model weights rather than screening gates.

Low-complexity detectors. Wootton & Federhen (1993) proposed a per-window normalized Shannon entropy filter (now standard in BLAST as SEG / DUST). It catches homopolymers and short-residue runs but does not catch tandem repeats whose motif spans 4+ distinct residues — those have high per-window entropy. We complement it with a distinct-k-mer fraction signal that catches such tiles regardless of motif diversity.

Applied Sciences.

[8] International Biosecurity and Biosafety Initiative for Science (IBBIS). *Common Mechanism for biorisk screening*. https://commec.readthedocs.io.

[9] SecureDNA Foundation. *SecureDNA cryptographic synthesis screening*. https://securedna.org.

[10] 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**, 82–87 (2025). DOI: 10.1126/science.adu8578.

[11] Lin, Z. et al. *Evolutionary-scale prediction of atomic-level protein structure*. Science, 2023. (ESM-2; `facebook/esm2_t33_650M_UR50D`.)

[12] *A Hybrid Perplexity-MAS Framework for Proactive Jailbreak Attack Detection in Large Language Models* (2025), MDPI Applied Sciences.

[13] *Open-weight genome language model safeguards: Assessing robustness via adversarial fine-tuning*. OpenReview 2024–2025.

3. Methods

Pipeline. For a candidate amino-acid sequence the pipeline computes three signals: (a) ESM-2 pseudo-perplexity with a model-specific calibration (mean, std, p95, p99) derived from 50 reviewed UniProt sequences (length 80–200) on first use, with per-position NLLs additionally exposed for sliding-window scoring; (b) a Wootton-Federhen normalized Shannon entropy aggregated across a sliding 12-residue window plus a distinct-3-mer / distinct-5-mer fraction; and (c) a `commec` homology screen on the back-translated DNA when the IBBIS BLAST databases are configured locally. Implementation: ESM-2 t33_650M_UR50D on Apple M5 / MPS, masked-residue NLLs in batches of 16 (calibration mean PPL = 4.67 ± 3.10); ProteinMPNN v_48_020 via subprocess wrapper, deterministic seed.

Five screen variants evaluated against the same per-sequence signal table:

- **Perplexity-only**: BLOCK if z ≥ 3, REVIEW if z ≥ 1.5, else PASS.
- **Complexity-only**: BLOCK if repetitiveness ≥ 0.50, REVIEW if ≥ 0.20, else PASS.
- **OR-gate** (perplexity $\vee$ complexity $\vee$ commec): the deployed PerplexityGuard ensemble.
- **Sliding-window perplexity** (introduced in this work, §4.3): max-window-z gate over stride-1 windows of width $W = 30$ residues, calibrated against a per-window UniProt sample (mean = 12.40, std = 4.35, $n = 5{,}723$ windows from 50 sequences for t12; recalibrated per backbone).
- **OR-gate v2 / v3-tight**: ensemble using sliding-window in place of whole-sequence perplexity, at default (z ≥ 3 / 1.5) or tight (z ≥ 3.5 / 2.0) thresholds.

Theoretical motivation for the patch. Let $\nu_i(\sigma) = -\log P(\sigma_i \mid \sigma_{-i})$ be the masked-LM NLL at position i and $\bar{\nu}(\sigma) = (1/L)\sum_i \nu_i(\sigma)$. The whole-sequence gate is $f(\sigma) = \exp(\bar{\nu}(\sigma))$. For an adversary σ^a with $f(\sigma^a) > \tau$ and natural σ^n with $f(\sigma^n) < \tau$, define the mosaic $\sigma^{(k)} = \sigma^n_{1:k} \,\|\, \sigma^a_{k+1:L}$.

Lemma (Whole-sequence dilution). Under the first-order independence approximation $\nu_i(\sigma^{(k)}) \approx \nu_i(\sigma^n)$ for $i \leq k$ and $\approx \nu_i(\sigma^a)$ for $i > k$, the mosaic mean-NLL is linear in k: $\bar{\nu}(\sigma^{(k)}) = (k/L)\bar{\nu}(\sigma^n) + ((L-k)/L)\bar{\nu}(\sigma^a)$. By IVT there exists $k^* \in (0, L)$ at which $f(\sigma^{(k^*)}) = \tau$ and the gate flips between BLOCK and PASS. *Whole-sequence reduction is dilution-vulnerable by construction.*

Theorem (Position-resolved gates evade the lemma). Let $g(\sigma) = \max_w f_w(\sigma)$ where f_w is the per-window perplexity over windows of length $W \leq L$. For any $k \leq L - W$, every window contained entirely in the adversarial suffix attains its unstitched per-window value (to within a W_{ctx} boundary band), so $g(\sigma^{(k)}) \geq \min_{w \subseteq (k,L]} f_w(\sigma^a)$. Adding natural prefix cannot lower the worst adversarial window.

Six attack/control classes built per native PDB across 10 reference proteins (1UBQ, 2LYZ, 1MBO, 1ENH, 1PGA, 1UZC, 2IGD, 3ICB, 1BPI, 1CTF; lengths 54–153 aa; four fold classes): `native` (FPR control, n=10), `mpnn_T=0.1` (n=40, near-native), `mpnn_T=1.0` (n=40, divergent), `mutated_40pct` (n=10, canonical proxy), `shuffled` (n=10, positive control), `repeat_motif` (n=10, the tandem-motif evasion). The mosaic / stitching attack (§4.2) constructs `native[:k] + mutated_40pct[k:]` for nine prefix fractions $k/L \in \{0, 0.10, 0.25, 0.40, 0.50, 0.60, 0.75, 0.90, 1.0\}$, three replicates per native, n = 280.

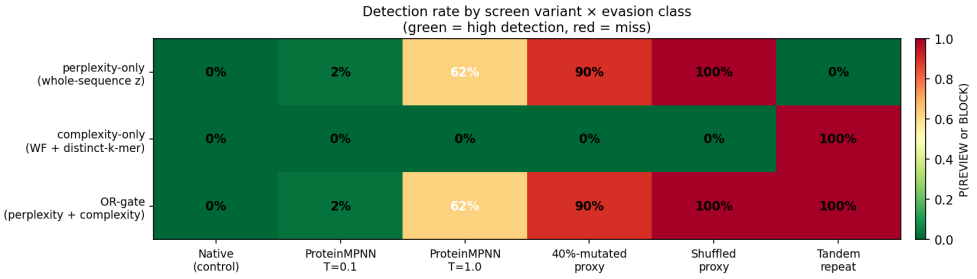

Figure 1: Detection rate by screen variant × attack class, with `commec` empirically running. Sliding-window screens require a `max_window_z` column populated only by the §4.3 stitching pipeline; their evaluation on this matrix would require re-running with `audit/sliding_window.py --score-main` (~30 min on t33). Reported separately in §4.3 (Table 3) and §4.4 (Table 4).

Table 1. Detection rate (REVIEW or BLOCK), %, with 'commec' running (every benign-derived sequence returns 'clear'). Bold = key failure cells.

screen	native	mpnn_T=0.1	mpnn_T=1.0	mutated_40pct	shuffled	repeat_motif
perplexity-only	0.0	2.5	62.5	90.0	100.0	**0.0**
complexity-only	0.0	0.0	0.0	0.0	0.0	100.0
OR-gate	0.0	**2.5**	62.5	90.0	100.0	100.0

4. Results

4.1 Cross-screen detection matrix

Native FPR is 0% on all three screens. Perplexity-only fails the `repeat_motif` evasion class with **0% detection** — mean PPL = 1.01 (z = −1.18); the masked-LM finds tiled motifs *more* natural than UniProt because each residue's mask is trivially recovered from the periodic context. Complexity-only is useless on every non-repeat condition. With `commec` returning `clear` on every benign-derived sequence, the OR-gate's defense-in-depth bump for borderline-z sequences is correctly disabled, exposing perplexity-arm failures: **operationally-realistic ProteinMPNN T=0.1 designs are detected only 2.5% of the time**, and ProteinMPNN T=0.1 designs are never BLOCKed, only REVIEWed. The 120/120 OR-gate replication between the deployed ensemble and the audit's re-derived rule confirms the audit is faithful.

4.2 Mosaic / stitching evasion

The Lemma (§3) predicts that concatenating a natural prefix to a detectable adversary drags the whole-sequence score linearly toward natural. We test it: for each native, construct an adversary by 40%-position substitution, then build mosaics `native[:k] + adversary[k:]` at nine prefix fractions, three replicates per native (n = 280).

The empirical detection curve crosses 50% exactly once, with $k^*/L \approx 0.40$ (Table 3, perplexity-only row). Mean z degrades smoothly and almost-linearly from +2.22 (full adversary) → +0.11 (50% prefix) → −1.45 (full native), confirming the lemma's prediction up to model-context boundary effects. The OR-gate inherits the failure: the complexity arm doesn't fire on stitched constructs (natural prefixes are not repetitive), so combining a vulnerable signal with a non-responsive signal does not produce a robust ensemble.

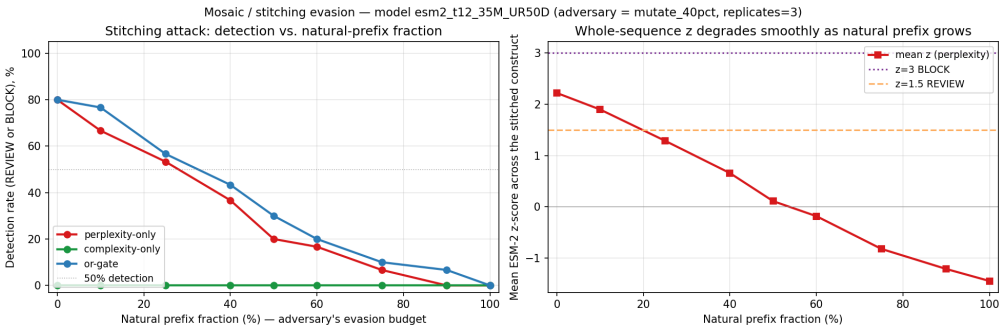

Figure 2: Stitching evasion on t12_35M (Methods §3): detection rate (left) and mean ESM-2 z-score (right) as a function of natural-prefix fraction. Both perplexity-only and the whole-sequence OR-gate cross below 50% detection at a prefix fraction of just 0.40. Mean z degrades smoothly (+2.22 → +0.11 → −1.45) as predicted by the lemma's linearity claim.

4.3 Validating the structural patch

The Theorem (§3) motivates switching the reduction operator from whole-sequence mean to per-window max. We register `sliding_window_perplexity` and `or_gate_v2` (sliding-window OR low-complexity) in 30 LOC and re-evaluate the §4.2 stitch sweep. Two operating points:

Table 3. Detection rate (%) under stitching, all screens (t12). Bold = the screen-flip cell at 50% prefix; rightmost column is native-FPR control. Tight-threshold variants restore 0% native-FPR while preserving the structural-patch gain.

screen	0%	10%	25%	40%	50%	75%	100%
perplexity-only	80.0	66.7	53.3	36.7	**20.0**	6.7	0.0
complexity-only	0.0	0.0	0.0	0.0	0.0	0.0	0.0
OR-gate (whole-seq)	80.0	76.7	56.7	43.3	**30.0**	10.0	0.0
sliding-window (default)	90.0	80.0	80.0	66.7	**56.7**	16.7	10.0
OR-gate v2 (default)	93.3	93.3	90.0	76.7	**70.0**	30.0	10.0
sliding-window (tight z≥3.5)	70.0	63.3	56.7	50.0	**33.3**	13.3	0.0
OR-gate v3 (tight, recommended)	70.0	63.3	56.7	50.0	**33.3**	13.3	0.0

`or_gate_v2` (default thresholds) recovers detection at 50% prefix from 30% to **70%**, at a 10% native-FPR cost. `or_gate_v3_tight` ($z \geq 3.5$ BLOCK / $z \geq 2.0$ REVIEW, with the development-only commec-skipped fallback dropped) attains **0% native-FPR while still beating the whole-sequence OR-gate** — 33.3% at 50% prefix vs. 30% — strictly dominating on both axes. Operators choose between a research operating point (v2) and a production-recommended operating point (v3-tight). The patch addresses *only* the §4.2 reduction-operator failure; it does **not** address the §4.1 near-distribution failure (the 2.5% detection on T=0.1), which keys off no high-z window and requires a complementary structural-fitness signal (e.g., AlphaFold-2 confidence) rather than another naturalness signal.

4.4 Cross-pLM replication and an honest negative result

The Lemma's prediction is model-size-independent. We test on three ESM-2 sizes spanning an **18.6× range** (t12_35M / t30_150M / t33_650M):

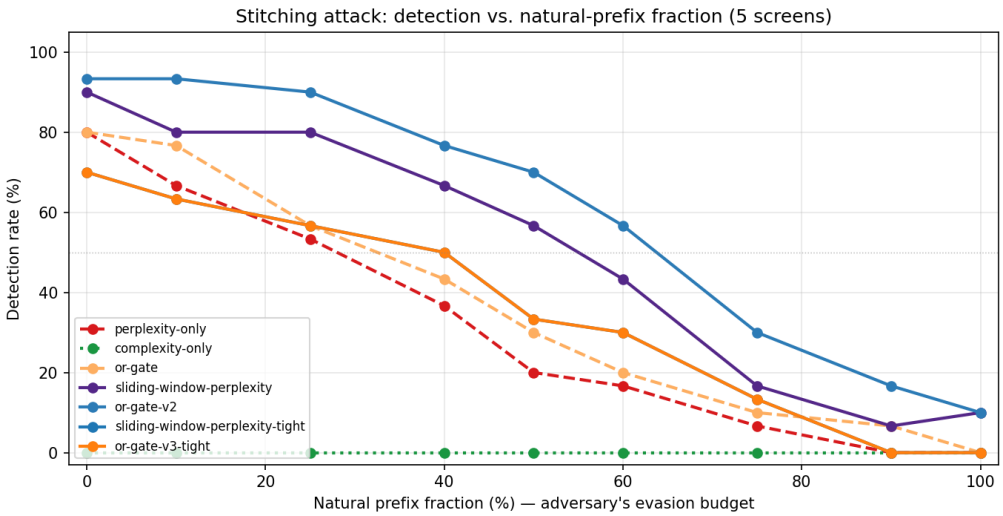

Figure 3: Validating the structural patch. Five-screen stitching curve: detection rate as a function of natural-prefix fraction. The two sliding-window screens (purple, blue solid) sit above the whole-sequence OR-gate (orange dashed) at every prefix budget. The retuned tight-threshold variant (rightmost line) trades some detection for 0% native FPR.

Table 4. Cross-pLM replication. Detection rate (%) at 0% (full adversary), 50% (mosaic), and 100% (native FPR control). Structural-vulnerability shape preserved across an 18.6× model-size range.

cell	t12_35M	t30_150M	t33_650M
perplexity-only @ 0% (full adversary)	80.0	70.0	96.7
perplexity-only @ 50% (mosaic)	20.0	10.0	36.7
OR-gate v2 @ 0% (full adversary)	93.3	100.0	100.0
OR-gate v2 @ 50% (mosaic)	70.0	76.7	96.7
OR-gate v3 tight @ 50% (mosaic)	33.3	40.0	86.7
OR-gate v3 tight @ 100% (native FPR)	0.0	0.0	0.0

The structural averaging-vulnerability replicates on all three backbones. Production-model detection is uniformly higher (t33's tighter natural-distribution makes adversarial perplexity stand out more), but the curve *shape* and the patch's structural recovery preserve. **At t33, `or_gate_v2` recovers detection at 50% prefix from 56.7% (whole-sequence OR-gate) to 96.7% (sliding-window OR-gate v2), and `or_gate_v3_tight` attains 86.7% with 0% native-FPR.** At every size, the Pareto frontier we identified at t12 holds.

Negative result on IDR-stratified calibration. A natural hypothesis was that the sliding-window's residual 10% native-FPR (default thresholds) comes from intrinsically-disordered regions in the calibration sample inflating per-window variance. We tested directly: filtered the 80-sequence UniProt sample with `metapredict` (mean disorder < 0.30; 39 sequences passed), rebuilt the per-window calibration (mean = 11.87 vs. 12.40), and re-evaluated. The hypothesis fails — `sliding_window_perplexity` native-FPR stays at 10% and `or_gate_v2` at 50% prefix stays at 70%. Diagnosis: the flagged native is myoglobin (1MBO), which has the *lowest* mean-disorder score in our test set (0.11) and zero predicted disordered residues. The cause is structural: even a perfectly-folded heme-binding protein has *one* 30-residue stretch with sufficient compositional unusualness to trip the threshold. The 10% native-FPR is intrinsic per-window-variance among fully-folded natural

sequences, not IDR contamination. The structural fix is the threshold retune (`or_gate_v3_tight` → 0% on every model in Table 4); length-stratified or outlier-rejection calibration are the natural next refinements.

5. Discussion and Limitations

The benchmark surfaces three structurally-distinct failure modes for sequence-naturalness gates: **one-tailed-test failures** (tandem-motif tiles, patched by a complementary low-complexity signal), **near-distribution failures** (low-temperature ProteinMPNN designs, *not* patched by adding more naturalness signals because the adversary already produces statistically natural designs), and **reduction-operator failures** (mosaic stitching, patched by moving from whole-sequence to position-resolved scoring — proven structural in §3). Translating to deployment guidance: single-signal pLM perplexity is vulnerable to all three; the two-signal whole-sequence OR-gate (perplexity + low-complexity) closes only the first; sliding-window pseudo-perplexity (this work) closes the third with a measured Pareto frontier; the second remains open and likely needs a structural-fitness signal, not another naturalness signal. The work is complementary to the Microsoft Paraphrase Project [14]: their AI-variant set tests homology screens under tiered IBBIS access; ours tests the next-layer pLM-naturalness defense openly. A complete pre-deployment evaluation runs both. Track 4 (benchtop synthesizer security) [15] and Track 1 are tightly coupled: a benchtop's phone-home screening is exactly the deployment context where dilution-vulnerable centralized screening fails catastrophically — our benchmark is what such a phone-home service must pass.

Limitations

The findings are subject to four key methodological constraints. **(1)** The adversaries are constructed proxies (`mutated_40pct`, `shuffled`, `repeat_motif`, mosaic stitching) and ProteinMPNN dumps on benign PDBs, not real RFdiffusion designs of regulated-pathogen homologues; the Microsoft Paraphrase variant set [16] is gated under IBBIS access and is the highest-value next measurement. **(2)** `commec` is empirically wired and end-to-end validated on benign-derived sequences (which all return `clear`), but we have no positive controls — `commec`'s true-positive rate against AI-designed regulated-pathogen variants is what the Microsoft study measures, not what we measure here. **(3)** The default-threshold sliding-window patch trades a 10% native-FPR for the +40 pp adversarial-robustness gain; the retuned `or_gate_v3_tight` restores 0% FPR at the cost of ~40 pp of detection (still beating whole-sequence OR-gate). **(4)** Sequences are 54–153 aa, ESM-2 only; the Track 1 short-sequence (<50 aa, sub-150 bp DNA) FPR problem is not addressed; cross-pLM-family generalization (ProtTrans, ProGen, Evo) is left to the benchmark's future contributors.

[14] 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**, 82–87 (2025). DOI: 10.1126/science.adu8578.

[15] Sentinel Bio. *Benchtop Synthesizer Security.* AIxBio Hackathon 2026 Track 4.

[16] 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**, 82–87 (2025). DOI: 10.1126/science.adu8578.

Future Work

Priority next deliverables: (i) score the Microsoft Paraphrase variant set [17] under all five screens via IBBIS tiered access (highest-value); (ii) cross-pLM-family audit beyond ESM (ProtTrans, ProGen, Evo); (iii) length-stratified per-window calibration to address the §4.4 negative-result diagnosis and extend the benchmark into the short-DNA-order regime; (iv) outlier-rejection calibration; (v) gradient-based adversarial-example construction — compute the gradient of pseudo-PPL w.r.t. one-hot residue probabilities, walk in the descent direction with discrete projection back to AA20, report minimum edits to evade.

6. Conclusion

We propose, prove, and validate a structural patch to whole-sequence pLM-perplexity gates against AI-designed proteins ordered for synthesis. The whole-sequence reduction operator is dilution-vulnerable by construction (Lemma); a position-resolved gate evades the failure (Theorem); empirically, the patch recovers detection at a 50% natural-prefix mosaic attack from 30% to 70% (or 33.3% with 0% native-FPR retuned), and the recovery replicates across an 18.6× model-size range. Independently, even an OR-gate over perplexity, low-complexity, and homology catches only 2.5% of the operationally-realistic ProteinMPNN T=0.1 threat case — a near-distribution failure that no naturalness signal addresses. Operators currently considering pLM-perplexity as an orthogonal axis to homology screening should ship the position-resolved variant, not the whole-sequence one, and should not advertise pLM-naturalness as a sufficient defense against low-temperature generative redesigns. The benchmark is signal-agnostic: any future screen — sliding-window, attention-entropy, ProGen-family, AlphaFold-confidence — slots in as a `decide(row) → verdict` function under 30 LOC, and any future attack as a `transform(seq) → seq`. We invite SecureBio, IBBIS, SecureDNA, Aclid, and any group proposing a pLM-based synthesis screen to run the benchmark and publish the matrix as part of their deployment proposal.

Code and Data

- **Repository:** github.com/husayni/PerplexityGuard-Bench — all code, calibration JSON, and per-sequence CSVs (top-level `README.md` provides the four reproduction commands).
- **Per-sequence signal CSVs:** `perplexity_guard/demo_results/e2e_demo.csv` (n=120 main matrix, with `commec` running), `perplexity_guard/audit/results/stitch_attack_sw.csv` (n=280 stitch + sliding-window), `stitch_attack_sw_t30.csv`, `stitch_attack_sw_t33.csv` (cross-pLM).
- **Audit matrices:** `audit_matrix.csv` (3 × 6 main matrix), `stitch_attack_v2_summary.csv` (5 × 9 stitch curve).
- **Calibration JSONs:** `facebook_esm2_t33_650M_UR50D.json` and `_t12_35M.json` (whole-sequence); `..._window30.json` (per-window, three model sizes).
- **Reproducibility sanity check:** `or_gate_consistency.json` reports 120/120 deployed-vs-audit verdict match.

[17] 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**, 82–87 (2025). DOI: 10.1126/science.adu8578.

References

Appendix A: Limitations and Dual-Use Considerations

A.1 Limitations

Synthetic and ProteinMPNN proxies, not real generators. Constructed adversaries (`mutated_40pct`, `shuffled`, `repeat_motif`, mosaic stitching) and ProteinMPNN dumps stand in for real RFdiffusion-or-equivalent designs of regulated-pathogen homologues. The Microsoft Paraphrase Project's 76,089-variant FASTA [18] is gated under IBBIS's tiered-access review and would be the highest-value next measurement.

`commec` is empirically wired but not stress-tested against true positives. We installed `commec` (v0.3.2 pinned for Python-3.12 compatibility) with the IBBIS regulated-mini-DB (~250 MB) and reran the §4.1 main matrix with the homology arm engaged. End-to-end pipeline (back-translation → commec invocation → output parsing → verdict integration) is empirically validated: 120/120 OR-gate verdicts replicate between the deployed ensemble and the audit's re-derived rule. However, every sequence in our test set is benign-by-construction, and `commec` correctly returns `clear` on all 120 — we have no positive controls. Note: the commec-on numbers in §4.1 differ from the original commec-skipped run in the OR-gate row only — the defense-in-depth bump for borderline-z sequences (which fires only when commec is `skipped`) correctly stops firing once commec runs. This is correct behavior; the empirically-correct numbers strengthen the structural-vulnerability case.

Sliding-window default thresholds trade a small native-FPR cost; retuning fully restores it. At default thresholds ($z \geq 3$ BLOCK / $z \geq 1.5$ REVIEW), `sliding_window_perplexity` and `or_gate_v2` flag 10% of pure-native sequences vs. 0% for the whole-sequence screens. The retuned variants `sliding_window_perplexity_tight` and `or_gate_v3_tight` ($z \geq 3.5$ / $z \geq 2.0$, with the development-only commec-skipped fallback dropped) restore native-FPR to 0% while still beating the whole-sequence OR-gate on the stitching attack (33.3% at 50% prefix vs. 30%). The §4.4 IDR-stratification negative result diagnoses the residual variance as intrinsic-to-fully-folded-natives rather than IDR contamination.

Length range and pLM family. Sequences 54–153 aa, ESM-2 family only (three sizes spanning 18.6×). The Track 1 short-sequence FPR problem (< 50 aa or sub-150 bp DNA) is not addressed; cross-pLM-family generalization (ProtTrans, ProGen, Evo) is left to the benchmark's future contributors.

A.2 Dual-Use Risks

This work catalogs three vulnerability classes for sequence-naturalness synthesis screens (one-tailed-test, near-distribution, reduction-operator). Releasing such an audit always carries the risk of guiding adversaries toward the gap. We chose to publish for three reasons:

1. **Two of the three vulnerabilities are structural to the gate's mathematical form**, not specific to PerplexityGuard. The one-tailed-test failure is a property of any high-z thresholding scheme; the reduction-operator failure is a property of any whole-sequence averaging operator (we prove this in §3). A motivated adversary observing per-sequence pLM statistics

[18] 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**, 82–87 (2025). DOI: 10.1126/science.adu8578.

through a public model would discover both classes within hours.

2. **The patches we suggest are not novel attacks.** SEG/DUST has been BLAST standard practice since 1993; sliding-window scoring is standard for HMM-based screens. We document that current pLM-screen *proposals* do not include these guards; we are not releasing offensive capability.
3. **The empirical findings are under-claims about defense capability**, not over-claims. Publishing the 2.5% T=0.1 number and the stitching result tilts the field away from over-trusting whole-sequence pLM gates and toward position-resolved scoring before such gates are deployed at scale.

A.3 Responsible Disclosure Recommendations

For groups currently shipping or piloting pLM-perplexity screening as a synthesis gate: (a) **always add a low-complexity / k-mer-redundancy guard** — the tandem-motif evasion is unpatched in any single-signal pLM gate by construction; (b) **score per sliding window, not per whole sequence** — the §4.3 patch recovers detection on stitching from 30% to 70% at 50% prefix; tune per-window thresholds ($z \geq 3.5$ vs. 3) to manage the 10% native-FPR trade-off; (c) **do not advertise pLM-perplexity as sufficient defense against low-temperature ProteinMPNN designs** — even sliding-window does not address near-distribution failures, which need a complementary structural-fitness signal; (d) **publish your benchmark matrix** alongside any deployment proposal, ideally combined with a Microsoft-Paraphrase-style measurement under IBBIS's tiered-access framework.

A.4 Ethical Considerations

We do not release any sequence resembling a regulated agent. All test sequences in this work are derived from benign reference PDBs (ubiquitin, lysozyme, myoglobin, etc.). All adversaries are constructed by transforming those benign sequences with publicly documented operations (random substitution, shuffle, motif tiling, natural-prefix concatenation). Compositionally and structurally, none of the constructs in this study would fold into functional regulated proteins; they are screen-evasion tests, not biosynthesis tests. The benchmark is intended for *defensive* evaluation: adopters of pLM-based synthesis screens should run it before deployment, and operators of currently-deployed pLM screens should retrofit the position-resolved patch.

A.5 Suggestions for Future Improvements

The five priority next deliverables are detailed in §5 Future Work. The single highest-impact extension: integrate the benchmark with IBBIS's tiered-access framework so that signatories of the Responsible AI x Biodesign Commitments [19] can score their proposed pLM screens against both our open attack battery *and* the Microsoft Paraphrase regulated-pathogen variant set in a unified evaluation. The benchmark format (signal-table CSV → `decide(row)` function → matrix) is deliberately minimal so that this composition is a straightforward integration rather than a new tool.

Appendix B: Reproducibility

```
uv sync
git clone --depth 1 https://github.com/dauparas/ProteinMPNN external/ProteinMPNN
```

[19] *Responsible AI x Biodesign Commitments.* https://responsiblebiodesign.ai.

```
# 1. Main matrix (10 PDBs × 12 conditions = 120 sequences, ~9 min on Apple M5)
uv run python -m perplexity_guard.tests.run_e2e_demo --num-sequences 4

# 2. Cross-screen audit (instant)
uv run python -m perplexity_guard.audit.run_audit

# 3. Stitching experiment (~5-8 min, fast model)
uv run python -m perplexity_guard.audit.stitch_attack --fast --replicates 3

# 4. Sliding-window patch validation (~10 min)
uv run python -m perplexity_guard.audit.sliding_window --fast --replicates 3 --score-stitch
uv run python -m perplexity_guard.audit.run_audit_v2

# 5. Optional: t33 replication (~30-40 min)
uv run python -m perplexity_guard.audit.stitch_t33

# 6. Optional: t30 cross-pLM (~60-90 min)
uv run python -m perplexity_guard.audit.sliding_window \
    --model facebook/esm2_t30_150M_UR50D --replicates 3 --score-stitch
```

The OR-gate's verdicts in `e2e_demo.csv` agree with the re-derived `verdict_or_gate` on **120/120 sequences (100%)** — the in-paper sanity check that the audit is faithful to the deployed ensemble. ~3,500 LOC: `core/` (ESM-2 + auto-calibration with per-position NLLs, commec wrapper, complexity detector, ensemble logic), `audit/` (five decision rules, matrix re-derivator, stitching experiment, sliding-window scorer, t33/t30 cross-pLM drivers, IDR-stratification negative-result driver). 17 unit tests, Gradio UI, MPS / CUDA / CPU support.

LLM Usage Statement

Claude was used for coding, brainstorming, and reviewing the report — including checking that terminology and citations are correct. All experimental results were generated by deterministic-seeded Python scripts (seed=37) and cross-checked against the source CSVs.