Quantifying BLAST's Sensitivity Floor for DNA Synthesis Screening

How Fragment Length and Adversarial Mutation Determine Detection

Utkarsh Singh *University of Illinois Urbana-Champaign*
Evan Coats *University of Illinois Urbana-Champaign*
Prithvi Vegesna *Purdue University West Lafayette*
Swetha Krishnamoorthy *University of Illinois Urbana-Champaign*

with **Apart Research**

Abstract

DNA synthesis screening is the principal safeguard against malicious orders of hazardous genetic material. The 2024 OSTP Framework requires screening at 200 base-pairs (bp), dropping to 50 bp by October 2026. We benchmark BLAST as a per-fragment screener across seven fragment lengths (20–200 bp) and six mutation rates (0–20%), under two threat models: pure evasion (orders entirely composed of hazardous fragments) and dilute evasion (one hazardous fragment hidden among benign filler). At the upcoming 50 bp threshold with 20% mutation, per-fragment sensitivity drops from 100% to 16%; under any-flag aggregation, only 60% of dilute attacks are caught at 1% false-positive rate. Dilute is the operationally meaningful failure regime — pure attacks remain catchable everywhere. A robustness check with diversified data preserves these findings and widens the gap. The 50 bp threshold is adequate against unmutated adversaries but inadequate against modestly sophisticated ones; policy choice must reference assumed adversary capability.

1. Introduction

DNA synthesis providers are an increasingly important defensive layer against the misuse of biological knowledge. When a customer orders a synthesized DNA sequence, the provider is expected to screen the order against curated lists of hazardous content before fulfilling it. We use “hazardous content” in the sense formalized by IGSC and OSTP frameworks: protein-coding sequences (CDSs) of regulated agents — toxin genes, virulence factors, pathogen markers, select agents — along with the regulatory and structural elements that enable their function. The International Gene Synthesis Consortium (IGSC) Harmonized Screening Protocol (v3.0, September 2024) formalizes this responsibility for industry members, and the 2024 OSTP Framework for Nucleic Acid Synthesis Screening makes adherence a condition of U.S. federal life-sciences research funding. Both frameworks currently require screening at a 200 base-pair (bp) granularity, with the OSTP Framework dropping the threshold to 50 bp effective October 2026. The implicit assumption: this threshold is biologically and algorithmically meaningful. We test that assumption.

An adversary who fragments a hazardous gene into pieces forces the screener into a regime where alignment-based detection becomes statistically marginal. An adversary who additionally mutates each fragment — e.g., through synonymous substitution or codon optimization — removes the exact-match signal that alignment tools rely on. The combination is the realistic threat model for DNA synthesis screening. The question is how far each axis can be pushed before the screener fails.

Theory of change

The OSTP threshold is a policy lever: it specifies a single number that determines the depth of every commercial DNA synthesis screening operation in the United States. The current 200 bp threshold reflects today's screening reality; the upcoming 50 bp threshold (effective October 2026) is what the field is preparing for now. Quantitative evidence about how that future threshold will perform against realistic adversarial inputs is the input policymakers need to refine the lever before it takes effect. Our pipeline produces exactly this evidence and is structured to be re-applied against alternative screening tools (commec, DIAMOND-based pipelines) by replacing one script. If the qualitative finding we surface — that per-fragment alignment screening with any-flag aggregation has a structural failure mode against dilute mutated attacks — generalizes to other tools, the policy implication is that the 50 bp threshold is insufficient unless paired with explicit assumptions about adversary capability or with screening capabilities beyond per-fragment alignment. If it does not generalize, our pipeline allows researchers to demonstrate that empirically. Either outcome moves the policy debate forward.

Contributions

- **A reproducible benchmark pipeline** (NEW) that takes a hazardous reference set as input and produces per-fragment and per-order detection statistics across a configurable grid of fragment lengths, mutation rates, and operating thresholds. Parameterized so the same evaluation applies to alternative screeners by replacing the screening step.
- **Quantitative measurements** (NEW) showing that BLAST per-fragment sensitivity degrades smoothly along both fragment length and adversarial mutation rate. Mutation rate matters at least as much as length: a 50 bp fragment at 20% mutation is harder for BLAST to detect than a 30 bp fragment at 5% mutation.
- **The dilute attack mode** (NEW framing) — one hazardous fragment hidden in benign filler — identified as the structural failure regime for any-flag aggregation. Pure attacks are caught reliably; dilute attacks defeat the OSTP 50 bp threshold under realistic adversarial mutation. To our knowledge, this asymmetry has not been quantitatively characterized in the open literature for DNA synthesis screening.
- **A robustness check** (NEW) under a more rigorous configuration (10 diversified honest sources spanning bacterial, viral, eukaryotic, plant, and synthetic origins, and an expanded curated hazardous database) confirming the qualitative findings hold and the quantitative gap widens.

Built on

BLAST (Altschul et al., 1990; Camacho et al., 2009) and its standard `blastn-short` task preset, used unmodified. Biopython (Cock et al., 2009) for sequence I/O. NCBI Entrez for the honest source genome. The IGSC Harmonized Screening Protocol v3.0 and OSTP framework define the policy context against which we benchmark. The commec project (IBBIS) provides the architectural pattern for layered DNA synthesis screening that we contrast against in our discussion of layered pipelines.

2. Related Work

DNA synthesis screening is governed in practice by the IGSC Harmonized Screening Protocol (current version v3.0, September 2024), which specifies the screening obligations of synthesis providers, including the use of BLAST-family alignment tools against curated regulated-sequence databases. The 2024 OSTP Framework for Nucleic Acid Synthesis Screening currently requires order screening at the 200 bp level, dropping to 50 bp by October 2026, motivated in part by the observation that smaller fragments carry less functional information per fragment but does not directly address the algorithmic detection limits of the screening tools at this threshold.

Among open-source implementations, commec (the Common Mechanism, IBBIS) is the most widely-cited reference pipeline. Commec is a layered system: DIAMOND-based protein search against a regulated-protein database, BLASTN against a regulated-nucleotide database, low-concern clearing via taxonomic filters, and biorisk-specific HMM scans. Each screening layer in commec operates per-record; a dedicated `commec split` utility breaks multi-record FASTAs into individual files for screening. Per-order or cross-order reasoning lies outside commec's scope, in the customer-screening layer that IBBIS treats as a parallel concern.

Prior work characterizing the limits of alignment-based screening has focused primarily on database curation and false-negative analysis under the assumption of intact target sequences. The operational regime addressed by our work — short fragments, mutated content, and hidden attacks under any-flag aggregation — is what an adversary actually controls and is, to our knowledge, not systematically characterized in the open literature. Our pipeline fills this gap.

3. Methods

3.1 Pipeline overview

The benchmark consists of four stages, each implemented as a standalone Python script. The honest source for the primary experiment is E. coli K-12 MG1655 (NC_000913.3, 4.6 megabases (Mb)), chosen as a single, well-annotated, uncontroversially benign reference. The hazardous reference set is a curated collection of toxin and pathogen-associated CDSs contributed by our team, aggregated from a folder of mixed-format (.fasta and .gb) files into a

single consolidated FASTA, then filtered to exclude records longer than 10 kilobases (kb) so that hits against the database reflect specific toxin coding sequences rather than incidental matches in larger genomic contexts. After filtering, the primary hazardous database contains 19 records ranging from 75 bp to roughly 7 kb. A robustness check (Appendix C) re-runs the entire experiment with a 10-source diversified honest corpus (~100 Mb) and an expanded curated hazardous database (40 records under a 50 kb cap).

3.2 Corpus generation

For each combination of fragment length $T \in \{20, 30, 50, 75, 100, 150, 200\}$ bp and mutation rate $M \in \{0, 1, 5, 10, 15, 20\}$%, we generate three corpora of 1,000 “orders” each. Order size varies uniformly between 5 and 20 fragments per order, modeling typical custom synthesis order shapes.

- **Honest corpus:** 5–20 random T-bp slices drawn from the honest source. No mutation. Models legitimate customer orders.
- **Pure evasion corpus:** 5–20 random T-bp slices drawn from a single hazardous parent record. Each fragment is mutated independently at rate M before being written. Models an adversary who synthesizes a fragmented copy of a hazardous gene.
- **Dilute evasion corpus:** exactly one mutated hazardous fragment plus 4–19 unmutated honest fragments per order. Models an adversary who hides one hazardous fragment among benign cover content within a single order.

Mutation is applied as independent per-base random substitution. Honest fragments and dilute-corpus filler fragments are never mutated, modeling the asymmetry that legitimate customers do not modify their orders to evade screening. Fragment IDs encode order membership and ground truth, allowing all downstream analysis to recover ground-truth labels without a separate metadata file.

3.3 Screening

Each corpus is screened with `blastn -task blastn-short -evalue 10 -num_threads 4` against the hazardous BLAST database. The reporting threshold of $e \leq 10$ is intentionally permissive; the screening decision threshold is swept downstream in the evaluation step. For each fragment, we record the minimum e-value and maximum bitscore. For each order, we apply any-flag aggregation: the order’s score is the minimum e-value across its fragments. Any-flag is the most charitable aggregation rule to the screener — the screener gets N independent chances to detect the hazardous content for an order with N fragments — so failures under any-flag are conservative results.

3.4 Evaluation

The per-fragment analysis (Figure 1) reports the fraction of hazardous and honest fragments that produce a hit at each of three e-value cutoffs (1e-3, 1e-9, 1e-30), as a function of T and M.

The per-order analysis (Figures 2–4) computes the receiver operating characteristic (ROC) under any-flag aggregation: the screener's e-value cutoff is swept and (false positive rate (FPR) on honest orders, true positive rate (TPR) on evasion orders) is recorded at each point. Area under the curve (AUC) is computed by trapezoidal integration; TPR at fixed FPR is reported at FPR ∈ {0.01, 0.05, 0.10}.

With 1,000 orders per cell, the 95% confidence interval on AUC is approximately ±1.5% near the tails and ±3% near 0.5. We use BLAST e-values as the screening decision criterion. E-values are database-size-dependent; alternative criteria (bitscore, percent identity) are not. Our results characterize BLAST under e-value thresholds specifically.

4. Results

4.1 Per-fragment sensitivity profile

Figure 1 shows BLAST's per-fragment detection rate as a function of fragment length T, stratified by adversary mutation rate (color) and BLAST e-value cutoff (rows of panels). Left column: TPR on hazardous fragments. Right column: FPR on honest fragments.

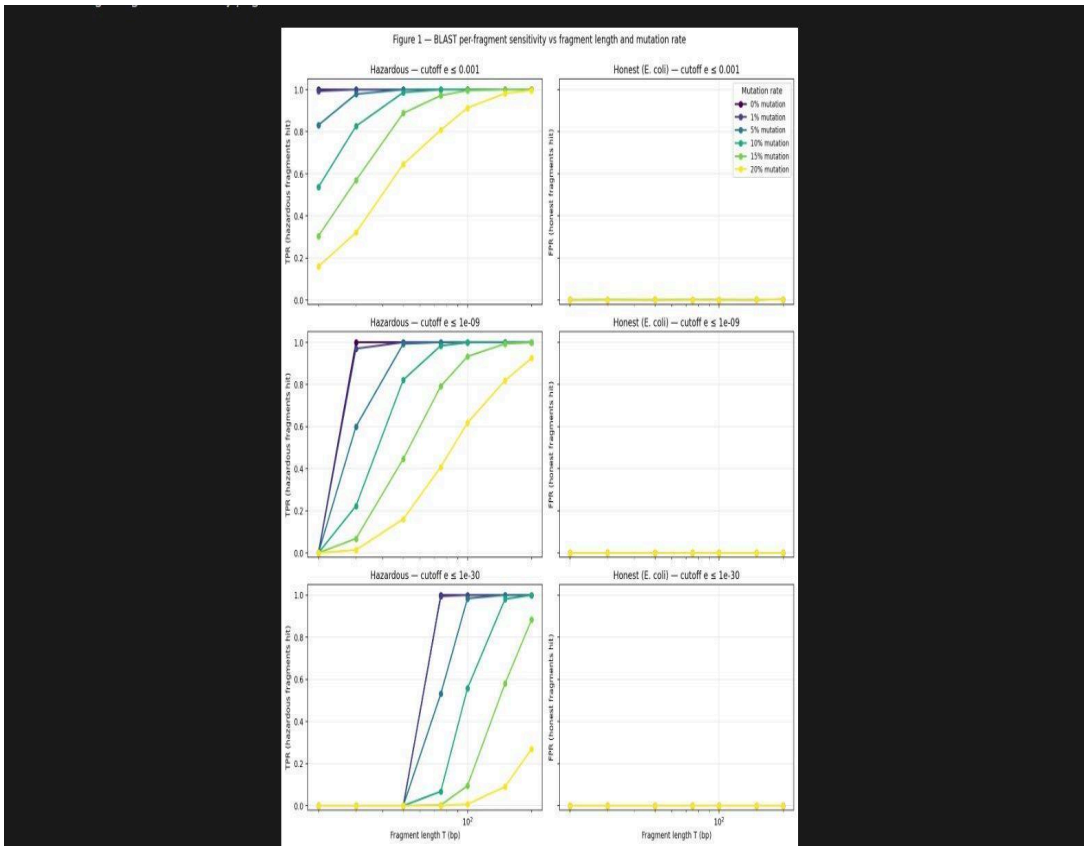

Figure 1. Per-fragment sensitivity. Left column: TPR on hazardous fragments; right column: FPR on honest E. coli fragments. Rows: e-value cutoffs (top: 1e-3; middle: 1e-9; bottom: 1e-30). Color: adversary mutation rate (0–20%). Hazardous TPR degrades smoothly with both shorter fragments and higher mutation; honest FPR remains near zero across all configurations.

Two qualitative observations. First, the curves are smooth and monotonic in both axes. There is no discontinuity at the OSTP 50 bp threshold or at any other fragment length. Second, mutation rate matters approximately as much as fragment length — at the standard biosecurity cutoff (e ≤ 1e-9, middle row), holding mutation at 20% requires fragments of ≥ 200 bp to recover near-perfect TPR; holding fragment length at 50 bp requires keeping mutation ≤ 5%. Numerically, at T = 50 bp and e ≤ 1e-9, TPR drops from 1.000 (0% mutation) to 0.993 (5%), 0.822 (10%), 0.447 (15%), and 0.161 (20%).

The honest FPR (right column) is essentially zero across all configurations. We initially anticipated diversifying the honest source would expose a TPR/FPR trade-off; our robustness check (Appendix C) found that this is not the case at our experimental scale. Whether this reflects database-bounded behavior (curated hazardous content limits chance hits) or insufficient honest diversity is an open question we cannot fully resolve within our timeline.

4.2 Per-order detection: the evasion gap

Figure 3 condenses the per-order ROC analysis into two heatmaps showing TPR at FPR = 1%.

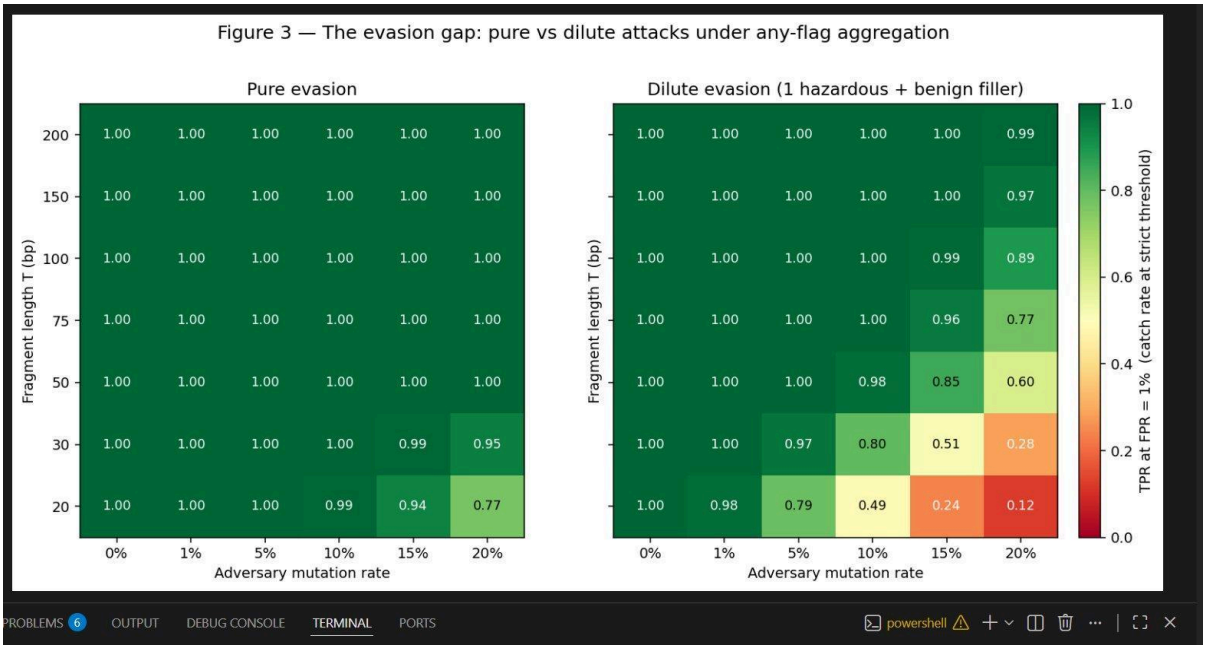

Figure 3. The evasion gap. TPR at FPR = 1% across fragment length (Y-axis) and adversary mutation rate (X-axis), under any-flag aggregation. Left: pure evasion. Right: dilute evasion. Pure evasion remains catchable across nearly all configurations; dilute evasion fails sharply in the bottom-right region (small T, high mutation).

The asymmetry between the two panels is the operationally meaningful finding. Pure evasion is caught reliably (TPR ≥ 0.94) at every cell except the most extreme. Dilute evasion shows a clear failure region: at T = 20 bp / M = 20%, only 12% of dilute attacks are caught at 1% FPR; at T = 30 bp / M = 20%, 28%; at T = 50 bp / M = 20%, 60%. The mechanism: pure orders contain 5–20 hazardous fragments, so the per-order failure probability is the per-fragment failure probability raised to roughly the order size. Dilute orders contain exactly one hazardous fragment, so

per-order detection equals per-fragment detection. Benign filler does not raise FPR but also does not help find the hazardous fragment.

4.3 Operating points at the OSTP threshold

Figure 4 summarizes catch rates at T = 50 bp across three adversary types: naive (0% mutation), realistic (10% mutation), and sophisticated (20% mutation).

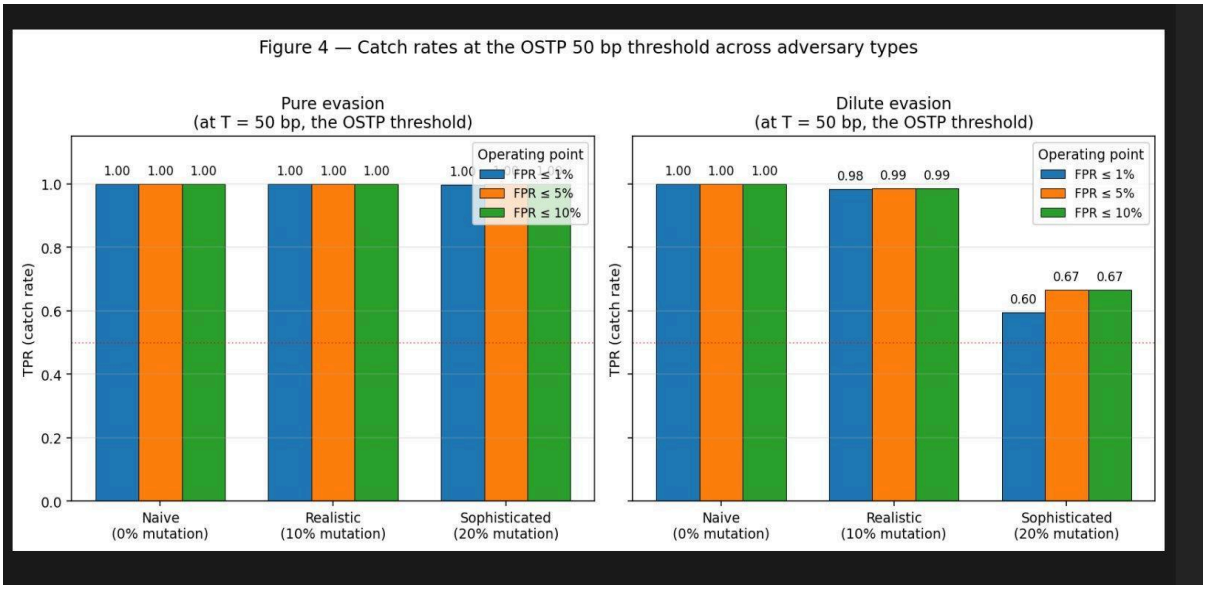

Figure 4. Catch rates at the OSTP 50 bp threshold across three adversary types. Bars: TPR at three FPR operating points (1%, 5%, 10%). Pure evasion caught reliably regardless of adversary sophistication; dilute evasion against the sophisticated adversary drops to 60% at 1% FPR.

Against pure evasion, BLAST holds up essentially perfectly across all three adversary types: ≥ 99.8% catch rate at 1% FPR even against the sophisticated adversary. Against dilute evasion, the picture changes: naive caught 100%, realistic (10% mutation) drops to 98%, sophisticated (20% mutation) drops to 60% at 1% FPR. Said differently: at the 50 bp threshold (which becomes the OSTP-mandated screening level in October 2026), four of every ten dilute attacks by a 20% mutation adversary succeed even with a permissive 1% FPR budget.

4.4 Robustness under expanded data

We re-ran the entire experiment under a more rigorous configuration: a diversified honest corpus (10 sources, ~100 Mb spanning bacterial, viral, eukaryotic, plant, and synthetic origins) and a curated hazardous database expanded to 40 records under a 50 kb cap. Full results are in Appendix C. Qualitative findings are preserved: BLAST per-fragment sensitivity degrades smoothly with both axes; the dilute attack failure region is reproduced; honest FPR remains ≤ 0.3%. Quantitatively, the catch rates in the failure region are slightly lower than in primary results (e.g., dilute T = 50 bp / 20% mutation: 60% → 48% at 1% FPR), strengthening rather than weakening the headline conclusion. Two confounded effects contribute: larger BLAST

databases increase e-value penalties for the same alignment, and the 50 kb cap admits records with proportionally more non-coding content. Disentangling these is left to future work.

5. Discussion

The combined picture is that there is no single “optimal” fragment-length threshold for screening. The right threshold depends on assumed adversary capability. Against an adversary who submits unmutated exact substrings, even T = 20 bp catches everything; the upcoming 50 bp threshold is appropriate against weak adversaries. Against an adversary who introduces 10–20% mutation — well within the range of realistic countermeasures like codon optimization — the threshold needs to be ≥ 100 bp to maintain high catch rates against dilute evasion. The 50 bp threshold (becoming OSTP-mandated in October 2026) sits in the middle: appropriate against weak adversaries, inadequate against modestly sophisticated ones. This shifts the policy framing from “find the magic threshold” to “specify the assumed adversary explicitly, and choose the threshold consistent with that assumption.”

Per-fragment screening with any-flag aggregation has a structural failure mode independent of the screener’s per-fragment sensitivity: dilute evasion. Because the screener gets only one shot at the single hazardous fragment in a dilute order, no improvement to aggregation can help — only per-fragment sensitivity matters. This argues that meaningful improvements to screening robustness require either (a) substantially better per-fragment sensitivity at small T (e.g., translated-protein search, or HMM-based detection as in commec), or (b) cross-fragment reasoning within an order that can detect that several fragments together reconstruct hazardous content even when none of them individually does. The latter is outside the scope of current sequence-level screening tools.

Our robustness check (Section 4.4) surfaces a third observation: BLAST’s per-fragment TPR shifts modestly downward when the curated hazardous database is enlarged. This is a property of e-value-based screening, not an algorithmic flaw — e-values scale with database size. Production screeners may mitigate via (a) bitscore- or identity-based criteria, (b) post-filtering layers that re-examine flagged hits, or (c) protein-level search. Our results characterize BLAST as a standalone screener with e-value thresholds and represent an upper bound on the alignment layer’s difficulty. They do not generalize to layered pipelines without additional measurement, which is the natural follow-up question.

Future work

The pipeline is structured so the screening step can be replaced with a commec-equivalent emitting the same per-order CSV format. A direct commec comparison would establish whether commec’s protein and HMM layers materially change the dilute-attack failure region. A second extension is implementing a more realistic mutation model: synonymous substitution for protein-coding regions, with reading-frame-aware sliding. A third extension is the cross-order experiment: disperse evasion fragments across N synthetic orders, measure per-order

detection, characterize the dispersion curve. A parametric study varying hazardous database size and honest source diversity independently would resolve the database-bounded vs diversity-bounded ambiguity surfaced in Section 4.1.

6. Conclusion

We built a reproducible benchmark of BLAST's sensitivity profile as a per-fragment DNA synthesis screener, evaluated across two adversary axes (fragment length and mutation rate) and two threat models (pure and dilute evasion) on a 7 × 6 × 2 grid of conditions with 1,000 orders per cell, with a robustness check under a more rigorous data configuration. BLAST is competent against unmutated adversaries at the upcoming 50 bp threshold but degrades sharply against dilute attacks under modest adversarial mutation: at T = 50 bp and 20% mutation, only 60% of dilute evasion attempts are caught at 1% FPR (48% under the more rigorous configuration). Policy threshold choices should be made with explicit reference to assumed adversary capability, and meaningful improvements to screening robustness against dilute attacks require capabilities beyond per-fragment alignment with any-flag aggregation.

References

- International Gene Synthesis Consortium. (2024). Harmonized Screening Protocol v3.0. September 2024. https://genesynthesisconsortium.org/
- White House Office of Science and Technology Policy (OSTP). (2024). Framework for Nucleic Acid Synthesis Screening. April 29, 2024 (revised September 30, 2024). https://aspr.hhs.gov/S3/Documents/OSTP-Nucleic-Acid-Synthesis-Screening-Framework-Sep2024.pdf
- International Biosecurity and Biosafety Initiative for Science (IBBIS). The Common Mechanism (commec). https://github.com/ibbis-screening/common-mechanism
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics, 10(1), 421.
- Cock, P. J. A., Antao, T., Chang, J. T., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422–1423.

Code and Data

- **Code repository:** https://github.com/utkarshsingh34/AIxBio
- **Honest source data (primary):** E. coli K-12 MG1655 (NCBI accession NC_000913.3), via `pipeline/fetch_data.py`.

- **Honest source data (robustness):** 10 sources, accessions in `pipeline/fetch_diverse_honest.py`.
- **Hazardous reference set:** Curated by the team; NOT included in the public repository for biosecurity reasons. A synthetic placeholder FASTA is committed for pipeline dry-running.
- **Generated artifacts:** Primary results in `results/v1_5/`. Robustness-check outputs in `results/v2_robustness/`.

Author Contributions

Utkarsh Singh, Evan Coats, and Prithvi Vegesna collaborated on project design, scope, and goals, discussing project direction and next steps at various stages. Utkarsh Singh built the pipeline, ran all experiments, curated the data, and drafted the writeup. Swetha Krishnamoorthy reviewed the writeup and provided feedback. Prithvi Vegesna followed up on review suggestions. Evan Coats finalized the writeup. All authors reviewed the final manuscript.

Appendix A. Per-order ROC grid

Figure 2 shows the full per-order ROC analysis for the primary configuration. Rows: T. Columns: M. Within each panel: pure (blue) and dilute (orange). Same data as Figure 3 in raw ROC form.

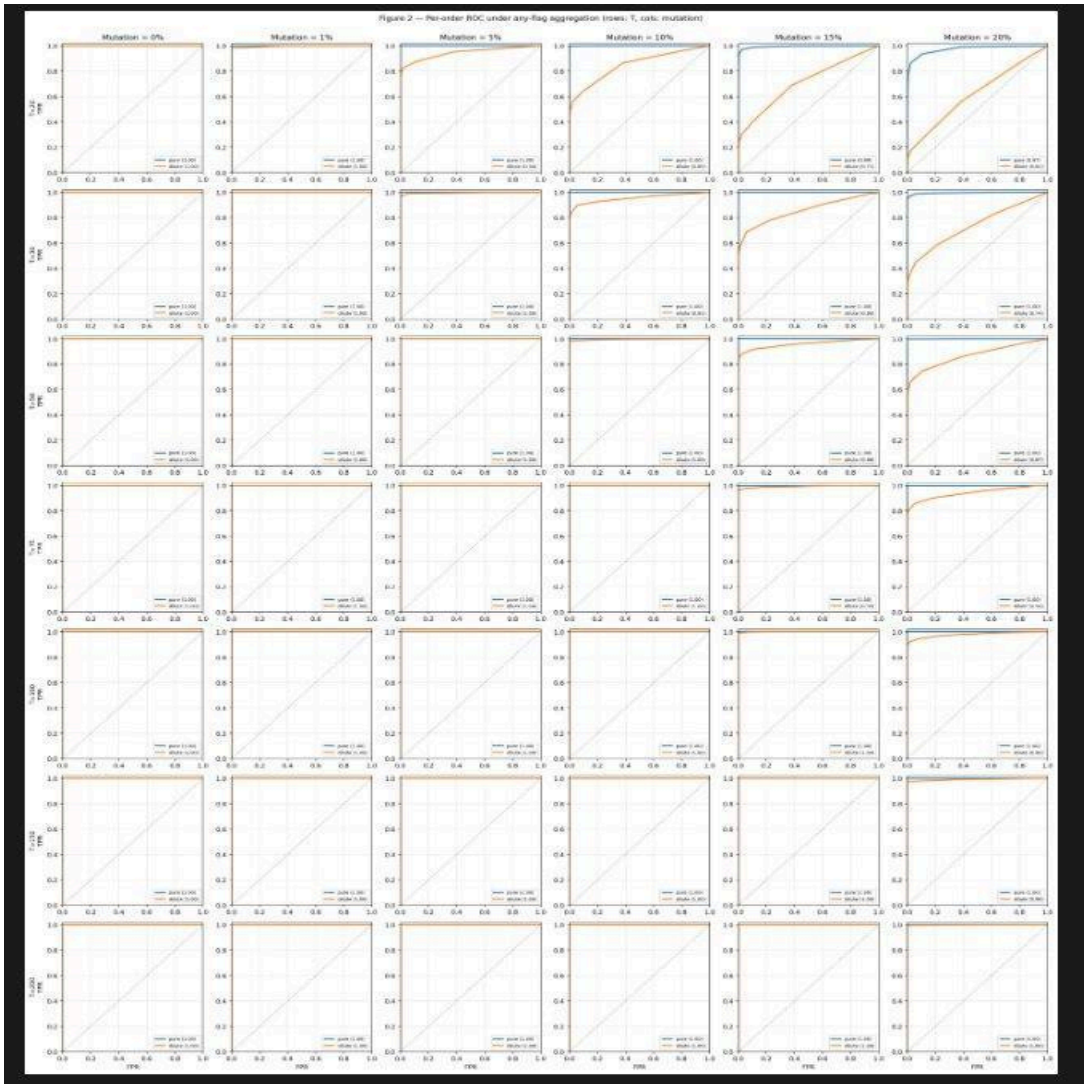

Figure 2. Per-order ROC under any-flag aggregation. Rows: fragment length (20–200 bp). Columns: adversary mutation rate (0–20%). Pure evasion (blue) and dilute evasion (orange) curves vs. honest baseline.

Appendix B. Summary table

T (bp)	Mutation	Type	AUC	TPR @ FPR=1%	TPR @ FPR=5%
50	0%	pure	1.000	1.000	1.000
50	0%	dilute	1.000	1.000	1.000
50	10%	pure	1.000	1.000	1.000
50	10%	dilute	0.996	0.984	0.986
50	20%	pure	1.000	0.998	0.999
50	20%	dilute	0.870	0.596	0.665
20	20%	pure	0.974	0.767	0.865

T (bp)	Mutation	Type	AUC	TPR @ FPR=1%	TPR @ FPR=5%
20	20%	dilute	0.619	0.123	0.174
100	20%	pure	1.000	1.000	1.000
100	20%	dilute	0.979	0.892	0.921

Table 1. Selected per-order operating points (primary configuration). Full table for all 84 cells in `results/v1_5/summary.csv`.

Appendix C. Robustness check

We re-ran the entire experiment under a more rigorous data configuration:

- **Honest source:** 10 sources, ~100 Mb. E. coli, B. subtilis, P. aeruginosa, M. tuberculosis, S. cerevisiae chromosome 1, A. thaliana chromosome 1, lambda phage, T4 phage, pUC19, human chromosome 22.
- **Hazardous database:** 40 records under a 50 kb cap (vs. 19 records under 10 kb in primary).
- **All other parameters unchanged:** same T grid, same mutation rates, same 1,000 orders per cell.

Figure C1: per-fragment sensitivity, robustness configuration.

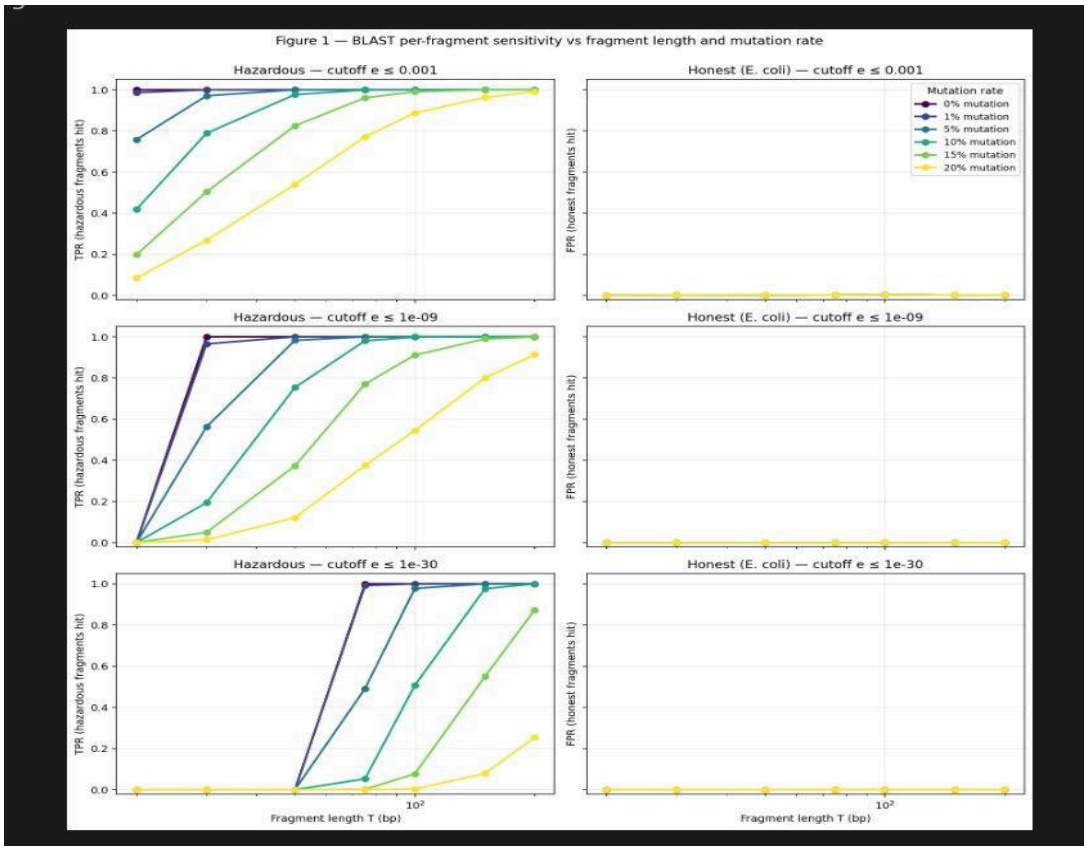

Figure C1. Per-fragment sensitivity under expanded data. Same axes as Figure 1; six mutation curves (vs. four in primary). Qualitative shape preserved.

Figure C2: per-order ROC grid, robustness configuration.

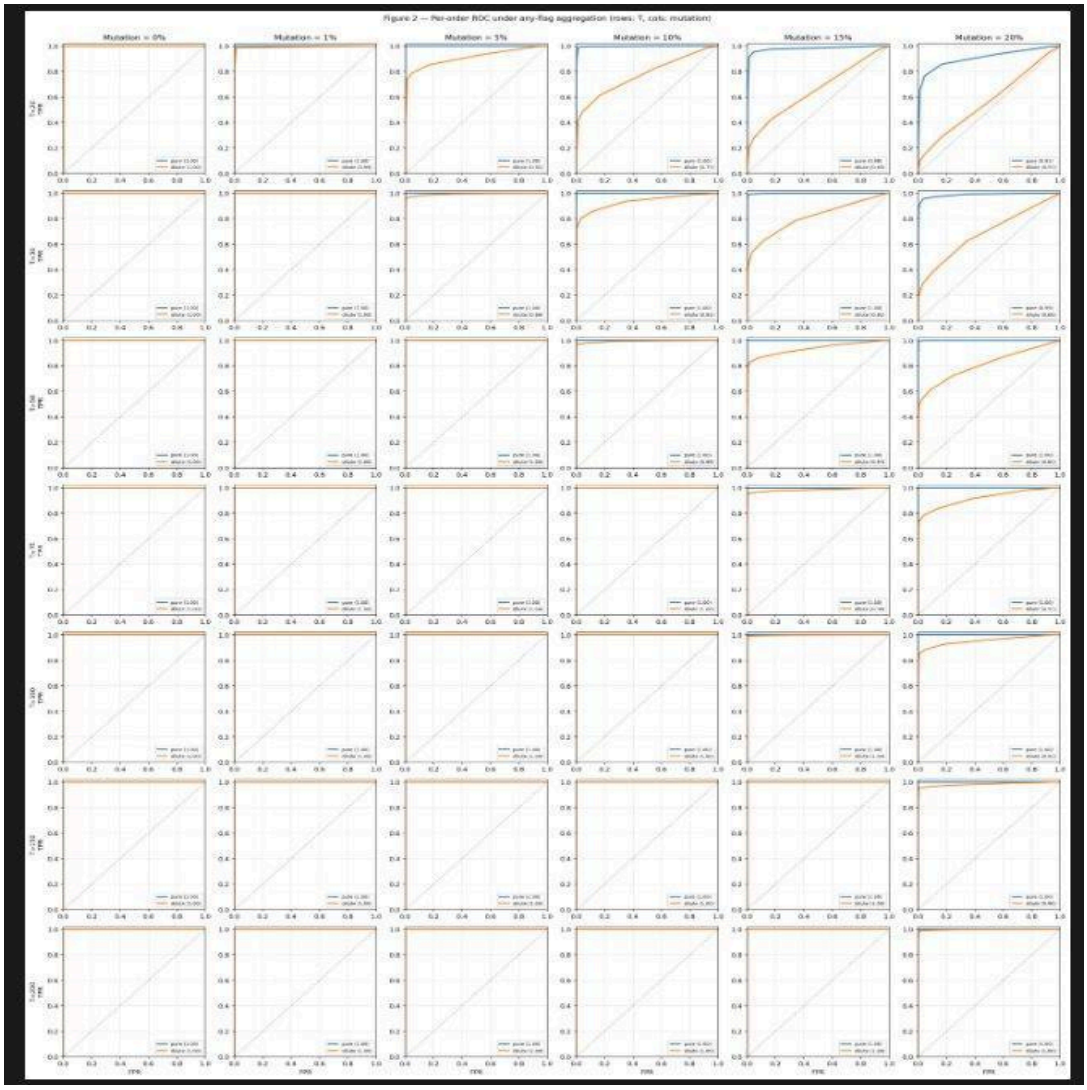

Figure C2. Per-order ROC grid under expanded data. Same layout as Figure 2.

Figure C3: evasion gap heatmap, robustness configuration.

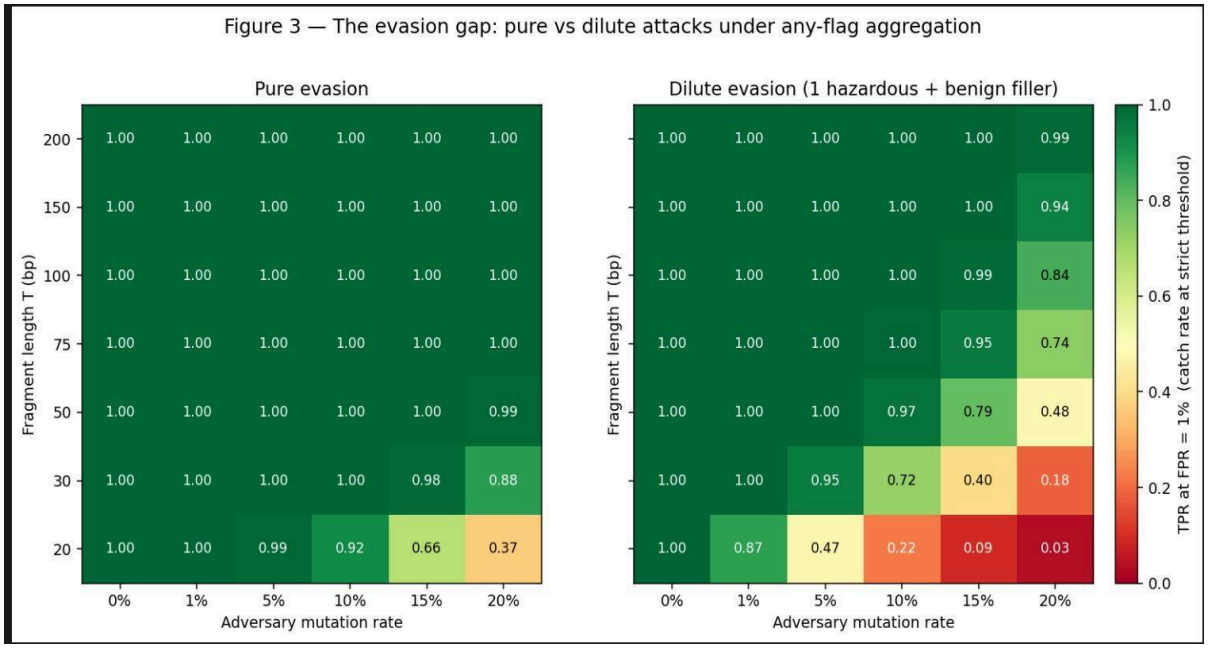

Figure C3. Evasion gap heatmap under expanded data. Dilute-attack failure region preserved and slightly more pronounced than Figure 3.

Figure C4: operating points at OSTP threshold, robustness configuration.

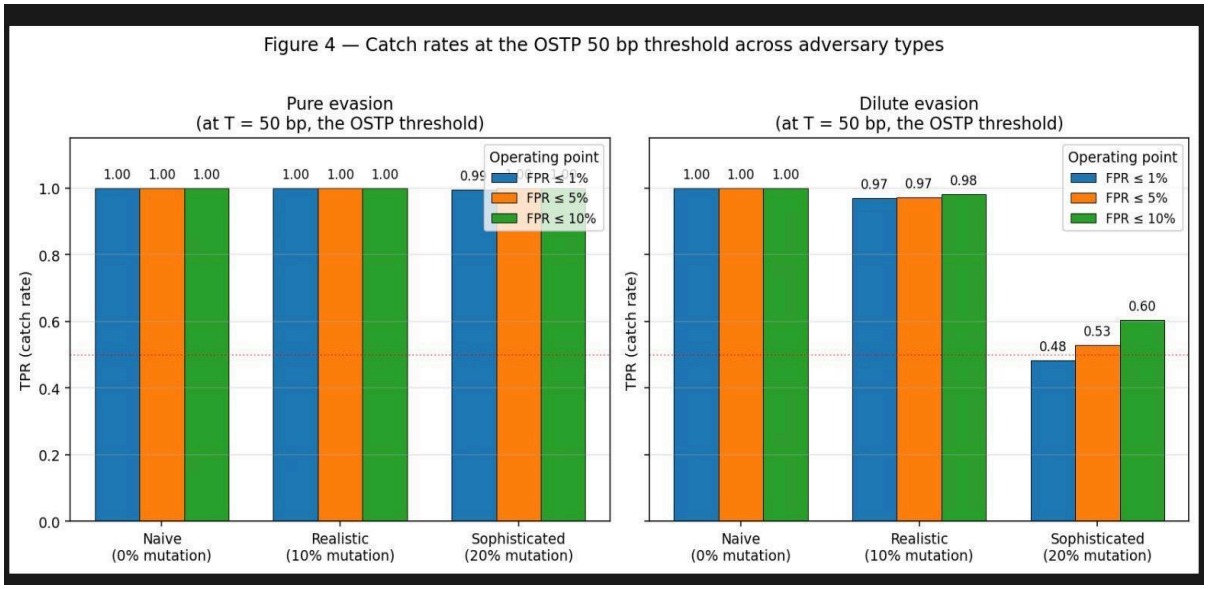

Figure C4. Catch rates at the OSTP 50 bp threshold under expanded data. Pure evasion still robustly caught; dilute evasion against sophisticated adversary drops to 48% at 1% FPR (vs. 60% in primary).

Quantitative shifts (primary → robustness): T = 50 bp, 20% mutation, dilute, TPR at 1% FPR: 0.596 → 0.483; T = 20 bp, 20% mutation, dilute: 0.123 → 0.033; T = 50 bp, 10% mutation, dilute: 0.984 → 0.969. Shifts concentrate in the small-T high-mutation cells. Two confounded effects: (1) larger BLAST databases increase e-value penalties for the same alignment, making weak hits harder to distinguish from noise; (2) the 50 kb cap admits records with proportionally

more non-coding content. We cannot disentangle these within our timeline. The qualitative findings are robust to both effects.

Appendix D. Limitations and Dual-Use Considerations

D.1 Limitations

Our results bound the alignment layer of DNA synthesis screening as exercised by BLAST under e-value thresholds. They do not directly bound the behavior of layered production pipelines.

- **False positives (Type I).** Honest FPR remained ≤ 0.3% across all configurations tested, including the diversified-honest robustness check. Whether this reflects inherent properties of curated-DB alignment or insufficient honest diversity at our experimental scale is open. Real customer order distributions span synthetic-bio constructs, primer-flank-decorated sequences, and poorly-annotated organism content that may have FPR profiles we did not measure.
- **False negatives (Type II).** False negatives are the central finding of our work — specifically the dilute-attack failure region at small T with high mutation. Our reported numbers are conservative bounds (random-substitution mutation is easier on BLAST than synonymous-substitution mutation a real adversary would use; layered pipelines may recover some sensitivity).
- **Edge cases.** Low-complexity sequences (homopolymer runs, simple repeats) trigger BLAST warnings about Karlin-Altschul parameter calculation; these are skipped by BLAST and contribute marginally to the per-fragment failure rate. Our pipeline does not pre-filter for these; their contribution is bounded by their natural frequency in the corpus (well under 1%).
- **Scalability constraints.** Our experiment runs on a curated hazardous database of 19–40 records (~250 kb to 1 Mb). Production screeners use thousands of records totaling 100+ Mb. The relative shape of our findings is expected to hold; absolute numbers may shift, particularly the e-value-dependent components of TPR.
- **Adversary model bounds.** Our adversary mutates uniformly at random per base. A real sophisticated adversary would synonymously substitute, codon-optimize, or insert primer flanks. Our 20% rate is therefore an upper bound on per-base mutation difficulty; realistic adversaries may achieve our failure region at substantially lower per-base rates.

D.2 Dual-use risks

Our work characterizes a screener's failure mode. This is the kind of finding that, taken in isolation, could be used by an adversary as a blueprint: "fragment to T ≤ 30 bp, mutate at ≥ 20%, dilute among benign filler." We weighed this risk explicitly in the project design and concluded:

- **The findings are not novel to a serious adversary.** The dynamics we measure are predictable from BLAST's statistical theory — e-values scale with database size and alignment quality, alignment quality decreases with mutation and length. A reader with a graduate-level bioinformatics background would derive these conclusions a priori. Our contribution is the quantitative measurement, not the qualitative direction.
- **The defensive value substantially exceeds the offensive value.** Quantitative measurements of where screeners fail are essential inputs for screener developers and policymakers refining thresholds. Without measurements like ours, screening guidance is set against unstated adversary models. Our work makes the implicit explicit.
- **We do not publish hazardous sequences.** Our hazardous reference set is not committed to the public repository, only a synthetic placeholder. The pipeline is reproducible against any user-supplied hazardous reference, and demonstrates the methodology rather than enabling specific attacks.

D.3 Responsible disclosure

We did not discover a vulnerability in a deployed system. We characterized a class of failure modes for an algorithm (BLAST per-fragment alignment with any-flag aggregation) that is publicly known to be used in DNA synthesis screening. The qualitative behavior is implicit in BLAST's published statistical model. We have therefore handled disclosure by direct publication; we are not aware of any specific deployment that uses BLAST in isolation (i.e., without the post-filtering layers that production pipelines include) and that would be uniquely vulnerable as a result of our findings. Should a synthesis provider determine that they are running a BLAST-only screening configuration that aligns with our threat model, we recommend they immediately layer additional capabilities (commec-style protein search and HMM detection) or raise their length threshold pending such layering.

D.4 Ethical considerations

DNA synthesis screening sits at the intersection of biosecurity, free scientific inquiry, and commercial sequence synthesis. Our work is intentionally framed to support, not undermine, the biosecurity infrastructure: we benchmark a tool widely used by responsible providers, identify a specific failure mode, and propose ways to address it. We do not provide adversary-actionable detail beyond what is implicit in BLAST's statistical theory and the published structure of commec. We acknowledge that any quantitative work on screener failure modes carries some informational risk; we believe the benefit to the screening community outweighs this.

D.5 Future improvements

- Direct comparison against commec and other layered pipelines (commec uses BLASTN as one of several layers; comparing the standalone-BLAST failure region against commec's would quantify the protein and HMM layers' contribution).

- Realistic mutation models: synonymous substitution, codon optimization, primer flank insertion, indel models.
- Cross-order experiments: dispersing fragments across multiple orders to characterize cross-order vulnerability under different splitting strategies.
- Parametric study independently varying hazardous-DB size and honest-source diversity to resolve the database-bounded vs diversity-bounded ambiguity.
- Bitscore- and identity-based decision criteria, alongside e-value, to characterize the trade-offs of different decision functions.

LLM Usage Statement

We used Claude (Anthropic) extensively as a coding and analysis collaborator throughout the project: discussing experimental design, scaffolding the four-stage pipeline, writing and debugging the screening and evaluation scripts, conducting methodological discussions including the robustness-check design, and producing draft text for this report. All experimental results were independently produced by running the pipeline on our team's machine; all numerical claims in the figures and tables are read directly from the generated CSVs. The final writeup was reviewed and edited by the team.