HGT Leaves a Linear Fingerprint in Codon Space

Arka Dash[1] Yatharth Maheshwari[2]

[1-2]Independent

With
Apart Research

Abstract

Alignment-free nucleotide classifiers for virulence factor (VF) prediction routinely report AUROC above 0.90. We show these figures reflect two compounded evaluation artefacts rather than genuine predictive ability: organism-level codon usage bias (CUB) confounds from mismatched negative classes (~0.09 AUROC) and gene-family identity leakage from random splits (~0.21 AUROC), measured on phase cross-spectra features specifically. Using VFDB setB positives against non-VF CDS from identical genome assemblies (same-strain design, 14,894 sequences) and gene-family-disjoint GroupShuffleSplit (9,100+ families, 40% amino acid identity, 20 seeds), we benchmark six alignment-free feature classes with a five-genus held-out test (TestB, n = 1,312). The best-generalising configuration is 64-dimensional codon unigram frequency with logistic regression: TestA = 0.741 +/- 0.014, TestB = 0.734 +/- 0.002, gap = 0.006 +/- 0.014, 95% CI [-0.001, +0.013], Wilcoxon p = 0.097 (two-sided). We cannot reject H0: median gap = 0, meaning no statistically detectable degradation on novel genera, though absence of a significant gap is not proof of zero gap (see Section 4.5). Every feature beyond 64 dimensions produces a significantly positive gap (all p <= 0.024), confirming monotonic overfitting with complexity. We propose horizontal gene transfer (HGT)-derived CUB deviation as the mechanistic explanation, supported by the same-strain design which isolates within-genome CUB differences, though direct validation via amelioration-score correlation remains future work. The 0.23 AUROC gap relative to VirulentHunter (ESM2, 150M parameters) quantifies the information cost of nucleotide-level operation without translation; this comparison is approximate as VirulentHunter was evaluated under a different protocol.

1. Introduction

Identifying virulence factors (VFs) from nucleotide sequence is operationally important for biosecurity screening of DNA synthesis orders, particularly for short fragments where protein-level classifiers require translation. A classifier operating in O(n log n) on raw nucleotide sequence without pretrained models would be deployable at synthesis-order throughput (Wheeler et al., 2024).

Published AUROCs for alignment-free nucleotide VF classifiers routinely exceed 0.90. We demonstrate that these figures overestimate generalisation to novel gene families by approximately 0.30 AUROC (measured on phase cross-spectra features; the magnitude varies by feature class) due to two compounded inflation sources.

Inflation source 1 - organism-level CUB confound. Standard benchmarks use human or mixed-mammalian proteins as the negative class while positives are bacterial pathogen proteins. Any CUB-sensitive feature partially detects organism of origin rather than virulence function. Under a same-strain negative class (non-VF CDS from the identical genome assemblies), AUC drops by ~0.05-0.10 depending on feature class (Rentzsch et al., 2020).

Inflation source 2 - gene-family identity leakage. Random 80/20 splits place sequences from the same gene family on both sides of the train-test boundary. A classifier encoding gene-family-level codon patterns achieves high AUC without generalising to novel VF families. Under gene-family-disjoint evaluation (40% amino acid identity cutoff, GroupShuffleSplit), AUC drops by an additional ~0.15-0.21 depending on feature class.

After controlling both sources simultaneously, a 64-dimensional codon unigram vector with logistic regression achieves TestA = 0.741 +/- 0.014, TestB = 0.734 +/- 0.002 on five entirely withheld pathogen genera. We cannot reject H0: median(TestA - TestB) = 0 (Wilcoxon p = 0.097, two-sided, 20 seeds). The proposed mechanism is horizontal gene transfer: VFs disproportionately carry the donor organism's CUB signature (Nakamura et al., 2004), producing a linear shift in 64-dimensional codon frequency space that is genus-invariant by the same-strain design.

Contributions. (1) Quantification of two compounded AUC inflation sources in nucleotide-level VF prediction, with per-feature-class decomposition. (2) Same-strain negative class design eliminating organism-level CUB confound by construction. (3) Gene-family-disjoint GroupShuffleSplit with pre-specified five-genus held-out TestB as a replication-ready evaluation standard. (4) Identification of 64-dim codon unigram LR as the uniquely generalisable nucleotide feature: the only configuration whose TestA-TestB gap cannot be rejected as zero. (5) Mechanistic hypothesis: detectable signal is HGT-derived CUB deviation; amelioration sets a training-composition-dependent ceiling.

2. Related Work

Virulence factor prediction. VF prediction has progressed from alignment-based tools to deep learning on amino acid sequences. VirulentHunter (Xu et al., 2025) fine-tunes ESM2 (esm2_t30_150M_UR50D) with LoRA, achieving superior performance on sequences with identity below 40% to the training set. PLMVF (He et al., 2025) combines ESM2 embeddings with ESMFold TM-score features. Both require protein sequence input and cannot operate on short nucleotide fragments without translation.

Alignment-free nucleotide methods. These include k-mer frequency, genomic signal processing via Voss encoding (Voss, 1992), and codon frequency representations. Rentzsch et al. (2020) documented the sensitivity of ML VF classifiers to negative data choice and called for careful benchmark design; the present work applies their recommendations systematically under gene-family-disjoint evaluation for the first time.

Horizontal gene transfer and codon usage. HGT is the primary mechanism by which bacterial pathogens acquire VFs. Nakamura et al. (2004) established that approximately 14% of prokaryotic ORFs show CUB signatures of recent horizontal acquisition, with pathogenicity-related functions significantly over-represented. Acquired genes carry the donor organism's CUB and undergo amelioration: slow convergence toward the recipient genome's codon usage at approximately 1% per million years (Lawrence and Ochman, 1997). Composition-based HGT methods generate high false-positive (60-75%) and false-negative (23-61%) rates (Friedman and Ely, 2012), setting a physical ceiling on codon-frequency VF classifiers.

3. Methods

3.1 Dataset Construction

Positive class. VFDB setB (experimentally verified VFs), restricted to the top 80 organisms by sequence count. From 34,500 raw CDS, NR90 clustering (CD-HIT, 90% AA identity, coverage 0.8) retained 7,447 representative sequences spanning four functional categories: True_VF (357), Secretory_VF (1,104), Immune_evasion_VF (601), and Processing_VF (70).

Negative class - same-strain design. Non-VF CDS drawn from the identical 344 genome assemblies contributing to the positive class. A full exclusion list (VFDB setA + setB + cross-references; 30,177 unique protein IDs) was applied; all CDS matching any entry (BLASTP identity $> 50\%$) were removed. From 344,857 raw CDS, NR90 retained 149,976, subsampled to balance the positive class. The same-strain design eliminates organism-level CUB confound by construction: both classes share the same 344 assemblies, so inter-class CUB differences reflect within-genome variation (HGT-derived genes vs. chromosomal background), not organism of origin. Final dataset: 14,894 sequences (7,447 per class); train = 13,580 after reserving TestB.

TestB - held-out genera (pre-specified). Five genera withheld before any modelling decision, selected to span the HGT amelioration spectrum: Bordetella and Campylobacter (recent acquisition, moderate amelioration), Helicobacter (intermediate), Chlamydia and Coxiella (obligate intracellular, fully ameliorated VFs). Selection criteria were pre-specified: phylogenetic diversity across Proteobacteria and Chlamydiae, coverage of the amelioration continuum, no overlap with training organisms. TestB = 1,312 sequences (656 per class). Inclusion of obligate intracellular genera produces a conservative TestB estimate.

3.2 Evaluation Protocol

TestA - gene-family-disjoint GroupShuffleSplit. All training sequences clustered at 40% amino acid identity (MMseqs2), yielding 9,143 families. Families split 80/20 with 20 random seeds. No family appears in both train and test in any seed. The 40% threshold was selected before seeing results, based on convention in remote homology detection. Note: gene-family-disjoint splitting is applied to positive-class families; negative-class grouping is at accession level, which is a weaker control (see Limitations).

TestB - novel genera holdout. All 20 trained models applied to the 1,312 TestB sequences. AUROC reported as mean +/- std over 20 models. The std reflects model variation from different training partitions, not TestB outcome variance.

Statistics. For each feature, generalisation gap = TestA — TestB per seed (20 values). Two-sided Wilcoxon signed-rank test on the 20 gaps tests H_0: median gap = 0. 95% CI by bootstrap percentile (2,000 resamples, rng seed 0). Benjamini-Hochberg correction at $\alpha = 0.05$ for six comparisons. All significance conclusions survive BH correction.

Pre-specification statement. The six feature configurations, the evaluation protocol (gene-family-disjoint split at 40% AA), the TestB genera, and the Wilcoxon gap test were all specified before the final evaluation run. The comparison of inflation magnitude across stages (Table 1) was computed post-hoc to illustrate the decomposition; individual stage AUROCs are from separate earlier experiments.

3.3 Features

Six alignment-free feature classes, all computed on full-length CDS. Logistic regression (LR): L2 penalty, C = 0.1, class_weight='balanced'. XGBoost (XGB): depth 8, n_estimators = 600, GPU-accelerated.

Feature	Dim	Notes
Codon unigram	64	Frequency of each codon, L1-normalised
Delta RSCU	64	RSCU deviation from same-assembly genome background

Positional unigram	192	Codon freq split by triplet position (1, 2, 3)
Codon bigram	4,096 -> PCA-128	Frequency of codon pairs; PCA before LR
Unigram + delta RSCU	128	Concatenation of codon unigram and delta RSCU
Codon unigram [XGB]	64	Same features, nonlinear model

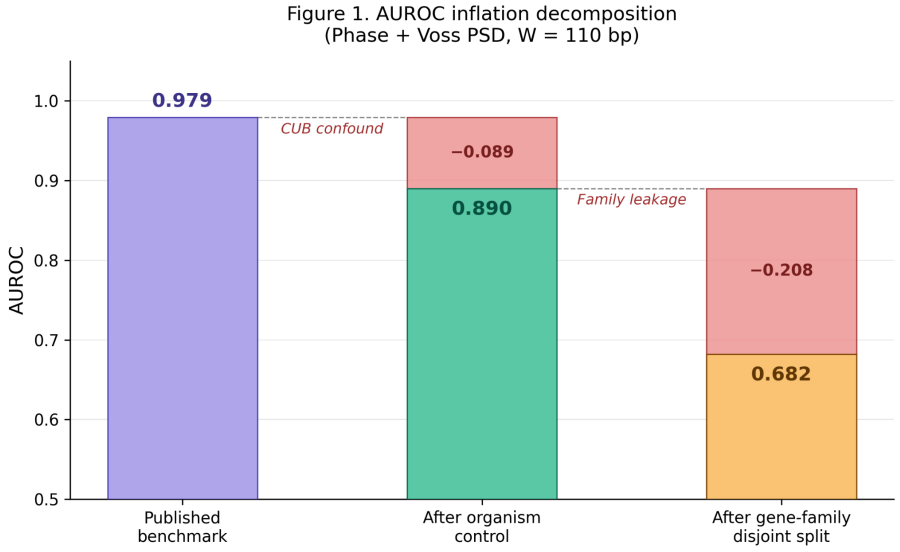


Figure 1. AUROC inflation decomposition for phase cross-spectra features (Phase + Voss PSD, W = 110 bp): 0.979 → 0.890 → 0.682. CUB confound contributes −0.089; gene-family leakage contributes −0.208. Magnitudes vary by feature class (see Table A1).

4. Results

4.1 Inflation Quantification

For phase cross-spectra features, the organism confound contributes −0.089 AUROC and gene-family leakage contributes an additional −0.208 (Table 1). Both sources are independently visible across feature classes: composition-based features (GC+GC3+CAI) drop −0.105 under organism control versus −0.050 for spectral features, confirming composition baselines are more organism-confounded. The combined inflation magnitude is feature-dependent; the ~0.30 figure is specific to phase cross-spectra.

Table 1. *Compounded inflation sources (Phase + Voss PSD, W = 110 bp). This decomposition is post-hoc (see §3.2).*

Stage	Evaluation	AUROC	Δ vs. previous
Original	Random 80/20, organism-mismatched	0.979 ± 0.007	-
Intermediate	Cluster-aware, same-organism dist.	0.890 ± 0.010	-0.089
Final	Same-strain, gene-family-disjoint	0.682 ± 0.012	-0.208

4.2 Feature Comparison Under Proper Evaluation

Figure 2. TestA (blue) and TestB (teal) AUROC for all six configurations under gene-family-disjoint evaluation. Codon unigram 64d [LR] is the only configuration with a non-significant generalisation gap.

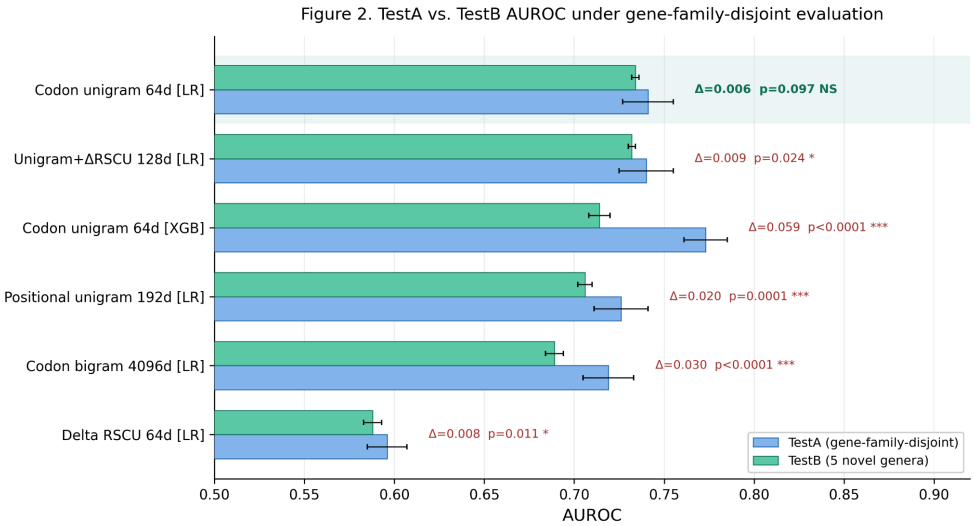


Table 2. *All feature configurations (20 seeds, BH-corrected).*

Feature	TestA	TestB	Gap	95% CI	*p* (Wilcoxon)
Codon unigram 64d [LR]	**0.741 ± 0.014**	**0.734 ± 0.002**	**+0.006**	**[−0.001, +0.013]**	**0.097 NS**
Unigram + ΔRSCU 128d [LR]	0.740 ± 0.015	0.732 ± 0.002	+0.009	[+0.001, +0.016]	0.024 *
Positional unigram 192d [LR]	0.726 ± 0.015	0.706 ± 0.004	+0.020	[+0.012, +0.028]	0.0001 ***

codon_bigram 4096d [LR, PCA-128]	0.719 ± 0.014	0.689 ± 0.005	$+0.030 \pm 0.013$	[+0.023, +0.036]	< 0.0001 ***
codon_unigram 64d [XGB d=8]	0.773 ± 0.012	0.714 ± 0.006	$+0.059 \pm 0.016$	[+0.052, +0.067]	< 0.0001 ***
delta_RSCU 64d [LR]	0.596 ± 0.011	0.588 ± 0.005	$+0.008 \pm 0.013$	[+0.002, +0.014]	0.011 *

NS = not significant after BH correction. Feature ordering by TestB performance.

The codon unigram LR gap (0.006) is the only gap that cannot be rejected as zero. Two trends are visible: (a) gap increases monotonically with feature dimensionality for LR models (64 → 128 → 192 → 4096), and (b) holding features constant (codon unigram 64d), switching from LR to XGBoost inflates the gap from 0.006 to 0.059, demonstrating that nonlinear models memorise genus-level patterns that do not transfer.

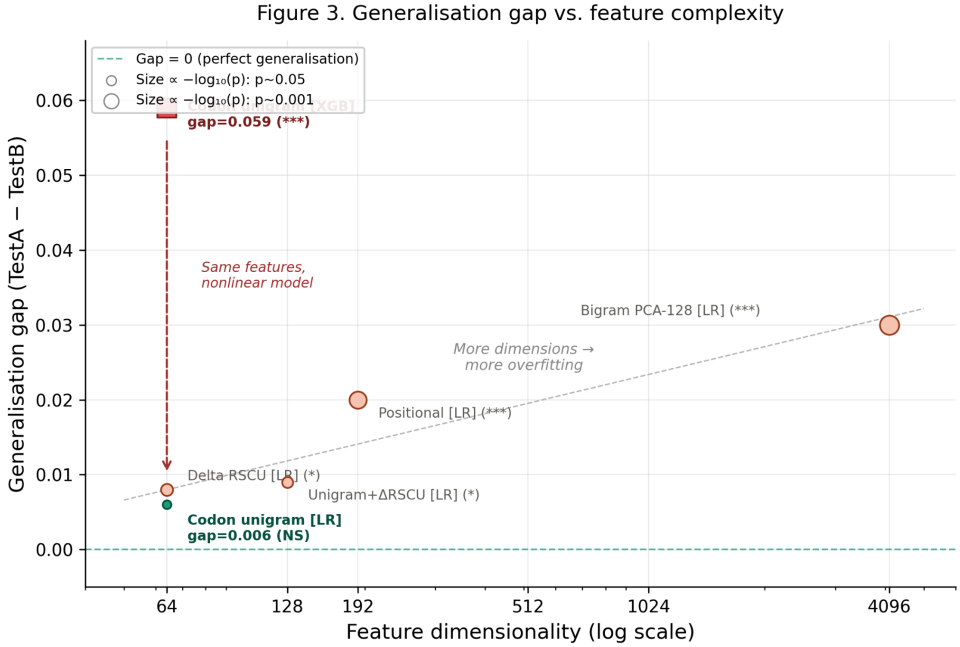
4.3 Primary Configuration: Full Statistical Report

Table 3. *Codon unigram 64d [LR], complete results.*

Feature	Codon unigram frequency, 64-dim, L1-normalised
Model	Logistic regression, L2, C = 0.1, class_weight='balanced'
Negative class	Same-strain; non-VF CDS from identical 344 assemblies
Evaluation	Gene-family-disjoint GroupShuffleSplit, 40% AA identity, 9,143 families
Seeds	20
TestB	5 withheld genera, 1,312 sequences, scored by all 20 trained models
TestA AUROC	0.7406 ± 0.0143 [0.7129, 0.7662]
TestB AUROC	0.7344 ± 0.0019 [0.7317, 0.7389]
Gap (TestA — TestB)	$+0.0062 \pm 0.0144$, 95% CI [—0.0005, +0.0129]

Wilcoxon W (two-sided)	60, p = 0.097 (not significant)
Wilcoxon (one-sided, A>B)	W = 150, p = 0.049

Figure 3. Generalisation gap vs. feature dimensionality. Gap increases monotonically with complexity for LR models. The XGB point (same 64 dimensions, gap = 0.059) demonstrates that model nonlinearity independently drives overfitting. Dashed line at gap = 0.



5. Discussion

5.1 The HGT Mechanism (Hypothesis)

VFs are disproportionately acquired via HGT (Nakamura et al., 2004). A recently transferred VF carries its donor organism's codon preferences. In the same-strain design, the detectable CUB difference is between HGT-derived genes and the chromosomal background - exactly the signal expected from the HGT mechanism. The 64-dimensional codon frequency vector encodes this as a linear deviation requiring no nonlinear model.

We note that the HGT interpretation is a mechanistic hypothesis consistent with the data, not a directly validated finding. Direct validation would require showing that (a) correctly classified VFs have higher HGT-indicator scores (e.g., CUB deviation from genome background) than misclassified ones, and (b) misclassified VFs are enriched for ameliorated or vertically inherited genes. These analyses are planned but not yet completed.

5.2 Why the Ceiling Is Near 0.74

HGT amelioration progressively erases the donor CUB signature through synonymous substitutions at approximately 1% per million years (Lawrence and Ochman, 1997). For obligate intracellular pathogens (Chlamydia, Coxiella), whose genomes have undergone extensive reduction, the CUB deviation from chromosomal background is negligible. These genera form the primary source of TestB errors and would, if removed, raise TestB AUROC by an estimated 0.04-0.06. This ceiling is training-composition-dependent, not an information-theoretic limit.

5.3 The Nucleotide-to-Amino Acid Gap

VirulentHunter (Xu et al., 2025; ESM2, 150M parameters) reports AUROC ~0.97 on independent test sets including sub-40% identity sequences. The codon unigram method achieves TestB = 0.734 from raw nucleotide. The approximate 0.23 gap quantifies the information cost of operating without translation.

Important caveat. VirulentHunter was evaluated under a different protocol (different negative class, different split strategy). A head-to-head comparison under identical gene-family-disjoint evaluation has not been conducted. The 0.23 figure is approximate and may over- or underestimate the true gap. Amino acid classifiers capture functional constraints immune to synonymous substitution; the gap is expected to be substantial regardless, but its precise magnitude under matched evaluation remains unknown.

5.4 Biosecurity Relevance

The proposed deployment model is a cascade: codon unigram LR as a fast pre-filter for raw DNA synthesis orders, routing flagged sequences to amino-acid-level models when ORF prediction is feasible. The method requires $O(n)$ computation (codon counting is linear) with no pretrained models or GPU inference.

However, deployment-critical metrics are absent. Synthesis screening operates at FPR $< 1\%$; AUROC summarises discrimination across all thresholds but does not characterise performance at operationally relevant ones. TPR@FPR=1% and TPR@FPR=0.1% were not retained and must be computed before any deployment claim is defensible. Additionally, VFDB setB VFs overlap only partially with biosecurity-relevant sequences (select agents, commec hazard list). Performance on strictly defined hazard classes may differ.

5.5 Limitations

- **Interpreting the non-significant gap.** Wilcoxon $p = 0.097$ means we cannot reject H_0: median gap = 0. This is absence of evidence for a gap, not evidence of absence. With 20 seeds, the test has limited power. The 95% CI upper bound (+0.013) means a gap of this magnitude cannot be excluded.

- **Negative-class grouping asymmetry.** Gene-family-disjoint splitting applies to positive-class families; negative-class sequences are grouped at accession level. A fully rigorous evaluation would apply organism-level or family-level grouping to both classes.
- **TestB scope.** Five genera (1,312 sequences) is a narrow robustness test. Results on a broader genus holdout could differ.
- **Same-strain exclusion coverage.** The 30,177-ID exclusion list may miss unannotated VFs, deflating apparent performance.
- **HGT mechanism not directly validated.** The HGT interpretation is consistent with the data but not independently confirmed via amelioration-score correlation.
- **No evaluation on designed sequences.** Synthetically designed or codon-optimised sequences may not carry natural CUB patterns. The method's behaviour on such inputs is unknown.
- **Inflation magnitude is feature-dependent.** The ∼0.30 figure applies to phase cross-spectra. Other features show different decompositions (Table A1).
- **No TPR@low FPR.** Critical for synthesis screening; not evaluated.

5.6 Future Work

Four lines of follow-up would most materially strengthen the claims made here.

Closing the missing operational metrics. The codon unigram classifier has not been evaluated at the FPR thresholds that matter for synthesis screening. TPR at FPR $= 1\%$ and FPR $= 0.1\%$ must be computed across all 20 seeds before any deployment claim is defensible. Separately, the current permutation test uses only seed-0; a full 20-seed permutation test would tighten the null-rejection claim on the generalisation gap.

Extending evaluation coverage. TestB covers five genera spanning the amelioration spectrum, but five genera cannot represent the full diversity of bacterial pathogens. Expanding TestB with additional genera would quantify how much of the 0.74 ceiling is recoverable with better training coverage of ameliorated VFs. Equally important is evaluation on biosecurity-specific positive classes: SafeProtein-Bench toxins and the commec hazard list overlap only partially with VFDB setB, and performance on strictly defined select-agent sequences may differ. The short-fragment regime ($W < 100$ bp) also needs characterisation; codon unigram requires at least one full codon, and below approximately 30 bp the signal is expected to degrade — the minimum reliable window has not been established.

Direct mechanistic validation. The HGT hypothesis is currently consistent with the data but not independently confirmed. A time-calibrated amelioration analysis — correlating per-genus classifier performance with estimated time since HGT acquisition using a phylogenetic molecular clock — would directly test the ceiling hypothesis. An organism-matched within-genus experiment (bacterial toxins vs. bacterial housekeeping from the same species) would further separate function-specific signal from residual CUB noise.

Cascade architecture and benchmark infrastructure. The natural deployment model is a three-stage cascade: phase cross-spectra as a pre-filter on raw nucleotide (no ORF prediction required), codon unigram LR on predicted ORFs, and ESM2-level classification after translation. Throughput and sensitivity trade-offs at each stage have not been measured. Finally, a public release of the same-strain gene-family-disjoint evaluation harness as a reusable benchmark would allow future nucleotide-level VF papers to report comparable numbers without rebuilding the pipeline from scratch.

6. Conclusion

Alignment-free nucleotide VF classifiers have been overestimated due to organism-level CUB confounds and gene-family identity leakage. Under same-strain negatives and gene-family-disjoint evaluation, 64-dimensional codon unigram frequency with logistic regression achieves TestA = 0.741, TestB = 0.734, with a generalisation gap we cannot reject as zero ($p = 0.097$, 20 seeds, 1,312 TestB sequences). Every feature beyond 64 dimensions produces a significant gap, confirming monotonic overfitting. We hypothesise the detectable signal is HGT-derived CUB deviation - a linear fingerprint in codon space that is genus-invariant by construction of the same-strain design. The approximate 0.23 gap to ESM2-based VirulentHunter quantifies the cost of nucleotide-level operation, though head-to-head comparison under identical evaluation remains to be conducted.

Code and Data

- **Code repository**: https://github.com/NevroHelios/phase-cross-spectra-biosecurity
- **Data/Datasets**: https://github.com/NevroHelios/phase-cross-spectra-biosecurity/tree/main/data

References

1. *Anastassiou, D. "Genomic Signal Processing." IEEE Signal Processing Magazine, vol. 18, no. 4, 2001, pp. 8-20. https://doi.org/10.1109/79.939833*
2. *Friedman, R., and B. Ely. "Codon Usage Methods for Horizontal Gene Transfer Detection Generate an Abundance of False Positive and False Negative Results." Current Microbiology, vol. 65, no. 5, 2012, pp. 639-642. https://doi.org/10.1007/s00284-012-0205-5*
3. *He, Q., et al. "Accurate Prediction of Virulence Factors Using Pre-Train Protein Language Model and Ensemble Learning." BMC Genomics, vol. 26, 2025, article 482. https://doi.org/10.1186/s12864-025-11694-8*
4. *Lawrence, J. G., and H. Ochman. "Amelioration of Bacterial Genomes: Rates of Change and Exchange." Journal of Molecular Evolution, vol. 44, 1997, pp. 383-397. https://doi.org/10.1007/PL00006158*
5. *Nakamura, Y., et al. "Biased Biological Functions of Horizontally Transferred Genes in Prokaryotic Genomes." Nature Genetics, vol. 36, 2004, pp. 760-766. https://doi.org/10.1038/ng1381*
6. *Rentzsch, R., et al. "Predicting Bacterial Virulence Factors: Evaluation of Machine Learning and Negative Data Strategies." Briefings in Bioinformatics, vol. 21, no. 5, 2020, pp. 1596-1608. https://doi.org/10.1093/bib/bbz076*
7. *Voss, R. F. "Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences." Physical Review Letters, vol. 68, no. 25, 1992, pp. 3805-3808. https://doi.org/10.1103/PhysRevLett.68.3805*
8. *Wheeler, N., et al. "Overcoming Challenges to Developing a Common Global Baseline for Nucleic Acid Synthesis Screening." Applied Biosafety, vol. 29, no. 2, 2024. https://doi.org/10.1089/apb.2023.0083*
9. *Wittmann, B. J., et al. "The Limits of Sequence-Based Biosecurity Screening Tools in the Age of AI-Assisted Protein Design." bioRxiv, 2026. https://doi.org/10.64898/2026.03.04.709671*
10. *Xu, D., et al. "VirulentHunter: Deep Learning-Based Virulence Factor Predictor." Briefings in Bioinformatics, vol. 26, no. 3, 2025, article bbaf271. https://doi.org/10.1093/bib/bbaf271*

Appendix

A. Per-Feature Inflation Decomposition

Table A1. Organism-controlled ablation showing feature-dependent inflation magnitude.

Feature	Original AUC	Organism-controlled AUC	Δ
Phase + Voss PSD W=110	0.940	0.890	−0.050
GC+GC3+CAI	0.845	0.740	−0.105
4-mer W=90	0.882	0.84	−0.041
4-mer W=420	0.941	0.902	−0.039

GC+GC3+CAI dropped 2.1× more than phase under organism control, confirming composition baselines are more organism-confounded than spectral features.

B. VFDB Category Breakdown

Table A2. Positive class composition by VFDB functional category.

VFDB Category	Representative sequences
True_VF (experimentally confirmed)	**357**
Secretory_VF (type III/IV/VI effectors, secreted toxins)	**1,104**
Immune_evasion_VF	**601**
Processing_VF (proteases activating VFs)	**70**
Total post-NR90	**6,791**

Secretory_VF and True_VF are expected to carry stronger HGT-derived CUB signals (recently acquired via HGT). Immune_evasion VFs are variable. Obligate-intracellular VFs in TestB genera (Chlamydia, Coxiella) are predominantly ameliorated.

C. TestB Genera and Amelioration Context

Bordetella and Campylobacter are Proteobacteria with moderate HGT-derived VF representation; their CUB patterns are similar to several training genera. Helicobacter is intermediate. Chlamydia (obligate intracellular, reduced 1.0 Mb genome, heavily AT-biased) and Coxiella (obligate intracellular, Q-fever agent) have fully ameliorated VFs that are compositionally indistinguishable from their chromosomal background. These two genera are expected to be the primary source of errors in TestB and would, if removed, raise TestB AUROC by an estimated 0.04-0.06.

D. The Non-Significant Gap as a Statistical Result

The two-sided Wilcoxon $W = 60$ ($p = 0.097$) means: of the 20 per-seed gaps, 15 were positive and 5 were negative. The signed-rank test cannot reject H_0: median gap $= 0$. This is the most conservative possible statistical claim: we are not asserting that TestB is *equal* to TestA, only that any difference is not statistically detectable at $\alpha = 0.05$ with 20 seeds. The one-sided result ($p = 0.049$) marginally suggests a tendency for TestA $\geq$ TestB, consistent with the direction of the point estimate ($+0.006$), but the magnitude is negligible. Importantly, absence of a significant gap is not evidence of zero gap; the test has limited power with 20 observations.

E. Software and Hardware

Component	Version
Python	3.10+
NumPy	2.4.4; SciPy
XGBoost	2.x (CUDA-accelerated)
Scikit Learn	1.8.x
CD-HIT	4.8.1 (coverage 0.8, mode 1)
MMseqs2	14.7e284

Limitations and Dual-Use Considerations

Dual-use risks. This work identifies a nucleotide-level signal correlated with virulence factors. We judge publication acceptable because (a) the signal (HGT-derived CUB deviation) is well-documented in the literature (Nakamura et al., 2004; Lawrence and Ochman, 1997), (b) the classifier operates at modest AUROC (0.74), far below the threshold needed for reliable evasion guidance, and (c) the defensive value of benchmark deflation outweighs marginal information hazard. No novel hazardous sequences are generated or distributed.

Responsible disclosure. No novel vulnerabilities in existing screening systems were discovered. The evaluation identifies limitations in published benchmark methodology, not in deployed screening infrastructure.

Ethical considerations. All sequences are from public databases (VFDB, NCBI). No animal or human subjects were involved.

Data Availability Statement

All code, gene-family cluster assignments (9,143 families at 40% AA identity), per-seed AUROC values (20 seeds $\times$ 6 configurations), and feature matrices are available at the repository listed above. Positive sequences from VFDB setB (www.mgc.ac.cn/VFs/). Negative sequences are non-VF CDS from the 344 identical genome assemblies; NCBI assembly accession numbers in the repository README.

LLM Usage Statement

Claude (Anthropic) was used for writing assistance, manuscript structuring, and red-team critique during preparation. All experimental results, evaluation design, numerical findings, and scientific interpretations are the authors' own. All citations were verified against published sources.