GENOTRACE: INHERITABLE WATERMARKS FOR GENOME FOUNDATION MODEL DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a genome model retain a detectable record of the synthetic sequences used to train it? We study watermark inheritance through distillation with GenoTrace, a codon-aware extension of green-list watermarking. Two token-level factors modulate the teacher's generation bias using codon position and organism-specific codon usage. The resulting sequences train a smaller student, whose outputs are audited without an active watermark processor. In a three-seed GenomeOcean-500M-to-100M experiment, the joint configuration achieves a mean audit score of 17.88 and 94.5% detection at a fixed threshold. It retains 49.0% detection after key-aware token substitution, compared with 0% for the available single-seed plain-watermark comparator, and 47.0% after combined mechanism-targeted nucleotide edits. Additional experiments establish inherited signal across five organism-conditioned datasets and teacher–student size ratios up to 40. Component ablations and computational sequence-quality assays reveal distinct operating points for detection strength and coding coverage. GenoTrace provides a practical token-level construction and an empirical account of how genomic structure shapes inherited watermark signals. The findings concern shared-tokenizer distillation and the tested editing procedures, with calibration and biological utility treated as separate evaluation requirements.

1 INTRODUCTION

Genome foundation models make synthetic DNA a reusable resource for downstream learning. Models such as GenomeOcean and Evo span a wide range of parameter scales and genomic contexts (Wu et al., 2025; Nguyen et al., 2024; Brixi et al., 2026). Their outputs can also become training examples for smaller models. Distillation transfers behavior from a teacher to a student (Hinton et al., 2015; Gou et al., 2021; Sanh et al., 2019), but the resulting model does not ordinarily reveal which synthetic corpus influenced its parameters. Detecting that influence would provide a useful component of model provenance.

Watermark inheritance supplies a possible trace. A teacher biases its output distribution using a secret key; a student trained on those outputs may learn a residual of the bias. Sander et al. (2024) establish this phenomenon in text models. Genomic data introduce a structural complication: models can share the four-letter DNA alphabet while using different token vocabularies, and subword tokens need not align with codons. A watermark that is easy to learn in token space may also change coding structure or sequence diversity. The design question is therefore how to allocate a watermark bias across genomic tokens, not simply how to make the bias stronger.

Consider a provider that supplies synthetic genomic sequences for downstream training. A second party trains a smaller generator on that corpus, then exposes only the student's outputs. The provider would like to test whether the marked distribution influenced the student without inspecting its parameters. This setting differs from detecting whether one sequence was directly generated by the teacher: the evidence must survive a learning step. It also differs from identifying a copied checkpoint, since the student can have a different architecture or parameter count.

GenoTrace addresses this setting by allocating generation bias using genomic structure. Its position factor measures the fraction of a token overlapping third codon positions. Its usage factor measures the fraction of complete codons inside the token belonging to an organism-specific preferred set. Their product modulates a keyed green-list bias, without changing the teacher architecture or student

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

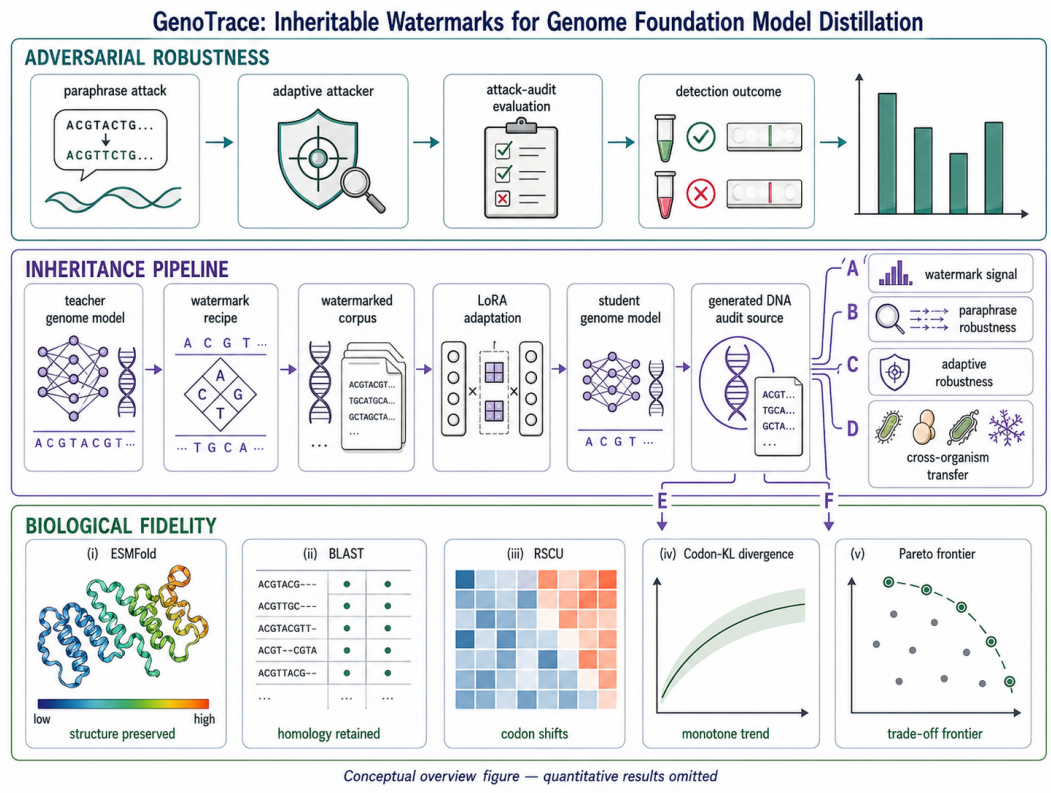

Figure 1: GenoTrace framework. A teacher generates watermarked DNA, a smaller student learns from these sequences, and the auditor scores student outputs without an inference-time watermark processor. The surrounding panels identify output-editing tests and computational quality assays. The schematic's preservation and constraint labels indicate design objectives; measured outcomes are reported in Section 5.

objective. A vocabulary-by-phase lookup connects these nucleotide-level preferences to variable-length model tokens. Figure 1 places this construction within the generation, learning, and audit pipeline.

We organize the study around three questions. Does the signal survive shared-tokenizer distillation, and how does inheritance vary with organism and model scale? How much signal remains after explicitly specified output edits? Which detection operating points preserve more sequence-level quality? These questions require different controls: an unwatermarked student tests background positives, an edited plain-watermark student tests a reference construction, and biological assays evaluate properties that the detector does not measure.

Our contributions are the codon-aware token-bias construction, a component-and-strength study with seed-level reporting, and a joint analysis of inherited signal and computational sequence quality. In the main GenomeOcean-500M-to-100M campaign, joint fractional weighting reaches mean audit score 17.88 and 94.5% detection. It retains 49.0% after key-aware token substitution and 47.0% after combined mechanism-targeted nucleotide edits. The extensions and quality analyses explain where these findings generalize and where additional calibration or validation is needed. We treat these as measured operating points, not guarantees of universal attribution or preserved biological function.

2 Related Work

Watermark design and inheritance. Green-list watermarking promotes a keyed subset of tokens during generation (Kirchenbauer et al., 2023). Reliability studies examine editing and partial retention (Kirchenbauer et al., 2024), while unbiased and distortion-free constructions target

sampling-distribution preservation (Hu et al., 2024; Kuditipudi et al., 2024). Cryptographic undetectability (Christ et al., 2024), provable edit robustness (Zhao et al., 2024), semantic partitions (Hou et al., 2024), and deployed sampling watermarks (Dathathri et al., 2024) address different design objectives. Inheritance concerns a further stage: a model trained on watermarked outputs. Sander et al. (2024) show that detection at this stage requires attention to repeated contexts and the statistical unit. We distinguish the observed inherited score from a calibrated provenance decision.

Attacks and evaluation. Paraphrasing, watermark stealing, and contrastive distillation expose distinct attack surfaces (Krishna et al., 2023; Jovanović et al., 2024; Yi et al., 2025). Pan et al. (2025) study removal before distillation and neutralization at inference; Diaa et al. (2025) optimize attacks against watermark objectives. SEEK addresses scrubbing and spoofing through vocabulary-side design (Shen et al., 2025). These studies motivate explicit access and quality budgets. Our tests apply specified transformations to student outputs rather than reproducing the complete text-attack training pipelines.

Genomic and biological provenance. DNAMark and CentralMark learn watermark logits from DNA or protein embeddings (Zhang et al., 2025). DNAMark constrains synonymous choices, whereas CentralMark modifies the second codon base to support detection after translation. FoldMark operates on protein structures (Zhang et al., 2024). GenoTrace instead asks whether generation bias is learned through model training. Genome models also differ substantially in tokenization and architecture (Nguyen et al., 2023; Zhou et al., 2024; Dalla-Torre et al., 2025; Sanabria et al., 2024). Codon degeneracy and usage variation motivate our factors (Crick, 1966; Plotkin & Kudla, 2011; Hershberg & Petrov, 2008), but neither factor guarantees functional neutrality. Appendix F situates the broader detection, biological-design, and statistical literature.

3 Codon-aware watermark inheritance

3.1 Setting and access model

The provider controls teacher sampling and the key used to mark its training outputs. The student trainer receives sequences rather than teacher logits and optimizes next-token prediction. The auditor observes student responses and knows the marking key and teacher tokenizer; student weights and inference logits are unnecessary for the score. Our principal experiments use compatible teacher and student vocabularies so that the same transition-level event can be reconstructed. Neither student evaluation nor auditing injects a new watermark.

An output editor sits between student generation and auditing. We separate key-aware token editing from keyless transformations targeted at codon position or usage. These capabilities are not nested: knowing the mechanism is not the same as knowing the key, and inserting bases is not a budget-matched substitute for token replacement. Accordingly, we report each transformation with its access, sampling rate, and source control. Training-time removal, adaptive detector-query optimization, and spoofing by a key holder are different experiments, not implied successes of this evaluation.

3.2 Generation and audit

For vocabulary $\mathcal{V}$ with 4096 entries, the key K and preceding token x_{t-1} determine a pseudorandom green set $\mathcal{G}_t$ of size $\gamma|\mathcal{V}|$, with $\gamma = 0.25$. A teacher logit $\ell_t(v)$ becomes

$$\ell'_t(v) = \ell_t(v) + \delta\,\mathbf{1}[v \in \mathcal{G}_t]a(v, p_t), \qquad a(v, p_t) \in [0, 1], \tag{1}$$

where p_t is the nucleotide offset relative to the estimated reading frame. Plain KGW uses $a = 1$. The codon-aware factors below are key-independent. We train the student on generated sequences using low-rank adaptation (Hu et al., 2022), then remove the generation processor from the evaluation path. Thus, any measured student signal comes from its learned distribution rather than an active inference-time watermark.

The auditor reconstructs green sets with the teacher's tokenizer and key. For L response tokens, it scores the $T = L - 1$ observed transitions:

$$z = \frac{G - \gamma T}{\sqrt{T\gamma(1-\gamma)}}, \qquad G = \sum_{t=2}^{L} \mathbf{1}[x_t \in \mathcal{G}_t]. \tag{2}$$

Detection denotes $z > 2.326$ unless stated otherwise. The corresponding Gaussian tail is nominally 1%; the empirical false-positive rate must be measured separately. Repeated contexts and shared prompts mean that a large aggregate score is not automatically a valid extreme-significance claim.

3.3 Position and organism-usage factors

Let $s(v)$ decode token v into $m(v)$ nucleotides. Fractional and binary position weights are

$$f_F(v,p) = \frac{1}{m(v)} \sum_{j=0}^{m(v)-1} \mathbf{1}[(p+j) \bmod 3 = 2], \qquad f_B(v,p) = \mathbf{1}[f_F(v,p) > 0]. \tag{3}$$

Empty or non-nucleotide tokens receive zero weight. The fractional factor approaches one third for long tokens; the binary factor is one whenever any third position is present. This token-level distinction matters for a subword vocabulary and cannot be replaced by a single nucleotide-position mask.

For organism o, let Ω_o contain codons at or above the upper-middle usage value within each synonymous family, including ties and single-codon families. The tables are based on Kazusa codon usage (Nakamura et al., 2000). If $\mathcal{C}(v,p)$ is the collection of complete in-frame codons contained within a token, its usage weight is

$$q_o(v,p) = \begin{cases} \dfrac{\sum_{c \in \mathcal{C}(v,p)} \mathbf{1}[c \in \Omega_o]}{|\mathcal{C}(v,p)|}, & |\mathcal{C}(v,p)| > 0, \\ 1/2, & |\mathcal{C}(v,p)| = 0. \end{cases} \tag{4}$$

Codons spanning token boundaries are not counted in this lookup. Position-only uses $a = f_F$ or f_B; usage-only uses $a = q_o$; joint weighting uses their product. The main joint configuration is fractional with $\delta = 8$ and the E. coli usage table. Usage weighting concentrates on preferred codons rather than matching the organism's entire distribution.

The token boundary is essential to interpreting these weights. For the nucleotide string `ATCG` starting at phase zero, one of four bases occupies a third codon position, so $f_F = 1/4$ while $f_B = 1$. Moving the start to phase two gives $f_F = 1/2$ without changing the binary factor. For a token containing one complete preferred codon and one nonpreferred codon, $q_o = 1/2$, not a hard exclusion. Thus fractional position weighting, binary position weighting, and usage weighting are different soft biases even at the same nominal δ.

For implementation, we precompute a $|\mathcal{V}| \times 3$ array $A_{v,r} = a(v,r)$ and a decoded-length vector $m(v)$. The next phase is the cumulative decoded length of the input minus the prompt's estimated frame, modulo three. Each generation call forms $\delta\, \mathbf{g}_t \odot A_{:,r_t}$, where $\mathbf{g}_t$ is the green mask, and adds it to the logits. The implementation caches masks by preceding token and transfers the batch bias tensor to the model device. Precomputation avoids repeated Python evaluation of candidate-codon features, while generation still requires a vocabulary-sized update and sequence-length bookkeeping. The lookup changes the implementation cost, not the scoring event used by the auditor.

The prompt frame is estimated from the longest stop-free forward run, not from a coding annotation. Moreover, a multi-base token may include non-wobble positions, and a complete codon spanning token boundaries is absent from the usage lookup. The construction therefore favors certain token continuations rather than enforcing synonymous editing or preservation of an existing protein. Those consequences motivate the separate quality measurements.

3.4 Teacher-side interpretation

At a fixed context before nucleus truncation, let P_0 be the unbiased teacher distribution and $b(v) = \mathbf{1}[v \in \mathcal{G}]a(v,p)$. Then

$$P_\delta(v) = \frac{P_0(v)e^{\delta b(v)}}{Z_\delta}, \qquad Z_\delta = \sum_{u \in \mathcal{V}} P_0(u)e^{\delta b(u)}. \tag{5}$$

Since b is nonnegative and supported on $\mathcal{G}$, $dP_\delta(\mathcal{G})/d\delta = (1 - P_\delta(\mathcal{G}))\mathbb{E}_{P_\delta}[b] \geq 0$. Appendix B derives this relation and states its scope. Increasing bias raises the teacher-side green probability under these conditions, but does not prescribe the student inheritance rate. That rate also depends on the training corpus, model capacity, and decoding; we measure it empirically.

4 EXPERIMENTAL DESIGN

Training chain. The default teacher is GenomeOcean-500M and the student is GenomeOcean-100M. For each defense and training seed, we generate 2,000 teacher examples, replace 25% with real-genome fragments, fit a student adapter, merge it into the student, and sample evaluation responses without a watermark processor. Replacement keeps the training-set size fixed; it is not an extra 25% of examples. The main real-genome source is E. coli K-12 MG1655. The objective learns from sequences rather than the teacher's probability vectors, so inheritance must arise through the sampled training corpus.

Student optimization uses batch size 4, 4,000 steps, maximum training length 512, LoRA rank 8 and scaling 16, learning rate 2×10^{-4}, and weight decay 0.01. A 200-step warm-up precedes cosine decay to a 0.05 learning-rate ratio. Teacher sampling uses batches of 32, temperature 1 and top-$p = 0.95$. Student evaluation uses vLLM with the same temperature and nucleus cutoff, a 512-token output budget, and ignored end-of-sequence stopping. It adds neither presence/frequency penalties nor a marking processor. Seeds are 44, 142, and 256; each main seed has 200 evaluation responses. Appendix A specifies processing order and exceptions.

Experimental comparisons. The eight codon-aware settings are position fractional/binary at strength 4, usage at strengths 2 and 4, and joint fractional/binary at strengths 4 and 8. This is a component-and-strength ablation, not a fully crossed search over every hyperparameter. The clean plain-KGW reference has three seeds, while its edited matrix contains only seed 44. The targeted-nucleotide campaign instead includes an unwatermarked negative control. We keep these baselines separate throughout.

Extensions test different explanations for transfer. A 4B teacher with 100M or 500M students probes model size; 500 versus 2,000 training examples probes corpus budget. Five organism-conditioned campaigns test plain-KGW inheritance under different sequence sources. Fine-tuned-teacher controls test a second training stage. A keyed synonymous-codon transform changes the layer at which the mark is applied, while HyenaDNA students and a GROVER retokenization probe examine representation dependence. These experiments complement the main defense matrix rather than enlarging its matched-comparison denominator.

Editing procedures. All tests here edit already generated student responses. CS performs frame-zero synonymous codon substitutions at probability 0.15. KT selects tokens from a corpus-derived bias dictionary and replaces them using a key-aware red-token fallback. KN replaces green tokens with red tokens at probability 0.5. DS uses a frequency-based discriminator to gate synonymous edits. These are local implementations, not equivalent reproductions of the original text paraphrasers or contrastive student-training attacks (Pan et al., 2025; Yi et al., 2025).

A separate campaign targets the codon factors. Phase inserts one nucleotide after the first base of a selected codon at probability 0.30. Usage resamples a synonymous codon from T4 usage at probability 0.40. Combined applies Usage at 0.30 followed by Phase at 0.20. These operations do not consult the watermark key; their cleaned outputs are retokenized for auditing. Operation probabilities do not guarantee matched realized edit budgets, and the procedures are not detector-optimized attacks (Diaa et al., 2025).

Statistical unit and quality assays. For each seed we compute the fraction of responses above the same score cutoff and the mean response score, then average these seed summaries. Equal response counts make the average rate numerically equal to the pooled proportion, but do not make 600 responses independent training replicates. Error bars, where shown, are sample standard deviations across the three seed means. A mean score describes signal magnitude; the detection proportion describes the fraction of outputs crossing the threshold. We report both because an uneven score distribution can change them differently. Unwatermarked positive rates are empirical controls, not assumed to equal the nominal Gaussian tail.

Quality is evaluated on cleaned DNA separately from detection. GC fraction and codon adaptation (Sharp & Li, 1987) summarize composition; overlapping 3-mer entropy and distinct 6-mer fraction summarize diversity. ORF coverage counts the union of complete forward-strand ATG-to-stop intervals of at least 60 bases across three frames. ESMFold (Lin et al., 2023) analyzes 150

index-paired translated outputs per condition, while BLAST (Altschul et al., 1990) searches 450 translated queries per condition against the E. coli K-12 proteome. Different selection procedures and denominators make these complementary assays, not interchangeable measures of function. Appendix E defines the metrics and reports the complete measurements.

5 RESULTS

5.1 CLEAN INHERITANCE AND DEFENSE COMPONENTS

Across three clean KGW seeds, the student reaches mean $z = 3.99$ and 92.3% detection. The joint fractional configuration at $\delta = 8$ reaches $z = 17.88$ and 94.5% detection without an inference-time processor. Higher mean score and higher detection rate are distinct outcomes: the score increase is much larger than the increase in the fraction crossing the threshold. This separation is useful when considering subsequent edits.

Table 1: Student detection rates (%) and clean mean scores. F/B denote fractional/binary position weighting. All codon-aware rows average three seeds; KGW* is the single-seed baseline available for the edit matrix. KT and KN have true-key access.

Scheme	δ	Clean $\bar{z}$	Clean	KT	KN
KGW*	2	4.47	97.5	0.0	0.0
Position-F	4	4.56	77.5	5.3	1.8
Position-B	4	8.59	92.8	8.3	8.5
Usage	2	8.34	90.8	14.2	6.3
Usage	4	13.47	89.0	44.8	20.0
Joint-F	4	4.61	71.8	10.7	2.3
Joint-B	4	13.58	91.8	42.0	19.3
Joint-B	8	14.98	90.8	47.2	25.8
Joint-F	8	17.88	94.5	49.0	35.8

Table 1 identifies several viable detection operating points. Usage weighting at $\delta = 4$ retains 44.8% after KT; joint fractional weighting at $\delta = 8$ retains 49.0%, compared with 0% for the available KGW baseline. Under KN, the joint configuration retains 35.8%. The result is not unconditional component synergy: joint fractional weighting at $\delta = 4$ retains 10.7% after KT, and binary joint weighting at $\delta = 8$ retains 47.2%. Factor choice and strength must be considered together. The single-seed baseline and differing starting scores also prevent interpreting these comparisons as a fully matched causal defense effect.

The complete output-edit matrix adds an important distinction between transformations. Joint fractional weighting retains 92.0% under CS and 93.3% under DS, close to its clean rate, while KT and KN produce much larger reductions. CS and DS act through synonymous substitutions in decoded DNA; KT and KN directly target keyed token preferences and need not preserve a protein sequence. Consequently, the ordering is informative about these implemented operations, not a ranking of attacker knowledge alone. The appendix reports both rates and mean scores for every seed and transformation so that retained tails can be distinguished from changes in average signal.

5.2 MECHANISM-TARGETED EDITS

The joint configuration retains 73.3% after Phase, 66.5% after Usage, and 47.0% after Combined (Figure 2). Combined detection spans 44.0–50.0% across seeds. Position-only and usage-only configurations retain less signal under these requested operations. The negative control remains near the nominal tail, at 1.0% after Combined. This is evidence of retained signal under the tested transformations, not a KGW-versus-joint adaptive comparison because the KGW-positive comparator is absent from this campaign.

The retokenized clean joint rate is 94.2%, versus 94.5% when auditing stored token IDs. The two rates use different tokenization paths and are analyzed separately. Likewise, 47.0% after Combined

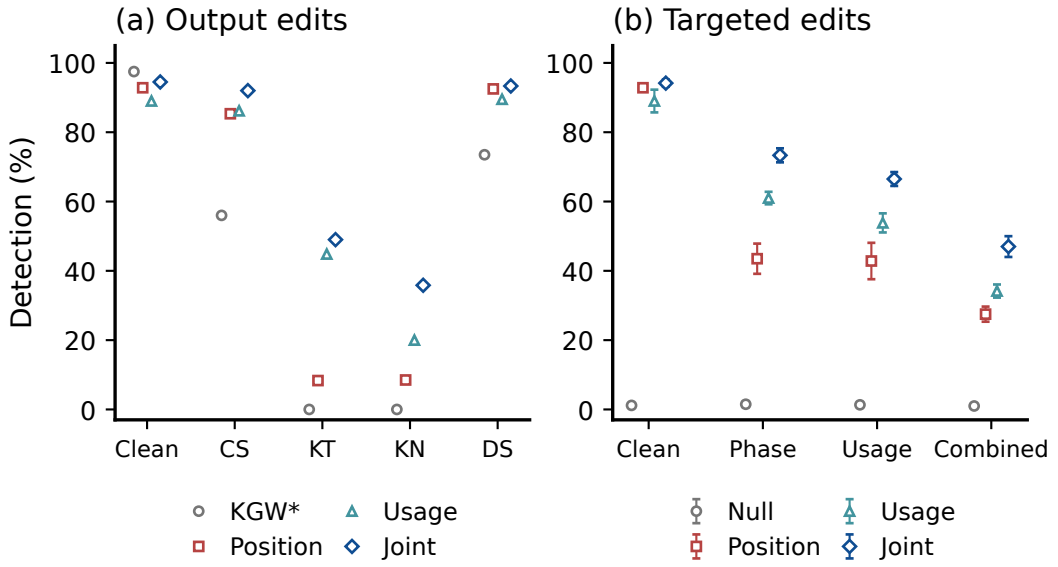

Figure 2: Inherited detection under two editing campaigns. Left: output edits; KGW* has one seed. Right: targeted nucleotide edits, with sample standard deviations over three seeds; Null is unwatermarked, not KGW. Series are offset within categories for visibility. Position and usage use $\delta = 4$; joint uses fractional weighting at $\delta = 8$.

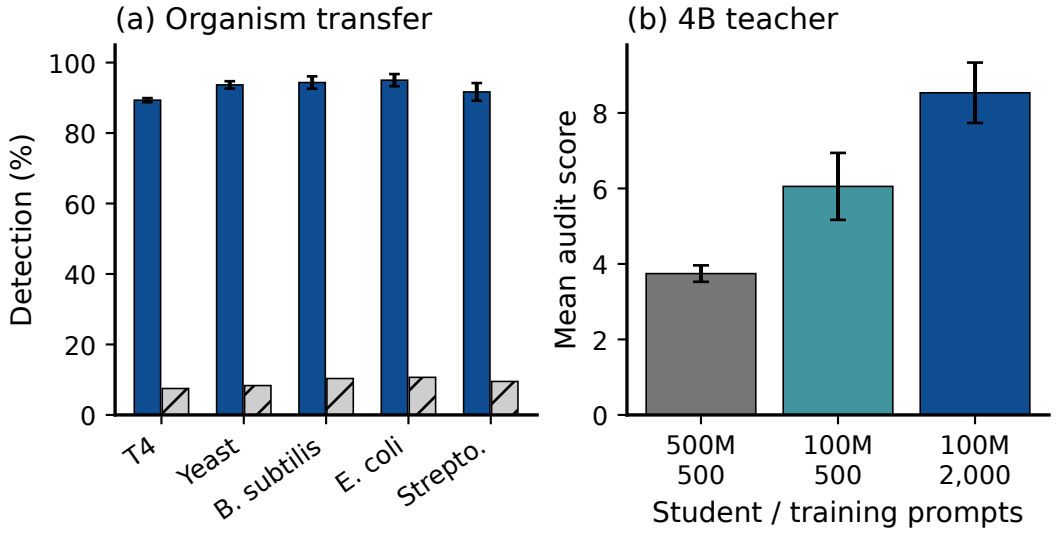

Figure 3: Extensions of plain-KGW inheritance. Left: five organism-conditioned datasets; blue bars are watermarked students and hatched gray bars are their controls. Right: a fixed 4B teacher with different student sizes and training budgets. Error bars show sample standard deviation across three seeds; control error bars are omitted for clarity.

and 49.0% after KT are not an equivalence test: adversary access and editing budgets differ. The complete matrices and denominators appear in Appendix C.

5.3 Organism and scale extensions

Inheritance appears across all five organism-conditioned datasets (Figure 3), with detection from 89.3% to 95.0% and mean scores from 5.76 to 7.88. Corresponding unwatermarked positive rates are 7.5–10.7%. Thus, the signal contrast generalizes across these datasets, but the nominal 1% Gaussian

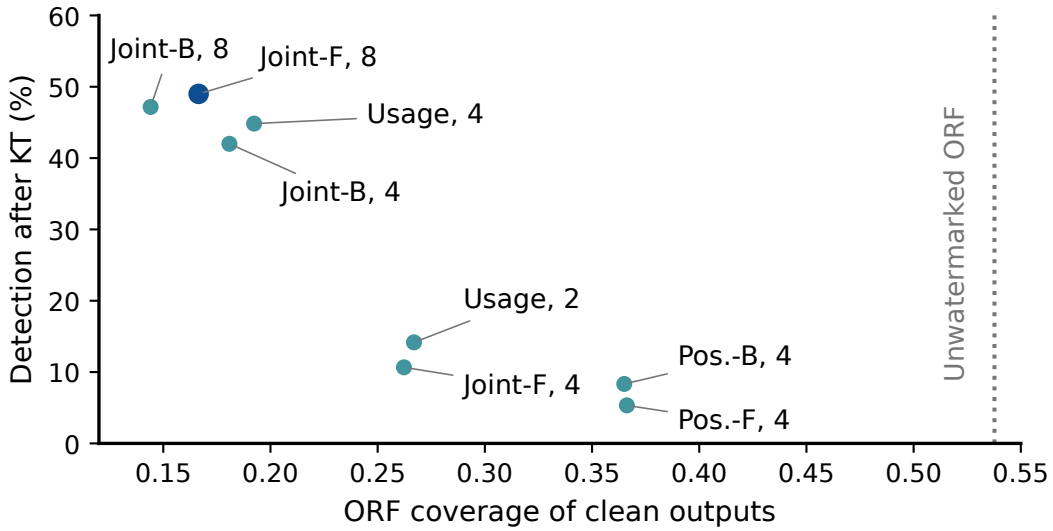

Figure 4: Detection after KT versus clean-output ORF coverage for all eight codon-aware configurations. Each point averages three seeds. The dotted line is the unwatermarked control's ORF coverage, not a matched KGW attack point. Labels give the weighting and bias strength.

tail does not supply a uniformly calibrated false-positive rate. These experiments use plain KGW at $\delta = 2$ and do not establish cross-organism robustness of the joint method.

With a fixed 4B teacher and 500 prompts, the 100M student reaches mean $z = 6.05$, versus 3.74 for the 500M student. Increasing the 4B-to-100M budget to 2,000 prompts yields $z = 8.53$. The first comparison changes student capacity at fixed teacher and corpus size; the second changes corpus size at fixed model pair. Both support inheritance at a 40-fold parameter ratio, without identifying a universal inverse-capacity law. The separate 4B joint-defense experiment reaches mean $z = 8.84$ and 85.2% detection using 500 prompts. Its smaller corpus prevents a direct teacher-size explanation for its difference from the main joint configuration.

5.4 Training controls and representation scope

A student trained from a teacher previously fine-tuned on marked outputs reaches mean $z = 8.43$ and 95.3% detection. The unmarked-teacher control reaches 0.38 and 9.8%. This contrast supports transmission through an additional training stage, while its control rate reinforces the need for empirical calibration. It does not imply that all teacher adaptation preserves the mark equally.

The alternative keyed synonymous transform reaches mean codon-detector scores 2.30 and 5.13 at application probabilities 0.3 and 0.5, with 51.7% and 67.5% detection. It shows that a surface-level codon mark can also leave a learned signal. This implementation is not the embedding-based DNAMark model (Zhang et al., 2025); its different counting events and null prevent a superiority claim based on dividing its score into GenoTrace's token score. Appendix D gives the transform and per-seed outcomes.

Tokenizer changes produce a different outcome. HyenaDNA-medium and -large probes have mean scores below one and cell-level positive rates of 3.0–15.0%, with 98–100 valid responses. A GROVER-tokenizer audit of joint outputs has mean score -5.47 but a 14.5% positive tail. The latter retokenizes existing outputs; it does not train a GROVER student. Changing token boundaries changes both the preceding-token hash input and the candidate token event, even when the nucleotide string is unchanged. These probes do not establish transferable detection across tokenizers, nor does their asymmetric score distribution justify calling every result a centered null.

5.5 DETECTION–QUALITY OPERATING POINTS

Figure 4 shows why detection should not be optimized in isolation. Joint fractional weighting at $\delta = 8$ has ORF coverage 0.166, compared with 0.365 for binary position weighting and 0.538 for the unwatermarked control. Its 3-mer entropy is 3.99 bits, compared with 5.60 for the control. The binary position setting preserves more coding coverage but retains 8.3% after KT. Usage weighting at $\delta = 4$ provides an intermediate operating point, with ORF coverage 0.192 and 44.8% KT detection. Choice among them depends on the desired quality constraint.

Codon-usage summaries also change across configurations, consistent with the intended weighting mechanism. Protein-level evidence is narrower: both control and joint outputs have mean predicted TM-score near 0.177, a low-confidence floor that cannot establish structural non-inferiority. BLAST yields 5 significant hits among 450 joint queries versus 12 among 450 controls at $E < 0.01$. These counts measure computational homology, not expression or function. Appendix E gives the full quality tables, structural distributions, and codon analysis.

6 DISCUSSION

The experiments separate three properties often conflated in provenance evaluation: injection of bias into a teacher, learning of that bias by a student, and calibration of an attribution decision. The fixed-context calculation addresses the first; the clean and edited student outputs address the second. Neither alone establishes the third. A deployment would need held-out unwatermarked models and keys, a chosen decision unit, and thresholds calibrated to the relevant distribution. Repeated token contexts and shared evaluation prompts make extremely small Gaussian tail probabilities particularly difficult to interpret.

The component ablation also clarifies the role of genomic structure. Position and usage factors change where bias is allocated, but do not act as biological validity constraints. Preferred-codon concentration can increase while ORF coverage and sequence diversity decrease. A quality-constrained choice should therefore specify the acceptable metric and compare configurations on that metric, rather than choosing the largest audit score alone. The current assay set does not measure expression, protein activity, or genomic function; low folding confidence is not evidence of non-inferiority.

Three next comparisons would most directly resolve the remaining uncertainty: an edited three-seed plain-KGW reference at matched starting detection, detector-optimized transformations under realized sequence-quality budgets, and an independently calibrated audit across keys and tokenizer families. A teacher/student capacity theory would additionally require matched teacher scores and a crossed size-and-data design. These are distinct missing measurements, not implications of the present scale sweep or a fitted inheritance coefficient.

7 CONCLUSION

GenoTrace provides a codon-aware token-bias construction and demonstrates a signature learned through genome-model distillation. Joint fractional weighting produces strong clean signal and retains measurable detection under the tested token and nucleotide edits. Component, organism, scale, and training controls identify both alternative operating points and the shared-tokenizer scope of the result. Reading these findings together with the quality assays yields the central conclusion: genomic structure can shape an inherited watermark, but detection strength, calibration, and biological utility must be evaluated separately.

AI USE STATEMENT

Writing assistance. Generative AI tools assisted with drafting, restructuring, and polishing the manuscript, and with literature checking and figure-generation code. AI-assisted research support also included implementation, methodological feedback, and mathematical exposition. The authors remain responsible for the final claims, citations, and artifacts.

Synthetic datasets. The genome foundation models described in Section 4 generated synthetic DNA used for student training and evaluation. Model choices, sampling settings, and data mixtures are specified in Appendix A; these synthetic sequences are not experimentally validated biological designs.

ETHICS STATEMENT

The work studies computational provenance for genome models. Watermarks may support responsible attribution, but false-positive accusations and covert tracking are potential misuse risks. Consequential attribution requires appropriate calibration and human oversight. Sequence homology and predicted structures are not wet-lab evidence of biological function.

REPRODUCIBILITY STATEMENT

Appendix A specifies the training and evaluation procedures. Appendices B–E contain derivations, complete detection tables, extension results, and computational quality analyses. Tables and quantitative figures are generated from machine-readable audit records. A compact supplementary export reproduces the reported aggregate values; it does not contain model checkpoints or claim to reproduce training.

REFERENCES

Scott Aaronson. Watermarking of large language models. Talk at the Simons Institute, UC Berkeley, 2023. URL https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 2024.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 1990.

Mihir Bellare, Ran Canetti, and Hugo Krawczyk. Keying Hash Functions for Message Authentication. In *Advances in Cryptology – CRYPTO '96*, 1996.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1995.

Andrew C Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 1941.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Jonathan C. Schmok, Ali

Taghibakhshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Michael T. Pearce, Elana Simon, Etowah Adams, Zachary J. Amador, Euan A. Ashley, Stephen A. Baccus, Haoyu Dai, Steven Dillmann, Stefano Ermon, Daniel Guo, Michael H. Herschl, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, John St. John, Jeremy Sullivan, Joseph Tey, Ben Viggiano, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Sudarshan Pinglay, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modelling and design across all domains of life with Evo 2. *Nature*, 2026.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *USENIX Security Symposium*, 2021.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable Watermarks for Language Models. In *Conference on Learning Theory*, 2024.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

J. Robert Coleman, Dimitris Papamichail, Steven Skiena, Bruce Futcher, Eckard Wimmer, and Steffen Mueller. Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. *Science*, 2008.

F.H.C. Crick. Codon—anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology*, 1966.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 2024.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 2025.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 2024.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 2022.

Abdulrahman Diaa, Toluwani Aremu, and Nils Lukas. Optimizing Adaptive Attacks against Watermarks for Language Models. In *International Conference on Machine Learning*, 2025.

Cameron L. M. Gilchrist, Milot Mirdita, and Martin Steinegger. Multiple protein structure alignment at scale with FoldMason. *Science*, 2026.

Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. *Journal of the ACM*, 1986.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 2021.

Ruth Hershberg and Dmitri A. Petrov. Selection on Codon Bias. *Annual Review of Genetics*, 2008.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased Watermark for Large Language Models. In *International Conference on Learning Representations*, 2024.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 2021.

Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark Stealing in Large Language Models. In *International Conference on Machine Learning*, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models. In *International Conference on Machine Learning*, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the Reliability of Watermarks for Large Language Models. In *International Conference on Learning Representations*, 2024.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust Distortion-free Watermarks for Language Models. *Transactions on Machine Learning Research*, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 2023.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *International Conference on Machine Learning*, 2023.

Yasukazu Nakamura, Takashi Gojobori, and Toshimichi Ikemura. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*, 2000.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. In *Advances in Neural Information Processing Systems*, 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 2024.

Leyi Pan, Aiwei Liu, Shiyu Huang, Yijian Lu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. Can LLM Watermarks Robustly Prevent Unauthorized Knowledge Distillation? In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. No Free Lunch in LLM Watermarking: Trade-offs in Watermarking Design Choices. In *Advances in Neural Information Processing Systems*, 2024.

Joshua B. Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 2011.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*, 2025.

Melissa Sanabria, Jonas Hirsch, Pierre M. Joubert, and Anna R. Poetsch. DNA language model GROVER learns sequence context in the human genome. *Nature Machine Intelligence*, 2024.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking Makes Language Models Radioactive. In *Advances in Neural Information Processing Systems*, 2024.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC^2)*, 2019.

Paul M. Sharp and Wen-Hsiung Li. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 1987.

Huanming Shen, Baizhou Huang, and Xiaojun Wan. Enhancing LLM Watermark Resilience Against Both Scrubbing and Spoofing Attacks. In *Advances in Neural Information Processing Systems*, 2025.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the Black Box of Deep Neural Networks via Information. *arXiv preprint arXiv:1703.00810*, 2017.

Samuel A. Stouffer, Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams, Jr. *The American Soldier: Adjustment During Army Life*, volume 1. Princeton University Press, 1949.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 2017.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images. In *Advances in Neural Information Processing Systems*, 2023.

Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 1927.

Weimin Wu, Zhihan Zhou, Robert Riley, Mohammad Abdulqader, Xuefeng Song, Satria Kautsar, Rob Egan, Steven Hofmeyr, Gefei Liu, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Yaping Liu, Andrei Stecca Steindorff, Fengchen Liu, Feng Chen, Rachael Morgan-Kiss, Lizhen Shi, Han Liu, and Zhong Wang. Uncovering the Genomic Manifold via Scalable Learning from the Global Microbiome. *bioRxiv*, 2025.

Xin Yi, Yue Li, Shunfan Zheng, Linlin Wang, Xiaoling Wang, and Liang He. Unified attacks to large language model watermarks: spoofing and scrubbing in unauthorized knowledge distillation. *arXiv preprint arXiv:2504.17480*, 2025.

Zaixi Zhang, Ruofan Jin, Kaidi Fu, Le Cong, Marinka Zitnik, and Mengdi Wang. FoldMark: Protecting protein generative models with watermarking. *arXiv preprint arXiv:2410.20354*, 2024.

Zaixi Zhang, Ruofan Jin, Le Cong, and Mengdi Wang. Securing the Language of Life: Inheritable Watermarks from DNA Language Models to Proteins. In *Advances in Neural Information Processing Systems*, 2025.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable Robust Watermarking for AI-Generated Text. In *International Conference on Learning Representations*, 2024.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. In *International Conference on Learning Representations*, 2024.

Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, Defne G. Ozgulbas, Natalia Vassilieva, James Gregory Pauloski, Logan Ward, Valerie Hayot-Sasson, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, and Arvind Ramanathan. GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *The International Journal of High Performance Computing Applications*, 2023.

A PROTOCOL AND IMPLEMENTATION

This appendix specifies the sequence of operations behind the main experiments. Appendix B separates fixed-context mathematical statements from assumptions required for statistical testing. Appendix C gives the complete output-edit matrix; Appendix D gives organism, scale, training-control, and tokenizer results. Appendix E defines the computational quality assays and supplies their detailed measurements. The final coverage section identifies which quantities the supplementary export can reproduce.

A.1 TRAINING AND GENERATION

The default teacher and student are GenomeOcean-500M and GenomeOcean-100M. The defended training runs use 2,000 prompt–response examples, batch size 4, 4,000 optimization steps, and a maximum training length of 512 tokens. Real-genome examples replace 25% of the training corpus rather than increasing its size. The student objective is next-token prediction on sequences, not matching teacher probability vectors. LoRA uses rank 8 and scaling 16. AdamW uses learning rate 2×10^{-4} and weight decay 0.01; the schedule has 200 warm-up steps, cosine decay, and a minimum learning-rate ratio of 0.05. The three training seeds are 44, 142, and 256.

Teacher sampling uses temperature 1 and top-$p = 0.95$ with batches of 32. Student evaluation uses vLLM, temperature 1, top-$p = 0.95$, a 512-token output budget, and `ignore_eos=True`. It does not add a watermark logits processor or override presence/frequency penalties. Each main condition has 200 evaluation responses. Training and inference were run in Apptainer on NVIDIA H200 accelerators. Model parameters are not updated during auditing.

A.2 DATA FLOW AND EXPERIMENTAL UNITS

Prompts are genomic nucleotide prefixes rather than natural-language instructions. The main source is E. coli K-12 substrain MG1655, reference assembly `GCF_000005845.2`. The main defense runs use the same training/evaluation prompt collections, truncated to their requested sizes. The three seed settings produce separate training runs; they are not three separately sampled reference genomes. The organism-conditioned studies change the sequence source, while the main usage factor uses the E. coli codon table.

For each condition, the pipeline has six stages:

1. Generate teacher continuations from the training prompts using the selected position/usage factors and bias strength.
2. Replace the prescribed fraction of synthetic examples with real-genome fragments, keeping the requested corpus size unchanged.
3. Fit the LoRA student on the resulting sequence corpus with the recorded optimization recipe.
4. Merge the adapter into the student before the evaluation generation stage.
5. Generate evaluation continuations without a marking processor. Preserve response strings and model token IDs for separate uses.
6. Audit stored token transitions for the clean matrix, apply the specified output edits for transformed conditions, and compute quality measurements from cleaned DNA.

Teacher training-example count, optimization steps, evaluation-response count, and valid translated-protein count describe different units. They should not be substituted for one another. The 512-token generation budget also does not imply a 512-nucleotide sequence because the vocabulary has variable-length tokens. Folding and homology require additional translation selection and therefore use their own denominators.

A.3 GREEN LISTS AND READING FRAMES

The green-list implementation hashes the key concatenated with a four-byte little-endian preceding-token identifier using SHA-256. The first eight digest bytes initialize a NumPy generator, whose permutation selects 1024 vocabulary entries. The first response token is not audited because its

Table 2: Protocol coverage. Counts denote training examples or audited responses, not independent genome samples. N/A indicates post-processing without a new training stage.

Campaign	Train examples	Eval per seed	Replication
Main defended matrix	2,000	200	8 settings, 3 seeds
Edited plain KGW reference	2,000	200	Seed 44 only
4B to 500M plain KGW	500	200	3 seeds
4B to 100M plain KGW	500 / 2,000	200	3 seeds per budget
4B joint defense	500	200	3 seeds
Targeted nucleotide edits	N/A	200	4 groups, 3 seeds
Clean folding comparison	N/A	50 pairs	3 seeds
Selected structure-edit pairs	N/A	50 pairs	3 seeds
Reference-proteome search	N/A	150 queries	3 seeds per group

preceding context token is not part of the stored response. This keyed construction is described operationally; it is not an implementation of HMAC or a proof of cryptographic pseudorandomness (Bellare et al., 1996; Goldreich et al., 1986).

For each prompt, the frame estimator chooses the longest stop-free codon run among the three forward phases. Generation tracks the cumulative length of the decoded nucleotide tokens relative to this frame. The position factor counts third-position nucleotides inside the candidate token. The usage factor counts only complete in-frame codons within that token. For a synonymous family with sorted usage values $u_1 \leq \cdots \leq u_m$, the preferred set includes values at least $u_{\lfloor m/2 \rfloor + 1}$. Single-codon families and ties are included. If a token contains no complete codon, its usage factor is $1/2$. These details matter when comparing the binary and fractional variants.

A.4 Lookup construction and generation state

The token map decodes each vocabulary entry independently, uppercases the result, and maps U to T. A decoded entry composed only of A, C, G, T, and N is reduced to its A/C/G/T characters. Other entries, including non-nucleotide special tokens, map to an empty string. The lookup assigns zero to empty entries whenever codon-aware factors are enabled. For each remaining token, all three possible start phases are evaluated once. No assumed average token length replaces this calculation.

Let $r \in \{0, 1, 2\}$ denote phase and d_v the mapped nucleotide string of token v. For each (v, r), position weighting examines $(r + j) \bmod 3$ for every base offset j. Usage weighting examines triplets beginning at offsets with $(r + j) \bmod 3 = 0$ and ending inside d_v. The chosen factors produce $A_{v,r}$. During generation, with current token sequence $x_{1:t-1}$ and estimated prompt frame r_0, the lookup column is

$$r_t = \left(\sum_{j=1}^{t-1} |d_{x_j}| - r_0 \right) \bmod 3. \tag{6}$$

Prompt frames are estimated separately for each batch row; padding tokens contribute no mapped nucleotide length. Ties in the longest stop-free run retain the first frame examined. The estimator does not require an initiating ATG and is not the same operation as the downstream ORF-coverage metric.

The implementation reconstructs cumulative length from the input token array, obtains the preceding-token green mask from a cache, and assembles the batch bias on the CPU before transferring it to the logits device. The lookup itself contains $3|\mathcal{V}|$ floating-point entries. This representation avoids recomputing synonymous-family membership for every candidate at every step, but mask storage and the vocabulary-sized batch update remain real costs. Runtime and throughput claims require a separate measured benchmark; they do not follow from the lookup shape alone.

A.5 Worked token examples

For `ATCG` at phase zero, the only wobble offset is two, giving fractional weight $1/4$ and binary weight one. At phase two, offsets zero and three are wobble positions, giving fractional weight $1/2$. Every

token of length at least three contains at least one such offset in every phase, so its binary position factor is one. This is why a nucleotide-level statement that only one third of positions are selected cannot be applied directly to binary token weighting.

At phase zero, `ATG` contains one complete codon and is preferred because methionine has a single codon. Its usage factor is one, whereas a token `A` has no complete codon and receives $1/2$. If a token contains two complete codons, one preferred and one not preferred, its usage factor is $1/2$. Codons crossing a token boundary do not contribute to either token's usage average. These examples show why usage weighting is neither a hard green-set intersection nor full matching of organism codon frequencies.

A.6 Output editing

CS decodes and cleans the DNA, attempts frame-zero synonymous substitutions at probability 0.15, and retokenizes the result. DS uses a logistic-regression discriminator over 64 token-frequency buckets. A 60-nucleotide local window gates synonymous editing at probability 0.30. It is a fixed post-processing procedure, not contrastive distillation or an optimized paraphraser.

KT estimates positive bigram log-frequency ratios from watermarked and unwatermarked samples. A current token is selected when its stored score exceeds 0.5. The dictionary contains only scores above 0.1, while the low-score candidate branch requires a score below 0.1. Thus this dictionary makes that candidate set empty, and the implemented fallback draws a red token using the true key. KN replaces a green token with a random red token with probability 0.5. Both edit using the original preceding token, while the detector scores the resulting transitions. Neither enforces a nucleotide-distance or protein-preservation constraint. The data therefore support the named token-edit tests, not a claim of replicating the original attacks of Pan et al. (2025) or Yi et al. (2025).

In the mechanism-targeted tests, Phase inserts a random base after the first nucleotide of a selected source codon. Usage draws from the synonymous choices weighted by T4 frequencies and may select the original codon. Combined applies Usage first, then Phase. Requested probabilities are 0.30 for Phase, 0.40 for Usage, and 0.30/0.20 for the two Combined components. These are not realized edit-rate guarantees. The transformations use a fixed random seed and do not optimize against detector feedback. Their audit path starts from cleaned response strings, which explains the small difference between clean rates in the two campaigns.

Table 3: Output-transformation access and units. No operation here updates the student parameters. A requested probability is not the observed nucleotide edit distance.

Operation	Key access	Selection and replacement
CS	None	Frame-zero synonymous substitutions, probability 0.15.
KT	True key in fallback	Corpus bigram score selects tokens; red-token fallback supplies replacements.
KN	True key	Replace a green token by a red token with probability 0.5.
DS	None	Frequency discriminator gates synonymous edits, probability 0.30.
Phase	None	Insert one base within a selected source codon, probability 0.30.
Usage	None	T4-weighted synonymous resampling, probability 0.40.
Combined	None	Usage at 0.30, then Phase at 0.20.

Replacing a token can alter its own green-membership event and the next event, because the detector hashes the resulting preceding token. The editing routines use the original predecessor for replacement selection, so a token chosen as red need not remain red under every newly formed transition. Conversely, a synonymous edit in nucleotide space can change multiple tokenizer boundaries. The reported rates incorporate this complete editing-and-retokenization path. They cannot be deduced by multiplying clean detection by the fraction of unchanged codons.

A.7 Evaluation coverage

The main matrix comprises eight defended configurations with three seeds each and one plain-KGW comparator. Each of these 25 conditions has five audits, including clean, for 125 condition-level audit records. The experimental unit is a trained student, with five audits per student. The nucleotide-edit campaign has four source groups, three seeds, and four conditions. Its negative control is an unwatermarked student rather than a KGW-positive student. Tables 4–9 give the clean references and edited results with their respective sample counts.

B Teacher bias and statistical interpretation

B.1 Fixed-context bias calculation

Let $I(v) = \mathbf{1}[v \in \mathcal{G}]$ and $b(v) = I(v)a(v)$, with $0 \le a(v) \le 1$. For the pre-truncation exponential tilt in Equation 5, differentiation gives

$$\frac{d}{d\delta}\mathbb{E}_{P_\delta}[I] = \mathbb{E}_{P_\delta}[Ib] - \mathbb{E}_{P_\delta}[I]\mathbb{E}_{P_\delta}[b] \tag{7}$$

$$= (1 - P_\delta(\mathcal{G}))\mathbb{E}_{P_\delta}[b] \ge 0. \tag{8}$$

The equality uses $Ib = b$. Under constant green weight one,

$$P_\delta(\mathcal{G}) = \frac{P_0(\mathcal{G})e^\delta}{1 - P_0(\mathcal{G}) + P_0(\mathcal{G})e^\delta}. \tag{9}$$

For general weights, the KL divergence relative to the same base token distribution is

$$\mathrm{KL}(P_\delta \| P_0) = \delta\mathbb{E}_{P_\delta}[b] - \log Z_\delta. \tag{10}$$

These relations hold at a fixed context and temperature before nucleus truncation. They neither imply component synergy after training nor bound codon KL relative to a different reference distribution. In particular, a token-uniform reference does not automatically map to a uniform distribution over the 64 codons.

B.2 Local response, divergence, and saturation

The fixed-context calculation permits a nonuniform base distribution. For any fixed function h on the vocabulary, differentiating the normalized exponential tilt gives

$$\frac{d}{d\delta}\mathbb{E}_{P_\delta}[h] = \mathrm{Cov}_{P_\delta}(h, b). \tag{11}$$

Writing $g_0 = P_0(\mathcal{G})$, a local expansion for green probability is therefore

$$P_\delta(\mathcal{G}) = g_0 + \delta(1 - g_0)\mathbb{E}_{P_0}[Ia] + O(\delta^2). \tag{12}$$

The coefficient is weighted by the teacher's token probabilities. A vocabulary-uniform average of the position/usage lookup is not generally equal to it. Even $g_0 = \gamma$ is an additional assumption about base probability mass, not a consequence of selecting a fraction γ of vocabulary entries.

Differentiating the KL expression gives

$$\frac{d}{d\delta}\mathrm{KL}(P_\delta \| P_0) = \delta\,\mathrm{Var}_{P_\delta}(b) \ge 0, \qquad \mathrm{KL}(P_\delta \| P_0) = \frac{\delta^2}{2}\mathrm{Var}_{P_0}(b) + O(\delta^3). \tag{13}$$

These expressions use natural logarithms. Division by $\ln 2$ gives bits. If $a = 1$, the limiting distribution is P_0 conditioned on $\mathcal{G}$, so the KL limit is $-\log g_0$. The special value $\log_2(1/\gamma)$ follows only if $g_0 = \gamma$. For unequal weights, the limit instead concentrates on the maximal-weight set $M = \mathrm{argmax}_{v:P_0(v)>0}\, b(v)$, with divergence $-\log P_0(M)$. This need not equal the constant-bias limit.

To see the last statement, divide the numerator and denominator of $P_\delta(v)$ by $e^{\delta \max b}$. Terms below the maximum vanish as δ grows, while the remaining terms retain their relative P_0 probabilities. This argument assumes a finite vocabulary and nonempty base support. Applying the same deterministic map to two token distributions contracts KL, but its reference becomes the mapped P_0, not an unrelated natural-codon distribution. Whole generated sequences, changing contexts, and nucleus truncation require additional analysis.

B.3 What a student-score relation would require

For a response with a fixed number T of audited transitions and green probability p_t at transition t, linearity of expectation gives

$$\mathbb{E}[z] = \frac{\sum_{t=1}^{T}(p_t - \gamma)}{\sqrt{T\gamma(1-\gamma)}}. \tag{14}$$

No independence assumption is needed for this mean identity. Independence matters for variance and tail probabilities. A single operational ratio between student and teacher green excess would require a matched measurement of both distributions under specified decoding and length conventions. It would be a description of those observations, not a derivation of student training dynamics. It need not lie in $[0, 1]$ without further assumptions and becomes unstable when teacher excess is near zero. We therefore do not infer it from a hypothetical uniform teacher.

Similarly, a fitted line through two strength settings has no residual degree of freedom with which to test a scaling law. The component matrix establishes observed contrasts between recorded strengths. The model-scale experiments establish contrasts at their stated corpus sizes. Neither justifies an exact inheritance law as a function of parameter ratio alone.

B.4 Aggregation, uncertainty, and calibration

The Stouffer form $Z_n = n^{-1/2}\sum_i z_i = \sqrt{n}\,\bar{z}$ is an identity (Stouffer et al., 1949). An interpretation of Z_n as a standard-normal test requires a suitable null and independence structure. Under the specific assumptions $z_i \sim N(0,1)$ under the null and $z_i \sim N(\mu, \sigma^2)$ under the alternative, independent sequence sampling would give the approximate requirement

$$n \geq \left(\frac{z_{1-\alpha} + \sigma z_{1-\beta}}{\mu}\right)^2 \tag{15}$$

for power $1-\beta$. This conditional calculation is not an experimentally established query budget for our dependent outputs. Finite-sample approximation also depends on absolute third moments (Berry, 1941); signed skewness cannot substitute for those moments.

Descriptive detection proportions and across-seed variation are the primary summaries in this paper. Wilson intervals (Wilson, 1927) concern a specified binomial sampling model; applying them to pooled repeated prompts does not capture between-training variability. Likewise, a standardized effect size (Cohen, 1988) does not by itself establish equivalence, and multiple-testing corrections (Benjamini & Hochberg, 1995) do not repair a miscalibrated null. The observed organism-control rates motivate explicit empirical calibration before high-consequence attribution.

For completeness, let k_s of n_s responses from seed s cross the cutoff, and let $\bar{z}_s$ be that seed's mean score. The reported rate is $S^{-1}\sum_s k_s/n_s$, and the reported mean score is $S^{-1}\sum_s \bar{z}_s$. The sample standard deviation of a seed-level quantity u_s is $\sqrt{\sum_s (u_s - \bar{u})^2/(S-1)}$. Within-seed response standard deviations and across-seed standard deviations answer different questions and are not interchanged.

If responses were independent Bernoulli observations, a Wilson interval with quantile c and $\hat{p} = k/n$ would have center and half-width

$$\frac{\hat{p} + c^2/(2n)}{1 + c^2/n}, \qquad \frac{c\sqrt{\hat{p}(1-\hat{p})/n + c^2/(4n^2)}}{1 + c^2/n}. \tag{16}$$

This conditional formula is provided to make the sampling assumption explicit, not to label pooled training replicates as independent. The equal-weight Stouffer statistic has variance $n^{-1}\sum_{i,j}\mathrm{Cov}(z_i, z_j)$; its familiar unit variance requires the covariance sum to equal n. A Berry–Esseen approximation for independent identically distributed scores with finite third absolute centered moment contains a term proportional to $\mathbb{E}|z-\mu|^3/(\sigma^3\sqrt{n})$. A small signed skewness is insufficient, because cancellation in a signed moment does not bound an absolute moment.

C COMPLETE DETECTION MATRICES

The following tables give clean references, aggregate editing results, and per-seed measurements. Every condition contains 200 responses. The edited plain-KGW reference has seed 44 only; its separate three-seed clean reference appears in Table 4.

Table 4: Clean three-seed plain-KGW and unwatermarked controls. Detection is the recorded positive rate at the nominal 1% score threshold.

Condition	Seed	n	$\bar{z}$	Detection (%)
Unwatermarked	44	200	0.21	0.5
Unwatermarked	142	200	0.09	1.5
Unwatermarked	256	200	0.51	2.0
KGW	44	200	4.47	97.5
KGW	142	200	3.53	85.5
KGW	256	200	3.96	94.0

Table 5: Complete output-edit matrix in percent. KGW uses one seed; all position, usage, and joint configurations use three. Labels and key access are defined in Section 4.

Scheme, δ	Clean	CS	KT	KN	DS
KGW, 2	97.5	56.0	0.0	0.0	73.5
Position-F, 4	77.5	44.0	5.3	1.8	64.7
Position-B, 4	92.8	85.3	8.3	8.5	92.5
Usage, 2	90.8	71.7	14.2	6.3	88.2
Usage, 4	89.0	86.2	44.8	20.0	89.5
Joint-F, 4	71.8	37.3	10.7	2.3	60.2
Joint-B, 4	91.8	88.7	42.0	19.3	91.8
Joint-B, 8	90.8	87.7	47.2	25.8	92.0
Joint-F, 8	94.5	92.0	49.0	35.8	93.3

Table 6: Per-seed mechanism-targeted detection rates (%), 200 responses per entry. Joint-F denotes strength 8; position and usage use strength 4.

Condition	Seed	Clean	Phase	Usage	Combined
Unwatermarked	44	0.5	1.0	1.5	0.5
Unwatermarked	142	1.5	2.5	0.5	1.0
Unwatermarked	256	1.5	1.0	2.0	1.5
Position-B, 4	44	92.0	41.5	37.5	26.0
Position-B, 4	142	93.0	40.5	43.0	26.5
Position-B, 4	256	93.5	48.5	48.0	30.0
Usage, 4	44	89.5	61.5	54.0	32.0
Usage, 4	142	92.0	62.5	51.0	35.5
Usage, 4	256	85.5	59.0	56.5	35.0
Joint-F, 8	44	94.5	74.5	68.5	47.0
Joint-F, 8	142	95.0	71.0	64.5	44.0
Joint-F, 8	256	93.0	74.5	66.5	50.0

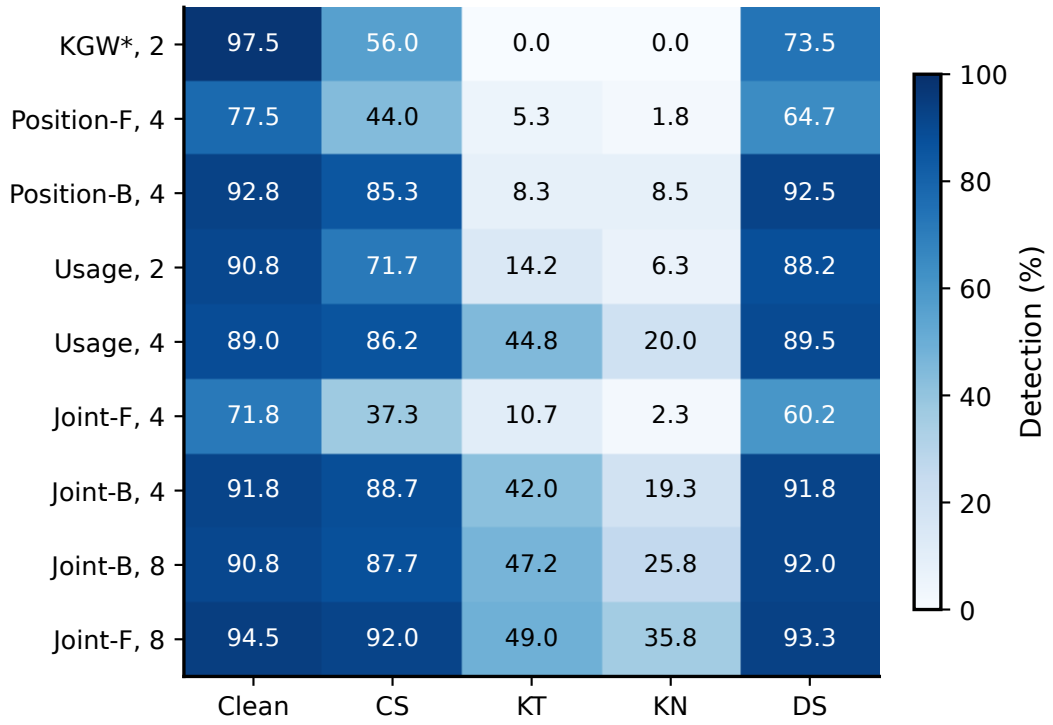

Figure 5: Complete detection matrix, using the same conditions and denominators as Table 5. The color scale is shared across rows and represents detection percentage, not statistical confidence.

Table 7: Mechanism-targeted editing campaign, mean detection percentages across three seeds. Null denotes an unwatermarked source. Position and usage have strength 4; joint fractional has strength 8.

Condition	Clean	Phase	Usage	Combined
Null	1.2	1.5	1.3	1.0
Position	92.8	43.5	42.8	27.5
Usage	89.0	61.0	53.8	34.2
Joint	94.2	73.3	66.5	47.0

Table 8: Mean audit scores for all 125 conditions, organized by student and transformation. Each entry summarizes 200 responses; edited KGW has seed 44 only.

Scheme, δ	Seed	Clean	CS	KT	KN	DS
KGW, 2	44	4.47	2.47	-1.72	-3.63	2.97
Position-F, 4	44	4.19	2.33	-1.48	-3.98	2.92
Position-F, 4	142	4.37	2.31	-1.57	-3.78	3.26
Position-F, 4	256	5.13	2.86	-1.48	-3.65	3.55
Position-B, 4	44	8.40	4.53	0.08	-2.22	6.23
Position-B, 4	142	8.38	4.73	0.21	-2.22	6.27
Position-B, 4	256	9.00	5.03	0.62	-1.86	6.07
Usage, 2	44	9.48	4.99	0.25	-2.05	7.57
Usage, 2	142	7.14	3.64	-0.01	-2.80	5.53
Usage, 2	256	8.40	4.37	0.81	-2.29	5.74
Usage, 4	44	13.41	6.81	3.73	-0.61	9.37
Usage, 4	142	13.67	6.77	3.87	-0.49	10.31
Usage, 4	256	13.32	6.88	3.50	-0.71	9.57

Table 8 (continued)

Scheme, δ	Seed	Clean	CS	KT	KN	DS
Joint-F, 4	44	4.43	2.39	-0.62	-3.71	3.10
Joint-F, 4	142	5.07	2.80	-0.62	-3.59	3.76
Joint-F, 4	256	4.32	2.14	-0.98	-3.81	3.22
Joint-B, 4	44	13.47	6.96	3.62	-0.63	10.17
Joint-B, 4	142	13.33	7.03	2.69	-0.74	10.28
Joint-B, 4	256	13.93	7.00	3.68	-0.37	10.13
Joint-B, 8	44	16.94	8.56	4.28	0.29	12.24
Joint-B, 8	142	12.82	6.29	4.26	-0.87	9.66
Joint-B, 8	256	15.17	7.52	4.45	0.02	12.01
Joint-F, 8	44	18.38	9.33	6.59	0.94	12.43
Joint-F, 8	142	17.62	9.22	5.93	0.68	12.52
Joint-F, 8	256	17.64	8.83	6.27	0.80	12.97

Table 9: Detection rates (%) for the same 125 conditions as Table 8, with 200 responses per entry.

Scheme, δ	Seed	Clean	CS	KT	KN	DS
KGW, 2	44	97.5	56.0	0.0	0.0	73.5
Position-F, 4	44	75.5	40.5	6.0	1.5	62.0
Position-F, 4	142	81.0	43.0	5.0	1.5	65.5
Position-F, 4	256	76.0	48.5	5.0	2.5	66.5
Position-B, 4	44	92.0	83.5	6.0	7.5	93.0
Position-B, 4	142	93.0	87.0	7.0	6.5	93.0
Position-B, 4	256	93.5	85.5	12.0	11.5	91.5
Usage, 2	44	92.5	75.5	13.0	8.0	91.0
Usage, 2	142	88.0	64.5	14.0	4.0	84.5
Usage, 2	256	92.0	75.0	15.5	7.0	89.0
Usage, 4	44	89.0	88.0	46.0	19.0	88.0
Usage, 4	142	92.5	88.0	49.5	18.0	89.0
Usage, 4	256	85.5	82.5	39.0	23.0	91.5
Joint-F, 4	44	67.5	33.0	10.5	3.5	54.5
Joint-F, 4	142	77.0	41.5	9.5	2.5	64.5
Joint-F, 4	256	71.0	37.5	12.0	1.0	61.5
Joint-B, 4	44	92.0	90.0	38.5	19.5	92.5
Joint-B, 4	142	89.5	84.5	40.5	18.5	88.0
Joint-B, 4	256	94.0	91.5	47.0	20.0	95.0
Joint-B, 8	44	92.0	89.5	41.5	32.0	93.0
Joint-B, 8	142	87.5	81.5	49.5	20.0	88.5
Joint-B, 8	256	93.0	92.0	50.5	25.5	94.5
Joint-F, 8	44	94.5	92.0	47.5	38.0	95.0
Joint-F, 8	142	95.5	92.5	49.0	32.5	91.5
Joint-F, 8	256	93.5	91.5	50.5	37.0	93.5

D EXTENSIONS AND ALTERNATIVE WATERMARK LAYER

The 500-prompt comparison varies student size at a fixed teacher, whereas comparing 500M-to-100M with 4B-to-100M at 2,000 prompts varies teacher size. The clean 500M-to-100M plain-KGW reference has mean $z = 3.99$. These observations are not a fully crossed model-capacity study. A separate 4B joint-defense experiment has mean student score about 8.84 and detection 85.2%, using 500 teacher prompts rather than the main joint experiment's 2,000. It supports transfer from the larger teacher, but does not isolate a teacher-size effect at fixed training budget.

Fine-tuned teachers. A teacher fine-tuned with watermarked training outputs produces a student with mean score 8.43 and detection 95.3%. The corresponding unwatermarked-teacher control has

Table 10: Plain-KGW organism-conditioned inheritance at $\delta = 2$, averaged over three seeds. Each row has its own unwatermarked control.

Organism	Positive (%)	Null (%)	$\bar{z}$
T4	89.3	7.5	5.76
Yeast	93.7	8.3	6.39
B. subtilis	94.3	10.3	7.51
E. coli	95.0	10.7	7.18
Strepto.	91.7	9.5	7.88

Table 11: Scale and training-budget experiments with a 4B teacher. SD is the sample standard deviation of three seed means.

Teacher to student	Prompts	$\bar{z}$	Seed SD
4B to 500M	500	3.74	0.22
4B to 100M	500	6.05	0.89
4B to 100M	2000	8.53	0.80

mean score 0.38 and positive rate 9.8%. The contrast is consistent with a training-dependent signature; the control rate again limits the nominal-threshold interpretation.

Table 12: Fine-tuned-teacher controls and 4B joint-defense extension, reported per seed. The 4B experiment uses 500 training prompts.

Condition	Seed	n	$\bar{z}$	Detection (%)
Marked teacher	44	200	8.44	97.0
Marked teacher	142	200	7.96	95.0
Marked teacher	256	200	8.88	94.0
Unmarked teacher	44	200	0.34	6.5
Unmarked teacher	142	200	0.23	8.5
Unmarked teacher	256	200	0.59	14.5
4B joint	44	200	9.09	82.5
4B joint	142	200	8.99	87.5
4B joint	256	200	8.45	85.5

Keyed synonymous-codon transform. An alternative baseline applies a deterministic keyed choice among synonymous codons to plain teacher responses before student training. It uses the preceding codon as the HMAC input and attempts substitutions with probability 0.3 or 0.5. At 0.3, the codon-level audit has mean score 2.30 and 51.7% detection; at 0.5, these values are 5.13 and 67.5%. These are results for the implemented keyed transform, not the learned embedding-based DNAMark model of Zhang et al. (2025). The detectors also use different null constructions, so their score magnitudes should not be interpreted as a calibrated cross-method superiority ratio.

Table 13: Keyed synonymous-transform baseline. The codon detector uses the recorded cutoff $z > 2.33$; these are not scores from the token-level detector.

Apply probability	Seed	n	Codon $\bar{z}$	Detection (%)
0.3	44	200	2.13	45.0
0.3	142	200	1.99	49.5
0.3	256	200	2.77	60.5
0.5	44	200	4.98	67.0
0.5	142	200	5.00	67.0
0.5	256	200	5.41	68.5

Tokenizer scope. HyenaDNA-medium and -large student probes yield mean scores below one and per-cell positive rates from 3.0% to 15.0%, with 98–100 valid sequences per cell. A smaller-model probe gives mean score 0.06 and a 4.0% positive rate on 100 sequences. A separate GROVER-tokenizer audit retokenizes joint outputs without training a GROVER student; its mean score is -5.47 and mean positive-tail rate is 14.5%. This is a distributional shift, not a centered null. The experiments do not establish positive watermark inheritance across tokenizers and restrict our main conclusions to compatible tokenization.

Table 14: Different-tokenizer student probes. Counts are the valid audited responses rather than assumed 100-response denominators.

HyenaDNA	Seed	n	$\bar{z}$	Detection (%)
medium	44	100	0.10	3.0
medium	142	100	0.99	15.0
medium	256	100	0.23	4.0
large	44	100	0.27	4.0
large	142	100	0.38	3.0
large	256	98	0.48	9.2

D.1 Replication and ablation coverage

The organism plot combines three training seeds per organism with their corresponding unwatermarked controls. Table 15 reports those seed-level rates. The scale table separates changes in student size from changes in training-set size; the grid is not fully crossed. Rank 8, the 25% real-genome fraction, and temperature 1 are held fixed, so these comparisons assess the listed model and data configurations rather than optimizing the training recipe.

Table 15: Organism-conditioned inheritance and controls, per seed. Rates are percentages.

Organism	Seed	$\bar{z}$	Positive (%)	Control (%)
T4	44	5.66	90.0	8.5
T4	142	5.81	89.0	6.5
T4	256	5.79	89.0	7.5
Yeast	44	6.40	92.5	11.0
Yeast	142	6.42	94.5	7.5
Yeast	256	6.35	94.0	6.5
B. subtilis	44	6.95	94.5	8.0
B. subtilis	142	8.33	96.0	12.0
B. subtilis	256	7.26	92.5	11.0
E. coli	44	7.66	96.0	9.0
E. coli	142	7.13	96.0	10.0
E. coli	256	6.76	93.0	13.0
Strepto.	44	8.20	92.0	10.5
Strepto.	142	7.05	89.0	7.5
Strepto.	256	8.39	94.0	10.5

The main component study varies position type, usage weighting, and strength jointly through its eight listed configurations. For example, the fractional joint comparison between strengths 4 and 8 is a within-construction change, whereas comparing position-only strength 4 to joint strength 8 also changes the bias allocation and magnitude. The results identify useful settings without attributing every difference to a single factor. The fine-tuned-teacher and synonymous-transform tables likewise use their own controls and detector definitions.

Table 16: All 4B-teacher scale cells. Train denotes training examples; n is the number of audited student responses.

Student	Train	Seed	n	$\bar{z}$	Detection (%)
500M	500	44	200	3.84	71.5
500M	500	142	200	3.50	69.5
500M	500	256	200	3.90	74.0
100M	500	44	200	5.08	88.0
100M	500	142	200	6.82	90.0
100M	500	256	200	6.27	89.5
100M	2000	44	200	9.45	94.0
100M	2000	142	200	8.19	93.5
100M	2000	256	200	7.96	94.0

E SEQUENCE AND PROTEIN-LEVEL ANALYSES

Table 17: Clean-output quality metrics, averaged over three seeds. ORF is forward-strand coding coverage; GC is a fraction; entropy is the overlapping 3-mer Shannon entropy in bits; diversity is distinct 6-mers divided by 6-mer windows.

Condition	ORF	GC	Entropy	Diversity
Unwatermarked, 0	0.538	0.560	5.595	0.508
Position-F, 4	0.366	0.653	4.955	0.343
Position-B, 4	0.365	0.641	4.894	0.346
Usage, 2	0.267	0.764	4.326	0.223
Usage, 4	0.192	0.763	4.051	0.174
Joint-F, 4	0.262	0.706	4.501	0.251
Joint-B, 4	0.181	0.791	4.116	0.181
Joint-B, 8	0.144	0.800	3.931	0.158
Joint-F, 8	0.166	0.745	3.989	0.162

ORF coverage includes complete ATG-to-stop intervals of at least 60 nucleotides in the three forward frames. It excludes incomplete terminal intervals and does not scan the reverse complement. Codon adaptation uses the geometric mean of synonymous-frequency weights normalized by the most frequent codon within each amino-acid family (Sharp & Li, 1987). These metrics measure different aspects of sequence composition and are not substitutes for expression or functional annotation.

E.1 EXACT SEQUENCE METRICS

Sequence-quality processing uppercases DNA, maps U to T, removes other characters, and excludes strings shorter than 60 nucleotides. Every reported main quality cell retains 200 strings. GC is the fraction of G/C bases. The ORF procedure marks the union of accepted intervals across the three forward frames, so overlapping intervals are not counted twice. Stop density is the number of frame-zero stop codons divided by nucleotide length, multiplied by 1,000.

For frame-zero sense codons $c_1, \ldots, c_m$, let $w(c) = u(c)/\max_{c' \in \mathrm{syn}(c)} u(c')$. Codon adaptation is $\exp(m^{-1} \sum_j \log w(c_j))$, with stop codons and nonpositive weights omitted. For overlapping k-mer frequencies p_j, entropy is $-\sum_j p_j \log_2 p_j$; diversity is the number of distinct 6-mers divided by the number of 6-mer windows. These values are computed per sequence, then averaged within a seed. They are not computed from one concatenated sequence, which would introduce artificial boundary windows.

Codon usage shifts under the weighting factors (Figure 6). Pooling codons yields large reference-frequency discrepancy statistics for many amino-acid families, but repeated sequence structure limits a simple independent-count interpretation. Usage weighting favors a preferred subset, so stronger

Table 18: Additional quality summaries: codon adaptation index and frame-zero stop density, averaged over three seeds.

Scheme, δ	CAI	Stops per kb
Unwatermarked, 0	0.637	10.133
Position-F, 4	0.617	7.647
Position-B, 4	0.627	8.432
Usage, 2	0.719	3.792
Usage, 4	0.728	4.571
Joint-F, 4	0.686	6.205
Joint-B, 4	0.727	3.535
Joint-B, 8	0.731	3.186
Joint-F, 8	0.719	5.273

Table 19: Quality measurements for every configuration and seed, with 200 valid sequences per row.

Scheme, δ	Seed	ORF	GC	Entropy	Diversity
Unwatermarked, 0	44	0.514	0.497	5.601	0.517
Unwatermarked, 0	142	0.549	0.552	5.611	0.506
Unwatermarked, 0	256	0.550	0.631	5.573	0.501
Position-F, 4	44	0.384	0.664	4.955	0.344
Position-F, 4	142	0.368	0.674	5.012	0.354
Position-F, 4	256	0.346	0.621	4.897	0.331
Position-B, 4	44	0.371	0.686	4.901	0.347
Position-B, 4	142	0.366	0.646	4.930	0.357
Position-B, 4	256	0.357	0.592	4.850	0.335
Usage, 2	44	0.233	0.794	4.231	0.206
Usage, 2	142	0.310	0.757	4.447	0.242
Usage, 2	256	0.258	0.743	4.300	0.221
Usage, 4	44	0.171	0.745	3.981	0.164
Usage, 4	142	0.229	0.796	4.138	0.183
Usage, 4	256	0.177	0.748	4.035	0.173
Joint-F, 4	44	0.262	0.706	4.429	0.241
Joint-F, 4	142	0.252	0.744	4.524	0.255
Joint-F, 4	256	0.272	0.668	4.549	0.257
Joint-B, 4	44	0.174	0.769	4.146	0.185
Joint-B, 4	142	0.166	0.817	4.112	0.177
Joint-B, 4	256	0.203	0.787	4.090	0.181
Joint-B, 8	44	0.118	0.824	3.847	0.149
Joint-B, 8	142	0.165	0.785	3.990	0.166
Joint-B, 8	256	0.149	0.790	3.957	0.160
Joint-F, 8	44	0.143	0.762	3.996	0.160
Joint-F, 8	142	0.183	0.757	4.025	0.167
Joint-F, 8	256	0.174	0.715	3.947	0.159

concentration does not imply closer agreement with the full natural organism distribution. Codon bias can affect biological behavior despite synonymous encoding (Coleman et al., 2008; Plotkin & Kudla, 2011; Hershberg & Petrov, 2008).

E.2 Codon frequencies and descriptive discrepancy

The codon analysis counts nonoverlapping frame-zero codons across all 600 responses per group. For amino-acid family a with m_a codons and observed counts N_c, its relative synonymous codon usage is

$$\mathrm{RSCU}(c \mid a) = \frac{m_a N_c}{\sum_{d \in \mathrm{syn}(a)} N_d}. \tag{17}$$

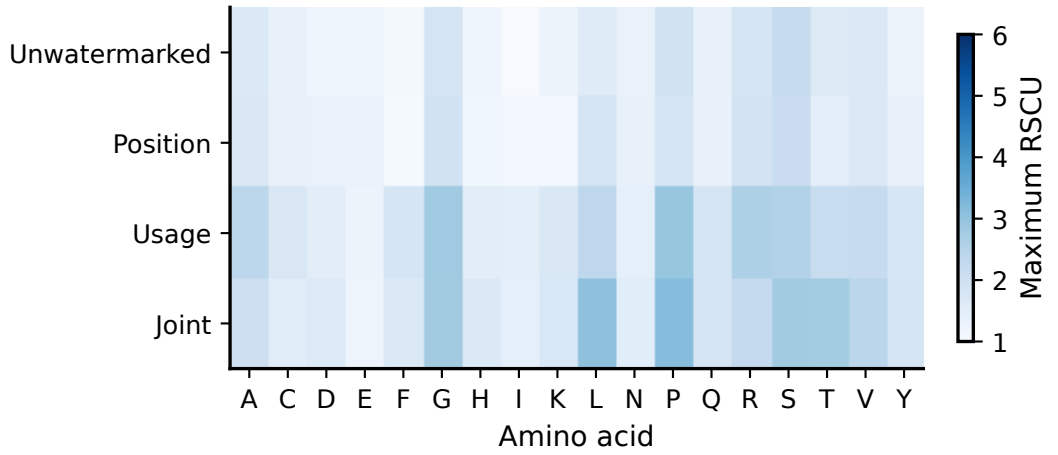

Figure 6: Maximum relative synonymous codon usage within each amino-acid family, pooled over the 600 responses per condition. Values above one indicate concentration relative to uniform synonymous usage. Position/usage use strength 4; joint fractional uses strength 8. This descriptive heatmap does not treat individual codons as independent model replicates.

RSCU values within a nonempty family sum to m_a. The heatmap shows their maximum; the complete table below retains every synonymous codon. Methionine, tryptophan, and stops are excluded from this multi-codon-family summary.

The recorded discrepancy against a reference frequency vector r_c is $X_a^2 = \sum_{c:r_c>0}(N_c - N_a r_c)^2/(N_a r_c)$, where N_a is the observed family total. Against the unwatermarked group, r_c is estimated from its pooled counts. This is a reference-frequency goodness-of-fit calculation, not a two-sample contingency test accounting for reference estimation. The same formula also uses the E. coli reference frequencies. Absolute values depend on count totals and correlated sequence composition; they are descriptive diagnostics here.

The joint group pools 547,106 codons versus 508,422 for the control. Its glutamate discrepancy is 1.38, unlike the large shifts in most other families. Seventeen of the eighteen families exceed the evaluator's approximate nominal 5% cutoffs, but these pooled comparisons do not constitute eighteen independent replicated biological findings. The summed E. coli-reference discrepancies are 312,893 for joint and 129,589 for control. Preferred-subset weighting therefore does not establish closer agreement with the full natural distribution.

Table 20: Per-family joint-versus-control discrepancy and maximum RSCU. X^2 uses pooled control frequencies as its reference; it is a descriptive rather than calibrated independent-codon statistic.

AA	Synonyms	X^2	Control max RSCU	Joint max RSCU
A	4	5546.71	1.696	2.071
C	2	1111.46	1.353	1.585
D	2	6264.74	1.205	1.665
E	2	1.38	1.226	1.243
F	2	2418.88	1.107	1.701
G	4	13099.67	1.892	2.809
H	2	1202.55	1.213	1.686
I	3	1560.12	1.030	1.457
K	2	1442.05	1.304	1.768
L	6	12441.44	1.587	3.053
N	2	251.58	1.351	1.530
P	4	20426.62	1.957	3.166
Q	2	4170.57	1.364	1.860
R	6	29062.56	1.822	2.257

Table 20 (continued)

AA	Synonyms	X^2	Control max RSCU	Joint max RSCU
S	6	6160.03	2.214	2.799
T	4	11933.39	1.661	2.777
V	4	9672.83	1.703	2.407
Y	2	1057.01	1.285	1.833

Table 21: Complete RSCU values for the 59 sense codons in multi-codon families. Position and usage have strength 4; joint fractional has strength 8. Single-codon families and stops are excluded.

AA	Codon	Control	Position	Usage	Joint
A	GCA	0.474	0.479	0.232	0.358
A	GCC	1.323	1.333	1.188	1.302
A	GCG	1.696	1.703	2.401	2.071
A	GCT	0.507	0.484	0.180	0.268
C	TGC	1.353	1.364	1.745	1.585
C	TGT	0.647	0.636	0.255	0.415
D	GAC	1.205	1.335	1.514	1.665
D	GAT	0.795	0.665	0.486	0.335
E	GAA	0.774	0.656	0.689	0.757
E	GAG	1.226	1.344	1.311	1.243
F	TTC	0.893	0.932	0.180	0.299
F	TTT	1.107	1.068	1.820	1.701
G	GGA	0.506	0.610	0.191	0.201
G	GGC	1.892	1.945	2.846	2.809
G	GGG	0.901	0.813	0.488	0.331
G	GGT	0.701	0.632	0.475	0.659
H	CAC	1.213	1.192	1.524	1.686
H	CAT	0.787	0.808	0.476	0.314
I	ATA	1.002	1.121	0.596	0.615
I	ATC	1.030	0.906	0.898	0.929
I	ATT	0.968	0.973	1.506	1.457
K	AAA	1.304	1.150	1.717	1.768
K	AAG	0.696	0.850	0.283	0.232
L	CTA	0.262	0.310	0.050	0.045
L	CTC	1.534	1.879	0.387	0.368
L	CTG	1.587	1.617	2.370	3.053
L	CTT	0.666	0.744	0.084	0.136
L	TTA	1.134	0.826	1.408	1.089
L	TTG	0.817	0.625	1.700	1.309
N	AAC	0.649	0.642	0.560	0.470
N	AAT	1.351	1.358	1.440	1.530
P	CCA	0.665	0.650	0.491	0.512
P	CCC	0.799	0.821	0.428	0.221
P	CCG	1.957	1.850	2.972	3.166
P	CCT	0.579	0.679	0.109	0.100
Q	CAA	0.636	0.636	0.171	0.140
Q	CAG	1.364	1.364	1.829	1.860
R	AGA	0.591	0.366	0.040	0.039
R	AGG	0.647	0.626	0.187	0.209
R	CGA	0.989	1.108	0.809	1.463
R	CGC	1.822	1.910	2.628	2.257
R	CGG	1.389	1.319	2.061	1.782
R	CGT	0.561	0.671	0.276	0.249
S	AGC	1.379	1.669	2.569	2.799

Table 21 (continued)

AA	Codon	Control	Position	Usage	Joint
S	AGT	0.388	0.319	0.171	0.157
S	TCA	0.589	0.627	0.337	0.391
S	TCC	0.870	0.803	0.506	0.586
S	TCG	2.214	2.147	2.239	1.711
S	TCT	0.559	0.436	0.177	0.357
T	ACA	0.663	0.781	0.150	0.119
T	ACC	1.221	1.386	1.610	1.061
T	ACG	1.661	1.470	2.189	2.777
T	ACT	0.455	0.363	0.051	0.042
V	GTA	0.477	0.428	0.151	0.178
V	GTC	1.703	1.690	0.995	0.723
V	GTG	1.225	1.439	2.229	2.407
V	GTT	0.595	0.443	0.625	0.693
Y	TAC	0.715	0.637	0.201	0.167
Y	TAT	1.285	1.363	1.799	1.833

E.3 Structure prediction and selection

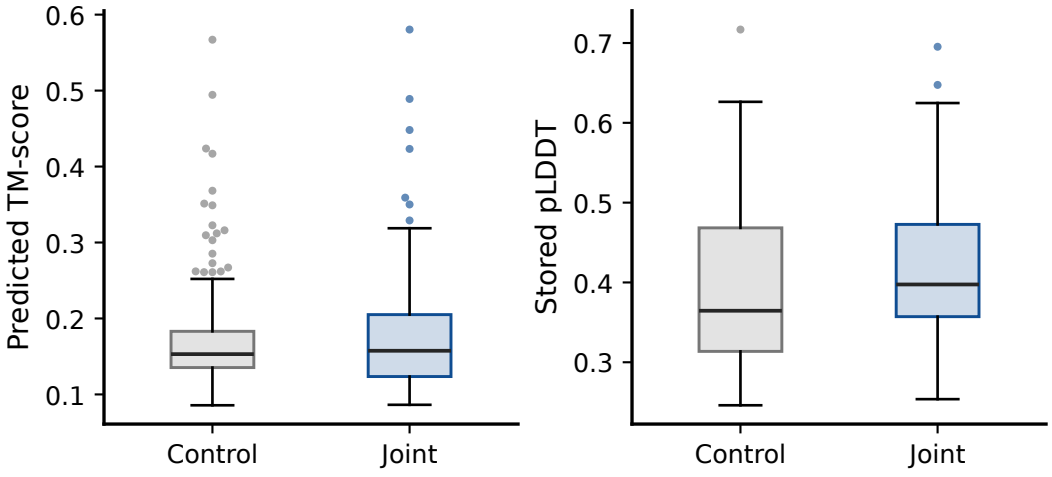

Figure 7: Computational structure summaries for 150 translated outputs per condition (50 index-paired outputs per seed). Boxes show medians and interquartile ranges; whiskers use the 1.5-IQR rule. All outliers are shown as translucent points, displaced only horizontally to prevent overlap. The right panel retains the stored pLDDT scale. Low-confidence scores cannot establish structural non-inferiority.

The folding procedure scans three forward frames, selects the longest stop-free translated segment with preference for an initial methionine, excludes segments shorter than 30 amino acids, and caps length at 400. It selects up to 50 index-paired examples per seed from an initial candidate pool. ESMFold predicts the two proteins separately. The implementation records NaN on a failed forward pass and has a zero-pTM fallback for selected internal exceptions, so retained scores are not interchangeable with validated structures. The displayed 150-pair archive contains finite scores. Mean pTM is about 0.177 for both conditions.

An additional structure-edit assay applies frame-zero synonymous substitutions to joint outputs at probability 0.30, choosing a least-used E. coli synonym other than the original codon. This is a separate editor from KT in the detection matrix. Both strings undergo the longest-segment translation procedure, and pairs with identical selected proteins are excluded. Although substitutions preserve frame-zero amino acids, the selected longest segment may originate in another frame; its protein

Table 22: Seed-level summaries of the control/joint structure comparison. pLDDT uses the stored evaluator scale, without a conversion to a presumed 0–100 range.

Condition	Seed	Pairs	Mean pTM	Stored pLDDT
Control	44	50	0.18532	0.41366
Joint	44	50	0.18110	0.41348
Control	142	50	0.17104	0.36524
Joint	142	50	0.17787	0.44195
Control	256	50	0.17531	0.39889
Joint	256	50	0.17159	0.41490

can change. The retained set is therefore conditional on successful translation and unequal selected proteins, not an unbiased sample of all output edits.

Table 23: Independent structure-edit assay, restricted to 50 nonidentical translated pairs per seed. Difference is original minus edited pTM; it is not a sequence-alignment or structural-distance metric.

Seed	Selected pairs	Original pTM	Edited pTM	Mean difference
44	50	0.1858	0.1367	+0.0490
142	50	0.1802	0.1517	+0.0285
256	50	0.1805	0.1643	+0.0161

The archived pair-level scores permit recomputation of these differences, but a change in model confidence does not measure a distance between two predicted structures. The filtering also prevents interpreting the selected-pair average as the average cost over all 200 evaluated outputs. This assay cannot establish the protein-preservation properties of KT, nor does it establish functional equivalence after synonymous editing.

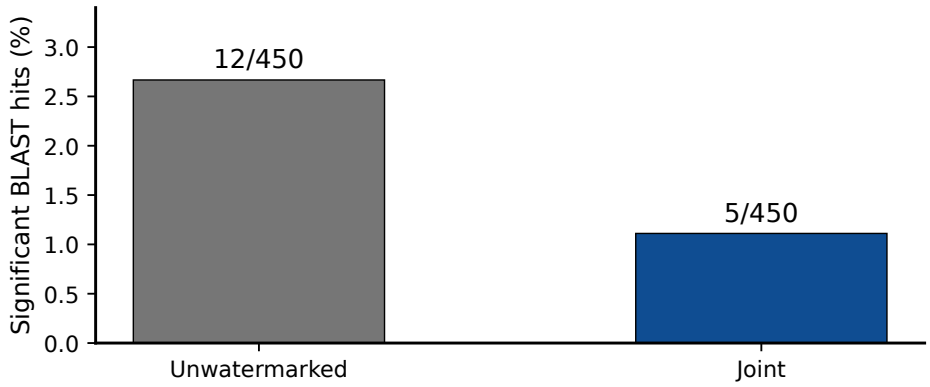

Figure 8: Significant BLAST hits against the E. coli K-12 reference proteome. Numerators and denominators are shown explicitly: 12 of 450 unwatermarked queries and 5 of 450 joint queries satisfy $E < 0.01$. No biological function claim follows from these sparse computational matches.

The BLAST query sets contain 150 translated proteins per seed. The ratio of significant-hit rates is $5/12 = 41.7\%$, but the absolute rates are 1.1% and 2.7%. The reported alignment-identity summaries, 31.71% and 31.65%, are averages of per-seed medians over returned alignments, not means over only significant hits. Structure-prediction methods (Jumper et al., 2021; Abramson et al., 2024) and structural alignment (Gilchrist et al., 2026) provide useful complementary assays, but were not run as additional validators in this study.

E.4 HOMOLOGY DENOMINATORS AND RETURNED ALIGNMENTS

BLAST queries are translated outputs rather than annotated genes. A query can yield no returned alignment, a returned alignment that misses the stated threshold, or a hit with $E < 0.01$. The denominator of a hit rate is all 150 queries per seed, not only those with returned alignments. In total, 427 control and 286 joint queries have returned alignments. At the looser $E < 1$ cutoff, the counts are 205/450 and 89/450, respectively. These values describe retrieval against the chosen reference proteome and depend on database and query composition.

Table 24: Complete homology counts. Returned denotes queries with any returned alignment; both threshold columns use all queries as the denominator.

Condition	Seed	Queries	Returned	$E < 0.01$	$E < 1$
Unwatermarked	44	150	140	7	73
Unwatermarked	142	150	143	0	63
Unwatermarked	256	150	144	5	69
Joint	44	150	97	1	30
Joint	142	150	90	2	24
Joint	256	150	99	2	35

Table 25: Per-seed alignment summaries over returned alignments. Length is median alignment length in amino acids; these medians are not conditioned on $E < 0.01$.

Condition	Seed	Median E	Median identity (%)	Length
Unwatermarked	44	0.92	29.748	60.0
Unwatermarked	142	1.20	31.667	58.0
Unwatermarked	256	1.20	33.534	48.0
Joint	44	2.60	31.111	56.0
Joint	142	2.45	32.026	52.5
Joint	256	2.10	32.000	53.0

The usual expectation-value form $E = Kmn \exp(-\lambda S)$ links a raw alignment score S to a null expectation under the search's scoring system and effective lengths. We use the values returned by BLAST, not an independently reconstructed distribution with assumed K and λ. Aggregate medians and hit counts cannot recover a full per-query histogram, and near-equal identity medians do not establish equal alignment quality for significant hits alone.

F BROADER RESEARCH CONTEXT

Detection, robustness, and attribution. Watermarking spans sampling-based constructions (Aaronson, 2023), logit bias, semantic partitions, and cryptographic designs. Passive detection examines other signals, including probability curvature (Mitchell et al., 2023); stress-testing shows why detection results depend on the adversary and distribution (Sadasivan et al., 2025). Watermark design choices impose different trade-offs (Pang et al., 2024). Provenance is also distinct from training-data extraction (Carlini et al., 2021). Image-domain watermarks (Wen et al., 2023) further illustrate that robustness claims are tied to a representation and transformation class. We do not transfer their guarantees to genomic tokenization.

Architectures and genomic representation. Transformer architectures (Vaswani et al., 2017) and their decoder implementations (Touvron et al., 2023) provide a foundation for genomic generators. DNABERT and DNABERT-2 (Ji et al., 2021; Zhou et al., 2024), the Nucleotide Transformer (Dalla-Torre et al., 2025), GROVER (Sanabria et al., 2024), and GenSLMs (Zvyagin et al., 2023) cover different tokenization and training choices. Geneformer and scGPT (Theodoris et al., 2023; Cui et al., 2024) address cellular representations rather than the DNA-generation setting considered here. A capacity or information-bottleneck interpretation (Shwartz-Ziv & Tishby, 2017) is a hypothesis for future work, not an explanation established by our scale sweep.

Biological utility and watermark layers. ProGen and ProteinMPNN study protein design and its experimental evaluation (Madani et al., 2023; Dauparas et al., 2022). Their results underline the distinction between computational similarity and demonstrated function. DNAMark, CentralMark, and FoldMark (Zhang et al., 2025; 2024) occupy sequence, translation, and structure-level watermark settings. Our study adds teacher-to-student transfer at the genome-token level. It does not claim superiority across these different tasks, nor does it infer wet-lab utility from codon statistics.

G REPRODUCIBILITY COVERAGE

The supplementary data contain the measurements underlying the tables and quantitative figures. The accompanying CPU-only script recomputes condition means, seed standard deviations, detection rates, and quality summaries. The data distinguish per-seed audit summaries, individual folding records, and pooled codon statistics. The edited KGW comparator has one training seed, while the defended settings have three; each record identifies its condition and denominator.

Table 26: Scope of the anonymous measurement export. Access to stored summaries supports recomputation, not rerunning model training.

Measurement layer	Exported content and verification scope
Detection	125 condition records: scheme, strength, seed, n, detected count and mean score.
Nucleotide edits	48 records across four groups, three seeds and four conditions.
Extensions	Organism, model-scale, teacher-control, synonymous-transform and tokenizer seed summaries.
Sequence quality	27 seed summaries, including composition, ORF, codon adaptation and diversity.
Codon analysis	Complete RSCU values, pooled count totals and per-family discrepancy values.
Structure	150 control/joint score pairs and 150 selected original/edited score pairs.
Homology	Per-seed query/hit counts and summaries of returned alignments.

Training corpora, exact evaluation prefixes, model weights, and watermark keys are not included in the results package. It therefore supports verification of the reported measurements, not an end-to-end rerun of training or a new calibration study. Such a rerun requires those inputs, the matching tokenizer and software environment, and compliance with the model and dataset licenses. Appendix A specifies the processing order, sample sizes, and hyperparameters needed to interpret the experiment design.