RESISTIA: REASONING-GUIDED AGENTIC EVALUATION OF SYNTHETIC METAL-RESISTANCE GENES FROM CONDITIONAL GENOMIC FOUNDATION MODELS

José Vásquez-Bastías* & **JS Stockle**
Faculty of Engineering, Universidad Autónoma de Chile
Av. Pedro de Valdivia 425, Providencia, Santiago, Chile.
jose.vasquez6@cloud.uautonoma.cl - juan.stockle@uautonoma.cl

ABSTRACT

Genomic foundation models can generate plausible coding DNA, but turning raw generations into scientifically useful candidates requires reliable post hoc evaluation and robust model-selection criteria. We present **ResistIA**, a *reasoning-guided agentic evaluation and refinement framework* for synthetic metal-resistance genes generated from a LoRA-adapted GenomeOcean-500M model.

ResistIA combines three complementary evaluators—(i) DIAMOND `blastx` protein homology, (ii) ORF/translation/physicochemical checks, and (iii) ChatNT-based semantic/regulatory Q&A with embedding similarity—under an auditable decision policy that selects tools based on runtime availability and evidence needs. We study three progressively stronger versions of the framework: **V1** (baseline multi-tool evaluation), **V2** (closed-loop reweighting over generated batches), and **V3** (validation-aware closed-loop optimization with score smoothing and early stopping).

Across representative and multi-seed benchmarks, V3 improves reliability of iteration selection and shows stable validation behavior (best validation composite score 0.2505 ± 0.0054, 95% CI: $[0.2458, 0.2553]$) with high ORF success (0.9662 ± 0.0058, 95% CI: $[0.9611, 0.9713]$). Validation DIAMOND pass rate remains lower and more variable (0.2217 ± 0.0459, 95% CI: $[0.1815, 0.2619]$), highlighting homology robustness as the main source of run-to-run variation. In paired 10-seed fixed-vs-adaptive comparisons, we do not observe robust adaptive superiority, motivating a conservative reasoning-agent framing.

These results position ResistIA as a compact, reproducible testbed for *reasoning over heterogeneous biological evidence* in agentic scientific workflows, and a practical framework for evaluator-guided refinement of genomic sequence generators.

1 INTRODUCTION

Heavy metal contamination from mining, industrial discharge, and legacy waste remains a persistent environmental and public health concern. Microbial bioremediation (Harshan et al., 2026) offers a scalable route to mitigate metal(loid) pollution, but practical deployment depends on microorganisms that can tolerate, transform, or export toxic ions through specialized resistance systems (e.g., efflux, sequestration, redox transformations, and regulation). Despite decades of study, the known catalog of metal resistance genes remains incomplete and biased toward a subset of cultivable organisms and well-studied metals, limiting the design space for engineering robust bioremediation strains.

Recent advances in *genomic foundation models* suggest a complementary strategy for resistance discovery: instead of identifying genes one by one, we can learn sequence-level priors from large DNA corpora and then steer generation toward specific functional regimes. In this work, we introduce **ResistIA**, a generative pipeline for designing candidate *metal resistance genes* at the DNA level. ResistIA builds on **GenomeOcean-500M**, a decoder-only nucleotide foundation model pretrained on diverse metagenomic assemblies, which provides broad compositional and contextual priors across

*For code and additional reproducibility details, see GitHub: https://github.com/Jose-240/ResistIA/

ecological niches (Zhou et al., 2025). We adapt this model to the metal-resistance domain through **parameter-efficient fine-tuning** with **LoRA**, preserving most pretrained weights while learning task-specific low-rank updates (Hu et al., 2021). This setup enables rapid experimentation on modest GPU resources and supports iterative refinement cycles without the cost of full-model retraining

We fine-tune on **MetalResistance** (Martiny et al., 2023), a curated collection of metal resistance coding sequences derived from experimentally confirmed genes (including BacMet resources). We consider two generation settings: (i) *metal-conditioned generation*, where prompts include a target metal (or metal group) to steer sampling toward relevant resistance families, and (ii) *unconditional generation*, which probes whether the model can rediscover resistance-like sequence structure without explicit conditioning.

Since wet-lab validation is expensive, ResistIA emphasizes computational triage that is fast, transparent, and reproducible. We first filter generated sequences using simple biological sanity checks (start codon, in-frame stop codon when required, length range, homopolymer limits, and GC range). We then perform a homology-based screen against reference resistance proteins using the DIAMOND aligner (Buchfink et al., 2015; 2021), retaining sequences with strong evidence of similarity under stringent alignment statistics. Finally, we summarize natural-vs-generated distributions using compact agent-side diagnostics and composition heatmaps, which help reveal overlap, bias, and persistent mismatches.

Reasoning-guided evaluation perspective. Beyond sequence generation itself, our focus is on how an agent can *reason over heterogeneous evidence* to prioritize candidates under practical constraints. In ResistIA, the agent must decide which evaluators are applicable at runtime (e.g., DIAMOND, ORF/physicochemical checks, ChatNT-based semantic probes), aggregate partially conflicting signals, and update generation preferences across iterations. This makes the problem a compact instance of *tool-augmented scientific reasoning*: the agent does not merely call tools, but uses their outputs to support selection, reweighting, and stopping decisions in an auditable loop.

Contributions. Our main contributions are:

- A **reasoning-guided agentic framework** (**ResistIA**) for evaluating and refining synthetic metal-resistance genes generated from a LoRA-adapted genomic foundation model.
- A **multi-tool evidence stack** that combines DIAMOND protein homology, ORF/translation integrity, physicochemical profiling, and ChatNT-based semantic/regulatory characterization under an auditable decision policy.
- A **closed-loop optimization strategy** with iterative reweighting (V2) and a **validation-aware, smoothed, early-stopped** controller (V3) for robust selection across generation rounds.
- An empirical comparison of **V1**$\rightarrow$**V3** showing improved reliability of selection and stronger validation-selected performance, together with explicit trade-offs between peak metric gains and iteration-to-iteration stability.

2 BACKGROUND

Genome language models (gLMs) treat nucleotide sequences as text and learn transferable priors for downstream genomic tasks. For autoregressive generation, tokenizer design (character-level, k-mer, or subword/BPE) strongly affects context efficiency and output characteristics, motivating task-aware tokenization choices in practical pipelines Shu et al. (2026); Lindsey et al. (2025).

3 METHODOLOGY

In this work, we follow the end-to-end pipeline summarized in Fig. 1. Starting from curated coding DNA sequences of experimentally supported metal-resistance genes (BacMet/MetalResistance v1.1), we adapt the pretrained genomic foundation model GenomeOcean-500M to the metal-resistance domain via parameter-efficient fine-tuning (LoRA). The fine-tuned model is then used for sequence generation in two modes: conditioned generation, where synthesis is guided by a target metal category, and unconditioned generation, where candidates are sampled without an explicit metal

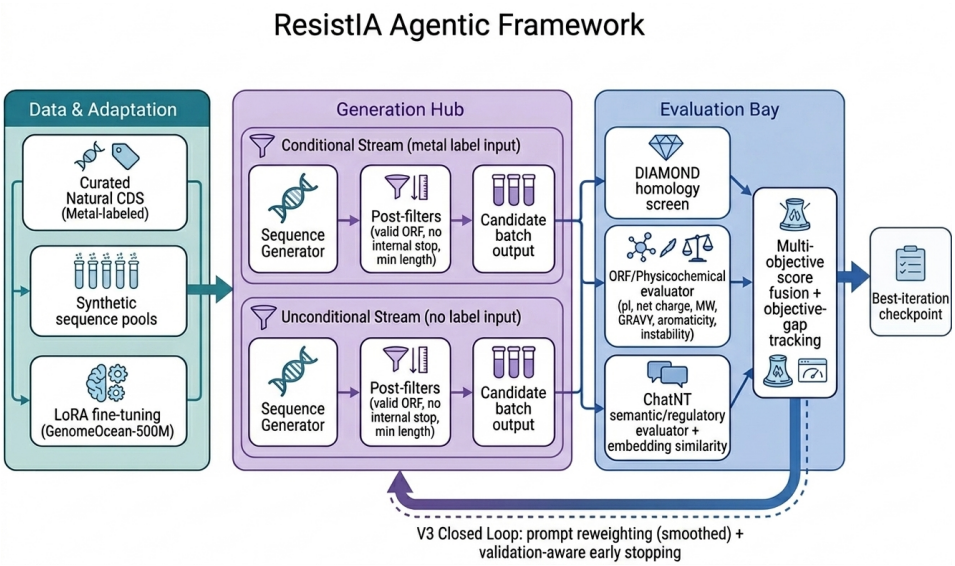

Figure 1: **ResistIA agentic framework.** Curated metal-labeled natural CDS and synthetic pools are used to LoRA-adapt GenomeOcean-500M and generate candidate genes through two streams: (i) metal-*conditioned* and (ii) *unconditioned*, followed by post-filters (valid ORF, no internal stop, minimum length). Candidate batches are evaluated with DIAMOND, ORF/physicochemical checks, and ChatNT semantic/regulatory similarity; their outputs are fused into a multi-objective score for best-iteration selection, metric/figure export, and V3 closed-loop refinement via smoothed prompt reweighting with validation-aware early stopping.

prompt. These generated genes constitute candidate resistance sequences for downstream screening, agentic evaluation, and comparative analysis.

3.1 RESISTIA DATASET

Source and provenance. We construct the *ResistIA* training corpus from the *MetalResistance* gene database released alongside the ARGprofiler pipeline. This collection was seeded from the BacMet v1.1 set of experimentally verified metal-resistance proteins, using BacMet GenBank accessions to extract the corresponding coding DNA sequences (Martiny et al., 2024). The authors then manually curated additional metal-resistance genes from the literature, with special focus on acquired resistance determinants for cobalt, zinc, copper, arsenite, mercury, cadmium, lead, and silver. The resulting database is publicly available on Zenodo (Martiny et al., 2023).

Sequence format and labels. The raw resource is distributed as a FASTA file of coding sequences (CDS). Each record header encodes a gene identifier and a metal-resistance label (e.g., `>merR1_BAB62429.1 Mercury resistance`). Sequences are DNA strings over `A,C,G,T`. In the dataset snapshot used here, we retain **578** natural coding sequences spanning **65** fine-grained metal labels. We additionally maintain an optional coarse mapping into broader metal categories for conditional generation.

3.2 GENOMIC MODEL CHARACTERIZATION: GENOMEOCEAN-500M

GenomeOcean-500M is a decoder-only Transformer genome foundation model trained with next-token prediction on large-scale metagenomic assemblies. It uses a SentencePiece BPE tokenizer with vocabulary size 4,096, built from frequent genomic substrings ($k \in [1, 12]$), and the 500M variant uses a 14-layer decoder (hidden size 1,536; intermediate size 6,144; 8 attention heads) with modern efficiency choices such as rotary position embeddings and optimized attention. The pre-training corpus spans diverse aquatic, terrestrial, and host-associated metagenomic coassemblies (Zhou et al., 2025).

3.2.1 LoRA fine-tuning of GenomeOcean on ResistIA

We fine-tune `DOEJGI/GenomeOcean-500M` on the ResistIA coding-sequence corpus using parameter-efficient adaptation with LoRA under a causal language modeling (CLM) objective. Training sequences are read from a curated FASTA file, uppercased and filtered to `A,C,G,T,N`, then tokenized and truncated/padded to max_length $= 1024$. We split the dataset into train/eval with a small held-out fraction (1%) for periodic evaluation.

Distributed training setup. We train with `torchrun` on two GPUs using DDP. To improve throughput and reduce memory pressure, we enable FlashAttention-2 when available, train in `bfloat16`, allow TF32 matrix multiplications, and use gradient checkpointing. Optimization uses a cosine learning-rate schedule with warmup ratio 0.03, learning rate 2×10^{-4}, batch size 8 per device with gradient accumulation 2 (effective batch size 32), and 3 epochs. We store the Hugging Face cache on shared storage (`HF_HOME`) to avoid redundant downloads across runs.

LoRA configuration. We inject low-rank updates into attention and feed-forward projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with rank $r = 16$, $\alpha = 32$, and dropout 0.05 while freezing base model weights. The resulting adapters and tokenizer are saved for downstream generation.

3.3 Agent-based quality evaluation (RES_AGENT)

To evaluate the quality of GenomeOcean-generated DNA with minimal assumptions and maximum auditability, we designed a compact agentic evaluation wrapper (`RES_AGENT`). The agent does *not* learn a policy; instead it applies a deterministic, human-written selection rule that chooses among several analysis tools depending on the software and databases available at runtime. Each run is logged to a timestamped output directory together with a machine-readable decision report.

Inputs and record collection. The agent assembles a joint pool of sequences by parsing FASTA files for: (i) curated natural resistance genes (`natural_fasta`), (ii) GenomeOcean metal-conditioned generations (`conditional_fasta`), and (iii) optional unconditional generations (`unconditional_fasta`). Optionally, it scans a directory of individual generation outputs (`generated_dir`) by discovering `*.fa` files up to `generated_limit`. All sequences are sanitized to the DNA alphabet $\{A, C, G, T, N\}$ before downstream analysis.

Deterministic tool selection (`auto` mode). If the user sets `--mode` to a specific tool, the agent uses it directly. Otherwise, in `auto` mode it selects the first applicable tool (requiring a non-empty record set) in the following priority order:

1. **Tool a: DIAMOND protein-level screen** if the `diamond` binary is detected in `PATH` and a valid DIAMOND database is provided (`--diamond_db`).
2. **Tools bcd: ORF → translation integrity → physico-chemical profiling** if Biopython and scikit-bio are importable (`Bio`, `skbio`).
3. **Tool e: ChatNT regulatory characterization + embeddings** if `transformers`, `sentence_transformers`, and `torch` are available.
4. **Fallback:** default to **tools bcd**.

The chosen tool and its rationale are saved to `agent_decision.json`.

Tool a: DIAMOND `blastx` homology screening. Tool `a` evaluates whether DNA candidates exhibit protein-level similarity to reference proteins by running DIAMOND `blastx` against a user-provided protein database. Records are searched with tabular output (query id, subject id, percent identity, alignment length, E-value, bit-score), keeping the best hit per query (`--max-target-seqs 1`). A record passes if it meets all thresholds: E-value $\leq$ `diamond_evalue`, aligned length $\geq$ `diamond_min_aln_len`, and bit-score $\geq$ `diamond_min_bitscore`. The merged screening table is exported as CSV/Parquet and summarized in JSON.

Tools bcd: ORF scan, translation integrity, and physico-chemical sanity checks. Tools `bcd` implement coding-structure validation and coarse biochemical characterization. We scan six reading frames for ORFs that start with `ATG` and terminate at `TAA/TAG/TGA`, discard ORFs shorter than `min_orf_nt`, select the longest ORF, translate it, and remove proteins with internal stop codons (premature $\star$). For translated proteins, we compute molecular weight, isoelectric point, net charge at pH 7, GRAVY, aromaticity, instability index (Biopython/ProtParam), plus an amino-acid entropy proxy (k=1, scikit-bio). We also track natural-vs-generated mean shifts in pI and net charge.

Tool e: ChatNT regulatory characterization and embedding similarity. Tool `e` characterizes sequences via natural-language regulatory descriptions. We subsample up to `chatnt_max_sequences` sequences (balanced between natural and generated when possible) and query ChatNT with up to 27 fixed questions about putative regulatory signals (promoters, Shine–Dalgarno sites, terminators, operator-like patterns, etc.). Responses are embedded with a sentence-transformer model (normalized vectors), and natural-vs-generated similarity is summarized by cosine similarity between group-centroid embeddings.

3.4 CLOSED-LOOP AGENTIC REFINEMENT: V1, V2, AND V3

We implemented ResistIA in three progressively stronger versions to study how agentic reasoning and evidence aggregation affect candidate quality and selection robustness.

V1: multi-tool baseline evaluation. V1 performs one-pass evaluation of natural and generated sequences using the available tools (DIAMOND, BCD, ChatNT) and exports per-tool summaries plus consolidated metrics. V1 establishes the baseline evidence profile and reveals dominant failure modes.

V2: closed-loop reweighting over generated batches. V2 introduces iterative refinement. After each iteration, the agent aggregates tool outputs into a composite score and updates generation preferences (prompt weights / sampling emphasis) to favor higher-scoring candidates in subsequent batches. This improves peak performance in several runs but can be sensitive to iteration-level variance.

V3: validation-aware reweighting with smoothing and early stopping. V3 extends V2 with (i) smoothed reweighting updates to reduce abrupt policy shifts across iterations and (ii) a separate validation split used to select the best iteration and trigger early stopping. This yields a more reliable controller that prioritizes generalization and reproducibility over peak-only optimization.

Composite evidence and trade-offs. The agent combines heterogeneous signals from protein homology (DIAMOND), coding plausibility and physicochemical sanity (BCD), and semantic/regulatory consistency (ChatNT). These signals are complementary and may conflict (e.g., improved ORF success but worse physicochemical gaps), motivating composite scoring with per-tool audit trails.

3.5 K-MER (1–6) COMPOSITION EMBEDDINGS

As an alignment-free compositional baseline, we embed each nucleotide sequence by counting overlapping k-mers for $k = 1, \ldots, 6$ over the DNA alphabet $\Sigma = \{A, C, G, T\}$ and converting counts to relative (length-normalized) frequencies. Such k-mer composition profiles provide strong alignment-free signals for genomic classification tasks, supporting their use as a lightweight baseline (Juneja et al., 2022).

For each order k, the feature space has size $|\Sigma|^k = 4^k$, and concatenating frequencies across $k = 1..6$ yields

$$\dim(\mathbf{x}) = \sum_{k=1}^{6} 4^k = 5{,}460.$$

These embeddings are used here for composition-aware diagnostics (including heatmap summaries and secondary geometry analyses).

Table 1: Summary of ResistIA versions (V1–V3) on representative runs. V2 emphasizes peak improvements via closed-loop reweighting, while V3 adds validation-aware selection, smoothing, and early stopping for more reliable iteration choice.

Version	Run	Selection criterion	Global score	DIAMOND pass	ORF success	pI gap	Net-charge gap	ChatNT cosine
V1	Baseline consolidated	single-pass summary	—	0.2103	0.8212	1.6029	6.5023	0.9723
V2	AB (closed loop)	best iteration (train composite)	0.2505	0.2150	0.9499	1.1200	4.3096	—
V2	ABE (closed loop + ChatNT)	best iteration (train composite)	0.2561	0.1875	0.9772	1.4477	5.8887	0.9936
V3	AB (val-aware + smoothing)	best validation iteration	0.2502	0.2583	0.9670	1.2496	6.8849	—
V3	ABE (val-aware + smoothing)	best validation iteration	0.2610	0.2833	0.9749	1.5064	6.4067	0.9702

4 EXPERIMENTS

4.1 GENERATION PROTOCOLS

We consider two generation modes:

- **Unconditional:** prompt with an in-frame start context (e.g., 60–120 nt beginning with `ATG`).
- **Conditional:** prompt with `<METAL_X>` followed by a start context, enabling per-metal sampling.

4.2 POST-GENERATION FILTERS

We apply lightweight biological sanity filters to reduce obvious artifacts: (i) locate the first `ATG`; (ii) optionally enforce an in-frame stop codon (`TAA/TAG/TGA`); (iii) constrain length; (iv) restrict ambiguous bases (`N`); (v) limit homopolymer runs; and (vi) keep GC content within a plausible range.

4.3 AGENTIC EVALUATION VERSIONS AND COMPARISON PROTOCOL

We report results from three versions of the evaluation/refinement framework: **V1** (baseline evaluation), **V2** (closed-loop reweighting), and **V3** (validation-aware closed-loop with smoothing and early stopping). For V2/V3, we run iterative loops over generated batches and log per-iteration summaries for each enabled tool. We compare versions using per-tool metrics (e.g., DIAMOND pass rate, ORF success, physicochemical gaps, ChatNT centroid cosine) and a composite score used to guide iteration selection. Importantly, V3 selects the reported iteration using validation performance rather than training-only peaks.

5 RESULTS

5.1 FROM BASELINE EVALUATION (V1) TO VALIDATION-AWARE CLOSED-LOOP REFINEMENT (V3)

Table 1 summarizes the progression from V1 to V3 on representative runs. Relative to the baseline (V1), V2 substantially improves structural plausibility (ORF success) and can reduce physicochemical gaps in some runs, indicating that iterative evidence-guided reweighting is effective. However, V2 can be sensitive to iteration-level fluctuations. V3 addresses this by selecting iterations using a held-out validation signal with smoothed updates and early stopping, which improves selection robustness and yields the strongest validation-selected composite score in our ABE setting. We observe a practical trade-off: V2 may achieve sharper pointwise gains in specific gaps (e.g., pI), whereas V3 provides a more defensible basis for reproducible reporting.

5.2 MULTI-SEED REPRODUCIBILITY AND V2–V3 COMPARISON

To complement representative single-run traces, we evaluated V3 across five independent seeds. The best validation composite score is 0.2505 ± 0.0054 (95% CI: $[0.2458, 0.2553]$), with high validation ORF success (0.9662 ± 0.0058, 95% CI: $[0.9611, 0.9713]$). In contrast, validation DIAMOND pass rate is more variable (0.2217 ± 0.0459, 95% CI: $[0.1815, 0.2619]$), indicating that homology-related behavior remains the dominant source of residual instability.

Figure 2 summarizes V3 multi-seed means and confidence intervals. Under matched multi-seed reporting, V3 yields slightly higher primary composite and ORF means, while V2 retains higher

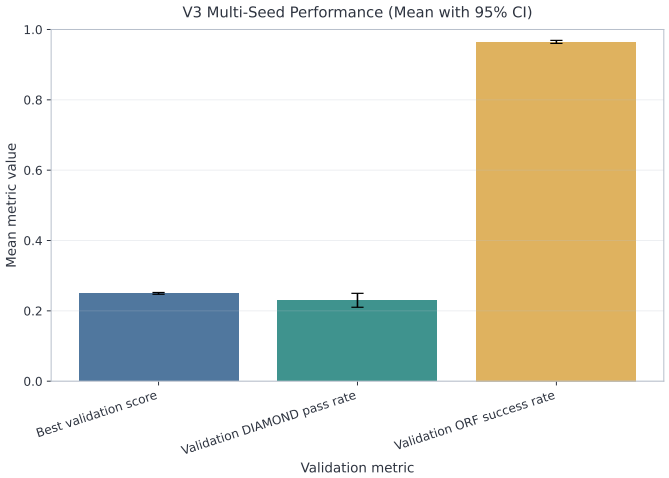

Figure 2: **V3 multi-seed reproducibility (5 seeds).** Mean and 95% CI across independent seeds for the primary validation metrics (best composite score, DIAMOND pass rate, and ORF success rate). Composite and ORF metrics remain stable with relatively tight intervals, while DIAMOND pass rate shows the largest cross-seed variability.

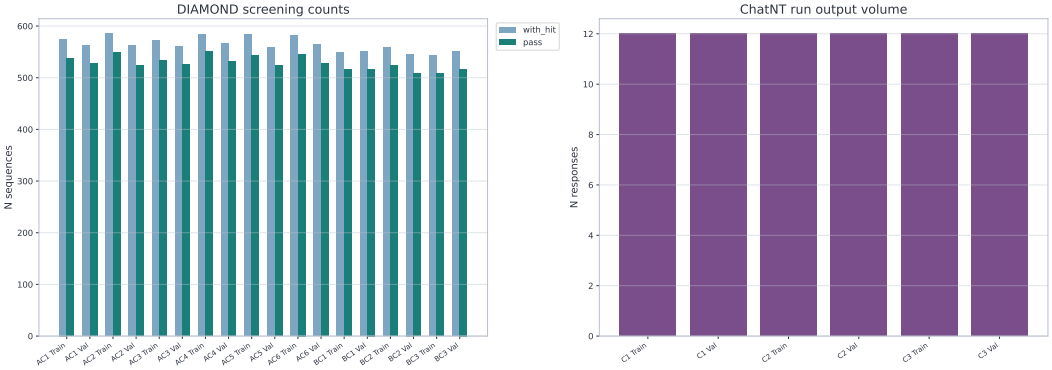

Figure 3: **Closed-loop evaluation footprint across cycles.** Left panel: DIAMOND screening outcomes per cycle, reporting sequences with any homology hit and high-confidence pass. Right panel: ChatNT semantic-evaluation throughput per cycle. Together, these traces provide an auditable view of evidence generation throughout the loop without relying on run-specific timestamps.

DIAMOND pass rate in this configuration. This pattern supports our claim that V3 improves selection reproducibility, but does not uniformly dominate every metric component.

5.3 Operational traces make the agentic loop auditable

Figure 3 summarizes the execution trace of representative V2/V3 runs, including DIAMOND screening counts (queries, hits, and pass counts) and ChatNT output volume across iterations. This view is useful for debugging and for interpreting the availability of evidence during the closed-loop process: score changes across iterations can be contextualized by actual tool outputs, not only by scalar summaries.

Additional diagnostics. Detailed per-tool comparison panels and composition heatmaps are provided in Supplementary Material, where they are reported at larger scale for visual interpretability.

5.4 Compositional heatmaps reveal metal-group-specific shifts

Beyond scalar summaries, we inspect composition-level differences across metal groups using amino-acid and biochemical-class heatmaps (Fig. 4). These diagnostics show where generated

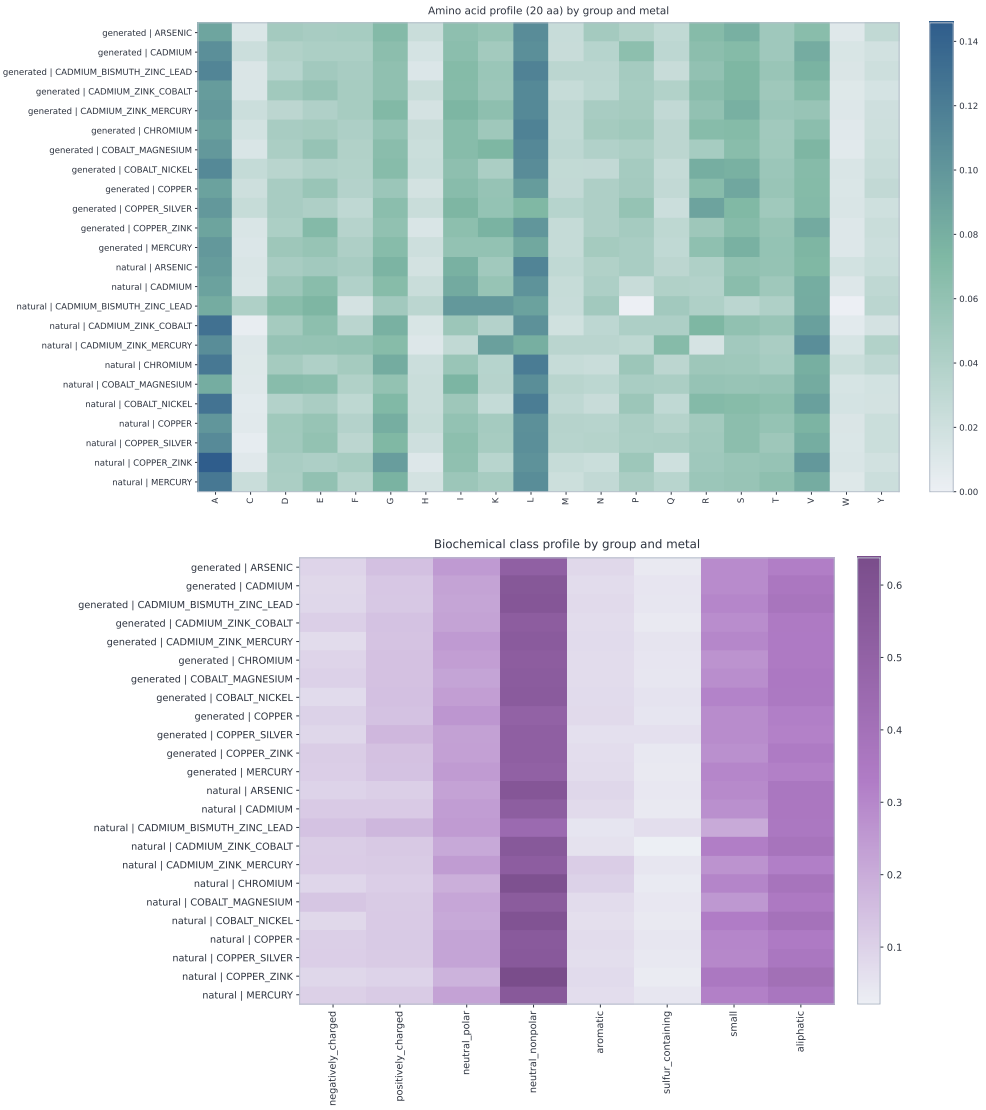

Figure 4: **Protein composition profiles by group and metal.** Top: heatmap of 20-amino-acid frequencies; bottom: heatmap of biochemical-class frequencies, each stratified by metal group for natural and generated sequences. Generated proteins broadly preserve the global compositional structure of natural references, but show metal-dependent local shifts in specific residues and classes, indicating partial conditioning fidelity and clear targets for further refinement.

proteins align with natural distributions and where systematic shifts persist, complementing DIAMOND/ORF/semantic scores with interpretable aggregate composition profiles.

5.5 Conditional vs. unconditional diagnostics

To address the conditional-versus-unconditional anomaly (reviewer concern R3), we ran a dedicated diagnostic analysis with confidence intervals and alignment-quality summaries. Under the same DIAMOND screening criteria, unconditional generations show a substantially higher pass rate than conditional generations (41.3% vs. 19.8%), and this gap remains when accounting for uncertainty (Table 2; Fig. 5). At the same time, both groups yield high-confidence hits among passing sequences (median bitscore 160.5 for conditional and 185.0 for unconditional), indicating that unconditional generations are not merely low-complexity artifacts.

Table 2: Conditional vs. unconditional DIAMOND screening diagnostics.

Setting	N	PASS	PASS (%)	95% CI (%)	Median bitscore
Conditional	3400	674	19.8	[18.5, 21.2]	160.5
Unconditional	543	224	41.3	[37.2, 45.4]	185.0

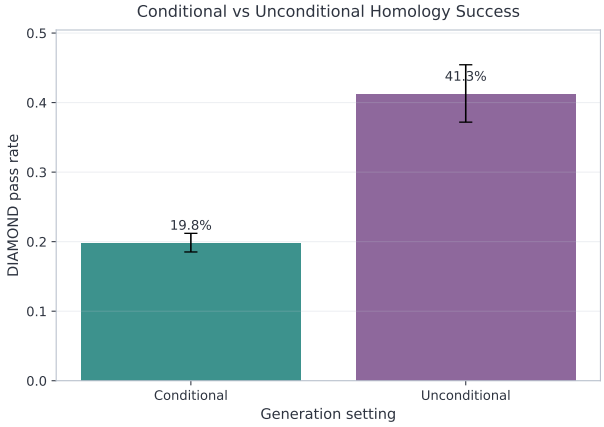

Figure 5: **Conditional vs. unconditional DIAMOND pass rate (95% CI).** Unconditional generation attains a substantially higher homology pass rate (41.3%) than conditional generation (19.8%). The 95% confidence intervals preserve this ordering, supporting a robust performance gap and highlighting conditioning fidelity as a primary target for improvement.

5.6 Adaptive-policy stress test: no robust superiority over fixed

To test a stronger agentic claim (reviewer concern R1), we compared fixed selection against adaptive policies under paired 10-seed evaluations and matched budget. In the final guarded-gap-UCB comparison, adaptive does not show robust gains over fixed on the primary validation composite (mean difference adaptive$-$fixed $= -0.00018$, bootstrap 95% CI: $[-0.00186, 0.00111]$, two-sided sign-flip $p = 0.875$). We observe the same pattern for validation DIAMOND pass rate (difference $= -0.00667$, 95% CI: $[-0.02083, 0.00500]$, $p = 0.5$) and ORF success (difference $= 0.00072$, 95% CI: $[-0.00387, 0.00731]$, $p = 1.0$).

Figure 6 summarizes the paired means and confidence intervals. These results improve transparency of the agentic analysis but do not justify a stronger superiority claim; we therefore keep a conservative reasoning-agent framing.

Secondary geometry diagnostics (reported in supplementary). We also computed k-mer (1–6) UMAP projections to inspect natural-vs.-generated overlap and coarse metal-category structure. These analyses provide complementary composition-space diagnostics, but are secondary to the main agentic results and are reported in Supplementary Material.

6 Discussion

What does the agent improve? The main gain from V1$\rightarrow$V3 is not only better raw metrics, but *better selection logic*. V2 demonstrates that evaluator-guided closed-loop updates can improve ORF success and other quality proxies, yet peak improvements are not always stable. Multi-seed analyses show that V3 provides a more reproducible validation-selection criterion through smoothing and early stopping, while still exposing metric-level trade-offs (notably in homology-related behavior).

Reasoning over heterogeneous evidence. ResistIA is a compact but realistic scientific reasoning setting: the agent must route between tools, combine partially conflicting signals (homology vs. physicochemical alignment vs. semantic similarity), and make auditable decisions under runtime constraints. This tool-augmented reasoning perspective is central to practical scientific workflows, where no single metric captures candidate usefulness.

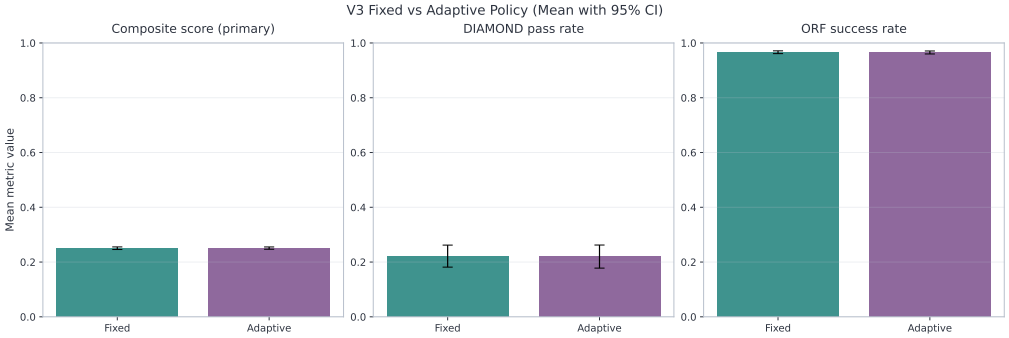

Figure 6: **Paired 10-seed fixed vs. adaptive policy comparison (mean with 95% CI).** We compare fixed and adaptive policies on three primary validation metrics: composite score, DIAMOND pass rate, and ORF success rate. Across all metrics, mean performance is closely matched and confidence intervals substantially overlap, indicating no robust adaptive advantage under this evaluation setting.

Limitations. All biological conclusions in this study are proxy-based. The evaluation stack prioritizes computational indicators of plausibility (ORF integrity, physicochemical consistency, and DIAMOND homology signals), but does not constitute direct functional validation of resistance phenotypes. Biological function requires stronger orthogonal evidence (e.g., domain consistency, structure-aware analyses, genomic context, and family-level annotation) and downstream assays. In addition, paired fixed-vs-adaptive tests do not yet provide statistically robust adaptive gains, so stronger agentic claims remain out of scope for this version.

A practical next-step roadmap is: (i) expanded in silico filtering (domain/motif consistency, structural plausibility, and orthogonal homology checks), (ii) multi-criterion ranking that balances ORF, homology, and diversity, and (iii) wet-lab confirmation via expression and phenotype assays under controlled metal-stress conditions.

Ethics and safety. We focus on methodological aspects of evaluation, refinement, and auditability for resistance-gene candidates in environmental bioremediation contexts. We do not provide wet-lab protocols or actionable guidance for engineering harmful organisms.

7 Conclusion

We presented ResistIA, a reasoning-guided agentic framework for evaluating and refining synthetic metal-resistance genes generated from a LoRA-adapted genomic foundation model. Rather than relying on a single proxy, ResistIA aggregates heterogeneous evidence from protein homology (DIAMOND), coding plausibility and physicochemical sanity (BCD), and ChatNT-based semantic/regulatory characterization, and uses this evidence to support iterative candidate refinement.

Our V1→V3 study shows that closed-loop evaluator-guided optimization improves generated-sequence quality, while validation-aware smoothing and early stopping (V3) provide a stronger basis for reliable iteration selection and reproducible reporting than peak-only optimization. Multi-seed evidence supports this stability claim, but also shows that homology-related metrics remain the main source of variability. Additional paired adaptive-policy stress tests improve transparency yet do not show robust superiority over fixed selection in the current setup.

Beyond this specific application, ResistIA provides a compact testbed for tool-augmented reasoning in scientific workflows, where an agent must combine partially conflicting signals, operate under runtime tool constraints, and make auditable selection decisions. Future work will extend the evidence stack with domain/family-level protein annotation, novelty-aware retrieval checks, and stronger metal-conditioned fidelity metrics.

REFERENCES

Benjamin Buchfink, Chao Xie, and Daniel H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 2015. doi: 10.1038/nmeth.3176.

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

K. Harshan, R. R. Pednekar, V. Priyadharshini, S. Jose, and A. P. Rajan. Deciphering microbial approaches for heavy metal removal: an overview on their mechanisms, innovations and future horizons. *Archives of Microbiology*, 208(3):143, 2026. doi: 10.1007/s00203-025-04685-6.

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. *arXiv*, 2021. doi: 10.48550/arXiv.2106.09685. URL https://arxiv.org/abs/2106.09685.

Sapna Juneja, Annu Dhankhar, Abhinav Juneja, and Shivani Bali. An approach to DNA sequence classification through machine learning: DNA sequencing, K mer counting, thresholding, sequence analysis. *Int. J. Reliab. Qual. E-Healthc.*, 11(2):1–15, August 2022.

LeAnn M. Lindsey, Nicole L. Pershing, Anisa Habib, Keith Dufault-Thompson, W. Zac Stephens, Anne J. Blaschke, Xiaofang Jiang, and Hari Sundar. The impact of tokenizer selection in genomic language models. *Bioinformatics*, 41(9):btaf456, 2025. doi: 10.1093/bioinformatics/btaf456.

Hannah-Marie Martiny, Nikiforos Pyrounakis, Oksana Lukjančenko, Thomas N. Petersen, Frank M. Aarestrup, Philip T. L. C. Clausen, and Patrick Munk. Metalresistance - collection of metal resistance genes (1.0.0). *Zenodo*, 2023. doi: 10.5281/zenodo.8108201. URL https://zenodo.org/records/8108201. Dataset version 1.0.0.

Hannah-Marie Martiny, Nikiforos Pyrounakis, Thomas N. Petersen, Oksana Lukjančenko, Frank M. Aarestrup, Philip T. L. C. Clausen, and Patrick Munk. Argprofiler-a pipeline for large-scale analysis of antimicrobial resistance genes and their flanking regions in metagenomic datasets. *Bioinformatics*, 40(3):btae086, 2024. doi: 10.1093/bioinformatics/btae086.

Liyuan Shu, Jiao Tang, Xiaoyu Guan, and Daoqiang Zhang. A comprehensive survey of genome language models in bioinformatics. *Briefings in Bioinformatics*, 27(1):bbaf724, 2026. doi: 10.1093/bib/bbaf724.

Zhihan Zhou, Robert Riley, Satria Kautsar, Weimin Wu, Rob Egan, Steven Hofmeyr, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Fengchen Liu, et al. Genomeocean: An efficient genome foundation model trained on large-scale metagenomic assemblies. *bioRxiv*, 2025. doi: 10.1101/2025.01.30.635558. URL https://www.biorxiv.org/content/10.1101/2025.01.30.635558v1.

SUPPLEMENTARY MATERIAL

S1. Additional cross-version reproducibility panel. Figure 7 provides a direct multi-seed comparison between V2 and V3 to contextualize the main-text stability claims for V3. We report the same primary validation endpoints used in the core analysis (composite score, DIAMOND pass rate, and ORF success rate), aggregated as mean with 95% confidence intervals across independent seeds. This panel is intended as an uncertainty-aware cross-version reference: instead of focusing on a single best run, it highlights how much performance varies under repeated stochastic training/generation conditions and whether apparent V2–V3 differences remain robust once interval estimates are considered.

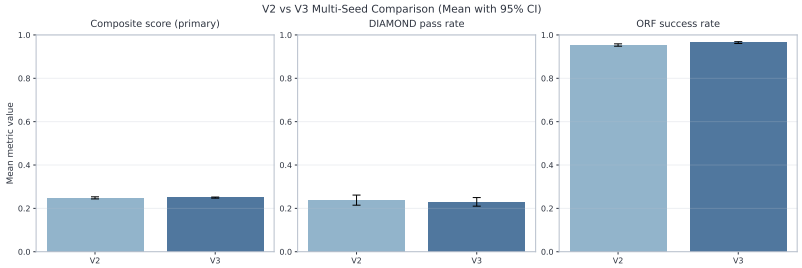

Figure 7: **Supplementary V2–V3 multi-seed comparison (mean with 95% CI).** Confidence intervals are shown for the primary composite objective and key validation metrics (DIAMOND pass rate and ORF success rate). The panel complements the main V3 results by showing cross-version behavior under repeated seeds rather than single-run snapshots.

S2. Alignment-quality diagnostics for DIAMOND-passing sequences. To go beyond binary pass/fail outcomes, Figure 8 analyzes alignment quality within the subset of sequences that already pass DIAMOND filters. We compare conditional and unconditional generation using three complementary alignment descriptors: bitscore (overall match strength), percent identity (residue-level agreement), and alignment length (extent of matched region). This decomposition helps separate two effects that can be hidden in pass-rate summaries alone: (i) central tendency of alignment quality among accepted candidates, and (ii) dispersion/tail behavior that reflects heterogeneity in candidate quality across generation settings.

S3. Agent-side evaluation panels from the previous version (audit trail). To preserve reviewer-requested traceability across manuscript revisions, Figure 9 retains compact agent summaries from the earlier evaluation stack. The upper panel presents the homology-aware view (including DIAMOND-linked outcomes), while the lower panel reports the DIAMOND-free BCD-focused breakdown with ORF and physicochemical descriptors. Keeping both views in supplementary form makes the evidence trail explicit: readers can inspect both the homology-driven perspective and the broader biochemical profile shifts between natural and generated sequences.

S4. Embedding geometry diagnostics for DIAMOND-passing sequences. To complement binary pass/fail outcomes, Figure 10 analyzes the representation geometry within the subset of sequences that already pass DIAMOND filters. We construct a joint 3D UMAP from k-mer embeddings of natural and DIAMOND-passing synthetic sequences, and quantify local cross-domain mixing using a kNN-based statistic. The observed mixing at $k = 15$ is 0.304, substantially lower than the permutation baseline (0.498 ± 0.005; $z = -40.88$), indicating that local neighborhoods remain strongly domain-structured despite partial global overlap in the projection. Including this panel in the supplementary material preserves the focus of the main text while documenting the manifold-level behavior of accepted candidates.

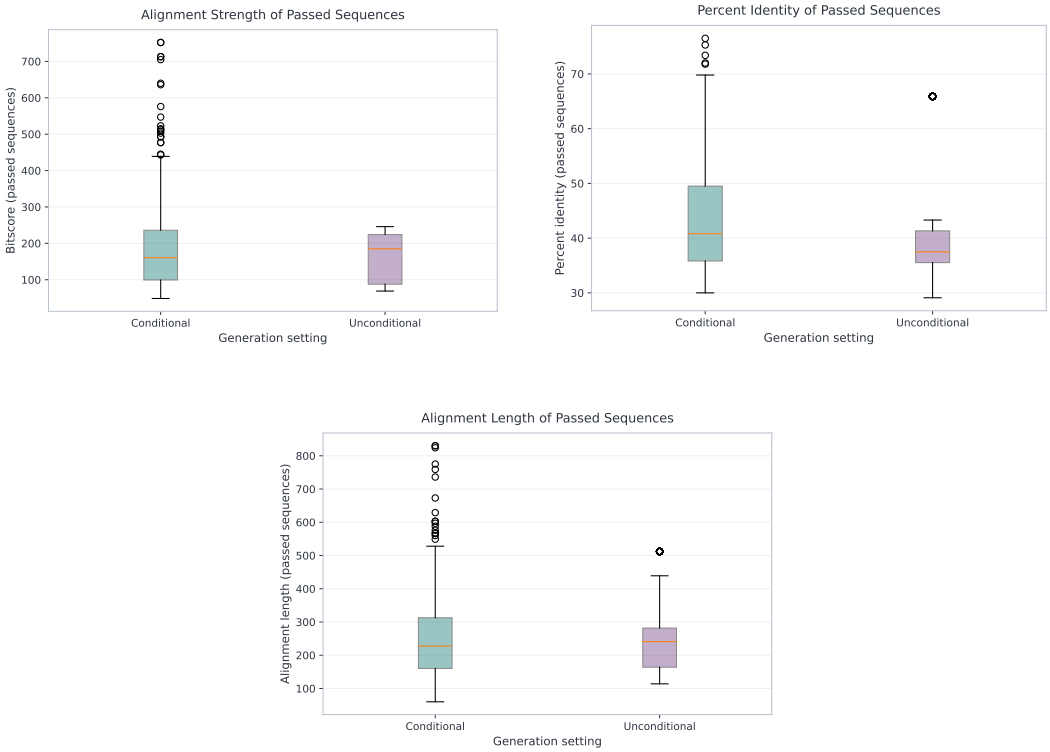

Figure 8: **Supplementary alignment-quality diagnostics among DIAMOND-passing hits.** Top row: bitscore (left) and percent identity (right). Bottom row: alignment length (centered). The distributions indicate how accepted candidates differ not only in pass frequency but also in the strength, identity, and span of their homology signal.

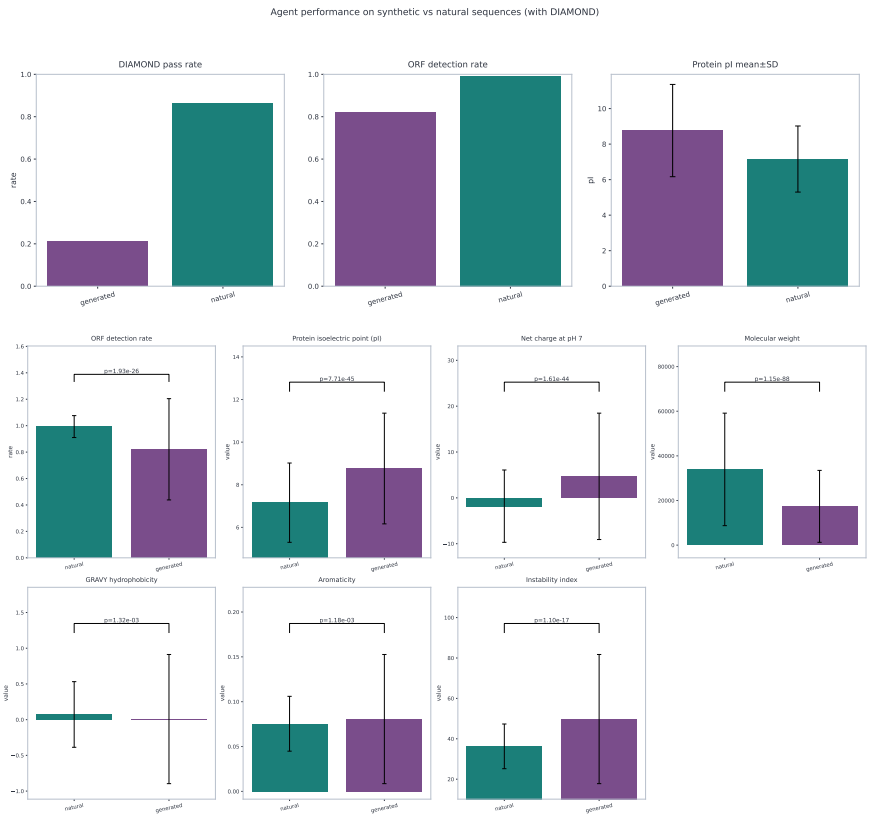

Figure 9: **Agent-side natural-vs-generated evaluation with and without DIAMOND.** Upper panel: compact homology-aware summary. Lower panel: expanded BCD-based comparison including ORF detection and protein-level physicochemical descriptors (e.g., pI, net charge, molecular weight, GRAVY, aromaticity, instability index), with statistical annotations where available. Together, the panels show which differences persist across evaluation modes.

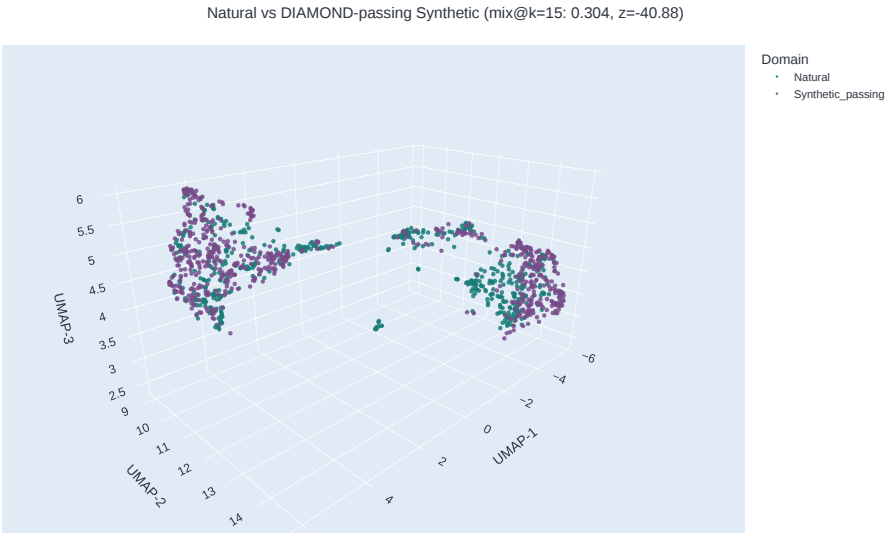

Figure 10: **3D UMAP.** Projection of k-mer embeddings for natural sequences ($n = 578$) and DIAMOND-passing synthetic sequences ($n = 674$). Colors denote domain. Local domain mixing at $k = 15$ is 0.304, substantially lower than the permutation baseline (0.498 ± 0.005; $z = -40.88$), indicating strong neighborhood-level separation between natural and synthetic manifolds despite partial global overlap.