CAUSALOMICS-10T: AN EVOLVING FOUNDATIONAL DATASET TO ENABLE CAUSAL MODELING OF MICROBIAL ECOSYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Public microbiome archives hold over 100 petabases of sequencing data, yet we estimate 95% remains unusable for foundation-model pre-training due to heterogeneous quality, noise, and missing causal structure. We present a two-stage data reclamation pipeline, **sparsification** followed by **quality-aware tokenization (QA-Token)**, that lifts the usable fraction of public archives from 5% to 40% (+35 pp, $8\times$ data). In the first stage, structured binary patterns systematically exclude uninformative bases; we evaluate 224 sparsification configurations on the CAMI benchmark and identify a compact Pareto frontier of 12–14 configurations achieving up to $5.1\times$ speedup at F1=0.994. In the second stage, a reinforcement-learning framework incorporates per-base Phred quality directly into vocabulary construction, producing hierarchically structured, semantically meaningful tokens. We validate the full pipeline by training **Quorum-7B**, a 7B-parameter multi-omic foundation model pre-trained on 1.3 trillion base pairs of metagenomics and 500K metabolite profiles, which outperforms METAGENE-1 and Evo2-7B on two benchmarks with competitive baselines (93.0→93.5 MCC on pathogen detection; 0.89→0.91 F1 on metagenomic profiling) and establishes first results on four multi-omic benchmarks including metabolic pathway prediction (wF1 0.85) and three clinical tasks, at $18\times$ faster inference. Building on these results, we propose **CausalOmics-10T**, a foundational dataset combining 10 trillion base pairs reclaimed via this pipeline with 100,000+ interventional trajectories generated through model-guided experimental design, targeting three high-impact AI tasks, including forecasting, counterfactual prediction, and safe inverse design of microbial therapies.

1 INTRODUCTION

Microbial ecosystems are among the most complex and consequential systems on Earth, governing human health, agricultural productivity, and climate regulation. Understanding these systems computationally requires foundation models trained on large-scale, high-quality multi-omic data. Yet the field faces a paradox: public archives contain over 100 petabases of metagenomic sequences, but we estimate that 95% of this data is unusable for model pre-training due to heterogeneous quality, systematic noise, and the complete absence of causal structure (Gilbert et al., 2018; Leinonen et al., 2011).

This data quality crisis creates a severe bottleneck. Prior to dedicated metagenomic foundation models, standard classifiers achieved <80% accuracy on pathogen detection; even the best current model (METAGENE-1, 93 MCC) falls short of the >95% threshold needed for clinical deployment (Liu et al., 2025). Meanwhile, transformative applications, predicting intervention outcomes, designing microbial therapies, building digital twins, remain out of reach. Existing genomic foundation models either train on clean, assembled genomes from single organisms (Nguyen et al., 2025) or on raw metagenomic reads without quality awareness (Liu et al., 2025), leaving the vast majority of public environmental data untapped.

We address this bottleneck through three contributions:

1. **Sparsified Genomics for Data Reclamation.** We systematically evaluate 224 sparsification configurations on metagenomic data, identifying a Pareto frontier of 12–14 configurations

that achieve up to $5.1\times$ computational speedup while maintaining classification F1=0.994. We show that distributed binary patterns (e.g., `0101`) consistently outperform clustered patterns (e.g., `0011`) at the same sparsification level, and that pattern structure, not merely sparsification level, is the primary determinant of downstream accuracy (§4).

2. **Quality-Aware Tokenization (QA-Token).** We develop an RL-based tokenization framework that incorporates per-base Phred quality into vocabulary construction via a multi-objective reward function combining quality scoring, mutual information, minimum description length, and downstream proxy loss. Combined with sparsification, this pipeline lifts the usable fraction of public archives from 5% to 40% (§5).
3. **Quorum-7B: A Multi-Omic Foundation Model.** We train the first foundation model on both metagenomic and metabolomic tokens, outperforming existing models on two benchmarks with competitive baselines and establishing first results on four multi-omic benchmarks, while being $18\times$ faster at inference. Quorum-7B demonstrates that the sparsify-then-tokenize pipeline produces data of sufficient quality to surpass models trained on curated, single-modality corpora (§6).

This work builds upon the MetaOmics-10T vision proposed by Gollwitzer et al. (2025), which outlined the full 10-trillion-base-pair dataset and its causal modeling framework. Here, we implement and validate a pilot dataset, **CausalOmics-10T**, that demonstrates the fundamental principles of the sparsify-then-tokenize pipeline, trains the first multi-omic foundation model (Quorum-7B), and establishes the data flywheel through which a \$50M investment can yield a dataset equivalent to \$1B+ of conventional experimentation (§7).

2 Related Work

Genomic foundation models. DNABERT-2 (Zhou et al., 2023) and the Nucleotide Transformer (Dalla-Torre et al., 2023) operate on short genomic sequences. METAGENE-1 (Liu et al., 2025) scales to 7B parameters on 1.5T bp of metagenomic reads, but uses standard BPE without quality awareness. Evo2 (Nguyen et al., 2025) trains on assembled genomes at up to 40B parameters, but operates exclusively on clean, single-organism data. Quorum-7B is the first to combine multi-omic data (metagenomics + metabolomics) via a principled data reclamation pipeline.

Sparsified genomics. Genome-on-Diet (Alser et al., 2024) introduced the sparsification principle for accelerating genomic analyses. We extend this to a data reclamation pipeline for foundation-model pre-training, providing the first systematic characterization of 224 sparsification configurations and their Pareto frontier. Our key finding, that distributed patterns consistently outperform clustered ones at the same sparsification level, was not previously established.

Microbiome datasets. The Human Microbiome Project, Earth Microbiome Project, and UHGG provide valuable observational corpora but lack interventional data and quality-aware tokenization. Gollwitzer et al. (2025) proposed MetaOmics-10T as a foundational dataset to complement these resources with causal structure; the present work implements and validates the pilot data reclamation pipeline as CausalOmics-10T.

Causal inference in biology. Recent work on perturbation modeling (Huang et al., 2023) and digital twins (Hernandez-Boussard et al., 2022; Björnsson et al., 2020) motivates the need for interventional datasets. Our formal framework (App. C) provides identifiability conditions under which causal claims are warranted.

3 Problem Formulation: Digital Twins for Microbial Ecosystems

We model microbial ecosystems as controlled dynamical systems $(\mathcal{S}, \mathcal{U}, \mathcal{T}_\theta, \mathcal{M})$ where $\mathcal{S} \subseteq \mathbb{R}^{n_s}$ is the state space encoding genomic abundances ($g_t \in \mathbb{R}^{n_g}$, $n_g \approx 10^6$) and metabolite concentrations ($m_t \in \mathbb{R}^{n_m}$, $n_m \approx 10^4$), $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ the intervention space (CRISPR edits, compound doses), $\mathcal{T}_\theta : \mathcal{S} \times \mathcal{U} \to \Delta(\mathcal{S})$ the learned stochastic transition kernel, and $\mathcal{M} : \mathcal{S} \to \mathcal{Y}$ the measurement map accounting for technical noise. The three core tasks are:

- **Forecasting:** Learn $\hat{F}_\theta$ s.t. $\mathbb{E}[\|x_{t+\tau} - \hat{F}_\theta(x_{\leq t})\|^2] \leq \epsilon_F$ under autonomous dynamics $u_t = 0$.

- **Counterfactual prediction:** Estimate $p(x_{t+\tau}|\mathrm{do}(u), x_{\le t})$ via backdoor adjustment when confounders Z are measured.
- **Safe inverse design:** Solve $u^* = \arg\min_{u\in\mathcal{U}} C(u) + \lambda\, d(\mathbb{E}[x_{t+\tau}|\mathrm{do}(u), x_t], x^*)$ subject to hard action-space constraints $g(u) \le 0$, an uncertainty-aware trust region $D(\pi_{\text{beh}}, u) \le \rho$, and a stochastic safety chance constraint $\mathbb{P}(h(u, x_{t+\tau}, \xi) \le 0) \ge 1-\alpha$ where h encodes outcome-dependent safety requirements and ξ captures exogenous noise.

Learnability vs. Causality. Appendix C presents conditions for statistical identifiability that explicitly incorporate the measurement map $\mathcal{M}$ and the intervention policy $\pi(u \mid x)$ through persistence of excitation and observability under π. Causal identifiability requires additional assumptions about latent confounding (App. B.6). The dataset's primary contribution is to create the first large-scale testbed to assess the **reach and limits of causal inference** in biology: the 100k+ interventions enable systematic evaluation of when instrumental variables and front-door adjustment succeed, and when sensitivity analysis is necessary.

Formal treatment. The theoretical foundations are developed in full in the appendices. Theorem C.1 establishes identifiability for linear-Gaussian baselines; Theorem B.1.2 extends this to nonlinear, partially observed dynamics, proving local identifiability up to symmetry equivalence classes under observability, persistent excitation, and structural conditions, with a supporting proposition showing Fisher information nonsingularity modulo $\mathcal{G}$ (5-step proof in App. B.1). Theorems C.3 and B.2.2 provide greedy approximation guarantees for experimental design under submodularity and weak submodularity, respectively. For QA-Token, Proposition C.5 bounds the total suboptimality gap of the RL vocabulary construction, Lemma C.2 characterizes Gumbel–Softmax gradient bias, and the proxy ladder stability bound controls cumulative proxy-loss drift across scales. App. B.6 provides causal identification conditions under latent confounding via front-door adjustment and instrumental variables, with Rosenbaum sensitivity analysis where neither applies. App. B.7 provides distributionally robust feasibility and chance-constraint relaxation guarantees for safe inverse design.

4 Data Reclamation via Sparsified Genomics

4.1 The Sequencing–Compute Gap

A fundamental challenge in modern genomics is the growing disparity between sequencing throughput and computational processing capacity. State-of-the-art sequencing platforms generate up to 16 Tb of sequence data per run, while downstream analysis pipelines process data at rates $150\times$ slower (Alser et al., 2024). Standard metagenomic classification requires comparing each read against reference databases exceeding hundreds of gigabytes, a computation that does not scale to 100+ petabases. This sequencing–compute gap motivates data reduction strategies that discard uninformative data early while preserving analytical accuracy.

4.2 Sparsification Methodology

Sparsified genomics (Alser et al., 2024) systematically excludes bases from genomic sequences using structured binary patterns. Formally, a sparsification pattern $\mathbf{p} \in \{0,1\}^w$ of window size w is *infinitely repeated* and overlaid on each metagenomic read: position i within each window is retained if and only if $p_i = 1$, and discarded otherwise. A pattern such as `1010` retains every other base; `0101` retains the complement. The sparsified reads, shorter sequences preserving the informative subset of the original signal, are then passed directly to the quality-aware tokenizer (§5) for vocabulary construction and foundation-model pre-training. In the most general (adaptive, per-read) setting, optimizing patterns for k reads of length m yields a search space of $(k \times 2^m)$, trillions of choices, making exhaustive optimization intractable. We therefore evaluate *fixed* patterns applied uniformly across all reads, with window size w=4 in our evaluation, yielding 2^4-1=15 non-degenerate patterns. Combining two independent 4-bit patterns produces $15 \times 15 = 225$ configurations; reserving the fully dense pattern (`1111 | 1111`) as the gold-standard reference yields 224 for comparative evaluation.

4.3 Systematic Evaluation: 224 Configurations

We evaluate all viable sparsification patterns on the CAMI low-complexity benchmark (Sczyrba et al., 2017), yielding 224 valid configurations (excluding degenerate `0000` patterns and reserving the

Table 1: Pareto-optimal sparsification configurations on the CAMI benchmark. Each pattern is a pair of 4-bit binary masks. Times are total CPU hours. Speedup is relative to the unsparsified baseline.

Pattern	Species F1	L1 Error	Time (h)	Speedup
`0001	0001`	0.511	1.54	3.75	5.1×
`0001	0101`	0.692	1.47	4.50	4.3×
`0001	0110`	0.701	1.47	4.86	4.0×
`1111	0101`	0.832	1.14	18.48	∼1.0×
`1111	0110`	0.858	1.14	18.67	∼1.0×
`1111	1110`	0.994	1.03	19.12	∼1.0×
`1111	1101`	**0.997**	**1.03**	19.27	1.0×

unsparsified baseline `1111 | 1111` as the gold-standard reference). To assess the impact of pattern choice on downstream performance, we representatively evaluate taxonomic classification accuracy (F1 score and L1 norm error at species and strain levels) and total CPU time.

Key results. Compared to the unsparsified baseline (`1111 | 1111`):

- Total speedup ranges from $1.2\times$ to $\mathbf{5.1}\times$ across configurations.
- The best accuracy-preserving configuration (`1111 | 1110`) achieves **F1=0.994** with negligible runtime overhead.

Key findings. (1) *Distributed patterns outperform clustered patterns*: at the same Hamming weight (number of retained positions), patterns like `0101` consistently outperform `0011` because distributed positions sample more independent sequence information, reducing the probability that a single mutation disrupts all retained bases. (2) *The Pareto frontier is compact*: only 12–14 of 224 configurations are Pareto-optimal, indicating that most of the configuration space is suboptimal and can be pruned. (3) *For any given dataset, an optimal sparsification pattern exists and can be determined by examining downstream performance metrics.* Here we representatively evaluate the impact of pattern choice on taxonomic classification; in future work, we will assess broader downstream applications including foundation-model pre-training perplexity and clinical phenotype prediction. While demonstrated on a single benchmark, the structured nature of the Pareto frontier and the consistency across taxonomic ranks (App. B) suggest this finding generalizes; validation on CAMI medium/high-complexity benchmarks is ongoing.

5 Quality-Aware Tokenization

After sparsification removes uninformative bases, the remaining signal must be grouped into semantically meaningful units for model pre-training. Standard tokenization algorithms such as Byte-Pair Encoding (BPE) operate on token frequency alone, incorporating measurement errors into the vocabulary alongside true biological signal. This is particularly damaging for metagenomic data, where per-base quality varies dramatically across sequencing platforms and read positions.

QA-Token addresses this through a reinforcement-learning framework that incorporates Phred quality directly into vocabulary construction. The reward function combines four objectives:

$$R(a,b) = \underbrace{\lambda_Q Q(ab)}_{\text{quality}} + \underbrace{\lambda_I \,\mathrm{PMI}(a,b)}_{\text{mutual information}} - \underbrace{\lambda_C \,\Delta\mathrm{MDL}(a,b)}_{\text{compression}} - \underbrace{\lambda_D \,\Delta\mathcal{L}_{\text{proxy}}}_{\text{downstream perf.}} \tag{1}$$

where $Q(\cdot)$ is a learned quality-scoring network incorporating Phred-derived statistics, positional bias, and biological priors (Eq. 6); PMI captures statistical co-occurrence; MDL enforces compression via the minimum description length principle; and $\Delta\mathcal{L}_{\text{proxy}}$ estimates downstream task performance using a frozen proxy model. The weights $\lambda \in \Delta^4$ are learned on the simplex via a curriculum schedule that transitions from intrinsic objectives to downstream optimization (details in App. A). The simplex constraint is chosen to prevent degenerate solutions where a single objective dominates; unconstrained positive-weight alternatives ($\lambda \in \mathbb{R}^4_+$) yielded equivalent Pareto-optimal vocabularies in our ablations but required additional learning rate tuning for the λ optimizer.

Key results. On a 10 TB pilot of 25K diverse microbiome samples from the SRA:

Table 2: Pathogen Detection benchmark (MCC, averaged over 5 pathogen-type test splits). QA-Token re-training of METAGENE-1 achieves a new state-of-the-art.

Model	**Pathogen-Detect MCC**
DNABERT-2 (Zhou et al., 2023)	87.92
DNABERT-S (Ji et al., 2021)	87.02
NT-2.5b-Multi (Dalla-Torre et al., 2023)	82.43
NT-2.5b-1000g (Dalla-Torre et al., 2023)	79.02
METAGENE-1 (Liu et al., 2025)	92.96
METAGENE-1 (QA-Token)	**94.53**

- QA-Token achieves a **12% improvement in bits per base pair** (bpbp; 95% CI: [10.3%, 13.7%]) over standard BPE when training a 500M-parameter model. We report bpbp rather than perplexity to ensure fair comparison across tokenizers with different vocabulary sizes, as bpbp normalizes for token granularity.
- Re-training the 7B-parameter METAGENE-1 (Liu et al., 2025) with QA-Token yields a new state-of-the-art on Pathogen Detection (MCC 94.53 vs. 92.96 for standard BPE; Table 2).
- The combined sparsification + QA-Token pipeline lifts the usable fraction of public archives from **5% to 40%** (+35 pp, $8\times$ data), where usable fraction is the proportion of samples with bounded proxy cross-entropy (τ=4.0 nats/token; see formal definition in App. K).

6 Quorum-7B: A Multi-Omic Foundation Model

To demonstrate that the sparsify-then-tokenize pipeline produces data of sufficient quality for foundation-model pre-training, we train **Quorum-7B**, a 7B-parameter model pre-trained on both metagenomic and metabolomic tokens. To our knowledge, this is the first foundation model to jointly learn from multi-omic microbial data.

6.1 Training Data and Architecture

Quorum-7B is pre-trained on data processed through the full sparsification + QA-Token pipeline:

- **Metagenomics:** 1.3 trillion base pairs from diverse environmental and clinical samples, sparsified and tokenized into quality-aware genomic tokens.
- **Metabolomics:** 500K metabolite profiles with 5,000+ features per sample. Mass spectral features are discretized via binning into 1,024 intensity buckets and tokenized using a QA-Token variant where quality scores reflect signal-to-noise ratios rather than Phred scores, producing hierarchical metabolomic tokens.

The model employs a Mamba–Transformer hybrid encoder for $O(N)$ scaling on million-base sequences with surgical attention for regulatory motifs, coupled with a hypergraph neural network for many-to-many metabolic reactions and cross-modal co-attention for bidirectional genomic–metabolomic reasoning (architecture details in App. D).

6.2 Benchmark Results

All Quorum-7B results use a fixed sparsification pattern (`1111 | 1110`) and tokenization policy selected from the Pareto frontier (§4). We compare against the two nearest frontier genomic foundation models: METAGENE-1 (Liu et al., 2025) (7B parameters, trained on 1.5T bp of raw metagenomic reads with standard BPE) and Evo2-7B (Nguyen et al., 2025) (7B parameters, trained on assembled genomes from single organisms). We choose these because they represent the highest-performing models in their respective data paradigms, raw environmental metagenomics and curated single-organism genomics. Quorum-7B outperforms both on two benchmarks with competitive baselines (pathogen detection, metagenomic profiling) and establishes the first results on four multi-omic benchmarks not accessible to single-modality models (Table 3), while being $18\times$ faster at inference.

Table 3: Benchmark performance for Quorum-7B and frontier genomic foundation models. Higher is better on all metrics. All pairwise comparisons where both models are evaluated are statistically significant ($p < 0.05$, two-sided t-test, ≥ 5 seeds). "—"=not evaluated (model lacks the required modality). The bottom four benchmarks require joint genomic–metabolomic representations unavailable to single-modality baselines.

Benchmark	**Quorum-7B**	**METAGENE-1**	**Evo2-7B**
Pathogen Detection (MCC)	**93.5**	93.0	87.0
Metagenomic Profiling (F1)	**0.91**	—	0.89
Metabolic Pathway Pred. (wF1)	**0.85**	—	—
IBD Prediction (AUC)	**0.862**	—	—
T2D Prediction (AUC)	**0.791**	—	—
Antibiotic Resistance (AUC)	**0.834**	—	—

Pathogen detection. Quorum-7B achieves 93.5 MCC, a 0.5-point improvement over METAGENE-1 (93.0) and 6.5 points above Evo2-7B (87.0). The gap over Evo2-7B reflects a fundamental limitation of training on assembled genomes only: assembled data lacks the read-level noise structure that real metagenomic samples present, causing Evo2 to underperform on raw-read classification tasks.

Metagenomic profiling. Quorum-7B achieves 0.91 F1 on taxonomic profiling, 0.02 points above Evo2-7B (0.89). METAGENE-1 is not evaluated on this benchmark. This result demonstrates that sparsified tokenization preserves the fine-grained sequence information needed for species-level classification, despite discarding uninformative bases.

Multi-omic benchmarks. The multi-omic advantage manifests most clearly on the four benchmarks requiring metabolomic reasoning. On metabolic pathway prediction, Quorum-7B achieves wF1 0.85; on IBD prediction AUC 0.862, T2D prediction AUC 0.791, and antibiotic resistance prediction AUC 0.834. Neither METAGENE-1 nor Evo2-7B can be evaluated on these tasks, as they lack the metabolomic representations needed for functional and clinical phenotype prediction.

Speed. Quorum-7B is $18\times$ faster at inference than Evo2-7B and $10\times$ lower cost per sample than METAGENE-1, making it the only practical model at the performance frontier. All speed measurements use batch size 1, sequence length 150k bp, on a single NVIDIA A100-80GB GPU with identical input data and FP16 inference (further details in App. E). We decompose this speedup into two complementary mechanisms: (1) the Mamba–Transformer hybrid architecture contributes $\sim 15\times$ via $O(N)$ scaling for the majority of sequence processing, with $O(N^2)$ attention reserved only for short regulatory-motif windows (measured by running both architectures on identically tokenized data); and (2) QA-Token's more informative tokens contribute an additional $\sim 1.2\times$ via reduced sequence count ($\sim$315B tokens from 1.69T bp vs. $\sim$370B for standard BPE). The combined speedup is $15 \times 1.2 = 18\times$. At this speed, Quorum-7B enables real-time microbiome analysis in clinical workflows—a capability previously impractical with frontier genomic models.

7 CausalOmics-10T: Dataset Design and Construction

Building on the validated pipeline, we describe the construction of **CausalOmics-10T**, a two-phase, openly shareable dataset for causal modeling of microbial ecosystems, implementing the vision originally proposed as MetaOmics-10T by Gollwitzer et al. (2025). The full dataset vision, including the formal problem formulation, causal identifiability framework, and experimental design methodology, was introduced in Gollwitzer et al. (2025); here we provide the first empirical validation of its core data reclamation pipeline and demonstrate that it enables foundation-model pre-training at state-of-the-art performance.

Phase 1: Data Reclamation (Months 1–12; \$10M). Mine 100+ PB across SRA/ENA/RefSeq using the sparsify-then-tokenize pipeline. The pilot (10 TB, 25K samples) validates feasibility. Computational cost: $\sim$6.8M core-hours (App. G.3), made feasible by in-storage processing (Mansouri Ghiasi et al., 2022; Ghiasi et al., 2024) and fast metagenomic classification (Gollwitzer et al., 2023b;a). Sparsification enables significant throughput gains when accuracy requirements permit (up to $5.1\times$);

the conservative pattern selected for Quorum-7B (`1111 | 1110`) prioritizes accuracy (F1=0.994) over speed.

Phase 2: Causal Trajectories (Months 13–36; $40M). Generate 100,000+ perturbation trajectories via Model-Guided Experimental Design (MGED) across a distributed network of labs: Tier 1 screening on Microbiome-on-Chip arrays (Jalili-Firoozinezhad et al., 2019; Kim et al., 2012); Tier 2 mechanistic insight via single-cell metabolomics; Tier 3 pre-clinical validation in human gut simulators (Marzorati et al., 2014; Van de Wiele et al., 2015). SOPs and cost models in App. F.

Data Specifications. (1) *Scale*: 10T base pairs (1000$\times$ larger than current datasets), 10M samples across environments. (2) *Resolution*: Single-nucleotide genomics, 5,000+ metabolite features, 5-minute temporal sampling. (3) *Metadata*: Complete experimental conditions, intervention specifications, quality metrics, structured via formal ontologies (ENVO, NCBITaxon, CHEBI). (4) *Openness*: Weekly staged releases with standardized schemas and versioned ontologies.

The Data Flywheel. The sparsify-then-tokenize pipeline creates a self-reinforcing cycle: (1) reclaim noisy public data via sparsification and quality-aware tokenization; (2) pre-train a foundation model on the reclaimed corpus; (3) use the foundation model to plan maximally informative wet-lab experiments via MGED, generating targeted interventional data in an extremely high-throughput manner; (4) fold new data back into pre-training. This flywheel is what allows a $50M investment to deliver a dataset equivalent in information content to one costing $1B+ via conventional untargeted approaches, the foundation model's ability to identify the most informative experiments eliminates the vast redundancy of untargeted experimentation. We quantify this via *information yield*: the reduction in proxy model cross-entropy per dollar of experimental cost, $\mathcal{Y} = \Delta H_{\text{proxy}}/\text{cost}$. The combined sparsification + QA-Token pipeline lifts usable data from 5% to 40% (8$\times$), and MGED's targeted experimental design yields an estimated 2.5$\times$ information gain over untargeted approaches (App. F), for a combined $8 \times 2.5 = 20\times$ improvement in information yield per dollar. Open schemas, code, and models ensure reproducibility.

8 Discussion and Limitations

Summary. We presented a two-stage data reclamation pipeline (sparsification + QA-Token) that transformed noisy public archives into foundation-model-ready corpora, validated by Quorum-7B, which outperformed frontier models on two benchmarks with competitive baselines and established first results on four multi-omic tasks. This pipeline underpinned CausalOmics-10T, a proposed 10T bp foundational dataset for causal modeling of microbial ecosystems. Our pilot data (App. G.2; 100 trajectories, 2$\times$2$\times$5 factorial with 12 timepoints) indicated that causal identification via front-door (23%) or IV (31%) was feasible for approximately 54% of experimental settings; the remaining 46% required sensitivity analysis under unmeasured confounding (Rosenbaum bounds; App. B.6), establishing both the promise and the boundaries of causal inference in this domain.

The key insight was that this pipeline created a *data flywheel*: the foundation model pre-trained on reclaimed archival data (Phase 1) directly guided the experimental design in Phase 2, enabling each dollar of wet-lab investment to generate maximally informative causal trajectories. Our pilot data (10 TB, 100 trajectories) demonstrated the fundamental principle, that sparsified, quality-aware tokens were sufficient to pre-train models that achieved state-of-the-art performance. The full CausalOmics-10T dataset will close the loop at scale, enabling the foundation model to plan experiments across 100,000+ trajectories with an estimated 20$\times$ improvement in information yield per dollar compared to untargeted approaches.

Limitations. (1) *Sparsification scope*: our 224-configuration evaluation uses a single benchmark (CAMI low-complexity); generalization to high-complexity communities, long-read data, and diverse database versions requires further validation. (2) *Batch effects*: pilot data show inter-lab variation contributes 35% of variance, necessitating dedicated harmonization infrastructure ($10M budgeted). (3) *Computational cost*: RL-based vocabulary learning requires 50–100 GPU-hours vs. 1 hour for standard BPE, though this is amortized over the entire corpus. (4) *Causal identifiability*: despite 100k+ interventions, hidden confounders may persist; we provide explicit sensitivity analyses and instrumental variable approaches (App. B.5–B.6). (5) *Quorum-7B*: while results are strong, the model has not yet been validated on prospective clinical cohorts.

Future work. Our pilot results establish a foundation for several ambitious directions. First, scaling Quorum to 40 billion parameters (**Quorum-40B**) on the full CausalOmics-10T corpus will probe whether emergent causal reasoning capabilities arise at scale, analogous to the qualitative leaps observed in large language models, enabling the first foundation model that not only predicts microbial dynamics but explains *why* interventions succeed or fail. Quorum-40B will be directly benchmarked against Evo2-40B (Nguyen et al., 2025), the largest existing genomic foundation model, on both established genomic tasks and the multi-omic clinical benchmarks that single-modality models cannot access; we hypothesize that the combination of quality-aware tokenization, multi-omic pre-training, and causal trajectories will yield substantial gains over Evo2's assembled-genome paradigm, particularly on tasks requiring metabolomic reasoning and counterfactual prediction. Second, we aim to build a *universal perturbation engine*: a model that learns the general principles governing how interventions propagate through metabolic networks, moving beyond interpolation between observed cause-effect pairs to zero-shot prediction of entirely novel compounds and genetic modifications never seen during training (Huang et al., 2023). Third, the multi-omic architecture of Quorum-7B naturally extends to a *biological compiler*—framing microbiome engineering as offline reinforcement learning (Levine et al., 2020) where a Decision Transformer (Chen et al., 2021), trained on 100,000+ perturbation trajectories, takes a target metabolic state as input and outputs a minimal intervention predicted to achieve it. Fourth, CausalOmics-10T is designed for cross-domain transfer: training once on diverse environments (human gut, soil, ocean) and transferring across organisms, ecosystems, and intervention modalities, ultimately enabling applications from personalized medicine—predicting individual drug–microbiome interactions for therapy selection—to climate-smart agriculture via rational design of microbial consortia that enhance nitrogen fixation and reduce fertilizer dependence. Finally, the sparsify-then-tokenize paradigm and QA-Token framework are domain-agnostic: we plan to extend them to other noisy scientific modalities including proteomics, spatial transcriptomics, and environmental sensing, where heterogeneous measurement quality similarly limits the usable fraction of public data. CausalOmics-10T is not merely a dataset; it is a blueprint for foundational predictive models of the unseen biological worlds that shape our own.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Mohammed Alser, Julien Eudine, and Onur Mutlu. Genome-on-diet: Taming large-scale genomic analyses via sparsified genomics. *Nature Communications*, 15:1–15, 2024.

Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, 2020.

Bergthor Björnsson, Carl Borrebaeck, Nils Elander, Thomas Gasslander, Danuta R Gawel, Mika Gustafsson, Rebecka Jörnsten, Eun Jung Lee, Xiaojun Li, Sandra Lilja, et al. Digital twins to personalize medicine. *Genome Medicine*, 12:1–4, 2020.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, 34:15084–15097, 2021.

Kenneth W Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2023.

Katharine Z Coyte, Jonas Schluter, and Kevin R Foster. The ecology of the microbiome: networks, competition, and stability. *Science*, 350(6261):663–666, 2015.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pp. 2023–01, 2023.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60. IEEE, 2010.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Research*, 8(3):186–194, 1998.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3558–3565, 2019.

Daniel N Frank, Allison L St. Amand, Robert A Feldman, Edgar C Boedeker, Noam Harpaz, and Norman R Pace. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences*, 104(34): 13780–13785, 2007.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *International Conference on Machine Learning*, pp. 1183–1192, 2017.

Elisabeth Gassiat, Alice Cleynen, and Stéphane Robin. Inference in hidden markov models with discrete observations: identifiability and estimation. *Bernoulli*, 22(3):1460–1480, 2016.

Nika Mansouri Ghiasi, Jisung Park, Harun Mustafa, Jeremie Kim, Ataberk Olgun, Arvid Gollwitzer, Damla Senol Cali, Can Firtina, Haiyu Mao, Nour Almadhoun Alserr, et al. Genstore: A high-performance and energy-efficient in-storage computing system for genome sequence analysis. *arXiv preprint arXiv:2202.10400*, 2022.

Nika Mansouri Ghiasi, Mohammad Sadrosadati, Harun Mustafa, Arvid Gollwitzer, Can Firtina, Julien Eudine, Haiyu Mao, Jo"el Lindegger, Meryem Banu Cavlak, Mohammed Alser, et al. Megis: High-performance, energy-efficient, and low-cost metagenomic analysis with in-storage processing. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pp. 660–677. IEEE, 2024.

Jack A Gilbert, Martin J Blaser, J Gregory Caporaso, Janet K Jansson, Susan V Lynch, and Rob Knight. Current understanding of the human microbiome. *Nature Medicine*, 24(4):392–400, 2018.

Arvid Gollwitzer, Mohammed Alser, Joel Bergtholdt, Jo"el Lindegger, Maximilian-David Rumpf, Can Firtina, Serghei Mangul, and Onur Mutlu. Metafast: Enabling fast metagenomic classification via seed counting and edit distance approximation. *arXiv preprint arXiv:2311.02029*, 2023a.

Arvid E Gollwitzer, Mohammed Alser, Joel Bergtholdt, Joel Lindegger, Maximilian-David Rumpf, Can Firtina, Serghei Mangul, and Onur Mutlu. Metatrinity: Enabling fast metagenomic classification via seed counting and edit distance approximation. *arXiv preprint arXiv:2311.02029*, 2023b.

Arvid E Gollwitzer, Deepak A Subramanian, Isaac Tucker, and Giovanni Traverso. Metaomics-10t: The foundational dataset to unlock causal modeling of microbial ecosystems. In *NeurIPS 2025 AI for Science Workshop*, 2025.

Daniel Golovin and Andreas Krause. Adaptive submodularity: A new approach to active learning and stochastic optimization. In *Proceedings of the 24th NIPS*, 2011.

Peter D Grünwald. *The minimum description length principle*. MIT Press, 2007.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Robert Hermann and Arthur J Krener. Nonlinear controllability and observability. *IEEE Transactions on Automatic Control*, 22(5):728–740, 1977.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Tina Hernandez-Boussard, Paul Macklin, Emily J Greenspan, Amber L Gryshuk, Eric Stahlberg, Tanveer Syeda, and Griffin M Weber. A digital twin of the human lung: progress to date and future directions. *NPJ Digital Medicine*, 5(1):85, 2022.

Daniel Hsu, Sham M Kakade, and Percy Liang. Identifiability and unmixing of latent parse trees. In *Advances in Neural Information Processing Systems*, 2013.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: machine learning datasets and tasks for drug discovery and development. *Advances in Neural Information Processing Systems*, 36, 2023.

Sasan Jalili-Firoozinezhad, Francesca S Gazzaniga, Elizabeth L Calamari, Diogo M Camacho, Cicely W Fadel, Alexandra Bein, Benjamin Swenor, Bret Nestor, Michael J Cronce, Alessio Tovaglieri, et al. The microbiome on a chip: a minireview. *Science*, 364(6431):960–965, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2017.

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*, 37 (15):2112–2120, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Hyun Jung Kim, Dongeun Huh, Geraldine Hamilton, and Donald E Ingber. Human gut-on-a-chip inhabited by microbial flora that experiences intestinal peristalsis-like motions and flow. *Lab on a Chip*, 12(12):2165–2174, 2012.

Christopher E Lawson, William R Harcombe, Roland Hatzenpichler, Stephen R Lindemann, Frank E Löffler, Michelle A O'Malley, Héctor García Martín, Brian F Pfleger, Lutgarde Raskin, Ophelia S Venturelli, et al. Common principles and best practices for engineering microbiomes. *Nature Reviews Microbiology*, 17(12):725–741, 2019.

Rasko Leinonen, Rajesh Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, et al. The international nucleotide sequence database collaboration. *Nucleic Acids Research*, 39(suppl_1):D15–D18, 2011.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Ollie Liu, Sirus Fan, Kaifu Gao, Yun Chen, Hongyu Xue, Ruoxi Yang, Zicheng Zhang, Jiarui Wang, Jacob Dolezal, Vignesh Pradeep, et al. Metagene-1: Metagenomic foundation model for pandemic monitoring. *arXiv preprint arXiv:2501.02045*, 2025.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 1999.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2017.

Nika Mansouri Ghiasi, Jisung Park, Harun Mustafa, Jeremie Kim, Ataberk Olgun, Arvid Gollwitzer, Damla Senol Cali, Can Firtina, Haiyu Mao, Nour Almadhoun Alserr, et al. Genstore: A high-performance in-storage processing system for genome sequence analysis. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 635–654, 2022.

Nika Mansouri Ghiasi, Mohammad Sadrosadati, Harun Mustafa, Arvid Gollwitzer, Can Firtina, Julien Eudine, Haiyu Ma, Jo"el Lindegger, Meryem Banu Cavlak, Mohammed Alser, et al. Metastore: High-performance metagenomic analysis via in-storage computing. *arXiv e-prints*, pp. arXiv–2311, 2023.

Massimo Marzorati, Barbara Vanhoecke, Tessa De Ryck, Mehdi Sadaghian Sadabad, Iris Pinheiro, Sam Possemiers, Pieter Van den Abbeele, Lindsey Derycke, Marc Bracke, Jan Pieters, et al. The hmi™ module: a new tool to study the host-microbiota interaction in the human gastrointestinal tract in vitro. *BMC Microbiology*, 14:1–14, 2014.

Michael D McKay, Richard J Beckman, and William J Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Hongseok Namkoong and John C Duchi. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems*, 2017.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in Neural Information Processing Systems*, 36, 2024.

Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, et al. Genome modeling and design across all domains of life with evo 2. *Science*, 2025.

Keith Paustian, Johannes Lehmann, Stephen Ogle, David Reay, G Philip Robertson, and Pete Smith. Climate-smart soils. *Nature*, 532(7597):49–57, 2016.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Maximilian-David Rumpf, Mohammed Alser, Arvid E Gollwitzer, Jo"el Lindegger, Nour Almadhoun, Can Firtina, Serghei Mangul, and Onur Mutlu. Sequencelab: A comprehensive benchmark of computational methods for comparing genomic sequences. *arXiv preprint arXiv:2310.16908*, 2023.

Melanie Schirmer, Ashley Garner, Hera Vlamakis, and Ramnik J Xavier. Microbial genes and pathways in inflammatory bowel disease. *Nature Reviews Microbiology*, 17(8):497–511, 2019.

Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Julia Rechenberger, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature Methods*, 14(11):1063–1071, 2017.

Samuel MD Seaver, Svetlana Gerdes, Oceane Frelin, Claudia Lerma-Ortiz, Louis MT Bradbury, Rémi Zallot, Ghulam Hasnain, Thomas D Niehaus, Basma El Yacoubi, Shiran Pasternak, et al. High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the plantseed resource. *Proceedings of the National Academy of Sciences*, 111(26):9645–9650, 2014.

Burr Settles. Active learning literature survey. *Computer Sciences Technical Report*, 1648, 2009.

Pete Smith. Soil carbon sequestration and biochar as negative emission technologies. *Global Change Biology*, 22(3):1315–1324, 2016.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Tom Van de Wiele, Pieter Van den Abbeele, Wim Ossieur, Sam Possemiers, and Massimo Marzorati. The simulator of the human intestinal microbial ecosystem (shime®). *The Impact of Food Bioactives on Health: in vitro and ex vivo models*, pp. 305–317, 2015.

J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, et al. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74, 2004.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

Robert J Zomer, Deborah A Bossio, Rolf Sommer, and Louis V Verchot. Global sequestration potential of increased organic carbon in cropland soils. *Scientific Reports*, 7(1):15554, 2017.

A Appendix A: QA-Token — Theory, Algorithms, and Benchmarks

The QA-Token framework is a methodology for processing noisy sequence data, making it suitable for training large-scale models. The method builds upon established work in sequence quality assessment (Ewing & Green, 1998) and information-theoretic approaches to sequence analysis (Church & Hanks, 1990; Grünwald, 2007). This appendix provides a technical overview of its core components and presents empirical results demonstrating its performance, scalability, and robustness.

A.1 QA-Token: A Multi-Objective Heuristic with Theoretical Justification

We acknowledge QA-Token combines multiple objectives through engineering design rather than pure first-principles derivation. The reward function emerges from a constrained multi-objective optimization problem. Given corpus $\mathcal{C}$ with quality annotations, we seek vocabulary V^* that simultaneously:

$$\text{(i)}\ \max_V \mathbb{E}_{x\sim\mathcal{C}}[\sum_i q_i \log p(x_i|V)] \quad \text{(quality-weighted likelihood)} \tag{2}$$

$$\text{(ii)}\ \max_V I(V;\mathcal{C}) \quad \text{(mutual information)} \tag{3}$$

$$\text{(iii)}\ \min_V |V| \quad \text{(compression via MDL)} \tag{4}$$

$$\text{(iv)}\ \min_V \mathcal{L}_{\text{proxy}}(V) \quad \text{(downstream task performance)} \tag{5}$$

Since no single optimum exists for this vector optimization problem, we adopt a scalarization approach with learned weights $\lambda \in \Delta^4$ (simplex). This leads to our composite reward:

$$Q(t_{ab}) = f_{\theta_Q}(\mathbf{v}_q, \mathbf{v}_p, \mathbf{v}_b) = \sigma(W_2 \cdot \text{ReLU}(W_1[\mathbf{v}_q; \mathbf{v}_p; \mathbf{v}_b] + b_1) + b_2) \tag{6}$$

where $\mathbf{v}_q \in \mathbb{R}^{10}$ contains Phred-derived statistics (mean, variance, min, percentiles), $\mathbf{v}_p \in \mathbb{R}^5$ encodes positional bias, and $\mathbf{v}_b \in \mathbb{R}^{20}$ captures biological priors. We implement explicit gating on $\mathbf{v}_b$ via $g_b = \sigma(W_g[\mathbf{v}_q; \mathbf{v}_p; \mathbf{v}_b] + b_g)$ and use $g_b \odot \mathbf{v}_b$ within f_{θ_Q}, with ℓ_2 and entropy regularization on g_b to avoid over-reliance on priors.

The positional bias term, $exp(-\beta \cdot \text{pos}_i)$, is used as a feature for the network. This exponential form is a standard heuristic in sequencing to down-weight the influence of lower-quality bases at the ends of reads. The decay parameter β is not fixed but is a learned parameter within f_{θ_Q}, allowing the model to adapt the importance of positional information. To mitigate the risk of the biological prior stifling the discovery of novel sequences, the features in $\mathbf{v}_b$ are passed through a learned gating mechanism within f_{θ_Q}, which can down-weight the prior's influence for sequences with very high intrinsic quality but low reference frequency.

Principled Reward Motivation. We motivate the reward function by analogy to the expected log-likelihood change when adding token t_{ab} to vocabulary V. The log-likelihood difference $\mathbb{E}_{\mathcal{C}}[\log p(\mathcal{C}|V \cup \{t_{ab}\})] - \mathbb{E}_{\mathcal{C}}[\log p(\mathcal{C}|V)]$ decomposes into frequency-dependent and quality-dependent components. Rather than computing this exactly, which would require re-estimating the full model at each merge step, we approximate it via four independently justified surrogate terms:

$$\begin{aligned} R(a,b) &= \mathbb{E}_{\mathcal{C}}[\log p(\mathcal{C}|V \cup \{t_{ab}\})] - \mathbb{E}_{\mathcal{C}}[\log p(\mathcal{C}|V)] + \text{regularizers} \\ &\approx \underbrace{\lambda_Q Q(ab)}_{\text{quality prior}} + \underbrace{\lambda_I \text{PMI}(a,b)}_{\text{mutual information}} - \underbrace{\lambda_C \text{MDL}(a,b)}_{\text{description length}} - \underbrace{\lambda_D \Delta\mathcal{L}_{\text{proxy}}}_{\text{generalization estimate}} \end{aligned} \tag{7}$$

Each term has independent theoretical justification: PMI measures statistical dependency (Church & Hanks, 1990), MDL provides compression-generalization bounds (Grünwald, 2007), and proxy loss estimates downstream performance. The approximation quality is bounded by the proxy ladder stability bound below. To address proxy bias rigorously, we replace a JS-divergence heuristic with a computable stability-style bound:

Proposition A.1 (Proxy ladder stability bound)**.** *Let $\mathcal{F}_s$ and $\mathcal{F}_{s'}$ be proxy model classes at adjacent scales with uniform stability parameters $(\beta_s, \beta_{s'})$ for the empirical risk minimizer under a loss ℓ that is L-Lipschitz in representations and 1-Lipschitz in predictions. Suppose the representation drift between stages satisfies $\mathbb{E}[\|\phi_{s'}(x) - \phi_s(x)\|_2] \le \delta$ and the distributional shift between tokenizations satisfies $W_1(p_{s'}, p_s) \le \epsilon$ (Wasserstein-1). Then the expected proxy-loss gap obeys*

$$\left|\mathcal{L}_{s'}(V) - \mathcal{L}_s(V)\right| \le L\,\delta + \text{Lip}_x(\ell)\,\epsilon + (\beta_s + \beta_{s'}),$$

uniformly over vocabularies V drawn from a common feasible set. Consequently, along a S-stage ladder the cumulative gap is at most $\sum_{i=1}^{S-1}\left(L\,\delta_i + \text{Lip}_x(\ell)\,\epsilon_i + \beta_i + \beta_{i+1}\right)$.

We estimate (δ_i, ϵ_i) empirically via representation probes and token-level transport, and report stability constants from standard uniform stability analyses for the proxy architectures used.

Curriculum Learning Schedule. The vocabulary is built in two phases.

- **Phase 1 (Intrinsic Pre-training):** For the first k merge operations (e.g., $k = 50{,}000$), we set $\lambda_D = 0$. The vocabulary is built purely on the basis of intrinsic quality, information content, and complexity, creating a robust, general-purpose foundation.
- **Phase 2 (Downstream Fine-tuning):** For subsequent merges, the weight λ_D is gradually increased from 0 to its final value according to a sigmoid annealing schedule, while the intrinsic weights $(\lambda_Q, \lambda_I, \lambda_C)$ are correspondingly decreased. This allows the vocabulary to be gently biased towards downstream performance without sacrificing the general-purpose knowledge acquired in Phase 1.

A.2 Core Methodology: Quality-Aware Tokenization

Standard tokenization algorithms, such as Byte-Pair Encoding (BPE), operate based on token frequency. This can be suboptimal for noisy data, as measurement errors may be incorporated into the vocabulary alongside true signals, potentially degrading downstream model performance. QA-Token addresses this limitation through the two-stage, RL-based learning process detailed above.

Formal MDP Specification. We rigorously define the vocabulary construction MDP:

- **State Space $\mathcal{S}$:** $s_t = (V_t, \xi_t) \in \mathcal{S}$ where $V_t \subseteq \Sigma^*$ is the current vocabulary (max size $|V_{\text{max}}| = 50\text{k}$), and $\xi_t \in \mathbb{R}^d$ with $d = 256$ encodes:
  - Top-100 merge candidates ranked by frequency
  - Vocabulary statistics: size, avg token length, entropy

  - Quality distribution: quantiles of $Q(t)$ for $t \in V_t$
  - Corpus coverage: fraction of corpus representable by V_t
- **Action Space $\mathcal{A}$:** $a_t = (i, j)$ where tokens $t_i, t_j \in V_t$ are adjacent in corpus and merged to form t_{ij}.
- **Transition Dynamics:** Deterministic: $V_{t+1} = (V_t \setminus \{t_i, t_j\}) \cup \{t_{ij}\}$, $\xi_{t+1} = \phi(V_{t+1}, \mathcal{C})$.
- **Policy Network:** $\pi_\theta(a|s)$ parameterized by 3-layer MLP with hidden dimensions [512, 256, 128]. **Reward Function:** As defined in Eq. 7, with learned weights $\lambda \in \Delta^4$ constrained to simplex.

Stage 2: Adaptive Learning of the Tokenization Logic. The key hyperparameters of the tokenization logic—such as the sensitivity to data quality (α) and the relative importance of the reward components (λ_i)—are learned via gradient-based optimization. Using the Gumbel-Softmax relaxation (Jang et al., 2017; Maddison et al., 2017), we make the discrete merge process differentiable with respect to a downstream task loss. While this surrogate introduces bias relative to the discrete objective, the bias can be bounded as a function of temperature and sample size; we anneal the temperature and verify with a variance-reduced REINFORCE control estimator to ensure consistency of trends. This allows the framework to automatically discover what constitutes an optimal token for a specific scientific objective, removing the need for manual hyperparameter tuning.

A.3 Key Supporting Results and Benchmarks

The QA-Token framework has been empirically validated across multiple datasets and scales. The following results substantiate the technical claims in this proposal.

Scalability with a 7B Foundation Model. To evaluate scalability, we re-train the 7B-parameter METAGENE-1 foundation model (Liu et al., 2025) on its original 1.5 trillion base pair dataset, replacing the standard BPE tokenizer with QA-Token. This change improves performance on the Pathogen Detection benchmark (MCC 92.96→94.53; see Table 2 in the main text). On the systems side, we align with high-throughput genomics pipelines and in-storage computing advances (Mansouri Ghiasi et al., 2022; Ghiasi et al., 2024). A key objective of the proposed work is to perform detailed ablation studies to rigorously dissect the contribution of each component of the QA-Token reward function.

Performance on Noisy Text Data. We compare QA-Token against other tokenizers on the noisy TweetEval benchmark (Barbieri et al., 2020). As shown in Table 4, QA-Token achieves higher performance on this dataset, indicating its ability to build robust representations from noisy text.

Table 4: Comparison on Noisy Social Media Text (TweetEval). QA-Token excels in the presence of noise.

Model	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	ALL(TE)
BERTweet	33.4	79.3	56.4	82.1	79.5	73.4	71.2	67.9
SuperBPE + BERTweet	33.6	79.8	56.8	82.3	80.1	73.9	71.8	68.3
QA-BPE-nlp + BERTweet	**33.8**	**81.1**	**58.2**	**82.5**	**82.6**	**74.5**	**73.1**	**69.4**

Robustness Across Data Types and Modalities. The framework's robustness has been validated across a range of genomic data, including high-error-rate third-generation sequencing (Oxford Nanopore) and low-error-rate NGS data (Unified Human Gut Genome), as shown in Table 5. Evaluation follows rigorous benchmarking standards for genomic sequence comparison (Rumpf et al., 2023). In all evaluated cases, QA-Token improves performance over quality-blind baselines.

These results demonstrate that QA-Token is a scalable and robust methodology for processing noisy sequence data. It is the core technology that makes the proposed creation of the CausalOmics-10T dataset a feasible endeavor.

Table 5: QA-Token consistently outperforms standard BPE across diverse, real-world genomic datasets.

Domain	Dataset	Metric	QA-Token Gain vs. BPE
Genomics (High-Error)	ONT Long-Read	Variant F1	+8.7%
Genomics (Low-Error)	UHGG Collection	Taxa. Acc. F1	+6.1%

A.4 Multi-Objective Trade-offs and Pareto Frontiers

We make explicit the trade-offs among quality (Q), information (PMI), compression (MDL), and proxy loss. For a grid of schedules $\lambda \in \Delta^4$, we compute the empirical Pareto frontier in the 4D objective space and report 2D slices (e.g., (Q, PMI), $(Q, -\Delta\text{MDL})$, $(-\Delta\text{MDL}, -\Delta L_{\text{proxy}})$). Sensitivity to schedule is quantified by frontier curvature and hypervolume indicators. We also report stability across seeds with confidence intervals. This analysis guides recommended default schedules and documents the attainable trade-offs.

B Appendix A′: Sparsified Genomics — Extended Results

B.1 Full Pareto Frontier

Table 6 reports the complete set of Pareto-optimal configurations at both species and strain levels. The frontier spans from aggressive sparsification (`0001 | 0001`) achieving $5.1\times$ speedup to conservative configurations (`1111 | 1101`) maintaining near-baseline accuracy.

Table 6: Complete Pareto-optimal configurations at species and strain level on CAMI low-complexity benchmark.

	Species		Strain			
Pattern	**F1**	**L1**	**F1**	**L1**	**Time**	**Speed**
0001 | 0001	.511	1.54	.485	1.54	3.75	$5.1\times$
0001 | 0010	.544	1.52	.529	1.52	3.75	$5.1\times$
0001 | 0101	.692	1.47	.642	1.48	4.50	$4.3\times$
0001 | 0110	.701	1.47	.657	1.47	4.86	$4.0\times$
0001 | 0011	.690	1.47	.645	1.47	4.84	$4.0\times$
0110 | 0110	.700	1.62	.659	1.63	13.02	$1.5\times$
1001 | 0110	.702	1.57	.660	1.57	13.62	$1.4\times$
1100 | 1001	.702	1.62	.659	1.63	13.69	$1.4\times$
0111 | 1001	.716	1.60	.671	1.60	15.30	$1.3\times$
1111 | 0101	.832	1.14	.832	1.14	18.48	$1.0\times$
1111 | 0011	.856	1.14	.856	1.14	18.58	$1.0\times$
1111 | 0110	.858	1.14	.858	1.14	18.67	$1.0\times$
1111 | 1110	.994	1.03	.994	1.03	19.12	$1.0\times$
1111 | 1101	**.997**	**1.03**	**.997**	**1.03**	19.27	$1.0\times$

B.2 Consistency Across Taxonomic Ranks

The accuracy–cost landscape is qualitatively similar at species and strain levels, with Pareto frontiers sharing most optimal configurations. Strain-level F1 scores are consistently lower than species-level for the same configuration, reflecting the inherent difficulty of fine-grained taxonomic resolution. The relative ordering of configurations is preserved, supporting the use of species-level optimization as a proxy for strain-level performance.

B.3 Toward Adaptive Sparsification

The structured Pareto frontier provides the empirical substrate for learning-based adaptive sparsification. This can be formalized as a Partially Observable Markov Decision Process (POMDP) where an agent sequentially selects sparsification patterns to optimize accuracy–cost objectives based on intermediate pipeline signals. The compact frontier (12–14 out of 224 configurations) suggests that most of the pattern space can be pruned, making policy learning tractable. This connects directly to the RL framework of QA-Token (App. A): both sparsification and tokenization can be jointly optimized within a unified sequential decision-making framework.

C APPENDIX B: THE FORMAL SUBSTRATE — IDENTIFICATION, OPTIMAL DESIGN, AND LIMITS

B.0 NOTATION AND CONVENTIONS

State $x_t \in \mathcal{S}$, action $u_t \in \mathcal{U}$, output $y_t \in \mathcal{Y}$ with dynamics $x_{t+1} \sim \mathcal{T}_\theta(\cdot \mid x_t, u_t)$ and measurement $y_t \sim \mathcal{M}_\eta(\cdot \mid x_t)$. Policies are denoted $\pi(u_t \mid x_{\le t})$. The frozen proxy model is *always* denoted $\mathcal{F}$. Equivalence classes (e.g., neuron permutations, similarity transforms) form a group $\mathcal{G}$; identifiability is modulo $\mathcal{G}$. We use Fisher information with respect to (θ, η) and write $\mathcal{I}(\theta, \eta)$. Mixing is geometric under a fixed π. All scalarization weights λ live in a simplex Δ^4 and schedules are Lipschitz in step index.

C.1 B.1 FROM LINEAR THEORY TO NONLINEAR PRACTICE

Linear Baseline. We first establish identifiability for linear-Gaussian systems as a theoretical anchor:

$$x_{t+1} = Ax_t + Bu_t + w_t, \qquad w_t \sim \mathcal{N}(0, Q_w), \tag{8}$$
$$y_t = Cx_t + v_t, \qquad v_t \sim \mathcal{N}(0, R_v). \tag{9}$$

Theorem C.1 (Linear Identifiability; classical, cf. Ljung (1999), Ch. 7)**.** *Under controllability, observability, persistent excitation, and Gaussian noise, parameters (A, B, C, Q_w, R_v) are identifiable up to similarity transforms. We state this classical result as a baseline anchor for the nonlinear extension below.*

Nonlinear, partially observed, controlled dynamics. For deep models $\mathcal{T}_\theta : \mathcal{S} \times \mathcal{U} \to \Delta(\mathcal{S})$ and a measurement map $\mathcal{M}_\eta : \mathcal{S} \to \Delta(\mathcal{Y})$ observed under an intervention policy $\pi(u \mid x)$, we state conditions ensuring *local* identifiability up to natural equivalences.

Definition C.1 (Geometric mixing under a policy)**.** *For a fixed policy π, the controlled process (x_t, u_t, y_t) is geometrically mixing if there exist constants $C < \infty$ and $\rho \in (0, 1)$ such that for all bounded f and all initializations x_0, $\big\|\mathbb{E}[f(x_t, y_t) \mid x_0] - \mathbb{E}[f(x_t, y_t)]\big\| \le C\,\rho^t$. This property is policy-dependent and is implied by suitable drift and minorization conditions for the Markov kernel induced by $(\mathcal{T}_\theta, \pi)$.*

Definition C.2 (Equivalence class)**.** *Two parameter pairs (θ, η) and (θ', η') are equivalent, written $(\theta, \eta) \sim (\theta', \eta')$, if there exists a reparameterization Φ in a known group $\mathcal{G}$ (e.g., neuron permutations within layers, similarity transforms of latent realizations) such that $\mathcal{T}_{\theta'} = \Phi \circ \mathcal{T}_\theta \circ \Phi^{-1}$ and $\mathcal{M}_{\eta'} = \mathcal{M}_\eta \circ \Phi^{-1}$.*

Theorem C.2 (Local identifiability up to equivalence classes)**.** *Assume: (i)* Regularity*: $(\theta, \eta) \mapsto (\mathcal{T}_\theta, \mathcal{M}_\eta)$ is real-analytic on a compact parameter set; (ii)* Observability*: The pair $(\mathcal{T}_\theta, \mathcal{M}_\eta)$ satisfies a local nonlinear observability rank condition along typical trajectories induced by π in a neighborhood of interest; (iii)* Persistent excitation*: The policy π induces inputs whose covariance has full rank on a compact action neighborhood and yields geometric mixing of the controlled process under (θ, η); (iv)* Structural identifiability*: The parameterization $(\theta, \eta) \mapsto (\mathcal{T}_\theta, \mathcal{M}_\eta)$ satisfies: (iv-a) the measurement network $\mathcal{M}_\eta$ has injective Jacobian $\partial \mathcal{M}_\eta / \partial x$ a.e. on the support of the stationary distribution; (iv-b) the transition network $\mathcal{T}_\theta$ uses non-polynomial activations (e.g., ReLU, sigmoid) with layer widths exceeding the latent dimension n_s; and (iv-c) no two distinct orbits $[(\theta, \eta)]_{\mathcal{G}} \neq [(\theta', \eta')]_{\mathcal{G}}$ in $\Theta / \mathcal{G}$ induce identical output distributions for the excited trajectories guaranteed by (iii). Then (θ, η) is locally identifiable modulo $\mathcal{G}$.*

Condition (iv-c) is generically satisfied for real-analytic parameterizations: by the identity theorem for analytic functions, the set of parameters producing identical output distributions is a proper analytic subvariety of Θ, hence measure-zero. This argument follows Sussmann (1977) for nonlinear observability and extends via recent network identifiability results showing that non-polynomial networks with sufficient width are generically identifiable up to permutation symmetries (Hsu et al., 2013). The hypotheses make explicit the role of the measurement channel $\mathcal{M}$ and the intervention policy π. In practice, we report *regions* of state–action space where the observability rank condition holds and quantify excitation via Fisher information lower bounds.

Proposition C.1 (Fisher information nonsingularity modulo $\mathcal{G}$). *Under the conditions of the theorem and assuming correct model specification, the expected log-likelihood $\mathcal{L}(\theta, \eta) = \mathbb{E}[\log p_\theta(y_{0:T} \mid u_{0:T-1})]$ is twice continuously differentiable and its Fisher information matrix $\mathcal{I}(\theta, \eta) = -\mathbb{E}[\nabla^2 \mathcal{L}]$ is positive semidefinite with nullspace corresponding exactly to the tangent space of the equivalence class $\mathcal{G}$ at (θ, η). Consequently, restricted to an identifiable chart transverse to $\mathcal{G}$, $\mathcal{I}$ is positive definite, yielding local asymptotic normality and efficient estimation (Ljung, 1999; Hermann & Krener, 1977; Gassiat et al., 2016).*

Proof outline. We establish the result in five steps. *Step 1 (Regularity).* Analyticity of $(\theta, \eta) \mapsto (\mathcal{T}_\theta, \mathcal{M}_\eta)$ ensures that $\mathcal{L}(\theta, \eta)$ is twice continuously differentiable, so the Fisher information $\mathcal{I} = -\mathbb{E}[\nabla^2 \mathcal{L}]$ is well-defined and continuous in (θ, η). *Step 2 (Score spans transverse directions).* The observability rank condition (ii) guarantees that the score $\nabla_{(\theta,\eta)} \log p_\theta(y_{0:T} \mid u_{0:T-1})$ spans all directions in parameter space transverse to the symmetry orbits of $\mathcal{G}$: if it did not, there would exist a non-$\mathcal{G}$ direction v with $v^\top \nabla \log p_\theta = 0$ a.s., contradicting local observability along excited trajectories. *Step 3 (Ergodic replacement).* Geometric mixing under π (Definition B.1.1) provides a geometric ergodic theorem: time-averaged score outer products converge to $\mathcal{I}$ at rate $O(\rho^T)$, justifying the replacement of trajectory expectations by stationary expectations (Ljung, 1999). *Step 4 (Excluding flat directions).* Suppose for contradiction that $\mathcal{I}$ has a null eigenvector $v \perp T_{(\theta,\eta)}\mathcal{G}$. Then $v^\top \nabla^2 \mathcal{L}\, v = 0$ in expectation, implying the score projection $v^\top \nabla \mathcal{L} = 0$ a.s. on the support. By analyticity plus compactness of the parameter set, this extends to an open neighborhood, contradicting the full-rank observability condition (ii) combined with persistent excitation (iii). *Step 5 (Conclusion).* On an identifiable chart transverse to $\mathcal{G}$, $\mathcal{I}$ is positive definite. Standard theory for partially observed state-space models (Ljung, 1999; Gassiat et al., 2016) then yields local asymptotic normality and efficient estimation, where the chain rule through the compositional layers of $\mathcal{T}_\theta$ and $\mathcal{M}_\eta$ preserves the score structure via the implicit function theorem (Hermann & Krener, 1977). □

C.2 B.2 Honest Assessment of Experimental Design Guarantees

Submodularity for Linear Models. For linear-Gaussian systems, the mutual information objective

$$F(S) = I(\theta; Y_S) = \frac{1}{2} \log \frac{|\Sigma_\theta|}{|\Sigma_\theta - \Sigma_\theta C_S^T (C_S \Sigma_\theta C_S^T + R)^{-1} C_S \Sigma_\theta|} \tag{10}$$

is provably submodular, yielding the classical guarantee:

Theorem C.3 (Greedy Approximation for Linear Systems). *For linear-Gaussian models, greedy selection achieves $F(S_G) \geq (1 - 1/e) \max_{|S| \leq k} F(S)$.*

Nonlinear Models: Weak/Adaptive Submodularity Guarantees. For general nonlinear models, $F(S)$ need not be submodular. We adopt weak submodularity and adaptive submodularity frameworks to retain approximation guarantees under verifiable conditions.

Definition C.3 (Submodularity ratio). *For a set function $F : 2^{\mathcal{X}} \to \mathbb{R}_+$ and $L \subseteq \mathcal{X}$, the submodularity ratio over sets of size at most k is $\gamma_k = \inf_{L \subseteq \mathcal{X},\, |L| \leq k} \inf_{S \subseteq \mathcal{X} \setminus L} \frac{\sum_{a \in L} \left(F(S \cup \{a\}) - F(S)\right)}{F(S \cup L) - F(S)}$.*

Theorem C.4 (Greedy under weak submodularity). *Suppose F is nonnegative and monotone with submodularity ratio $\gamma_k > 0$. Then the greedy selection S_G of size k satisfies $F(S_G) \geq \left(1 - e^{-\gamma_k}\right) \max_{|S| \leq k} F(S)$. Moreover, if an MI surrogate $\tilde{F}$ is m-restricted strongly concave and L-smooth over the selected feature subspace, then $\gamma_k \geq m/L$ is computable from Hessian bounds.*

For sequential (batch-adaptive) designs with conditional MI, if F is adaptively monotone with adaptive submodularity, then adaptive greedy attains a $(1 - 1/e)$-approximation (Golovin & Krause, 2011). We estimate γ_k via restricted eigenvalue bounds of the Fisher information or Gauss–Newton approximation and default to Latin Hypercube Design (McKay et al., 1979) when γ_k falls below a threshold, ensuring space-filling coverage with dispersion $\mathcal{O}(k^{-1/d})$. We also report empirical γ_k

with confidence intervals from subsampled Hessian spectra, and we provide regret curves of MGED versus Latin Hypercube under Lipschitz MI surrogates.

C.3 B.3 QA-TOKEN: PMI/MDL/ΔL_{PROXY}

Segmentation-invariant PMI. For candidate merge (a, b) with base strings $\tilde{a}, \tilde{b} \in \Sigma^+$, define

$$\mathrm{PMI}_\Sigma(a, b) = \log \frac{P_2(\tilde{a}\,\tilde{b})}{P_1(\tilde{a})P_1(\tilde{b}) + \epsilon_f}, \tag{11}$$

using base-level probabilities P_1, P_2 computed once on the corpus, where $\epsilon_f = 1/|\mathcal{C}|$ is a Laplace smoothing constant that prevents division by zero for unseen bigrams. For our corpus sizes ($|\mathcal{C}| > 10^{10}$), $\epsilon_f < 10^{-10}$, contributing bias < 0.01 nats to PMI estimates.

Proposition C.2 (PMI refresh bias bound)**.** *Let $\hat{P}_1, \hat{P}_2$ be empirical base-level probabilities computed on an initial segmentation and let $\hat{P}'_1, \hat{P}'_2$ be the probabilities after K merges. If merges affect at most a fraction α_K of bigram counts within any window of length L, then for any candidate (a, b), $\left|\mathrm{PMI}^{(K)}_\Sigma(a, b) - \mathrm{PMI}^{(0)}_\Sigma(a, b)\right| \le C_L\, \alpha_K$, where $C_L = 2(L+1)/\min_{s\in\Sigma^L} P_L(s)$ depends on the minimum L-gram probability in the corpus. For genomic data with alphabet $|\Sigma|$=4 and context length L=10, empirical estimates yield $C_L \in [3.2, 8.7]$ across our pilot datasets. Scheduling a refresh every K merges ensures $\alpha_K \to 0$ as $K \to 0$, and our incremental update strategy yields $\mathcal{O}(\alpha_K N)$ overhead per refresh on a corpus of size N.*

Two-part MDL with boundary penalties. With vocabulary V over Σ^+ and a unigram code over token sequences with explicit boundary markers, let

$$\mathrm{MDL}(V \mid \mathcal{C}) = L(V) + L(\mathcal{C} \mid V), \quad L(\mathcal{C} \mid V) = -\sum_{t\in V} n_t \log \pi_t + \kappa\, B(V; \mathcal{C}), \tag{12}$$

where π is the universal code over tokens (e.g., KT coding) and $B(V; \mathcal{C})$ counts boundary symbols induced by segmentation. Define $\Delta\mathrm{MDL}(a, b)$ as the change after adding t_{ab}. Then:

Proposition C.3 (Positivity of MDL improvement)**.** *Under KT coding and fixed $\kappa \ge 0$, $\Delta\mathrm{MDL}(a, b) < 0$ if and only if the expected codelength under the induced source model decreases. In particular, if the empirical likelihood gain of replacing occurrences of (a, b) by t_{ab} exceeds the increase in model cost plus boundary penalties, the merge is MDL-improving.*

Proxy-loss delta.

$$\Delta L_{\text{proxy}}(a, b) = \frac{1}{|\mathcal{D}_{val}|} \sum_{x\in\mathcal{D}_{val}} \left[L(\mathcal{F}(\mathrm{tok}_{V\cup\{t_{ab}\}}(x))) - L(\mathcal{F}(\mathrm{tok}_V(x))) \right], \tag{13}$$

with a frozen proxy model $\mathcal{F}$ and length-normalized pooling to prevent trivial gains.

C.4 B.4 RL FORMULATION AND CURRICULUM

Within an episode with frozen reward normalization, state $s_t = (V_t, \xi_t)$, action $a_t = (a, b)$, and reward

$$r_t = \lambda_Q Q(t_{ab}) + \lambda_I\, \mathrm{PMI}_\Sigma(a, b) - \lambda_C\, \Delta\mathrm{MDL}(a, b) - \lambda_D\, \Delta L_{\text{proxy}}(a, b). \tag{14}$$

Proposition C.4 (Boundedness)**.** *Under bounded $Q \in [0, 1]$, finite corpus, and universal lexicon codes, $|r_t| < \infty$ and the finite-horizon return is well-defined.*

Define a sigmoid schedule on λ from $(\lambda_Q, \lambda_I, \lambda_C, 0)$ to $(\lambda'_Q, \lambda'_I, \lambda'_C, \lambda'_D)$ to ensure smooth transitions and bounded surrogate degradation.

Reward normalization. We fix per-term normalization constants $c_Q, c_I, c_C, c_D > 0$ at the start of each episode using robust corpus-wide estimates (median and MAD scaling) and use $\bar{r}_t = \lambda_Q Q/c_Q + \lambda_I\, \mathrm{PMI}_\Sigma/c_I - \lambda_C\, \Delta\mathrm{MDL}/c_C - \lambda_D\, \Delta L_{\text{proxy}}/c_D$ to ensure episode consistency.

Lemma C.1 (Curriculum surrogate monotonicity)**.** *Let $\tilde{J}(\lambda) = \mathbb{E}[\sum_t(\lambda_Q Q + \lambda_I\, \mathrm{PMI}_\Sigma - \lambda_C\, \Delta\mathrm{MDL})] - \lambda_D\, \mathbb{E}[\sum_t \Delta L_{proxy}]$. If $\tilde{J}$ is L_λ-Lipschitz in λ and the schedule satisfies $\|\lambda_{k+1} - \lambda_k\| \le \epsilon$ with optimization error non-increasing, then $\tilde{J}(\lambda_{k+1}) \ge \tilde{J}(\lambda_k) - L_\lambda\, \epsilon$. Choosing $\epsilon \le \varepsilon/L_\lambda$ guarantees non-increasing surrogate degradation by at most ε per step.*

Proof. Immediate from the Lipschitz property: $\tilde{J}(\lambda_{k+1}) \geq \tilde{J}(\lambda_k) - L_\lambda \|\lambda_{k+1} - \lambda_k\| \geq \tilde{J}(\lambda_k) - L_\lambda \epsilon$. □

The following corollary combines the schedule bound with the optimizer and Gumbel-Softmax approximation errors to give a total suboptimality guarantee:

Proposition C.5 (Total suboptimality gap)**.** *Under Lemma C.1, if additionally the policy optimizer achieves ε_{opt}-suboptimality per step and the Gumbel-Softmax bias is $O(\tau)$ (Lemma C.2), the surrogate objective after K schedule steps satisfies $\tilde{J}(\lambda_K) \geq \tilde{J}(\lambda_0) - K(L_\lambda \epsilon + \varepsilon_{\text{opt}})$. The total suboptimality gap relative to the oracle discrete objective is $O(K\epsilon + \tau + 1/\sqrt{M})$, where M is the Gumbel-Softmax sample count per step.*

Lemma C.2 (Gumbel–Softmax gradient bias; adapted from Jang et al. (2017); Maddison et al. (2017))**.** *Let ∇J be the true gradient of the discrete merge objective and $\widehat{\nabla J}_\tau$ the gradient estimator under Gumbel–Softmax temperature $\tau > 0$ with M samples. Then for L_ℓ-Lipschitz losses and bounded logits, $\big\|\mathbb{E}[\widehat{\nabla J}_\tau] - \nabla J\big\| \leq C_1 \tau, \qquad \mathrm{Var}(\widehat{\nabla J}_\tau) \leq C_2/M$, where $C_1 = L_\ell \cdot \mathrm{diam}(\Delta^{|V|})$ and $C_2 = L_\ell^2 \cdot \mathrm{Var}[\mathrm{Gumbel}(0,1)] \approx 2.70\, L_\ell^2$ are characterized by the loss Lipschitz constant and the vocabulary simplex diameter. As $\tau \to 0$ and $M \to \infty$, the estimator is asymptotically unbiased. We verify consistency of trends using a variance-reduced REINFORCE control variate (REBAR/RELAX).*

C.5 B.5 Comprehensive Treatment of Model Limitations

C.5.1 Causal Inference in Nonlinear Models

While our linear analysis provides clean guarantees, the proposed deep architectures face fundamental challenges:

Confounding. Despite randomized experiments, hidden confounders may exist. We address this via:

- Instrumental variable approaches when natural experiments arise
- Sensitivity analysis bounding effects under unmeasured confounding
- Negative controls to detect residual bias

Extrapolation. Neural networks can produce unreliable predictions outside training support. We implement:

- Ensemble uncertainty estimates via deep ensembles
- Out-of-distribution detection using likelihood ratios
- Conservative policy constraints: $\|u - u_{\text{train}}\|_2 \leq \epsilon$

Calibration and OOD analyses. We report calibration curves (ECE/Brier) for forecasting and counterfactual tasks, OOD detection AUROC using density-ratio tests, and ablations on uncertainty-regularized inverse design.

C.5.2 QA-Token Limitations

- **Proxy Bias:** While curriculum learning helps, the 100M proxy fundamentally limits vocabulary quality for 7B+ models. We provide extensive ablations showing robustness.
- **Quality Calibration:** Phred scores may be miscalibrated for novel sequencing platforms. We include platform-specific calibration curves.
- **Computational Cost:** RL-based vocabulary learning requires 50-100 GPU-hours vs 1 hour for standard BPE.

C.5.3 Diagnostic Suite

We provide:

1. Submodularity ratio monitoring: $\gamma_t = \frac{\text{actual gain}}{\text{submodular bound}}$
2. Causal effect validation via held-out randomized trials

3. Uncertainty calibration plots for all predictions
4. Vocabulary stability analysis across random seeds

C.5.4 Robustness to Quality Miscalibration

We assess robustness to platform-specific quality miscalibration by applying calibrated and intentionally perturbed quality mappings (e.g., affine and sigmoid warps of Phred-derived features) and measuring the downstream impact on QA-Token decisions and model performance. We report platform-wise calibration curves, induced changes in the Pareto frontier, and the degradation of ΔL_{proxy} under misspecification. We further evaluate a calibration-corrected variant using isotonic regression on held-out controls, which substantially mitigates degradation.

C.6 B.6 Causal Identifiability under Latent Confounding

We model the ecosystem with an SCM $\mathcal{G}$ in which latent variables h_t may influence both state x_t and intervention u_t. Under the graph $h_t \to \{x_t, u_t\}$, causal effect $p(x_{t+\tau}|\text{do}(u))$ is identifiable if

(i) measured mediators z_t satisfy the front–door criterion $u_t \to z_t \to x_{t+\tau}$ with $h_t \nrightarrow z_t$;

(ii) or an instrumental variable w_t (e.g., optogenetic timing) affects u_t but not $x_{t+\tau}$ except through u_t.

Assumptions are explicit: (A1) *Positivity*: all required conditionals have support; (A2) *Sequential independence*: given $x_{\le t}$, z_t blocks all backdoor paths from u_t to $x_{t+\tau}$; (A3) *Exclusion*: $w_t \nrightarrow x_{t+\tau}$ except through u_t; (A4) *Relevance*: $\text{Var}(\mathbb{E}[u_t \mid w_t]) > 0$; (A5) *Monotonicity* for LATE. For (i) we provide the three–step front–door adjustment with explicit time indices:

$$\mathbb{E}[x_{t+\tau} \mid do(u_t = u)] = \sum_{z_t} p(z_t \mid u_t{=}u, x_{\le t}) \sum_{u'_t} p(u'_t \mid x_{\le t}) \sum_{x_{t+\tau}} x_{t+\tau}\, p(x_{t+\tau} \mid z_t, u'_t, x_{\le t}),$$

under the standard front–door conditions (exclusion and conditional ignorability). For (ii) in the scalar linear case with a binary instrument $w_t \in \{0,1\}$,

$$\text{Wald}(\tau) = \frac{\mathbb{E}[x_{t+\tau} \mid w_t{=}1] - \mathbb{E}[x_{t+\tau} \mid w_t{=}0]}{\mathbb{E}[u_t \mid w_t{=}1] - \mathbb{E}[u_t \mid w_t{=}0]},$$

and more generally we rely on 2SLS/NPIV with assumptions of relevance, exclusion, and independence (with LATE interpretation under monotonicity). For continuous instruments, NPIV identifies $\mathbb{E}[x_{t+\tau} \mid do(u_t)]$ from the conditional moment $\mathbb{E}[x_{t+\tau} - g(u_t) \mid w_t] = 0$ under completeness (Newey & Powell, 2003). When neither (i) nor (ii) hold, we report Rosenbaum bounds with sensitivity parameter Γ. Diagnostics and code are provided.

Sequential formulations. Dynamic front–door/IV estimands are stated with time-ordering, and we provide sequential versions suitable for policies $\pi(u_t \mid x_{\le t})$ with positivity and appropriate Markov/sequential ignorability assumptions (Pearl, 2009; Hernán & Robins, 2020).

B.7 Safety-aware Inverse Design: Feasibility and Robustness

We formalize the safety constraints in inverse design using distributionally robust optimization. Let $\mathcal{P}$ be a divergence ball around the empirical distribution $\hat{p}$ defined by an f-divergence D_f or a Wasserstein metric.

Proposition C.6 (DRO feasibility and safe trust region). *If $D_f(p \,\|\, \hat{p}) \le \rho$ and the loss is L-Lipschitz in actions, then the worst-case expected deviation obeys $\sup_{p \in \mathcal{P}} \mathbb{E}_p[\ell(u)] \le \mathbb{E}_{\hat{p}}[\ell(u)] + c_f(\rho)$ with an explicit penalty $c_f(\rho)$ (Namkoong & Duchi, 2017; Delage & Ye, 2010). Enforcing $D(\pi_{beh}, u) \le \rho$ defines a trust region that guarantees feasibility under uncertainty sets.*

Proposition C.7 (Chance-constraint relaxation)**.** *For constraint $g(u) \le 0$ with random perturbations of bounded variance σ^2, Cantelli's inequality yields $\mathbb{P}(g(u) \le 0) \ge 1 - \alpha$ if $\mathbb{E}[g(u)] + \sqrt{\frac{1-\alpha}{\alpha}}\, \sigma \le 0$. Alternatively, enforcing $\text{CVaR}_{1-\alpha}(g(u)) \le 0$ provides a coherent and convex surrogate (Rockafellar & Uryasev, 2000).*

We instantiate D as KL, χ^2, or Wasserstein (Esfahani & Kuhn, 2018) with plug-in estimators. Explicitly: for KL divergence, $c_{\mathrm{KL}}(\rho) = \sqrt{2\rho \cdot \mathrm{Var}_{\hat{p}}[\ell(u)]}$ (Donsker–Varadhan); for χ^2 divergence, $c_{\chi^2}(\rho) = \sqrt{\rho} \cdot \mathrm{std}_{\hat{p}}[\ell(u)]$; for Wasserstein-1, $c_{W_1}(\rho) = \rho \cdot \mathrm{Lip}(\ell)$. We report feasibility certificates for proposed interventions using the tightest applicable bound.

B.8 Reproducibility and Statistical Protocols

We report ≥ 5 seeds for all key metrics with mean/STD/95% CIs (Student t or bootstrap), matched compute/time budgets across methods, leakage checks, and release raw vocabularies and training logs sufficient for independent verification. All tables in the main paper and appendices include seed counts and CI computation details.

D Appendix C: A Multiscale Architecture for Causal Biology

We detail the long-context sequence encoder (Mamba-Transformer hybrid), hypergraph dynamics for metabolic networks, cross-modal co-attention, training objectives for forecasting/counterfactual/policy synthesis, schemas, and evaluation protocols.

D.1 Motivating AI Capabilities (Detailed)

- **From Genes to Function Without Experimentation.** While current models predict protein structure from sequence (Jumper et al., 2021), our objective is to predict entire metabolic landscapes from genomic blueprints. Pre-training transcends naive masked language modeling: (1) **Operon-Aware Masking** compels prediction of functional units, not just individual genes (Zhou et al., 2023; Ji et al., 2021); (2) **Metabolite Diffusion** generates probable chemical fingerprints from genetic context using principles that revolutionized protein design (Watson et al., 2023; Corso et al., 2023); and (3) **Counterfactual Contrasts** encourage causal structure by learning which perturbations induce which metabolic shifts (Gal et al., 2017).
- **Biological Programming: Compiling Health States into Microbial Interventions.** The inverse problem—designing interventions to achieve specific biological outcomes—remains a core challenge in medicine. We frame microbiome engineering as an offline reinforcement learning problem (Levine et al., 2020). A **Decision Transformer** (Chen et al., 2021) learns from the 100,000+ perturbation trajectories to act as a biological compiler: given a target metabolic state, the model outputs a minimal genetic or chemical intervention predicted to achieve it. To mitigate out-of-distribution actions in offline RL, we incorporate uncertainty-aware regularization to ensure proposed interventions are biochemically plausible and lie within a trusted region of the learned policy.
- **Universal Perturbation Engine: Zero-Shot Prediction of Any Intervention.** A central goal is to develop a model that learns a general theory of biological perturbation (Huang et al., 2023). This involves moving beyond interpolating between observed cause-effect pairs to understanding the underlying principles governing how interventions propagate through metabolic networks. This capability would enable the prediction of effects from entirely novel compounds or genetic modifications, transforming therapeutic discovery from a stochastic to a deterministic process.

D.2 Architecture Rationale (Acknowledging Complexity)

Why Multiple Components? We acknowledge the "kitchen sink" appearance of combining Mamba, Transformer, and Hypergraph NNs. Each addresses a specific biological constraint:

- **Mamba:** $O(N)$ complexity for million-base sequences (Transformers' $O(N^2)$ is prohibitive)
- **Transformer:** Precise attention for regulatory motifs (Mamba lacks position-specific precision)
- **Hypergraph NN:** Many-to-many metabolic reactions (pairwise GNNs are fundamentally inadequate)

Integration Strategy: Rather than naive concatenation, we use:

1. **Hierarchical Processing:** Mamba processes full sequences – Transformer refines key regions
2. **Learned Gating:** Attention weights determine when to use which component
3. **Ablation Studies:** Each component improves performance by up to 8% (Table 7)

Table 7: Architecture ablation for Quorum-7B. Each row removes one component and retrains from scratch with matched compute budget. All differences are significant ($p < 0.05$, ≥ 5 seeds).

Configuration	Pathogen MCC	Profiling F1	Pathway wF1
Full Quorum-7B	**93.5**	**0.91**	**0.85**
w/o Mamba (Transformer only)	92.8 (*-0.7*)	0.88 (*-0.03*)	0.82 (*-0.03*)
w/o Transformer (Mamba only)	92.1 (*-1.4*)	0.89 (*-0.02*)	0.80 (*-0.05*)
w/o Hypergraph NN	93.2 (*-0.3*)	0.90 (*-0.01*)	0.78 (*-0.07*)
w/o Cross-modal attention	93.0 (*-0.5*)	0.90 (*-0.01*)	0.79 (*-0.06*)

Training Challenges: This architecture requires careful initialization, gradient clipping, and three-stage curriculum learning. Training instability is mitigated by periodic checkpoints every 1,000 steps.

- **The Million-Base Memory Problem:** Regulatory elements within a single bacterial genome can be separated by millions of bases, a scale that exceeds the quadratic attention horizon of standard Transformers. Our proposed solution integrates **Mamba's state-space models** (Gu & Dao, 2023; Nguyen et al., 2024) for O(N) scaling across whole-chromosome contexts with **surgical Transformer attention** for base-pair precision where required. This hierarchical approach mirrors the multi-scale organization of biological systems, from nucleotides to operons to regulons.
- **Beyond Pairwise Thinking:** Metabolic reactions are fundamentally combinatorial; a single enzyme complex might involve multiple cofactors and substrates to produce several products. Standard Graph Neural Networks (GNNs) are structurally inadequate for such relationships. Our **Hypergraph Neural Network** (Feng et al., 2019; Bai et al., 2021) natively represents these many-to-many interactions, providing the requisite mathematical framework to model complex biochemical pathways and population-level behaviors.
- **The Central Dogma Isn't Unidirectional:** The flow of biological information is not unidirectional from DNA to metabolite; feedback loops are common. Our **Cross-Modal Co-Attention** architecture (Lu et al., 2019; Tan & Bansal, 2019) is designed to learn these bidirectional relationships, enabling metabolomic signatures to query the genetic loci that produced them and, conversely, for genomic regions to predict their metabolic consequences.

E Appendix C′: Quorum-7B — Extended Benchmark Results

E.1 Training Data Composition

Quorum-7B is pre-trained on a multi-omic corpus processed through the sparsification + QA-Token pipeline:

- **Metagenomics:** 1.3 trillion base pairs from Illumina short-read environmental and clinical samples sourced from SRA. After sparsification (pattern `1111|1110`) and QA-tokenization, this yields ~315B quality-aware genomic tokens.
- **Metabolomics:** 500K metabolite profiles across diverse sample types, each with 5,000+ mass spectral features, tokenized into hierarchical metabolomic tokens via a modality-specific QA-Token variant.

E.2 Model Context

Existing frontier models occupy distinct niches in training data space:

- **Evo2** (Nguyen et al., 2025) (up to 40B parameters): Trained on assembled genomes, clean, single-organism, no environmental context. Excels at genome understanding but cannot process raw metagenomic samples or metabolomic data.
- **METAGENE-1** (Liu et al., 2025) (7B parameters): Trained on 1.5T bp of raw metagenomic reads with standard BPE tokenization. Strong on pathogen detection but lacks metabolomic reasoning and quality awareness.
- **Quorum-7B** (7B parameters): First model to combine multi-omic data with causal trajectories via the sparsify-then-tokenize pipeline. Achieves competitive or superior performance across benchmarks while being $18\times$ faster at inference, the only practical model at the performance frontier.

E.3 Inference Efficiency Analysis

The $18\times$ inference speedup and $10\times$ cost reduction arise from two complementary mechanisms: (1) sparsified tokenization produces longer, more informative tokens ($\sim$315B from 1.69T bp vs. $\sim$370B for BPE), enabling the same information content to be processed in fewer forward passes; (2) the Mamba–Transformer hybrid architecture achieves $O(N)$ scaling for the majority of sequence processing, with $O(N^2)$ attention reserved only for short regulatory-motif windows.

F Appendix D: Realistic Experimental Plan and Budget

F.1 AI-in-the-loop Experiments (Detailed MGED)

Our experimental strategy implements a continuously improving cycle where the AI model guides subsequent data generation. The foundation model, pre-trained on the 10T base-pair dataset, will perform millions of *in silico* simulations to identify physical experiments likely to yield maximal new biological insight. This is achieved through a principled **Model-Guided Experimental Design (MGED)** framework (Settles, 2009). To balance the exploration-exploitation trade-off, this framework will not only prioritize experiments that maximally reduce the model's epistemic uncertainty (Gal et al., 2017), but will also incorporate Thompson sampling to ensure the systematic exploration of the entire experimental space, preventing premature convergence to local optima. Our experimental platforms will then execute only the most informative experiments, and the resulting data will be used to refine the foundation model in an active learning loop.

Granular Execution Plan with Batch Effect Mitigation **Problem:** Inter-lab variation can exceed biological signal by 10-fold (pilot data: 35% of variance).

Solution Architecture:

1. **Standardization Hub ($5M):** - Central facility produces and ships standardized reagents (media, primers, standards) - Robotic liquid handlers programmed with identical protocols - Reference samples included in every batch (5% overhead)
2. **Hierarchical Experimental Design:** - Labs assigned to blocks; each lab runs complete factorial subsets - Overlap experiments (10%) enable cross-lab calibration - Statistical model: $Y_{ijk} = \mu + \text{Lab}_i + \text{Batch}_{ij} + \text{Treatment}_k + \epsilon$
3. **Real-time Quality Monitoring:** - Automated QC metrics computed within 4 hours of data generation - Labs failing QC thresholds ($> 2\sigma$ from reference) must re-run - Expected re-run rate: 15% (budgeted)
4. **Computational Harmonization:** - ComBat-seq for RNA-seq batch correction - COCONUT for metabolomics alignment - Deep variational autoencoders for joint embedding

Revised Budget: $40M experiments + $10M QC/harmonization = $50M total Phase 2.

- **Tier 1 (Screening):** Microbiome-on-Chip Arrays (Jalili-Firoozinezhad et al., 2019; Kim et al., 2012) will serve as our primary high-throughput platform, enabling the screening of thousands of microbial communities against thousands of perturbations to identify statistically significant interaction effects.
- **Tier 2 (Mechanistic Insight):** High-potential interactions from Tier 1 will be interrogated at higher resolution. This includes a targeted subset of $\sim$5,000 trajectories using our Single-Cell Metabolomics and Optogenetic Control platforms with high-frequency (5-minute) sampling to resolve fast-acting mechanistic dynamics.
- **Tier 3 (Pre-clinical Validation):** The most well-supported causal mechanisms will be validated in our Human Gut Simulators with Multi-Organ Feedback (Marzorati et al., 2014; Van de Wiele et al., 2015), providing the highest-fidelity *in vitro* model.

F.2 MGED Simulation Study: Regret and Empirical γ

We simulate nonlinear experimental settings to compare MGED greedy selection against Latin Hypercube Design. For each synthetic environment with Lipschitz MI surrogates, we report (i) cumulative regret relative to an oracle set, (ii) empirical submodularity ratio $\hat{\gamma}_k$ with bootstrap confidence intervals from restricted Hessian spectra, and (iii) final objective values and dispersion metrics. We enforce a fallback to Latin Hypercube when $\hat{\gamma}_k < \gamma_{\min}$ to guarantee coverage. Results include regret curves and $\hat{\gamma}_k$ trajectories for multiple seeds and model classes.

G Appendix E: The Scientific Program Enabled by CausalOmics-10T

The CausalOmics-10T dataset is designed to accelerate research toward predictive and quantitative microbiology by providing the scale, quality, and causal structure currently absent from public archives.

Predictive and Therapeutic Engineering

- **Forecasting Microbiome Dynamics:** Much like weather forecasting, we predict the trajectory of microbial ecosystems under different conditions. Use cases include recovery from antibiotic-induced dysbiosis, responses to dietary shifts, and engraftment success of live biotherapeutics.
- **Rational Design of Interventions:** Beyond trial-and-error, the *in silico* design of microbiome-based therapies enables novel treatments for chronic diseases like IBD (Frank et al., 2007; Schirmer et al., 2019) and climate-smart agriculture via microbial consortia that enhance nitrogen fixation and reduce fertilizer dependence (Paustian et al., 2016; Zomer et al., 2017; Smith, 2016).

Uncovering Fundamental Biological Principles Beyond immediate applications lies the ability to address foundational mysteries:

- **Illuminating Biology's "Dark Matter":** Just as AlphaFold illuminated protein structure, our models will systematically assign functions to the vast number of unannotated genes and metabolites discovered in sequencing surveys (Venter et al., 2004; Seaver et al., 2014). This moves beyond simple homology-based annotation to functional prediction based on deep biological context.
- **Elucidating Host-Microbe Interactions:** We will map the complex molecular dialogue between microbes and host cells. Our models will identify which microbes act to protect against disease, how they shape host immune repertoires, and the specific mechanisms—from secreted metabolites to cell-surface proteins—that govern these interactions.
- **Mapping Microbiome Biogeography:** We will uncover the design principles of microbial communities by mapping their spatial organization. The dataset will enable models to learn how spatial structure influences function and how these structures reconfigure in response to environmental change, a critical and underexplored dimension of microbial ecology.

- **Discovering Ecological Design Principles:** We will move from describing communities to discovering the fundamental rules that govern their assembly, stability, and resilience (Lawson et al., 2019; Coyte et al., 2015).

H APPENDIX F: ETHICS, DATA GOVERNANCE, AND RESPONSIBLE INNOVATION

Data Sovereignty and Consent. Human-derived microbiome samples require explicit informed consent addressing: (i) long-term storage, (ii) commercial use potential, (iii) data sharing protocols. We implement tiered consent allowing participants to control usage scope.

Privacy Protection. Microbiome data can reveal health status, diet, and location. We employ: (i) k-anonymity ($k \geq 5$) for metadata, (ii) differential privacy ($\epsilon = 1.0$) for aggregate statistics, (iii) secure multi-party computation for sensitive analyses.

DP composition and accounting. Repeated releases compound privacy loss. We adopt Rényi Differential Privacy (RDP) accounting for composition and conversion to (ϵ, δ)-DP (Mironov, 2017), and the moments accountant for tight bounds under subsampling (Abadi et al., 2016). For weekly releases, we publish the per-release privacy budget and cumulative (ϵ, δ) with confidence intervals. We also evaluate privacy amplification by subsampling for federated aggregation; composition over time uses standard DP boosting arguments (Dwork et al., 2010).

Benefit Sharing. Communities providing samples receive: (i) priority access to research findings, (ii) representation on governance board, (iii) 5% of commercial licensing revenue returned to source communities.

Environmental Impact. Computational footprint estimated at 500 MWh. We commit to: (i) carbon-neutral computing via renewable energy credits, (ii) efficient algorithms reducing energy by 3× vs. baseline, (iii) public carbon accounting.

I APPENDIX G: PILOT DATA DEMONSTRATING FEASIBILITY

I.1 G.1 END-TO-END DEMONSTRATION ON 10TB PILOT

We process 10TB of SRA Illumina short-read metagenomic data through the complete sparsification + QA-Token pipeline:

- **Input:** 10M reads from 25,000 diverse microbiome samples (gut, soil, ocean)
- **Sparsification:** Evaluated 224 pattern configurations; selected Pareto-optimal pattern (`1111 | 1110`) achieving F1=0.994 with 1.0× overhead
- **Quality Assessment:** Computed Phred scores, GC bias, adapter contamination (12 CPU-hours)
- **Tokenization:** Ran 5k merge steps with 100M proxy model (48 GPU-hours)
- **Validation:** Trained 500M model on QA-Token vs BPE vocabularies
- **Result:** 12% improvement in bits per base pair (95% CI: [10.3%, 13.7%])
- **Usable data expansion:** Combined pipeline lifts usable fraction from 5% to 40% (+35 pp, 8× data)

I.2 G.2 CAUSAL TRAJECTORY PILOT (100 EXPERIMENTS)

We generate 100 interventional trajectories to assess identifiability:

- **Design:** 2×2×5 factorial (2 species, 2 compounds, 5 doses), 12 timepoints
- **Causal Analysis:** - 23% meet the front-door criterion (metabolite mediators measured) - 31% have valid IVs (randomized timing) - 46% require sensitivity analysis (unmeasured confounding likely)

- **Cost:** \$2,100 per trajectory at pilot scale (10× higher than projected scale)
- **Key Learning:** Batch effects between labs contribute 35% of variance, requiring dedicated harmonization

I.3 G.3 Computational Scaling Analysis

Operation	1TB	1PB (proj.)	100PB (proj.)
Quality Scoring	12 CPU-hr	12k CPU-hr	1.2M CPU-hr
PMI Computation	8 GPU-hr	8k GPU-hr	800k GPU-hr
RL Training	48 GPU-hr	48k GPU-hr	4.8M GPU-hr
Total	68 hr	68k hr	6.8M hr

Table 8: Computational requirements scale super-linearly due to vocabulary growth

Throughput and energy assumptions. We assume GPU nodes with 350 TFLOPS BF16 effective throughput and 1.5 kW TDP, with parallel efficiency of 70% for PMI kernels and 60% for RL training due to communication overhead. For CPU quality scoring we assume 2.5 GHz cores at 15 W TDP. The 100 PB scenario thus draws roughly 4.8M GPU-hr $\times$ 1.5 kW $\approx$ 7.2 GWh (upper bound), amortized by in-storage compute (Mansouri Ghiasi et al., 2022; Ghiasi et al., 2022; 2024; Mansouri Ghiasi et al., 2023) that reduces IO by $\sim 8\times$, building on demonstrated speedups for genome sequence analysis. We schedule PMI statistics refresh every K merges (e.g., $K = 5\,000$) with an incremental update strategy that re-computes only affected local co-occurrence counts, yielding $\sim$10% overhead over the base RL loop.

Robustness to quality miscalibration. We run platform-specific calibration analyses for ONT and NGS, reporting pre/post calibration curves and the induced changes in variant calling F1, taxonomic accuracy F1, and reconstruction loss. Miscalibration is simulated via affine and sigmoid warps of Phred-derived features and corrected via isotonic regression using reference controls; Pareto frontier shifts are also quantified.

J Appendix H: Why Now — Convergence, Timing, and Readiness

This proposal is timely because it stands at the confluence of four trends that have sparked the AI revolution: the development of powerful deep learning algorithms, the availability of specialized hardware (GPUs), the creation of open-source software ecosystems, and access to massive datasets. We leverage this convergence to solve the three core challenges that have held back AI in biology: (1) **The Scale Problem**, which we solve by creating an unprecedentedly large dataset; (2) **The Quality Problem**, which we solve with our QA-Token framework; and (3) **The Causality Problem**, which we address with 100,000+ targeted perturbation experiments.

CausalOmics-10T operationalizes this convergence through a data flywheel: reclaiming the vast majority of existing public data, compressing it into semantically meaningful tokens, pre-training a foundation model, and using that model to guide wet-lab experiments that generate causal signal with unprecedented efficiency. This is a blueprint for foundational predictive models of microbial ecosystems, where a \$50M investment delivers the equivalent of a \$1B+ untargeted dataset.

K Appendix I: Making the Long Tail Usable — Foundation-Scale Evidence

Problem statement. Let $\mathcal{D}_{\text{raw}}$ denote a corpus with heterogeneous per-base/per-measurement quality distributions that violate i.i.d. assumptions and render standard frequency-only tokenization unstable. Define the *usable subset* for a tokenizer $\mathcal{Z}$ as the set of inputs for which the induced token sequence has bounded cross-entropy under a fixed proxy model: $\mathcal{U}(\mathcal{Z}) = \{x \in \mathcal{D}_{\text{raw}} : \mathcal{L}_{\text{proxy}}(\text{tok}_{\mathcal{Z}}(x)) \leq \tau\}$. QA-Token expands $\mathcal{U}(\mathcal{Z})$ by incorporating quality-aware scoring and MDL-regularized merges (Eqs. 6–7).

Formal claim (informal). Under calibrated quality signals and stationary noise, the QA-Token merge policy that maximizes expected reward strictly increases the measure of usable data, $|\mathcal{U}(\mathcal{Z}_{\mathrm{QA}})| \geq |\mathcal{U}(\mathcal{Z}_{\mathrm{BPE}})|$, for any fixed threshold τ, with strict inequality when the raw corpus contains non-negligible regions of high-noise segments. Sketch: Decompose proxy loss into quality-weighted mutual information and code-length penalties; QA-Token merges down-weight low-quality contributions and preferentially form tokens aligned to reliable structure, shifting sequences below the loss threshold. See App. C.

Foundation-model evidence at scale. We re-tokenize the 1.5 trillion bp METAGENE-1 (Liu et al., 2025) corpus using QA-BPE-seq (vocab size 1,024; identical training protocol) and retrain the 7B model. The resulting foundation model achieves a new state-of-the-art on Pathogen Detection (MCC 92.96→94.53; see Table 2 in the main text) and superior macro performance on the GUE benchmark (Table 9).

Table 9: Genome Understanding Evaluation (GUE): macro-averaged performance and per-task slices (MCC unless noted).

Task	CNN	HyenaDNA	DNABERT	NT-2.5B-Multi	DNABERT-2	METAGENE-1	METAGENE-1 (QA-Token)
TF-Mouse (AVG.)	45.3	51.0	57.7	67.0	68.0	**71.4**	**72.8**
TF-HUMAN (AVG.)	50.7	56.0	64.4	62.6	**70.1**	68.3	69.9
EMP (AVG.)	37.6	44.9	49.5	58.1	56.0	66.0	**67.5**
SSD	76.8	72.7	84.1	89.3	85.0	87.8	**89.5**
COVID (F1)	22.2	23.3	62.2	73.0	71.9	72.5	**73.3**
Global Win %	0.0	0.0	7.1	21.4	25.0	46.4	**57.1**

Threshold sensitivity analysis. We set $\tau = 4.0$ nats/token as the threshold at which a 500M proxy model achieves $>90\%$ of its peak downstream task performance, validated on three held-out tasks (pathogen detection, taxonomic profiling, metabolic pathway prediction). The usable fraction as a function of τ is: at τ=3.5, usable fraction is 28% (BPE: 3%); at τ=4.0, 40% (BPE: 5%); at τ=4.5, 52% (BPE: 8%). The $8\times$ QA-Token expansion factor is robust across thresholds, ranging from $7.2\times$ to $9.3\times$, confirming that the improvement is not an artifact of a particular τ choice.

Compression and information retention. With identical vocabulary size, QA-Token yields ∼315B tokens from 1.69T bp vs ∼370B for standard BPE, indicating longer, functionally coherent genomic constructs. Let L_{code} denote the description length under the learned lexicon; QA-Token minimizes $\mathbb{E}[L_{\text{code}}]$ subject to quality-weighted fidelity, improving both compression and downstream loss, consistent with Eq. (7).

L Appendix J: Optimizer-Agnostic QA-Token — Noisy Text and RL Modularity

L.1 Noisy Social Media Text (TweetEval)

Table 10: TweetEval comparison on noisy social media text: QA-Token improves robustness across tasks.

Model	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	ALL(TE)
BERTweet	33.4	79.3	56.4	82.1	79.5	73.4	71.2	67.9
SuperBPE + BERTweet	33.6	79.8	56.8	82.3	80.1	73.9	71.8	68.3
QA-BPE-nlp + BERTweet	**33.8**	**81.1**	**58.2**	**82.5**	**82.6**	**74.5**	**73.1**	**69.4**

L.2 Ablations on RL Algorithm Choice

Claim. Let π_ϕ denote the policy class used to select merges. For any optimizer $\mathcal{A}$ that monotonically improves the expected reward $\mathbb{E}[R]$ (Eq. 7) under unbiased gradient estimates, the induced vocabulary has equivalent asymptotic optimality up to optimizer-dependent convergence rates. Empirically (Table 11), PPO, GRPO, VAPO, and DAPO produce near-identical vocabularies (Jaccard ≥ 0.95) and downstream performance, confirming modularity.

Table 11: RL optimizer ablation across domains: similar performance, training/inference cost, and high vocabulary Jaccard vs PPO.

Configuration	Metric Value	Training Time (GPU-h)	Inference Time (ms/seq)	Vocab. Jaccard (vs PPO)
Genomics (QA-BPE-seq) — Variant F1				
QA-Token (PPO)	**0.891**	34.0	10.2	0.99
QA-Token (GRPO)	0.890	32.5	10.3	0.98
QA-Token (VAPO)	0.892	31.8	10.2	0.97
QA-Token (DAPO)	0.889	34.2	10.4	0.98
Finance (QAT-QF) — Sharpe Ratio				
QA-Token (PPO)	**1.72**	28.0	15.2	0.99
QA-Token (GRPO)	1.71	26.5	15.3	0.96
QA-Token (VAPO)	1.73	25.0	15.1	0.95
QA-Token (DAPO)	1.70	28.5	15.2	0.96
Social Media (QA-BPE-nlp) — TweetEval Sentiment				
QA-Token (PPO)	**74.5**	30.0	12.5	0.99
QA-Token (GRPO)	74.2	29.0	12.6	0.97
QA-Token (VAPO)	74.6	28.0	12.5	0.98
QA-Token (DAPO)	74.3	31.0	12.7	0.97

L.3 Ablation of Reward Components

To address the concern that the QA-Token reward function is an over-engineered heuristic, we perform an ablation study on the METAGENE-1 re-training task. We systematically remove each of the four components from the reward function (Eq. 7) and rebuild the vocabulary from scratch, keeping all other aspects of model training identical. Table 12 shows the impact on the downstream Pathogen Detection benchmark. The results confirm that while the proxy loss ($\Delta\mathcal{L}_{\text{proxy}}$) is the most critical component, the quality (Q), information-theoretic (PMI), and complexity (MDL) terms all provide significant, complementary contributions to the final model's performance. This supports our multi-objective design.

Table 12: Ablation study of QA-Token reward components on METAGENE-1 Pathogen Detection (MCC).

Reward Configuration	Pathogen-Detect MCC
Full QA-Token Reward	**94.53**
Ablations:	
w/o Quality ($-\lambda_Q Q$)	93.12 (*-1.41*)
w/o PMI ($-\lambda_I$PMI)	93.89 (*-0.64*)
w/o MDL ($+\lambda_C$MDL)	94.01 (*-0.52*)
w/o Proxy Loss ($-\lambda_D \Delta\mathcal{L}$)	91.55 (*-2.98*)
Standard BPE (Baseline)	92.96