MetaOmics-10T: The Foundational Dataset to Unlock Causal Modeling of Microbial Ecosystems

Arvid E. Gollwitzer*
Broad Institute of MIT and Harvard
Cambridge, MA, USA
arvidg@mit.edu

Deepak A. Subramanian
Dept. of Chemical Engineering, MIT
Koch Institute for Integrative Cancer Research, MIT
Broad Institute of MIT and Harvard
Cambridge, MA, USA

Isaac Tucker
Broad Institute of MIT and Harvard
Cambridge, MA, USA

Giovanni Traverso*
Dept. of Mechanical Engineering, MIT, Cambridge, MA, USA
Div. of Gastroenterology, Hepatology and Endoscopy,
Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA
Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA, USA
Broad Institute of MIT and Harvard, Cambridge, MA, USA
cgt20@mit.edu

Abstract

We propose **MetaOmics-10T**—an openly shareable, foundational dataset to unlock AI-accelerated discovery in microbial ecosystems. The dataset directly enables three high-impact AI tasks: (1) forecasting ecosystem dynamics, (2) predicting counterfactual outcomes of interventions, and (3) inverse-design of microbial therapies under safety constraints. MetaOmics-10T combines **10 trillion base pairs** reclaimed from public archives using a Quality-Aware Tokenization (QA-Token) framework with **100,000+ interventional trajectories** generated via model-guided experimental design. The result is a first-of-its-kind, probabilistic, intervention-ready corpus that addresses the principal bottleneck for causal modeling in microbiome science and provides an empirical testbed to assess the reach and limits of causal inference at scale.

1 From Observation to Intervention: The Formal Contract for Digital Twins

Proposal at a Glance. *AI task:* forecasting, counterfactual prediction, and safe inverse design. *Rationale:* lack of interventional, quality-aware, multi-omic time series is the core bottleneck for causal modeling. *Dataset:* 10T bp reclaimed from archives + 100,000+ interventional trajectories with full metadata (protocols, doses, timings, quality). *Shareability:* weekly open releases with standardized schemas and ontologies. *Impact:* enables identifiable digital twins, robust counterfactuals, and principled policy synthesis. *Feasibility:* detailed cost, throughput, and experimental SOPs in Appendix D. **Digital Twin Definition.** 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 ($g_t \in \mathbb{R}^{n_g}$,

*Corresponding authors: arvidg@mit.edu, cgt20@mit.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Reach and Limits of AI for Scientific Discovery.

$n_g \approx 10^6$) 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 parameterized by deep neural networks, and $\mathcal{M} : \mathcal{S} \to \mathcal{Y}$ the measurement map accounting for technical noise. To avoid symbol collisions later with model classes, we denote any *frozen proxy model* used for evaluation by $\mathcal{F}$, never by $\mathcal{M}$. The three core tasks with formal specifications:

- **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 $u_t = 0$.
- **Counterfactuals:** Estimate $p(x_{t+\tau}|do(u), x_{\leq t})$ via backdoor adjustment when confounders Z are measured.
- **Inverse design:** Solve constrained optimization $u^* = \arg\min_{u \in \mathcal{U}} C(u) + \lambda\, d(\mathbb{E}[x_{t+\tau}|do(u), x_t], x^*)$ subject to safety constraints $g(u) \leq 0$ and an uncertainty-aware trust region $D(\pi_{\text{beh}}, u) \leq \rho$ (with D a divergence, e.g., Wasserstein or KL, ensuring safe extrapolation from observed actions), together with a chance constraint $\mathbb{P}(g(u) \leq 0) \geq 1 - \alpha$ under model uncertainty.

Learnability vs. Causality: Acknowledging the Abyss. Appendix B 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/mixing under π. Causal identifiability, which is even stronger, requires assumptions about latent confounding (App. B.6). Therefore, this 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 methods like IVs and front-door adjustment succeed, and when sensitivity analysis is necessary.

2 From Noisy Archives to Causal Signal: The Missing Substrate

Current Data Catastrophe. Microbiome data today resembles astronomy before telescopes—fragmented, noisy, and causally impoverished. Public archives contain 100+ petabases of sequences, yet 95% is unusable due to: (1) Heterogeneous quality making standard tokenization fail, (2) Missing metadata preventing biological interpretation, (3) Zero causal structure—only observational snapshots. This creates an insurmountable barrier: current models achieve <60% accuracy on basic tasks like pathogen detection, while transformative applications require >95%.

MetaOmics-10T Breaks Through via Two Innovations:

- **Quality-Aware Tokenization (QA-Token):** A reinforcement learning framework that incorporates Phred quality directly into vocabulary construction, expanding usable data by 15%. Formal MDP and guarantees in App. A.
- **100,000+ Causal Trajectories:** Systematic perturbation–response records (CRISPR/compounds; 12-point time series) enabling counterfactual inference.

Data Specifications: (1) *Scale*: 10T base pairs (1000× larger than current datasets), 10M samples across environments; (2) *Resolution*: Single-nucleotide genomics, 5000+ metabolite features, 5-minute temporal sampling; (3) *Metadata*: Complete experimental conditions, intervention specifications, quality metrics—structured via formal ontologies (ENVO, NCBITaxon, CHEBI).

3 Crossing the Scaling Wall: Causal Foundation Models for Biology

Current State. Microbiome AI is stuck in the pre-ImageNet era—starved of quality and causality. MetaOmics-10T changes this.

Immediate Model Development Acceleration:

- **10× Larger Models:** Enables 10B+ parameter foundation models with emergent capabilities.
- **Causal Reasoning:** First dataset for counterfactuals—predicting intervention outcomes, not just observations.
- **Generalization:** Train once; transfer across organisms, environments, and interventions.

Why Now? Convergence of scalable long-context models, identifiable causal ML, automated experimentation at 100k scale, and urgent biomedical/climate needs.

4 From Archive to Causality: A Two-Phase, Open Pipeline

Phase 1 (Months 1–12; $10M). Mine 100+ PB across SRA/ENA/RefSeq with QA-Token. **Realistic computational cost:** The full tokenization pipeline is a significant undertaking, projected to require ∼6.8 million core-hours (App. G.3), dominated by quality scoring (CPU-bound) and RL-based vocabulary training (GPU-bound). This is made feasible by leveraging massively parallel in-storage processing [1–4] to eliminate data movement costs, combined with fast metagenomic classification pipelines [5, 6]. **Phase 2 (Months 13–36; $40M).** Run **100k+ MGED-selected** perturbation trajectories across a distributed network of labs using standardized protocols (Microbiome-on-Chip screens; single-cell metabolomics/optogenetics for mechanism; gut simulators for validation). Throughput/cost models and SOPs in App. D.

Feasibility and openness. Total $50M yields a dataset otherwise costing $5B; QA-Token lifts usable fraction from 5% to 20% (+15 pp, 4× data); open schemas/code/models with staged weekly releases (Appendix, App. D).

Acknowledgments

The authors thank all members of the Traverso Lab for the scholarly environment they provide.

Arvid E. Gollwitzer especially thanks Dr. Deepak A. Subramanian and Prof. Giovanni Traverso for their unwavering support and mentorship.

Competing Interests

G.T. has or currently receives equity/stock/royalties/gifts or board/advisor/consulting roles from Exact Sciences, Horizon, Pavoda, Entrega, CBSET, Avaxia, Lyndra, Novo Nordisk, SNS Nano, Hoffman la Roche, Janssen, Egalet, Synlogic, Suono Bio, Merck, Verily, Eagle Pharmaceuticals, Vivtex, Celero Systems, Bilayer Therapeutics, Teal Bio, Wired Consulting, Avadel Pharmaceuticals, Moderna, Syntis Bio, Vitakey, Absco Therapeutics, GEM-Bioscience, Bill and Melinda Gates Foundation, JHU technology transfer office, MIT technology licensing office and the MGB technology licensing office.

A.E.G. is a co-founder of Anto Biosciences (YC F25).

D.A.S. and I.T. declare no competing interests.

References

[1] 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*, pages 635–654, 2022.

[2] 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: In-storage filtering of genomic data for high-performance and energy-efficient genome analysis. In *2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 283–287. IEEE, 2022.

[3] 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)*, pages 660–677. IEEE, 2024.

[4] 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 preprint arXiv:2311.12527*, 2023.

[5] 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*, 2023.

[6] 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*, 2023.

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

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

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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*, pages 2023–01, 2023.

[16] 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*, pages 1644–1650, 2020.

[17] 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.

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

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

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

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

[22] 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.

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

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

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

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

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

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

[41] 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.

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

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

[44] 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.

[45] 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.

[46] 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.

[47] 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*, pages 305–317, 2015.

[48] 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.

[49] 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.

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

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

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

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

[58] 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*, pages 308–318, 2016.

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

[60] 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.

[61] 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*, pages arXiv–2311, 2023.

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 [7] and information-theoretic approaches to sequence analysis [8, 9]. 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:

$$\text{(i) } \max_V \mathbb{E}_{x\sim\mathcal{C}}[\sum_i q_i \log p(x_i|V)] \quad \text{(quality-weighted likelihood)} \tag{1}$$

$$\text{(ii) } \max_V I(V;\mathcal{C}) \quad \text{(mutual information)} \tag{2}$$

$$\text{(iii) } \min_V |V| \quad \text{(compression via MDL)} \tag{3}$$

$$\text{(iv) } \min_V \mathcal{L}_{\text{proxy}}(V) \quad \text{(downstream task performance)} \tag{4}$$

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:

$$Q(t) = f_{\theta_Q}(\mathbf{v}_q, \mathbf{v}_p, \mathbf{v}_b) = \sigma(W_2 \cdot \text{ReLU}(W_1[\mathbf{v}_q; \mathbf{v}_p; \mathbf{v}_b] + b_1) + b_2) \tag{5}$$

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 Derivation. The reward function emerges from maximizing expected log-likelihood under quality-weighted data distribution:

$$R(a,b) = \mathbb{E}_{\mathcal{C}}[\log p(\mathcal{C}|V \cup \{t_{ab}\})] - \mathbb{E}_{\mathcal{C}}[\log p(\mathcal{C}|V)] + \text{regularizers} \tag{6}$$

$$= \underbrace{\lambda_Q Q(ab)}_{\text{quality prior}} + \underbrace{\lambda_I \text{PMI}(a,b)}_{\text{mutual information}} - \underbrace{\lambda_C \text{MDL}(a,b)}_{\text{description length}} + \underbrace{\lambda_D \Delta\mathcal{L}_{\text{proxy}}}_{\text{generalization estimate}} \tag{7}$$

Each term has theoretical justification: PMI measures statistical dependency [8], MDL provides compression-generalization bounds [9], and proxy loss estimates downstream performance. 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 $|V_{\text{max}}| = 50$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 [10, 11], 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-trained the 7B-parameter METAGENE-1 foundation model [12] on its original 1.5 trillion base pair dataset, replacing the standard BPE tokenizer with QA-Token. As shown in Table 1, this change improved performance on the Pathogen Detection benchmark. On the systems side, we align with high-throughput genomics pipelines and in-storage computing advances [1, 3]. 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.

Table 1: Results on the Pathogen Detection benchmark, comparing the original METAGENE-1 [12] with a version re-trained using QA-Token. QA-Token enables a new state-of-the-art foundation model (Metric: MCC).

Task	**DNABERT-2** [13]	**DNABERT-S** [14]	**NT-2.5b-Multi** [15]	**NT-2.5b-1000g** [15]	**METAGENE-1** [12]	**METAGENE-1 (QA-Token)**
Pathogen-Detect (avg.)	87.92	87.02	82.43	79.02	92.96	**94.53**

Performance on Noisy Text Data. We compared QA-Token against other tokenizers on the noisy TweetEval benchmark [16]. As shown in Table 2, QA-Token achieves higher performance on this dataset, indicating its ability to build robust representations from noisy text.

Table 2: 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 3. Evaluation follows rigorous benchmarking standards for genomic sequence comparison [17]. In all evaluated cases, QA-Token improves performance over quality-blind baselines.

Table 3: 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%

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 MetaOmics-10T dataset a feasible endeavor.

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 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 $\pi(u_t \mid x_{\leq t})$. 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.

B.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 B.1 (Linear Identifiability)**.** *Under controllability, observability, persistent excitation, and Gaussian noise, parameters (A, B, C, Q_w, R_v) are identifiable up to similarity transforms.*

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 B.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, $\left\|\mathbb{E}[f(x_t, y_t) \mid x_0] - \mathbb{E}[f(x_t, y_t)]\right\| \leq C\,\rho^t$. 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 B.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 B.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)* Generic injectivity up to symmetries*: If (θ, η) and (θ', η') produce identical finite-dimensional distributions over $(y_{0:T}, u_{0:T-1})$ for all T under π, then they are related by a reparameterization in a known equivalence class (e.g., neuron permutations, similarity transforms of latent realizations). Then (θ, η) is locally identifiable modulo this equivalence class.*

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 B.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 [18–20].*

Sketch. The observability rank condition ensures local injectivity of the output map with respect to parameters along excited trajectories; geometric mixing under π yields ergodicity to replace time averages by expectations; analyticity plus compactness excludes pathological flat directions beyond $\mathcal{G}$. Standard arguments for partially observed state-space models transfer to the deep parameterization by smoothness, establishing nonsingularity of the Fisher information transverse to symmetry orbits.

B.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 B.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 B.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* $\gamma_k = \inf_{L\subseteq\mathcal{X},\,|L|\leq k}\ \inf_{S\subseteq\mathcal{X}\setminus L} \frac{\sum_{a\in L}\big(F(S\cup\{a\}) - F(S)\big)}{F(S\cup L) - F(S)}$.

Theorem B.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* $F(S_G) \geq \big(1 - e^{-\gamma_k}\big)\max_{|S|\leq k} F(S)$*. 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 [21]. We estimate γ_k via restricted eigenvalue bounds of the Fisher information or Gauss–Newton approximation and default to Latin Hypercube Design [22] 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.

B.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.

Proposition B.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)*,* $\big|\mathrm{PMI}_\Sigma^{(K)}(a,b) - \mathrm{PMI}_\Sigma^{(0)}(a,b)\big| \leq C_L\,\alpha_K$*, for a constant* C_L *depending only on local context length. 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 B.3 (Positivity of MDL improvement)**.** *Under KT coding and fixed $\kappa \geq 0$, $\Delta\mathrm{MDL}(a,b) < 0$ 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.

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

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

B.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 B.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.

Theorem B.5 (Curriculum surrogate monotonicity with trust region)**.** *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 $\|\lambda_{k+1} - \lambda_k\| \leq \epsilon$ with optimization error non-increasing, then $\tilde{J}(\lambda_{k+1}) \geq \tilde{J}(\lambda_k) - L_\lambda\, \epsilon$. Choosing $\epsilon \leq \varepsilon / L_\lambda$ guarantees non-increasing surrogate degradation by at most ε per step.*

Theorem B.6 (Gumbel–Softmax gradient bias)**.** *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 Lipschitz losses and bounded logits there exist constants C_1, C_2 such that $\big\| \mathbb{E}[\widehat{\nabla J}_\tau] - \nabla J \big\| \leq C_1\, \tau, \quad \mathrm{Var}(\widehat{\nabla J}_\tau) \leq C_2/M$. As $\tau \to 0$ and $M \to \infty$, the estimator is asymptotically unbiased. We verify trends using a variance-reduced REINFORCE control variate (e.g., REBAR/RELAX) [10, 11].*

B.5 B.5 Comprehensive Treatment of Model Limitations

B.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.

B.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.

B.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

B.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.

B.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 $x_{\leq t}$, 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:

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

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 [23]. 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 $\pi(u_t \mid x_{\leq t})$ with positivity and appropriate Markov/sequential ignorability assumptions [24, 25].

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 B.5 (DRO feasibility and safe trust region)**.** *If $D_f(p \,\|\, \hat{p}) \leq \rho$ and the loss is L-Lipschitz in actions, then the worst-case expected deviation obeys $\sup_{p \in \mathcal{P}} \mathbb{E}_p[\ell(u)] \leq \mathbb{E}_{\hat{p}}[\ell(u)] + c_f(\rho)$ with an explicit penalty $c_f(\rho)$ [26, 27]. Enforcing $D(\pi_{beh}, u) \leq \rho$ defines a trust region that guarantees feasibility under uncertainty sets.*

Proposition B.6 (Chance-constraint relaxation)**.** *For constraint $g(u) \leq 0$ with random perturbations of bounded variance σ^2, Cantelli's inequality yields $\mathbb{P}(g(u) \leq 0) \geq 1 - \alpha$ if $\mathbb{E}[g(u)] + \sqrt{\frac{1-\alpha}{\alpha}}\,\sigma \leq 0$. Alternatively, enforcing $\mathrm{CVaR}_{1-\alpha}(g(u)) \leq 0$ provides a coherent and convex surrogate [28].*

We instantiate D as KL, χ^2, or Wasserstein [29] with plug-in estimators, and report feasibility certificates for proposed interventions.

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.

C 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.

C.1 Motivating AI Capabilities (Detailed)

- **From Genes to Function Without Experimentation.** While current models predict protein structure from sequence [30], 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 [13, 14]; (2) **Metabolite Diffusion** generates probable chemical fingerprints from genetic context using principles that revolutionized protein design [31, 32]; and (3) **Counterfactual Contrasts** encourage causal structure by learning which perturbations induce which metabolic shifts [33].
- **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 [34]. A **Decision Transformer** [35] 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 [36]. 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.

C.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 5-8% (Table in App. C)

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** [37, 38] 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** [39, 40] 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 [41, 42] 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.

D Appendix D: Realistic Experimental Plan and Budget

D.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 [43]. To balance the exploration-exploitation trade-off, this framework will not only prioritize experiments that maximally reduce the model's epistemic uncertainty [33], 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 [44, 45] 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 ∼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 [46, 47], providing the highest-fidelity *in vitro* model.

D.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.

E Appendix E: The Scientific Program Enabled by MetaOmics-10T

The MetaOmics-10T dataset will not just accelerate existing research; it will enable a new paradigm of inquiry, transforming biology into a truly predictive and engineering discipline.

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 [48, 49] and climate-smart agriculture via microbial consortia that enhance nitrogen fixation and reduce fertilizer dependence [50–52].

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 [53, 54]. 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 [55, 56].

F 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 [57], and the moments accountant for tight bounds under subsampling [58]. 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 [59].

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 $3\times$ vs. baseline, (iii) public carbon accounting.

G Appendix G: Pilot Data Demonstrating Feasibility

G.1 G.1 End-to-End Demonstration on 1TB Subset

We processed 1TB of SRA data (0.001% of target) through complete QA-Token pipeline:

- **Input:** 10M reads from 1000 diverse microbiome samples (gut, soil, ocean)

- **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 held-out perplexity (95% CI: [10.3%, 13.7%])

G.2 G.2 Causal Trajectory Pilot (100 Experiments)

Generated 100 interventional trajectories to assess identifiability:

- **Design:** 2×2×5 factorial (2 species, 2 compounds, 5 doses), 12 timepoints
- **Causal Analysis:** - 23% met front-door criterion (metabolite mediators measured) - 31% had valid IVs (randomized timing) - 46% required sensitivity analysis (unmeasured confounding likely)
- **Cost:** $2,100 per trajectory at pilot scale (10× higher than projected scale)
- **Key Learning:** Batch effects between labs contributed 35% of variance—requires dedicated harmonization

G.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 4: 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 [1, 60, 3, 61] 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.

H 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.

MetaOmics-10T is not merely a dataset; it is a blueprint for building **foundational predictive models of the unseen biological worlds** that shape our own. It is an engine for a new paradigm of AI-driven, automated scientific discovery.

I 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. 5–7).

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, $|\mathcal{U}(\mathcal{Z}_{\text{QA}})| \geq |\mathcal{U}(\mathcal{Z}_{\text{BPE}})|$, 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. B.

Foundation-model evidence at scale. We re-tokenized the 1.5 trillion bp METAGENE-1 [12] corpus using QA-BPE-seq (vocab size 1,024; identical training protocol) and retrained the 7B model. The resulting foundation model achieves a new state-of-the-art on Pathogen Detection (Table 5) and superior macro performance on the GUE benchmark (Table 6).

Table 5: Pathogen Detection benchmark: METAGENE-1 with standard BPE vs QA-Token (MCC).

Task	DNABERT-2	DNABERT-S	NT-2.5b-Multi	NT-2.5b-1000g	METAGENE-1	METAGENE-1 (QA-Token)
Pathogen-Detect (avg.)	87.92	87.02	82.43	79.02	92.96	**94.53**

Table 6: 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**

Compression and information retention. With identical vocabulary size, QA-Token yielded $\sim$315B tokens from 1.69T bp vs $\sim$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).

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

J.1 Noisy Social Media Text (TweetEval)

Table 7: 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**

J.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 8), PPO, GRPO, VAPO, and DAPO produce near-identical vocabularies (Jaccard ≥ 0.95) and downstream performance, confirming modularity.

Table 8: 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

J.3 Ablation of Reward Components

To address the concern that the QA-Token reward function is an over-engineered heuristic, we performed an ablation study on the METAGENE-1 re-training task. We systematically removed each of the four components from the reward function (Eq. 7) and rebuilt the vocabulary from scratch, keeping all other aspects of model training identical. Table 9 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 9: 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