The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling

Andre Cornman[*,1] Jacob West-Roberts[1] Antonio Pedro Camargo[2] Simon Roux[2]
Martin Beracochea[3] Milot Mirdita[4] Sergey Ovchinnikov[5] Yunha Hwang[*,1]

[1]Tatta Bio, USA
[2]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[3]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI),
Wellcome Genome Campus, Hinxton, Cambridge, UK
[4]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea
[5]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
[*]Correspondence: {yunha,andre}@tatta.bio

Abstract

Biological language model performance depends heavily on pretraining data quality, diversity, and size. While metagenomic datasets feature enormous biological diversity, their utilization as pretraining data has been limited due to challenges in data accessibility, quality filtering and deduplication. Here, we present the Open MetaGenomic (OMG) corpus, a genomic pretraining dataset totalling 3.1T base pairs and 3.3B protein coding sequences, obtained by combining two largest metagenomic dataset repositories (JGI's IMG and EMBL's MGnify). We first document the composition of the dataset and describe the quality filtering steps taken to remove poor quality data. We make the OMG corpus available as a mixed-modality genomic sequence dataset that represents multi-gene encoding genomic sequences with translated amino acids for protein coding sequences, and nucleic acids for intergenic sequences. We train the first mixed-modality genomic language model (gLM2) that leverages genomic context information to learn robust functional representations and coevolutionary signals in protein-protein interfaces. Furthermore, we show that deduplication in embedding space can be used to balance the corpus, demonstrating improved performance on downstream tasks. The OMG dataset is publicly hosted on the Hugging Face Hub at https://huggingface.co/datasets/tattabio/OMG and gLM2 is available at https://huggingface.co/tattabio/gLM2_650M.

1 Introduction

Biological language models present an effective avenue for leveraging large amounts of unstructured sequence data and learn functionally meaningful representations. Similar to natural language processing (NLP) models ([47]; [11]), the quality and diversity of pretraining data dictate the behavior and performance of biological language models ([10]). To date, the most widely used datasets for biological language models ([13]; [23]; [25]; [28]) are derived from curated data repositories such as UniProt ([49]), UniRef ([45]) and GTDB ([30]). However, biological sequence diversity is immense and the above-mentioned data repositories cover only a small fraction of the full sequence diversity found in nature. In order for biological language models to improve, the size and diversity of pretraining data must also scale with the size of the model.

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

Metagenomic sequences are partial genomic sequences derived from direct sequencing of environmental (e.g. soil, ocean) or biological samples (e.g. human skin, gut). Because metagenomic sequencing circumvents the need for cultivation and isolation of biological organisms, metagenomes typically feature sequences derived from uncultivated and novel microorganisms ([48]), encoding high levels of molecular diversity ([15]). To date, metagenomic sequences have not been fully utilized in biological language models due to following limitations:

1. **Metagenomic sequences are not readily downloadable in a single archive.** To date, the download of raw contigs (assembled genomic segments) from the two main public repositories, Joint Genome Institute (JGI)'s IMG ([26]) and European Molecular Biological Laboratory (EMBL)'s MGnify ([35]), requires a large number of database queries and/or rate-limited web API calls.

2. **Metagenomic sequences require extensive pre-processing.** Raw metagenomically assembled contigs first undergo gene calling in order to identify protein coding sequences and extract translated sequences. Additional quality filtering is critical, as many metagenomes include poor or mis-assembled contigs.

3. **Metagenomic sequences are difficult to deduplicate and balance.** Like most biological sequence datasets, metagenomes feature biases. Unlike protein databases that can be deduplicated and balanced using computationally efficient clustering algorithms (e.g. MMseqs2 ([42])), clustering of a large dataset comprising genomic sequences of arbitrary region and length is computationally costly.

Here, we document the collection and preprocessing steps of the OpenMetaGenome (OMG) corpus. We then train the first mixed-modality genomic language model (gLM2) trained on OMG, that leverages genomic context information to learn contextualized functional representations of genomic elements. By training on mixed-modality data, gLM2 can perform both protein and DNA downstream tasks, and outperforms ESM2 ([23]) on most protein tasks. Additionally, training on multi-protein contexts enables gLM2 to predict protein-protein interfaces through co-evolutionary signal. Finally, we show that embedding-based deduplication of the OMG dataset leads to improved functional representations, especially for underrepresented sequences.

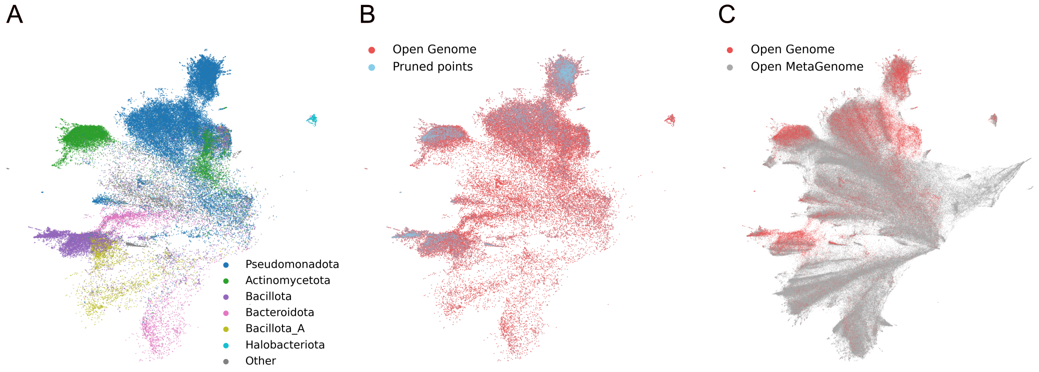


Figure 1: **(A)** UMAP visualization of the OG dataset examples, colored by taxonomic phylum, using embeddings from the 150M parameter gLM2 model. Distinct clusters form for different phyla in embedding space. **(B)** Semantic deduplication of the OG dataset, with pruned points highlighted in blue. Deduplication primarily removes samples from dense clusters corresponding to overrepresented phyla. We visualize the semantic deduplication on OG dataset to highlight taxonomic phyla most heavily pruned, and apply the same pruning process to the OMG dataset for model training. **(C)** Comparison of the OG and OMG datasets using a random 0.1% subset of each. Notably, the metagenomic data (OMG) exhibits higher diversity.

2 The Open MetaGenome corpus

Here, we document the construction of the OMG corpus. The OMG is a 3.1T base pair (bp) pretraining dataset comprising EMBL's MGnify database[1] and JGI's IMG database[2]. We utilize the gene predictions conducted by the databases; the gene calling protocols for IMG and MGnify are detailed in ([16]) and ([35]) respectively. The combined dataset is pre-processed into a mixed-modality dataset (described in G.1) upon sequential element-by-element quality-filtering (described in G.4 and G.5). The mixed-modality dataset of Open Metagenomes is made available as the OMG dataset (Fig. 1) containing 3.3 billion protein coding sequences (CDS) (Appendix. C). We also make available a 10x smaller subset of OMG that only consists of prokaryotic and viral genomes from INSDC[3] as the Open Genome mixed-modality dataset OG (Appendix C). Finally, we make available a protein-only dataset OMG_prot50, consisting of protein sequences derived from the OMG dataset, clustered at 50% sequence identity (Appendix F). OMG_prot50 contains 207M representative sequences from clusters with at least two members, representing >3-fold increase in sequence diversity compared to UniRef50 ([45]). All three datasets are available for download from the Hugging Face Hub as Hugging Face datasets and all dataset processing scripts are available at `https://github.com/TattaBio/OMG`. As more metagenomic data becomes available, we plan on regular updated releases of the corpus in the future.

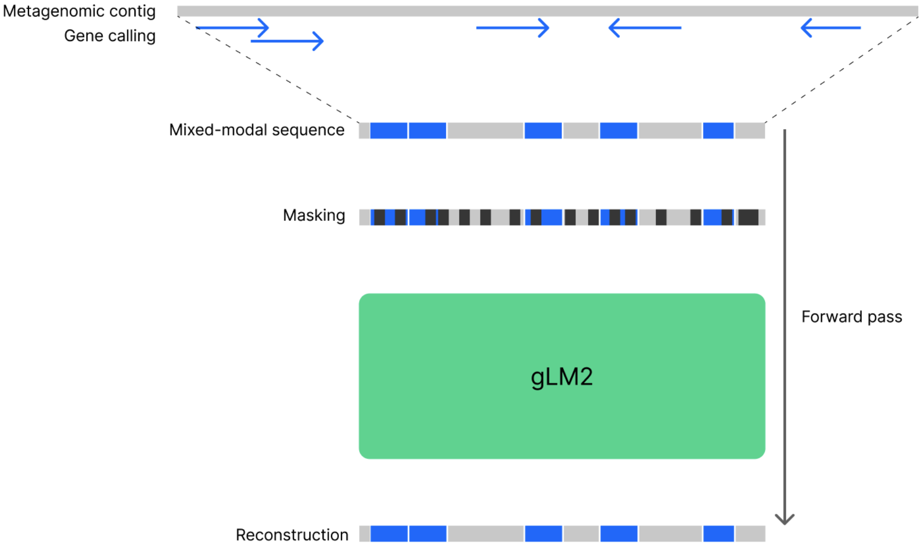


Figure 2: **Mixed-modality sequence processing and gLM2 masked language modeling.** A gene-called metagenomic contig is first preprocessed into a mixed-modality sequence consisting of CDS elements (blue) and IGS elements (grey). The mixed-modality sequence then undergoes masking at 30% and gLM2 is trained with a masked token reconstruction objective.

3 Experiments

3.1 GLM2: A Mixed-modality genomic language model

To showcase the efficacy of the OMG dataset for pretraining, we introduce gLM2: a mixed-modality genomic language model pretrained on OMG. gLM2 learns contextualized representations of genomic contigs, which are represented as sequences of CDS and IGS elements. In order to tokenize the mixed-modality sequence, CDS elements are tokenized using per-amino acid tokens, and IGS elements are

[1] Snapshot date 2022-11-23 (excluding all embargoed/restricted metagenomic samples, see database statistics in Appendix A)

[2] Snapshot date 2023-08-27 (excluding all embargoed/restricted metagenomic samples and including IMG genomes dataset derived from NCBI.)

[3] `https://www.insdc.org`, retrieved from IMG/M, metadata available in Appendix P

tokenized using byte-pair encoding (BPE) ([37]) of the nucleotide sequences, with a vocabulary size of 4,096. To distinguish strand orientation for CDS elements, we introduce two special tokens: <+> and <->, which are prepended to each genomic element to indicate the positive and negative strands, respectively. gLM2 is trained using the masked language modeling objective, where 30% of both CDS and IGS tokens are masked. Cross-entropy loss is applied only on the masked tokens. gLM2 is trained at two scales: 150M and 650M parameters. Both models are trained on the semantically deduplicated OMG dataset (Section 3.2) for 600k steps. We train gLM2 using a context window of 2048 tokens to allow for multiple (6.7 ± 2.7) CDS and IGS elements to appear in each example. For model architecture and training hyperparameters, refer to Appendix H.

3.2 OMG corpus balancing with genomic Semantic Deduplication

Biological datasets exhibit significant biases that can influence the performance and generalizability of trained models ([10]; [52]). Unlike protein databases, where short sequence lengths allow for clustering-based deduplication, (meta)genomic sequences have highly variable lengths (**??**B), making sequence-based clustering challenging. To address this challenge, we perform deduplication in embedding space by pruning examples with small cosine distance, following Semantic Deduplication (SemDeDup) ([1]). SemDeDup previously showed efficacy in removing semantically similar examples over web-scale text and image datasets, demonstrating significant speed up in convergence for downstream tasks. For genomic semantic deduplication, we first trained a 150M gLM2 on the tokenized OMG dataset for 600k steps. We then embed the entire OMG dataset, by extracting a mean-pooled, per-example representation from the model's last hidden layer. The example-level embeddings correspond closely to the taxonomic classification available for the OG dataset (Fig. 1A). We prune the OMG dataset at 42% (i.e. 42% of the original data is removed) at the deduplication threshold 1e-3 (where examples with embeddings <1e-3 in cosine distance are deduplicated) (Appendix I). The pruned examples are saturated in highly dense clusters (Fig. 1B) which results in taxonomic balancing (Appendix J) , measured by increased entropies of distribution across taxonomic levels (Appendix K). We then trained a 150M gLM2 on the pruned OMG dataset for an equal number of steps, and compared its performance against the un-pruned version on DGEB ([52]). While pruning results in only a slight increase in the aggregate DGEB score (0.49 vs 0.48), we observe consistent improvements in tasks that feature underrepresented taxa (e.g. ArchRetrieval, BacArch BiGene mining, RpoB Arch phylogeny) (Appendix L). This improved performance for underrepresented taxa appears to come at the cost of small regressions on tasks that are biased towards overrepresented taxa. Genomic SemDeDup presents a tunable method for effectively pruning unstructured genomic data without reliance on taxonomic labels.

3.3 GLM2 performance on DGEB

We compare the performance of the 150M and 650M gLM2 models trained on the pruned OMG dataset against the ESM2 series trained on the UniRef50/D dataset (Fig. 3). Overall, we observe more efficient scaling on DGEB amino acid (AA) tasks for gLM2 compared to the ESM2 series. In particular, gLM2's performance scales with pretraining floating point operations (FLOPs) on protein tasks where ESM2 plateaus in performance with scaling (i.e. Operon pair classification tasks, ModAC paralogy task) (Appendix M). Such improved functional representation learning is likely due to gLM2 ability to leverage genomic context information, and thereby learn relationships between genomic elements. gLM2, being a mixed-modality model, also learns intergenic sequence representations. We compare gLM2's performance on DGEB nucleic acid (NA) tasks against the Nucleotide Transformer series (Appendix N). gLM2 performs similarly on NA tasks when compared to Nucleotide Transformers, despite only a small fraction of the training tokens consisting of DNA sequences.

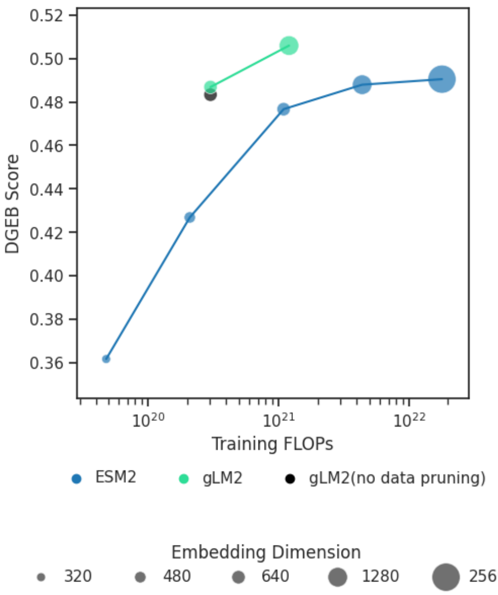


Figure 3: **Scaling performance on DGEB amino acid tasks for gLM2 and ESM2, relative to pretraining floating point operations (FLOPs).** gLM2_150M trained with no data pruning is shown in black.

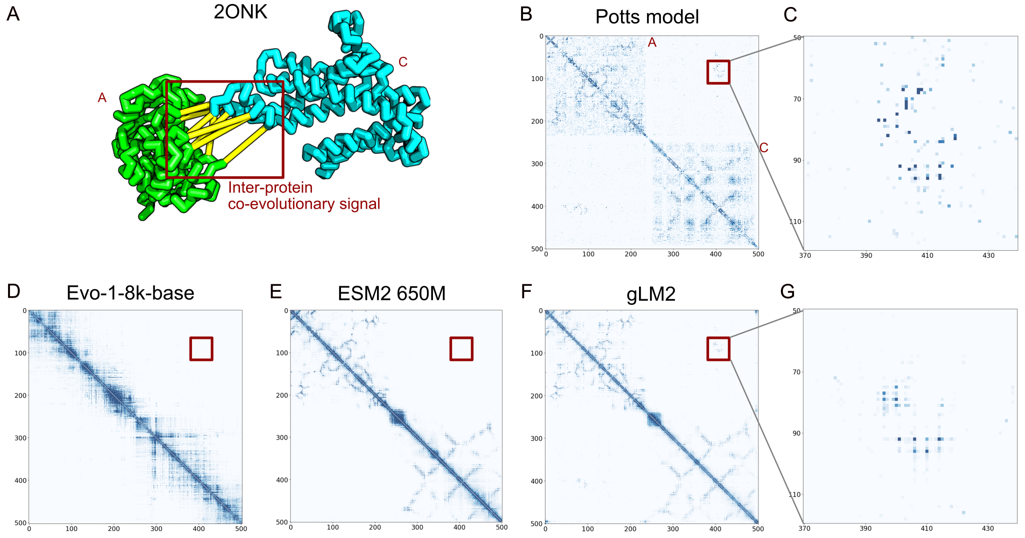


Figure 4: **gLM2 learns protein-protein interface co-evolutionary signal in the 2ONK (ModAC) complex.** **(A)** ModA and ModC forms a structural complex with co-evolutionary signal between residues (in yellow). **(B)** Co-evolutionary signal extracted from multiple sequence alignment of 2ONK, calculated and visualized using GREMLIN (`PDB_benchmark_alignments/2ONK_A2ONK_C.fas`). The region of inter-protein co-evolutionary signals are highlighted with a red box. **(C)** Zoomed-in region of inter-protein coevolutionary signal in B. **(D)** Categorical Jacobian calculated for Evo on the DNA sequence encoding 2ONK_A and 2ONK_C (from 89,891 to 91,376 of genomic sequence NC_000917.1). The L2 norm was computed over the (3,4,3,4) tensor for every pair of codon positions to generate the contact map. **(E)** Categorical Jacobian calculated for ESM2 650M on the concatenated 2ONK_A_2ONK_C sequence. No inter-protein co-evolutionary signal is detected. **(F)** Categorical Jacobian calculated for gLM2_650M on the concatenated 2ONK_A_2ONK_C sequence. **(G)** Zoomed-in region of inter-protein coevolutionary signal in G.

3.4 GLM2 learns protein-protein interaction interfaces

We test gLM2's ability to learn coevolutionary signals between proteins in protein-protein interaction interfaces ([29]). Previous studies have shown that pLMs learn within-protein co-evolutionary information that can be extracted with a supervised contact prediction head ([23]) using an unsupervised "categorical Jacobian" calculation ([55]). However, pLMs trained on individual proteins or protein families cannot learn co-evolutionary information across proteins. We calculate the categorical jacobian values from gLM2_650M on the concatenated sequence of 2ONK_A (ModA) and 2ONK_C (ModC) (Appendix O). We demonstrate that gLM2 leverages multi-protein context to learn protein-protein interfaces from a single concatenated sequence that closely matches the co-evolutionary signal that can be learned from multiple sequence alignment (MSA) based Potts model (GREMLIN ([18])) (Fig. 4). Such protein-protein interface signals cannot be extracted in existing language model methods such as ESM2 650M and Evo-1-8k-base (Fig. 4E and F). The ability to extract interacting residues without supervision nor MSA presents an opportunity to predict novel protein-protein interactions from sequence information alone.

4 Acknowledgements

The OMG dataset is derived from metagenomes deposited at JGI's IMG and EBI's MGnify. We first acknowledge the teams that maintain these public data repositories. Secondly, we acknowledge the work done by individual research labs who have collected and processed the original samples. We list the DOIs associated with the individual datasets/projects in Appendix P. We also thank Amro Abbas for engaging in helpful discussions regarding the SemDeDup methodology.

5 Data and model availability

The three datasets introduced in this study are publicly hosted on the Hugging Face Hub at `https://huggingface.co/datasets/tattabio/OMG` and the two gLM2 models are available at `https://huggingface.co/tattabio/gLM2_150M` and `https://huggingface.co/tattabio/gLM2_650M`. We make the data preprocessing and corpus generation code available at `https://github.com/TattaBio/OMG`. We make model inference and categorical jacobian script available at `https://github.com/TattaBio/gLM2`.

6 Competing interest

The authors declare no competing interest.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. March 2023.

[2] Mohammad Bahram, Tarquin Netherway, Clémence Frioux, Pamela Ferretti, Luis Pedro Coelho, Stefan Geisen, Peer Bork, and Falk Hildebrand. Metagenomic assessment of the global diversity and distribution of bacteria and fungi. *Environ. Microbiol.*, 23(1):316–326, January 2021.

[3] Adair L Borges, Yue Clare Lou, Rohan Sachdeva, Basem Al-Shayeb, Petar I Penev, Alexander L Jaffe, Shufei Lei, Joanne M Santini, and Jillian F Banfield. Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. *Nat Microbiol*, 7(6):918–927, June 2022.

[4] Tomas Bruna, Alexandre Lomsadze, and Mark Borodovsky. A new gene finding tool GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *bioRxiv*, April 2024.

[5] Antonio Pedro Camargo, Lee Call, Simon Roux, Stephen Nayfach, Marcel Huntemann, Krishnaveni Palaniappan, Anna Ratner, Ken Chu, Supratim Mukherjeep, T B K Reddy, I-Min A Chen, Natalia N Ivanova, Emiley A Eloe-Fadrosh, Tanja Woyke, David A Baltrus, Salvador Castañeda-Barba, Fernando de la Cruz, Barbara E Funnell, James P J Hall, Aindrila Mukhopadhyay, Eduardo P C Rocha, Thibault Stalder, Eva Top, and Nikos C Kyrpides. IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata. *Nucleic Acids Res.*, 52(D1):D164–D173, January 2024.

[6] Antonio Pedro Camargo, Stephen Nayfach, I-Min A Chen, Krishnaveni Palaniappan, Anna Ratner, Ken Chu, Stephan J Ritter, T B K Reddy, Supratim Mukherjee, Frederik Schulz, Lee Call, Russell Y Neches, Tanja Woyke, Natalia N Ivanova, Emiley A Eloe-Fadrosh, Nikos C Kyrpides, and Simon Roux. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Research*, 51(D1):D733–D743, 11 2022.

[7] Ryan Cook, Andrea Telatin, George Bouras, Antonio Pedro Camargo, Martin Larralde, Robert A Edwards, and Evelien M Adriaenssens. Driving through stop signs: predicting stop codon reassignment improves functional annotation of bacteriophages. *ISME Commun*, 4(1):ycae079, January 2024.

[8] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. September 2023.

[9] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023.

[10] Frances Ding and Jacob Steinhardt. Protein language models are biased by unequal sequence sampling across the tree of life. March 2024.

[11] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. April 2021.

[12] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through Self-Supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):7112–7127, October 2022.

[13] Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raúl Santiago Molina, Neil Thomas, Yousuf A Khan, Chetan Mishra, Carolyn Kim, Liam J Bartie, Matthew Nemeth, Patrick D Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. July 2024.

[14] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. December 2017.

[15] Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A Relman, Kari M Finstad, Ronald Amundson, Brian C Thomas, and Jillian F Banfield. A new view of the tree of life. *Nat Microbiol*, 1:16048, April 2016.

[16] Marcel Huntemann, Natalia N Ivanova, Konstantinos Mavromatis, H James Tripp, David Paez-Espino, Kristin Tennessen, Krishnaveni Palaniappan, Ernest Szeto, Manoj Pillay, I-Min A Chen, Amrita Pati, Torben Nielsen, Victor M Markowitz, and Nikos C Kyrpides. The standard operating procedure of the DOE-JGI metagenome annotation pipeline (MAP v.4). *Stand. Genomic Sci.*, 11:17, February 2016.

[17] Yunha Hwang, Andre L Cornman, Elizabeth H Kellogg, Sergey Ovchinnikov, and Peter R Girguis. Genomic language model predicts protein co-regulation and function. *Nat. Commun.*, 15(1):2880, April 2024.

[18] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, 2013.

[19] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. January 2020.

[20] Senying Lai, Shaojun Pan, Chuqing Sun, Luis Pedro Coelho, Wei-Hua Chen, and Xing-Ming Zhao. metaMIC: reference-free misassembly identification and correction of de novo metagenomic assemblies. *Genome Biol.*, 23(1):242, November 2022.

[21] Alla L Lapidus and Anton I Korobeynikov. Metagenomic data assembly - the way of decoding unknown microorganisms. *Front. Microbiol.*, 12:613791, March 2021.

[22] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[23] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[25] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, James S Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.*, 41(8):1099–1106, August 2023.

[26] Victor M Markowitz, I-Min A Chen, Krishna Palaniappan, Ken Chu, Ernest Szeto, Yuri Grechkin, Anna Ratner, Biju Jacob, Jinghua Huang, Peter Williams, Marcel Huntemann, Iain Anderson, Konstantinos Mavromatis, Natalia N Ivanova, and Nikos C Kyrpides. IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.*, 40(Database issue):D115–22, January 2012.

[27] Daniel R Mende, Alison S Waller, Shinichi Sunagawa, Aino I Järvelin, Michelle M Chan, Manimozhiyan Arumugam, Jeroen Raes, and Peer Bork. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One*, 7(2):e31386, February 2012.

[28] Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, Stefano Ermon, Stephen A Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D Hsu, and Brian L Hie. Sequence modeling and design from molecular to genome scale with evo. March 2024.

[29] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife*, 3:e02030, May 2014.

[30] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, 50(D1):D785–D794, January 2022.

[31] Donovan H Parks, Fabio Rigato, Patricia Vera-Wolf, Lutz Krause, Philip Hugenholtz, Gene W Tyson, and David L A Wood. Evaluation of the microba community profiler for taxonomic profiling of metagenomic datasets from the human gut microbiome. *Front. Microbiol.*, 12:643682, April 2021.

[32] Georgios A Pavlopoulos, Fotis A Baltoumas, Sirui Liu, Oguz Selvitopi, Antonio Pedro Camargo, Stephen Nayfach, Ariful Azad, Simon Roux, Lee Call, Natalia N Ivanova, I Min Chen, David Paez-Espino, Evangelos Karatzas, Ioannis Iliopoulos, Konstantinos Konstantinidis, James M Tiedje, Jennifer Pett-Ridge, David Baker, Axel Visel, Christos A Ouzounis, Sergey Ovchinnikov, Aydin Buluç, and Nikos C Kyrpides. Unraveling the functional dark matter through global metagenomics. *Nature*, 622(7983):594–602, October 2023.

[33] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. June 2024.

[34] Kim D Pruitt, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M Farrell, Jennifer Hart, Melissa J Landrum, Kelly M McGarvey, Michael R Murphy, Nuala A O'Leary, Shashikant Pujar, Bhanu Rajput, Sanjida H Rangwala, Lillian D Riddick, Andrei Shkeda, Hanzhen Sun, Pamela Tamez, Raymond E Tully, Craig Wallin, David Webb, Janet Weber, Wendy Wu, Michael DiCuccio, Paul Kitts, Donna R Maglott, Terence D Murphy, and James M Ostell. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, 42(Database issue):D756–63, January 2014.

[35] Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, Tom Curtis, Alejandra Escobar-Zepeda, Tatiana A Gurbich, Varsha Kale, Anton Korobeynikov, Shriya Raj, Alexander B Rogers, Ekaterina Sakharova, Santiago Sanchez, Darren J Wilkinson, and Robert D Finn. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.*, 51(D1):D753–D759, January 2023.

[36] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.*, 118(15), April 2021.

[37] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. August 2015.

[38] Noam Shazeer. Glu variants improve transformer, 2020.

[39] Ben Sorscher, Robert Geirhos, Shashank Shekhar, S Ganguli, and Ari S Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Adv. Neural Inf. Process. Syst.*, abs/2206.14486, June 2022.

[40] Martin Steinegger, Milot Mirdita, and Johannes Söding. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods*, 16(7):603–606, July 2019.

[41] Martin Steinegger and Steven L Salzberg. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.*, 21(1):115, May 2020.

[42] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, 35(11):1026–1028, November 2017.

[43] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nat. Commun.*, 9(1):2542, June 2018.

[44] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.

[45] Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, May 2007.

[46] Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S Morcos. D4: Improving LLM pretraining via document De-Duplication and diversification. *Adv. Neural Inf. Process. Syst.*, abs/2308.12284, August 2023.

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. February 2023.

[48] Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, March 2004.

[49] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47(D1):D506–D515, January 2019.

[50] John Vollmers, Sandra Wiegand, and Anne-Kristin Kaster. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLoS One*, 12(1):e0169662, January 2017.

[51] Patrick T West, Alexander J Probst, Igor V Grigoriev, Brian C Thomas, and Jillian F Banfield. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.*, 28(4):569–580, April 2018.

[52] Jacob West-Roberts, Joshua Kravitz, Nishant Jha, Andre Cornman, and Yunha Hwang. Diverse genomic embedding benchmark for functional evaluation across the tree of life. July 2024.

[53] Jacob West-Roberts, Luis Valentin-Alvarado, Susan Mullen, Rohan Sachdeva, Justin Smith, Laura A Hug, Daniel S Gregoire, Wentso Liu, Tzu-Yu Lin, Gabriel Husain, Yuki Amano, Lynn Ly, and Jillian F Banfield. Giant genes are rare but implicated in cell wall degradation by predatory bacteria. November 2023.

[54] Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019.

[55] Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Matteo Dal Peraro, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. January 2024.

Appendix A Data sources

	Type	Snapshot date	# Samples	# contigs*	Total bps	# CDS
IMG	Metagenomes	2023-08-27	36,273	182M	1.70T	1.84B
	Genomes	2023-08-27	131,744	6.2M	0.4T	0.4B
MGnify	Metagenomes	2022-11-23	33,531	82M	1.03T	1.03B

*Number of contigs after filtering and preprocessing.

Appendix B Related Works

B.1 Pretraining corpora preprocessing in NLP

A number of previous studies have developed methods to improve the diversity and quality of pretraining corpora in NLP. For instance, raw snapshots of Common Crawl (collection of webtext crawls) contain undesirable data (e.g. hate speech, placeholder text). Studies have demonstrated that careful deduplication and rule-based filtering of Common Crawl ([11]) improves overall model performance ([33]). More recently, efforts have been made to prune and balance pre-training data in semantic embedding space to achieve increased training efficiency ([39]; [46]; [1]). Dataset preprocessing presents an important opportunity to minimize training resources, given the power-law nature of LLM scaling (i.e. exponentially increasing compute requirement for diminishing returns in performance improvement) ([14]; [19]).

B.2 Biological sequence language models and their training datasets

Biological sequence language models are self-supervised models trained on discrete protein sequences or genomic segments. Protein language models (pLMs) ([23]; [25]; [12]) are typically trained on high quality and curated publicly available datasets such as UniRef ([45]). UniRef is convenient for pLM training because it has been deduplicated using sequence similarity-based clustering (i.e. UniRef50 is deduplicated using 50% sequence identity). Previous efforts to increase the diversity of the pretraining data includes cluster-balanced sampling (e.g. UniRef50/D for ESM models ([36])) and sequence identity-based clustering of compiled protein databases beyond curated databases (e.g. BFD ([40])) ([12]). Genomic language models (gLMs) are trained on genomic sequences chunked at predefined length thresholds. Diversification efforts for genomic datasets include pretraining on MGnify's metagenomic contigs ([17]) and balancing efforts in genomic pretraining datasets include taxonomy-aware sampling ([8]; [28]) of curated genomic databases such as RefSeq ([34]), IMG/VR ([6]), IMG/PR ([5]) and GTDB ([30]).

B.3 Metagenomic datasets

In this study, we define metagenomic datasets as collections of genomic contigs (contiguous genomic segments) computationally assembled from either short-read or long-read raw sequence libraries. Typically, metagenomic datasets are sequenced from mixed community samples, which consist of multiple species, ranging from hundreds to thousands of distinct species ([2]). Complete genomes are rarely obtained from metagenomic assemblies. Therefore, metagenomic assemblies require extensive taxonomic profiling ([31]) and partial genome reconstruction through contig clustering (i.e. binning). Because metagenomes are sequenced from diverse environments without the need for cultivation, their sequences feature the highest level of molecular diversity amongst publicly available sequence datasets ([32]). Metagenomic datasets also vary in quality depending on sequencing depth and sample type, where low quality metagenomes feature computational assembly errors, short contig lengths, and truncated protein sequences ([27]; [20]). Furthermore, while most metagenomic datasets are predominantly analyzed with a focus on microbial (archaea, bacteria, viruses) communities, eukaryotic genomic material can comprise a substantial portion of the raw library ([51]). Many standard metagenomic post-processing steps (e.g. gene calling) fail on eukaryotic sequences, resulting in poor quality protein sequence predictions. Critically, quality filtering and dataset deduplication of metagenomes require domain-specific knowledge, yet there is little documentation of preprocessing steps needed to make these datasets suitable for biological language model pretraining.

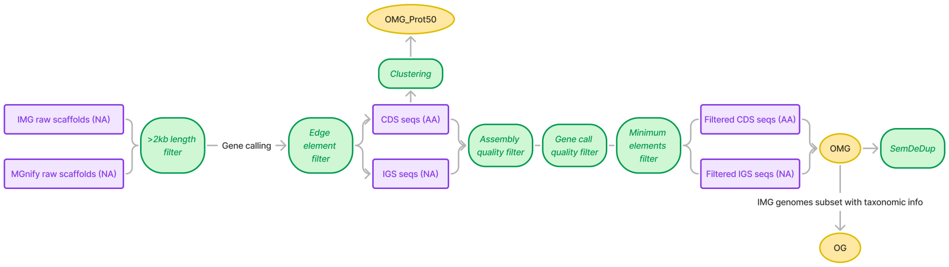
Appendix C OMG dataset statistics

Statistics for the datasets made available in this study. CDS: Coding sequences, IGS: Intergenic sequences. For reference, UniRef50 consists of 66M proteins.

	# CDS	# IGS	Total (bps)	# Contig	Size (TB)	Description
OMG	3.3B	2.8B	3.1T	271M	1.25	Filtered mixed-modality genomic sequences featuring multiple protein coding genes (represented in AAs) interleaved with intergenic sequences (represented in NAs).
OG	0.4B	0.3B	0.4T	6.2M	0.16	Fraction of the IMG data that consist of prokaryotic genomes and associated taxonomic metadata.
OMG_prot50	207M	–	–	–	0.05	Protein coding sequences in AA, clustered at 50% sequence identity. Singleton clusters were removed from the database. Clustering detail is found in Appendix B

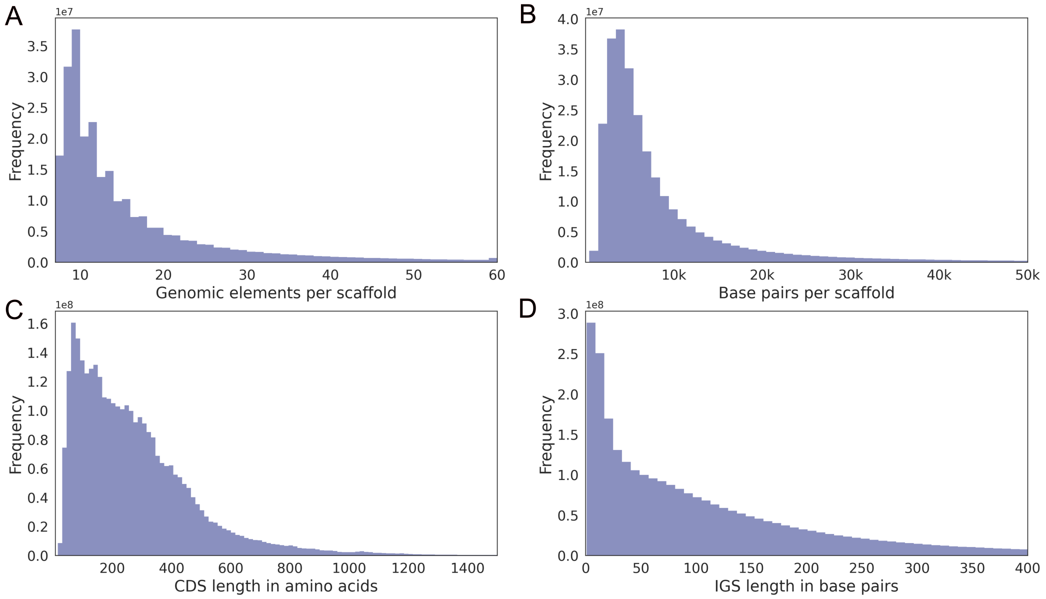
Appendix D OMG preprocessing flow chart

Sequences (purple) undergo filtering steps (green), yielding three Hugging Face datasets (yellow) made available with this paper. ‘NA’ and ‘AA’ refer to nucleic acid and amino acid data modalities respectively.



Appendix E Length distributions of the OMG corpus.

(A) Distribution of contig lengths in the number of genomic elements (CDS and IGS). **(B)** Distribution of contig lengths in base pairs. **(C)** Distribution of CDS lengths in amino acids. **(D)** Distribution of IGS lengths in base pairs.



Appendix F OMG_prot50 clustering method

A total of 4.2B protein sequences were first clustered to remove fragments using MMseqs2 linclust ([43]) (commit f6c98, parameters:–min-seq-id 0.9 -c 0.9 –cov-mode 1). Subsequently, the resulting sequences were clustered at 50% sequence id and 90% sequence coverage using MMseqs2 `linclust -min-seq-id 0.5 -c 0.9`. Singleton clusters (only one sequence in the cluster across the full dataset) were removed and remaining 207M cluster representatives were uploaded as the Hugging Face dataset.

Appendix G OMG processing methods

G.1 Multi-modal data processing

Metagenomic contigs often encode multiple genes on either strand of the sequence. A genomic language model can be trained on raw nucleic acid sequences (e.g. Evo ([28]), Nucleotide Transformers ([8])) or by representing each genomic sequence as an order- and orientation-preserved list of translated coding sequences in amino acids (e.g. ([17])). For the former method, the context length needed to encode genomic sequences in nucleic acids can result in unfeasibly large compute requirements. Furthermore, a recent study comparing nucleic acid (NA) models against amino acid (AA) models on protein functional representations demonstrated that NA may not be the most efficient input format for learning translated protein functions ([52]). The latter method, while benefiting from the compressed sequence length and more expressive AA sequences for proteins, does not leverage the information stored in intergenic regions. These intergenic regions contain important, yet, lesser characterized sequence patterns involved in transcription regulation and cellular function such as ncRNA, microRNA, promoters, and transcription factor binding sites. We developed a mixed-modality dataset that represents a genomic contig as a list of elements where an element is either a coding sequence (CDS) or an intergenic sequence (IGS) (see Fig. 2). CDS elements are represented in translated AA sequences and IGS elements are represented in NA sequences. We also store the strand information (+/-) of CDS elements and the order of all elements in the contig.

G.2 Edge-element removal

Metagenomic contigs are not complete genomic sequences, therefore, both edges of the sequences are more likely to contain gene-calling errors. In our pre-processing, we remove edge CDS elements to address miscalled open reading frames (ORFs) and fragmented protein sequences at the beginning and end of the metagenomic contigs ([41]). Specifically, if a scaffold starts/ends with an interrupted CDS, we remove that CDS element. If a scaffold starts/ends with a non-coding region, we remove the IGS element and the CDS adjacent to the IGS element.

G.3 Contig length-based filtering and preprocessing

Assembly of shotgun metagenomic libraries results in many short contigs that are often low in quality. To limit the impact of the fragmented nature of metagenome assemblies, we first remove all metagenomic contigs that are shorter than 2kb from the raw databases. Secondly, we enrich the corpus with contigs that contain multiple genes by removing contigs that contain less than seven elements in total or less than three CDS elements. In preprocessing these contigs into Hugging Face datasets ([22]), we found that extremely large contigs resulted in process hanging errors and inefficient storage. To address this issue, we chunk large contigs into 1000 elements. Appendix E visualizes the distributions of contig lengths in number of elements (Appendix E A) and in base pairs (Appendix E B) in the OMG corpus. We also document the distributions of element sequence lengths of CDS (Appendix E C) and IGS (Appendix E D) elements.

G.4 Assembly quality (N/X-frequency) filtering

Due to the computational nature of the metagenomic assembly, misassembled contigs comprise a nontrivial fraction of the data. The quality of the assembly differs significantly across samples, depending on the biological community composition, sample type, and sequencing depth ([50]; [21]). Notably, the quality of assembly may vary across the contig, where a section of the contig may contain assembly gaps due to shallow sequencing depth. One way to determine poorly assembled sequences is by identifying the fraction of Ns (gaps or ambiguous bases) in the raw DNA sequence (or Xs in the translated AA sequence). For OMG, we process each contig sequentially element-by-element, and if an element comprises >20% in invalid characters, we discard the element and start a new contig. Importantly, only contigs that meet the length requirement (>3 CDS, >7 total elements) are added to the dataset. This sequential processing allows high quality regions of the contigs to be preserved, while low quality stretches are discarded.

G.5 Element length-based filtering

A nontrivial portion of the metagenome can be eukaryotic, however, most metagenomic gene-calling software tools are not optimized for eukaryotic ORF prediction ([4]). Additionally, metagenomes can contain sequences from organisms that utilize alternative genetic codes ([3]; [7]), which may not all be correctly predicted by common tools. A salient pattern observed for poor gene prediction is low coding density, (i.e. long stretches of IGS) or presence of very long CDS sequences. To identify these, we process each contig sequentially element-by-element and remove any CDS element >15,000 AAs or IGS element >4000 bps in length, and start a new contig, as described in Section G.4. These thresholds are designed to exclude regions of questionable gene calls, such as long intergenic regions where no genes are predicted, and giant protein sequences, which are prone to assembly errors and require careful curation to verify ([53]).

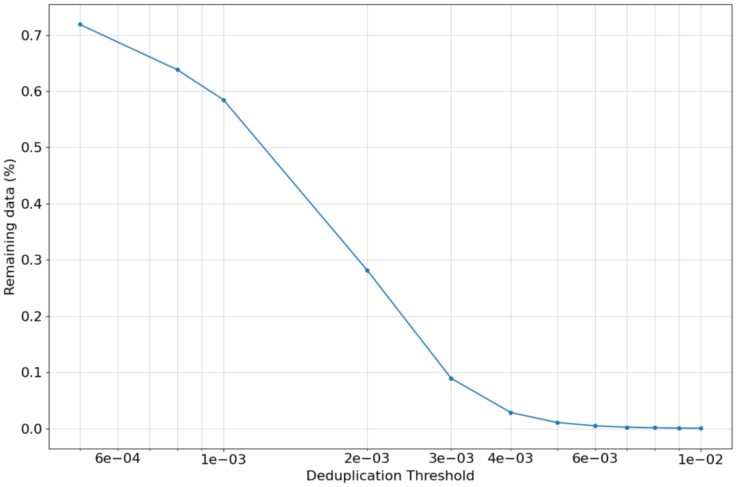
Appendix H GLM2 model parameters

gLM2 is a transformer encoder optimized using AdamW ([24]) and trained in mixed precision bfloat16. We set the AdamW betas to (0.9, 0.95) and weight decay of 0.1. We disable dropout throughout training. The learning rate is warmed up for 1k steps, followed by a cosine decay to 10% of the maximum learning rate. gLM2 uses RoPE ([44]) position encoding, SwiGLU ([38]) feed-forward layers, and RMS normalization ([54]). We leverage Flash Attention 2 ([9]) to speed up attention computation over the sequence length of 2048.

	Dim	Num heads	Num layers	Context length	Learning rate	Batch size	Pretraining tokens
gLM2-150M	640	10	30	2048	5e-4	224	275B
gLM2-650M	1280	20	33	2048	5e-4	224	275B

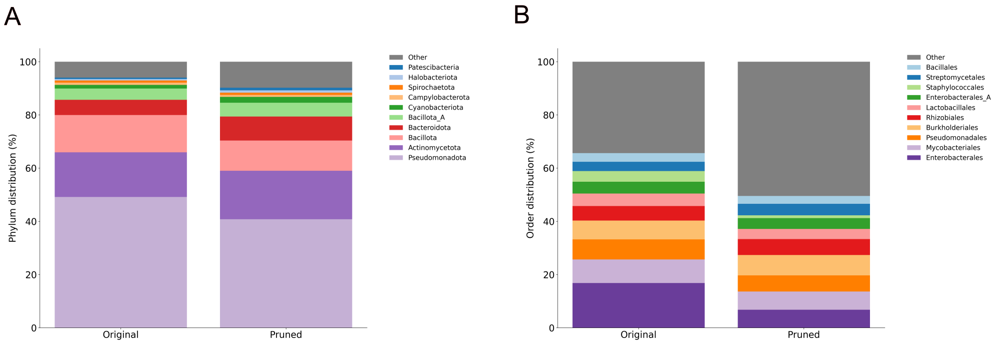
Appendix I Semantic deduplication distance threshold

The percentage of remaining training examples as a function of the embedding distance threshold. Examples within the distance threshold in embedding space are deduplicated.



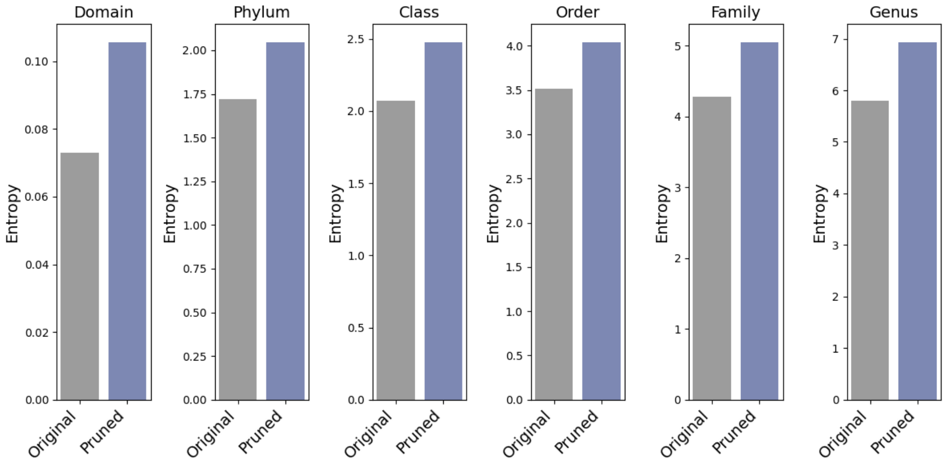
Appendix J Taxonomic distribution of the OG dataset before and after pruning

Data pruning through semantic deduplication reduces dataset bias toward overrepresented phyla and orders.



Appendix K Taxonomic entropy of the OG dataset before and after pruning

Semantic deduplication of the OG dataset consistently increases the taxonomic entropy across all taxonomic ranks, indicating a more even distribution.



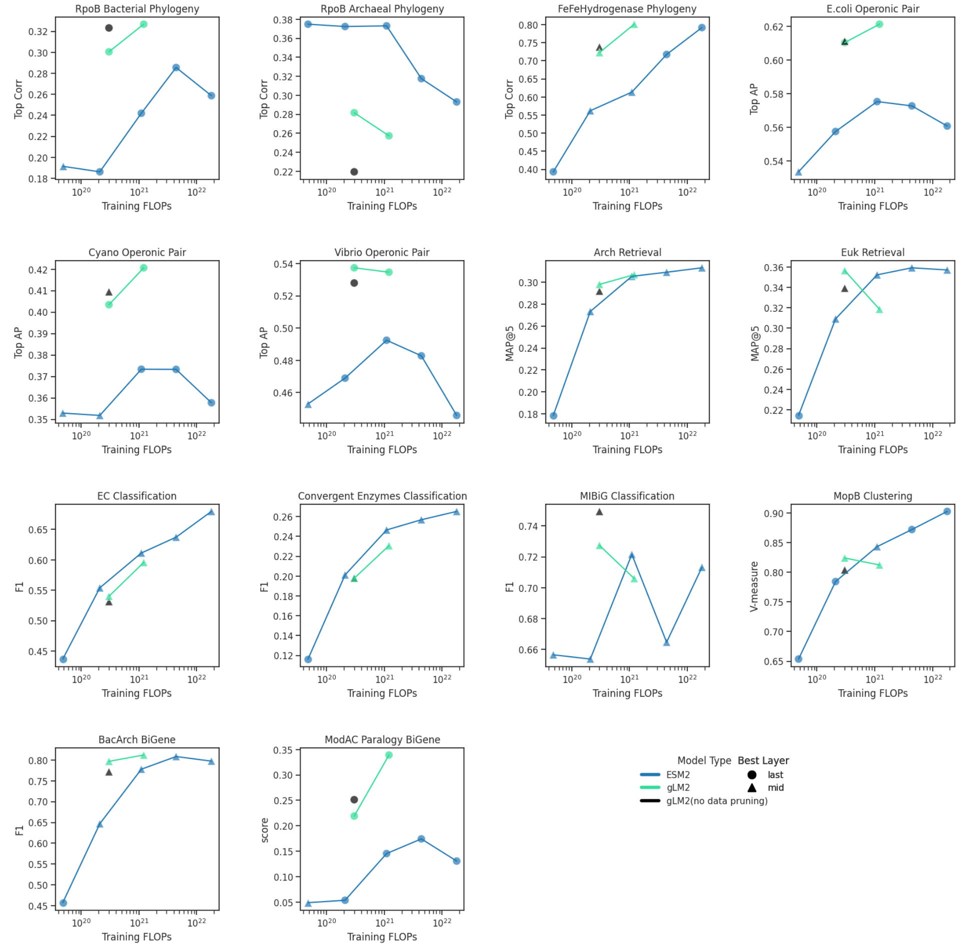
Appendix L Ablation of semantic deduplication

We train two 150M parameter gLM2 models on the original and pruned OMG dataset, each for 600k steps. Both models are evaluated on the DGEB benchmark. Pruning improves performance, especially for tasks with under-represented sequences.

	Bac Arch BiGene	Arch Retrieval	Cyano Oper. Retri eval	ModAC Paralogy BiGene	FeFe Hydrogenase Phylogeny	RpoB Archaeal Phylogeny	EC Classification	MIBiG Classifica tion	E.coli Operonic Pair	Euk Retrieval	Convergent Enzymes Classification	MopB Clustering	Vibrio Operonic Pair	RpoB Bacterial Phylogeny	DGEB Score
gLM2_150M (no data pruning)	0.77	0.29	**0.41**	**0.25**	**0.74**	0.22	0.53	**0.75**	0.61	0.34	0.20	0.80	0.53	**0.32**	0.48
gLM2_150M (with data pruning)	**0.80**	**0.30**	0.40	0.22	0.72	**0.28**	**0.54**	0.73	0.61	**0.36**	0.20	**0.82**	**0.54**	0.30	**0.49**

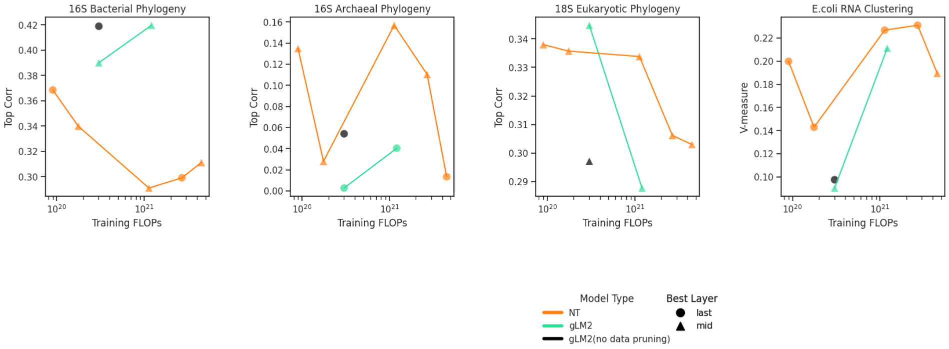
Appendix M Per task DGEB scaling with FLOPs for ESM2 and gLM2 models in amino acid tasks

Primary metric from the best scoring layer (between mid, and last) is reported for each task. To account for model-specific patterns in learning task-relevant functional information across different layers in the network ([52]), DGEB calculates model performance for both mid and last layer and reports the best score between the two.



Appendix N Per task DGEB scaling with FLOPs for Nucleotide Transformers and gLM2 models in nucleic acid tasks.

Primary metric from the best scoring layer (between mid, and last) is reported for each task. To account for model-specific patterns in learning task-relevant functional information across different layers in the network ([52]), DGEB calculates model performance for both mid and last layer and reports the best score between the two.



Appendix O ModA and ModC sequence concatenation

This concatenated sequence was derived from the 2ONK_A_2ONK_C alignment used in [29].

```
MFLKVRAEKRLGNFRLNVDFEMGRDYCVLLGPTGAGKSVFLELIAGIVKPDRGEVRLNGADITPLPPERGIGFV
PQDYALFPHLSVYRNIAYGLRNVERVERDRRVREMAEKLGIAHLLDRKPARLSGGERQRVALARALVIQPRLLLLDEPLSAV
DLKTKGVLMEELRFVQREFDVPILHVTHDLIEAAMLADEVAVMLNGRIVEKGKLKELFSAKNGEVAEFLSARNLLLKVSKIL
DMRLLFSALLALLSSIILLFVLLPVAATVTLQLFNFDEFLKAASDPAVWKVVLTTYYAALISTLIAVIFGTPLAYILARKSF
PGKSVVEGIVDLPVVIPHTVAGIALLVVFGSSGLIGSFSPLKFVDALPGIVVAMLFVSVPIYINQAKEGFASVDVRLEHVAR
TLGSSPLRVFFTVSLPLSVRHIVAGAIMSWARGISEFGAVVVIAYYPMIAPTLIYERYLSEGLSAAMPVAAILILLSLAVFV
ALRIIVGREDVSEGQG
```

Appendix P Additional Files

Additional Files are found in urlHiddenForAnonymity

Additional File 1. OG sample ID to original NCBI metadata. A JSON file mapping OG sample ID (taxon_oid) to NCBI metadata (accessions, collection dates).

Additional File 2. DOIs for MGnify samples. DOIs for MGnify samples that were included in this study, where available.

Additional File 3. DOIs for IMG samples, DOIs for IMG samples that were included in this study, where available.