
Omics Integrated Library (OIL): An Open Multi-domain Codon-resolved Nucleotide Framework for Genomic Language Modeling

Ivan Garibay

University of Central Florida, Orlando, FL
DeepBio Scientific, Winter Park, FL
igaribay@ucf.edu

Mehdi Yazdani-Jahromi

DeepBio Scientific, Winter Park, FL
yazdani@deepbioscientific.com

Ozlem Ozmen Garibay

University of Central Florida, Orlando, FL
ozlem@ucf.edu

Sina Abdidizaji

University of Central Florida, Orlando, FL
sina.abdidizaji@ucf.edu

Ali Khodabandeh Yalabadi

University of Central Florida, Orlando, FL
yalabadi@ucf.edu

Abstract

Genomic language modeling depends on the size and diversity of its pretraining corpus, yet for protein-coding sequence the representational question, amino acid, nucleotide or codon, remains unsettled. It matters most for messenger RNA therapeutics, where synonymous codon usage governs expression, transcript stability and immunogenicity. No open corpus combines codon resolution with the biological diversity across which codon usage actually varies: codon corpora are confined to reference organisms and discard genomic context, while frame-agnostic and metagenomic corpora either omit frame information or store coding sequence as translated protein. In this paper, we introduce OIL, the Omics Integrated Library, an open framework that assembles a catalogue of 2.39 billion records (2.35 trillion nucleotides) across fifteen sources spanning metagenomics, eukaryotes, viruses, pathogens, regulatory RNAs, antibodies, and patents. Of these, 86% have passed cross-source deduplication, and release configurations are selected from that population by declarative query. The intergenic stream is maintained separately: 1.77 billion staged fragments reduced to 254 million representatives. Each coding sequence in OIL is represented in canonical reading frame at codon resolution, each intergenic sequence as a sequence of nucleotides, and each sequence has provenance in a catalogue that is queried via configuration in declarative form. Our inclusion policy restricts OIL to only observed nucleotides, with 99.98% of the deduplicated catalogue being native transcribed cDNA, and its coding portion about 6.7 times larger than the largest published codon corpus.

1 Introduction

Language models pretrained on biological sequence now surpass specialised algorithms across genomics and proteins, largely on the strength of better pretraining corpora [1, 2]. One design choice underlying those gains is rarely made explicit: at what resolution should protein-coding sequence be represented, amino acids, raw nucleotides or codons?

The question is sharpest for messenger RNA therapeutics, where design largely means choosing synonymous codons that set translational efficiency, transcript stability and immunogenicity [3, 4]. All such approaches need one thing: an understanding of how natural biology chooses codons.

Table 1: Composition of the OIL v1.0 catalogue by biological domain. Counts are physical records and include duplicate record identifiers. Per-source detail is in Appendix A.

Domain	Records	Share
Metagenomic coding (JGI IMG, MGnify)	1,890,800,696	79.19%
Metagenomic intergenic (IMG, IMG OID, MGnify)	253,528,927	10.62%
Eukaryotic reference (RefSeq R233, legacy RefSeq, RNAcentral)	172,666,812	7.23%
Plant and fungal (Phytozome, MycoCosm)	43,056,471	1.80%
Viral and pathogen (IMG/VR, BV-BRC, NCBI Virus)	27,101,444	1.14%
Antibody and patent (PLAbDab, Thera-SAbDab, WIPO)	377,528	0.02%
Total (catalogue)	2,387,531,878	

Those rules are not human-specific. They reflect GC content, tRNA pools and translational selection, each varying widely across life, most of it uncultivated, which is why ESM, trained largely on microbes and metagenomes, transfers zero-shot to human-specific tasks [5]. The substrate for codon design should therefore be the widest natural variation in codon use, mostly metagenomic, with host specificity acquired downstream by conditioning or fine-tuning. That signal is recoverable: Equi-mRNA learns GC content and tRNA pool size from symmetric synonymous-codon representations [6], and SynCodonLM shows a codon model falls back on amino-acid representation unless codons are explicitly kept [7].

No open corpus satisfies this standard. Frame-independent DNA corpora such as OpenGenome2/Evo 2 and GenomeOcean keep genes inside unfiltered contigs under six-frame and strand uncertainty, with tokenization crossing codon boundaries [2, 8]. Metagenomic corpora such as OMG hold coding sequence as translated amino acids, from which synonymous codons cannot be recovered [9, 10]. Codon-aware corpora do retain codons [7, 11–15], but all use curated reference genomes and carry no non-coding sequence. It is a matter of representation scope, not size.

Contributions. We present the Omics Integrated Library (OIL), an open nucleotide framework that assembles 2.349 trillion nucleotides from 15 sources. **(1) Codon-resolved metagenomic-scale framework:** To the best of our knowledge, OIL is the first framework to extend collection beyond curated reference organisms into metagenomics. Its deduplicated coding catalogue holds 289,485,932 representative coding sequences after identity clustering, about 6.7 times the 43.5M of SynCodonLM, the previous largest openly released; both figures are stated before semantic deduplication, which SynCodonLM did not apply. **(2) NT-Only Inclusion Policy:** All sequences are introduced in terms of observed nucleotides, none reverse translated or artificially produced, with source indicated for each individual entry at 99.98% original transcription from cDNA. **(3) Scope beyond curated reference data:** inclusion of metagenomic sequence, eukaryote-infecting viruses, pathogens, antibodies, and patents with retention of the intergenic parts. **(4) A Catalog-first reproducible resource:** in which training datasets are declarative queries from the catalog with provenance per record, provided with Evo 2 embeddings, a leakage filter, and a written methods ledger. The catalogue’s event table instruments per-record and per-chunk operations; corpus-scale mutations, release tagging included, are documented in the ledger rather than emitted as events. OIL does not claim to be the first cross-kingdom codon collection. Its novelty is integration: it complements rather than competes with the frame-agnostic corpora, and it uses Evo 2 embeddings for its semantic deduplication stage [2].

2 The Framework of Data Collection

The OIL catalogue holds 2,387,531,878 records and 2.349 trillion base pairs. It has three levels, and every figure below names which it counts: the catalogue as ingested; the 2,055,313,350 records that have passed cross-source deduplication; and the release configurations selected from those by query. The intergenic stream is maintained separately: 1,774,646,393 staged fragments (388 Gbp) reduced to 253,528,927 representatives (70.38 Gbp), which are included in the catalogue total above. Keeping the two streams apart makes genomic context available to a model without imposing it. The composition statistics for the domains have been presented in Table 1, while Appendix A provides the per-source table.

Metagenomic coding sequence is about 79% of the catalogue, so OIL’s scope lies in the long tail of the remaining ~243 million records, not the bulk. The metagenomic bulk supplies coding variety reference corpora omit; the smaller sources supply domain variety the metagenomes lack. Metagenomic data are drawn directly from JGI IMG and EMBL-EBI MGnify rather than from any existing corpus, since those repositories have already been used to build one storing coding sequence as translated amino acids [9]. Antibody and patent sequences enter at the size the inclusion policy admits (Appendix A). OIL-Earth is the broad pre-training set: every coding and non-coding RNA sequence meeting the inclusion criteria and carrying no evaluation-leakage flag, uncapped and in natural proportion. It selects 364,761,456 catalogue rows, of which 300,610,584 are distinct records; the difference is duplicate record identifiers, discussed in Appendix A. The intergenic stream is admitted to the training population but not to the r1 release tag, which was fixed before intergenic ingestion completed. OIL-Med (27,623,740 rows) intersects that predicate with the sources and molecular classes most relevant to mRNA and antibody modeling; two sources named in its definition, an HIV-specific database and NCBI Pathogen, contribute no records, so the realised configuration is narrower than its stated predicate (Appendix A).

3 Construction

The NT-Only Inclusion Policy. The OIL framework includes observed nucleotide sequence only. All coding sequences are taken from actual transcripts or genes, not from artificial back-translations of amino acids nor synthesized. Protein source databases are included only when nucleotides of record itself can be identified and validated by translation (see Appendix C). Every record has provenance flag, and default configurations filter out codon optimized sequence. Back-translation would remove exactly what a codon model reads: synonymous usage reflects GC content, species-specific bias and functional motifs, and an optimiser manufactures those statistics rather than recording them. Recovery was attempted for 52,056 amino-acid-only candidates and succeeded for 15,721. Of 2,237 therapeutic-antibody candidates almost none carried a recoverable, translation-verified nucleotide sequence; that set is retained as catalog-only. Across the deduplicated catalogue, 99.98% of records are native cDNA; every WIPO patent entry carries the codon-optimized flag and is catalog-only, and 21 lack any provenance flag. WIPO contributes 325,533 catalogue records of 395,882 acquired; the remaining 69,529 were never ingested (Appendix A).

Gene Calling and Filtering. Annotated sources arrive with sequence already defined. Unannotated ones, chiefly the metagenomic assemblies, arrive as raw contigs and require gene calling. Where the repository already supplies genes and intergenic regions we use them; otherwise we call genes with Pyrodigal in metagenomic mode [16, 17], keeping both the coding sequences and the intergenic regions between them. Our workflow follows the published OMG methodology, reimplemented independently with no reuse of their code or artifacts, and departs from it in two choices: codons remain untranslated, and two extraction paths are used, existing gene calls where the archive supplies them and de novo calling otherwise. No per-position frame annotation is needed, because the storage format fixes the frame: codon boundaries fall at every third nucleotide and each coding sequence is stored as a single gene in canonical frame, which the frameless corpora cannot offer. Filtering runs in three steps: contig-level criteria following [9], which truncate a contig at an implausible element rather than discarding it; sequence-level criteria reflecting what a codon model needs, itemised with per-criterion counts in Appendix C; and a corpus-level step testing only length, GC content and ambiguity, since codon criteria are irrelevant to non-coding RNA. No sequence is removed at intake. The outcome is a flag the configurations query at release time.

Catalog. Each sequence is stored once, with fields describing its source, molecular type, quality and cluster membership, in a single relational table of 136 columns (Appendix B). Configurations are predicates over that catalog, not builds: a custom configuration is a view on an already cleaned, deduplicated and filtered corpus, and the assumptions a fixed corpus would bake in are exposed as columns and predicates.

4 Deduplication

OIL’s sources overlap and some are internally redundant, so deduplication runs across all of them together. Clustering uses MMseqs2 linclust [18, 19] at 80% nucleotide identity, retaining cluster membership per record so the pre-deduplication corpus is recoverable and a different threshold is

a query. It is a huge reduction. The 2,055,313,350 records that have passed clustering reduce to 300,971,252 representatives, of which 289,485,932 are coding; the intergenic stream reduces from 1,774,646,393 fragments (388 Gbp) to 253,528,927 (70.38 Gbp). RefSeq Release 233, admitted to the catalogue after clustering had run, is not included in these figures. Identity filter only eliminates near-identical duplications but not redundancy among different sequences with the same biological function. Therefore, OIL uses the second, semantic level as a part of SemDeDup pipeline [20] using Evo 2 embeddings, after the identity-based clustering. Anisotropy in the embedding space is very high [21, 22]. In particular, mean cosine similarity between random pairs is 0.861, so thresholding with the original method’s threshold would reduce the whole corpus to almost zero. We use a whitening transform [23], fit once and frozen in the release manifest (see Appendix D).

Choosing the threshold against biology. Since OIL’s contribution is coverage, the threshold must be judged against the risk that deduplication removes the diversity the corpus exists to preserve. We evaluated it where that loss would occur, on metagenomic pairs: 1,423 pairs across similarity bins, against a control drawn from pairs falling just below the cut. At the source method’s 0.90 threshold the absorbed pairs are statistically indistinguishable from the control (0.85% versus 0.90% reaching 80% nucleotide identity), so it isolates no duplicate population among metagenomic sequence. Above it precision rises sharply: 63.3% of absorbed pairs reach 80% identity at cosine 0.95 (95% CI 57.9–68.2), and 94.3% at 0.97. We therefore use **0.95**, more conservative than the method we adapt, set where duplication becomes evident rather than at the conventional threshold. Risk and taxonomy preservation are in Appendix D.

5 Leakage Screening and Characterization

A shared pretraining corpus carries one distinctive risk: if it contains benchmark test samples, reported gains may reflect memorisation. Every record is matched against benchmark datasets including iCodon [24] and the Diverse Genomic Embedding Benchmark [25], and flagged. Configurations draw only on unflagged sequences; flagged ones remain in the catalog so that a user targeting a different benchmark can rescreen (Appendix E). There is minimal overlap: 7,126 coding records, and 39 intergenic hits falling on 12 distinct representatives. The intergenic hit rate is roughly three orders of magnitude below the coding rate, which is what should happen given that the benchmark sequences are coding.

Our characterization requires no trained model (Appendix J). Provenance is auditable record-by-record from the catalog, and codon-usage indices are available for 94.84% of the 300,610,584 distinct records in OIL-Earth, with the remainder being mostly non-coding RNA and viral sequence for which codon usage is undefined by construction and not merely missing. Directly relevant to coverage, third-position GC ranges from 0.27 to 0.88 in the metagenomic portion against 0.36 to 0.65 in the RefSeq-derived eukaryotic reference set, so the metagenomic portion covers codon-usage space the reference organisms do not. This follows from the composition of the data, not from any claim about what a model trained on it would learn.

6 Conclusion and Limitations

OIL combines codon resolution with metagenomic and multi-provenance sampling while retaining genomic context and per-record provenance, exposed as a queryable object. Two principles governed its construction, both verified rather than assumed: every codon was observed, and deduplication does not reduce diversity.

We make no modelling claims. Four limitations apply and are given in full in Appendix K: prokaryotic bias within the metagenomic majority, absent per-record taxonomy, predicted rather than observed coding sequence, and validation by stratified sampling.

Availability. Upstream licence terms prevent redistribution of the assembled sequences. We release the framework: the ingestion and gene-calling pipeline, the catalog schema, the quality and inclusion predicates, the deduplication configuration including the frozen whitening transform and its checksum, and the methods ledger. We measured what these rebuild rather than assuming it. The gene caller is deterministic across the release series used, so the coding stream reproduces from the pipeline alone given the same upstream snapshots; the intergenic stream also needs the extraction coordinate sidecars, which ship with the formal release.

References

- [1] 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, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature methods*, 22(2):287–297, 2025. 1, 14
- [2] Garyk Brixi, Matthew G Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A Gonzalez, et al. Genome modelling and design across all domains of life with evo 2. *Nature*, 652(8112):1349–1361, 2026. 1, 2, 14, 16
- [3] He Zhang, Liang Zhang, Ang Lin, Congcong Xu, Ziyu Li, Kaibo Liu, Boxiang Liu, Xiaopin Ma, Fanfan Zhao, Huiling Jiang, et al. Algorithm for optimized mrna design improves stability and immunogenicity. *Nature*, 621(7978):396–403, 2023. 1, 14
- [4] Adibvafa Fallahpour, Vincent Gureghian, Guillaume J Fillion, Ariel B Lindner, and Amir Pandi. Codontransformer: a multispecies codon optimizer using context-aware neural networks. *Nature Communications*, 16(1):3205, 2025. 1, 14
- [5] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. 2
- [6] Mehdi Yazdani-Jahromi, Ali Khodabandeh Yalabadi, and Ozlem Garibay. Equi-mrna: Protein translation equivariant encoding for mrna language models. *Advances in Neural Information Processing Systems*, 38:26839–26879, 2026. 2
- [7] James Heuschkel, Laura Kingsley, Noah Pefaur, Andrew Nixon, and Steven Cramer. Advancing codon language modeling with synonymous codon constrained masking. *Nucleic Acids Research*, 54(5):gkag166, 2026. 2, 14
- [8] Zhihan Zhou, Robert Riley, Satria Kautsar, Weimin Wu, Rob Egan, Steven Hofmeyr, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Fengchen Liu, et al. Genomeocean: An efficient genome foundation model trained on large-scale metagenomic assemblies. *bioRxiv*, 2025. 2, 14
- [9] Andre Cornman, Jacob West-Roberts, Antonio Pedro Camargo, Simon Roux, Martin Bera-cohea, Milot Mirdita, Sergey Ovchinnikov, and Yunha Hwang. The omg dataset: An open metagenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, pages 2024–08, 2024. 2, 3, 8, 12, 14
- [10] Chenrui Duan, Zelin Zang, Yongjie Xu, Hang He, Siyuan Li, Zihan Liu, Zhen Lei, Ju-Sheng Zheng, and Stan Z Li. Fgenebert: function-driven pre-trained gene language model for metagenomics. *Briefings in Bioinformatics*, 26(6):bbaf592, 2025. 2, 14
- [11] Carlos Outeiral and Charlotte M Deane. Codon language embeddings provide strong signals for use in protein engineering. *Nature Machine Intelligence*, 6(2):170–179, 2024. 2, 14
- [12] Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Lorenzo Kogler-Anele, Milad Miladi, Jacob Miner, Fabien Pertuy, Dinghai Zheng, et al. Codonbert large language model for mrna vaccines. *Genome research*, 34(7):1027–1035, 2024. 14
- [13] Logan Hallee, Nikolaos Rafailidis, and Jason P Gleghorn. cdsbert-extending protein language models with codon awareness. *bioRxiv*, 2023. 14
- [14] Bian Bian, Yiming Zhang, Hongmin Li, Jiuzhou Zhong, and Yutaka Saito. Large generative mrna language foundation model for efficient coding sequence generation and design with mrna-gpt. *bioRxiv*, pages 2025–12, 2025. 14
- [15] Sajad Darabi+, Fan Cao+, Mohsen Naghipourfar+, Sara Rabi, Ankit Sethia, Kyle Gion, Jasleen Grewal, Jonathan Cohen, William Greenleaf, Hani Goodarzi*, and Lakshman Sundaram*. Learning the language of codon translation with CodonFM. 2025. URL <https://research.nvidia.com/labs/dbi/assets/data/manuscripts/nv-codonfm-preprint.pdf>. 2, 14
- [16] Martin Larralde. Pyrodigal: Python bindings and interface to prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, 7(72):4296, 2022. 3
- [17] Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):119, 2010. 3

- [18] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017. 3
- [19] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1):2542, 2018. 3
- [20] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semd-edup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023. 4, 12
- [21] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65, 2019. 4, 12
- [22] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019. 4, 12
- [23] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*, 2021. 4, 12
- [24] Michay Diez, Santiago Gerardo Medina-Muñoz, Luciana Andrea Castellano, Gabriel da Silva Pescador, Qiushuang Wu, and Ariel Alejandro Bazzini. icodon customizes gene expression based on the codon composition. *Scientific reports*, 12(1):12126, 2022. 4, 13
- [25] Jacob West-Roberts, Joshua Kravitz, Nishant Jha, Andre Cornman, and Yunha Hwang. Diverse genomic embedding benchmark for functional evaluation across the tree of life. *bioRxiv*, pages 2024–07, 2024. 4, 13
- [26] Nuala A O’Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufu, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016. 8
- [27] Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic acids research*, 49(D1):D212–D220, 2021. 8
- [28] David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, 40(D1):D1178–D1186, 2012. 8
- [29] Igor V Grigoriev, Roman Nikitin, Sajeet Haridas, Alan Kuo, Robin Ohm, Robert Otillar, Robert Riley, Asaf Salamov, Xueling Zhao, Frank Korzeniewski, et al. Mycocosm portal: gearing up for 1000 fungal genomes. *Nucleic acids research*, 42(D1):D699–D704, 2014. 8
- [30] Robert D Olson, Rida Assaf, Thomas Brettin, Neal Conrad, Clark Cucinell, James J Davis, Donald M Dempsey, Allan Dickerman, Emily M Dietrich, Ronald W Kenyon, et al. Introducing the bacterial and viral bioinformatics resource center (bv-brc): a resource combining patric, ird and vipr. *Nucleic acids research*, 51(D1):D678–D689, 2023. 8
- [31] Antonio Pedro Camargo, Stephen Nayfach, I-Min A Chen, Krishnaveni Palaniappan, Anna Ratner, Ken Chu, Stephan J Ritter, TBK Reddy, Supratim Mukherjee, Frederik Schulz, et al. 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, 2023. 8
- [32] Eneida L Hatcher, Sergey A Zhdanov, Yiming Bao, Olga Blinkova, Eric P Nawrocki, Yuri Ostapchuck, Alejandro A Schäffer, and J Rodney Brister. Virus variation resource—improved response to emergent viral outbreaks. *Nucleic acids research*, 45(D1):D482–D490, 2017. 8, 9
- [33] World Intellectual Property Organization. WIPO PatentScope and patent sequence listings. <https://patentscope.wipo.int/>, 2024. Accessed: 2024. 8, 9
- [34] Brennan Abanades, Tobias H Olsen, Matthew IJ Raybould, Broncio Aguilar-Sanjuan, Wing Ki Wong, Guy Georges, Alexander Bujotzek, and Charlotte M Deane. The patent and literature antibody database (plabdb): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures. *Nucleic Acids Research*, 52(D1):D545–D551, 2024. 8, 9

- [35] Matthew IJ Raybould, Claire Marks, Alan P Lewis, Jiye Shi, Alexander Bujotzek, Bruck Taddese, and Charlotte M Deane. Thera-sabdab: the therapeutic structural antibody database. *Nucleic acids research*, 48(D1):D383–D388, 2020. [8](#), [9](#)
- [36] Carlos P Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. eggno-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution*, 38(12):5825–5829, 2021. [13](#)
- [37] Yanshuo Chen, Yuming Zhang, Joshua Li, Boxue Tian, and Heng Huang. Codontranslator: a conditional codon language model for codon optimization across life domains. *bioRxiv*, pages 2025–11, 2025. [14](#)
- [38] Yunshan Zhong, Weiqi Yan, Yiming Zhang, Kuanhuan Tan, Yutaka Saito, and Bian Bian. Large mrna language foundation modeling with nuwa for unified sequence perception and generation. *bioRxiv*, pages 2025–11, 2025. [14](#)
- [39] Ollie Liu, Sami Jaghouar, Johannes Hagemann, Shangshang Wang, Jason Wiemels, Jeff Kaufman, and Willie Neiswanger. Metagene-1: Metagenomic foundation model for pandemic monitoring. *arXiv preprint arXiv:2501.02045*, 2025. [14](#)
- [40] Yupeng Li, Fan Wang, Jiaqi Yang, Zirong Han, Linfeng Chen, Wenbing Jiang, Hao Zhou, Tong Li, Zehua Tang, Jianxiang Deng, et al. Deep generative optimization of mrna codon sequences for enhanced mrna translation and therapeutic efficacy. *Nature communications*, 16(1):9957, 2025. [14](#)

Appendices

A Data Sources

Table 2 presents per-source records and base counts. Three figures in it are not simple ingest totals and are worth naming. RNAcentral contributed 40,712,942 parsed records of which 39,684,585 reached the catalogue; the 1,028,357 difference is a confirmed ingestion gap, not a policy exclusion. WIPO and the two antibody sources are smaller than their acquired candidate sets because the NT-only policy admits a record only when its own nucleotides are recoverable and translation-verified. And record identifiers are not unique across rows, for two distinct reasons: a uniqueness constraint was absent during part of ingestion, affecting five sources; and the metagenomic stream writes some records under more than one extraction path. Within the OIL-Earth configuration the two together account for 64,150,872 rows, so its 364,761,456 rows carry 300,610,584 distinct identifiers.

Table 2: Per-source composition of OIL

Source	Records	Distinct	Bases	OIL-Earth	OIL-Med
<i>Metagenomic coding</i>					
OMG passthrough [9]	1,890,800,696	—	1.683 Tbp	306,858,773	0
<i>Metagenomic intergenic</i>					
IMG	96,430,941	—		0	0
IMG OID	69,936,946	—		0	0
MGNify	87,161,040	—		0	0
<i>Eukaryotic reference</i>					
RefSeq Release 233 [26]	78,689,601	78,689,601	225.4 Gbp	0	0
Legacy RefSeq snapshot [26]	54,292,626	49,762,650	165.9 Gbp	24,260,885	24,260,885
RNAcentral [27]	39,684,585	39,684,585	30.7 Gbp	11,199,043	3,357,805
<i>Plant and fungal</i>					
Phytozome [28]	22,162,825	20,897,307	28.4 Gbp	5,433,779	0
MycoCosm [29]	20,893,646	20,321,773	27.7 Gbp	11,925,633	0
<i>Viral and pathogen</i>					
BV-BRC [30]	21,460,901	21,460,901	34.9 Gbp	965,231	0
IMG/VR [31]	5,621,398	5,576,197	82.4 Gbp	4,096,023	0
NCBI Virus [32]	19,145	19,145	573 Mbp	17,039	0
<i>Antibody and patent</i>					
WIPO patents [33]	325,533	317,332	1.39 Gbp	0	0
PLAbDab [34]	51,993	51,993	121 Mbp	5,050	5,050
Thera-SAbDab [35]	2	2	767 Kbp	0	0
Total	2,387,531,878	—	2.349 Tbp	364,761,456	27,623,740

A.1 Per-source Detail

The metagenome-based reference set was compiled using the following sources. First, the OMG passthrough contains assembled coding and intergenic regions provided by JGI IMG and EMBL MGNify, as detailed in Section 3, and represents prokaryote content of OIL. Second, IMG/VR database [31] provides 5.6 million viral records from the same metagenomic environment.

Coding sequences of vertebrates and other eukaryotes are taken from two frozen databases instead of the live ingestion. First, it is the legacy snapshot based on RefSeq Release 208 (September 2021), consisting of 400 GenBank databases of vertebrates, mammals, invertebrates, fungi, and viruses, ingested as 54.29 million records. It includes only predicted RefSeq (XM and XR accessions). RefSeq Release 233 has since been ingested and admitted to the catalogue: 78,689,601 records, of which 531,412 are curated and the remainder computationally predicted. It arrived after cross-source clustering had run and has not yet been deduplicated, so it is present in the catalogue and absent from every deduplicated figure in this paper. The second database is RNAcentral [27] (39.68 million catalogue records) providing non-coding and regulatory RNAs. Plant and fungal coding sequences are provided by Phytozome [28] (22.16 million) and MycoCosm [29] (20.89 million).

Pathogen and any other viral sequences are retrieved from BV-BRC [30] (21.46 million sequences derived from 996,518 genomes, annotated similarly to those described in Section 3) and NCBI Virus

[32]. BV-BRC is a predominantly viral resource and not a bacterial one: more than three-quarters of the genomes in BV-BRC are SARS-CoV-2, and most of these genomes collapse under deduplication to produce, roughly speaking, 260,000 unique non-SARS viral genomes.

Antibody sequence data derives from PLabDab [34] and Thera-SAbDab [35], admitted only when the native nucleotide sequence was retrievable and matched the cataloged protein. The patent sequence data derived from WIPO [33], of which 325,533 of 395,882 acquired records reached the catalogue. Since patent nucleotide sequence is optimized for expression in a particular host rather than being transcribed by the cell, every WIPO entry is marked as codon optimized and cataloged but not released in either reference. Some of these domains are excluded in the large frame-agnostic corpora. For instance, OpenGenome2 does not include eukaryotic infecting viruses, and to our knowledge, no nucleotide corpus exists that includes antibody or patent sequences.

B Catalog Schema

The catalog is a relational table made up of 136 columns, each row being one sequence. The main columns in the table are listed in Table 3.

Table 3: Principal catalog fields by group.

Group	Fields	Notes
Identity and sequence	record_id, sequence, length_nt, gc_content, n_fraction	composition statistics used by quality control
Provenance	source, nt_origin, taxon_id	nt_origin realized 99.98% native_cdna; taxon_id null for the metagenomic majority
Molecular structure	mol_type, frame_class, cds_start, cds_end, cds_length, ncRNA_class	UTR lengths present
Quality control	qc_flags (including the release-selection flag)	evaluated per record, queried at release
Deduplication	mmseqs_cluster_id, is_cluster_representative	see Section 4
Configuration and evaluation	selected_for_training, release tags, eval_leakage	
Codon-usage indices	cai_vs_human, cai_vs_ecoli, cai_vs_corpus, enc, gc3, cpg_fraction	computed from faithful codons
Annotation (scaffolded)	75 columns across license-partitioned tiers, of which 35 are populated	schema in place;

The framework also considers including codon usage statistics per sequence such as: Codon Adaptation Index relative to the standard human and E. coli codon usage tables, effective number of codons, GC content at the third position of the codons, and CpG dinucleotide fraction.

C Inclusion Policy, Gene Calling, Preprocessing, and Quality Control

The metagenomic construction path is shown end to end in Figure 1. Quality control follows the catalog-first design of Section 3, wherein no record is excluded or shortened on input. Every record gets assessed once according to a list of inclusion criteria and is marked with a boolean flag that is queried by the reference configurations upon their release. These are a minimum length of 60 nucleotides (twenty codons), a GC ratio from [0.15, 0.85], and an ambiguous base fraction lower than 0.05. If the record does not pass any of these checks, it remains in the catalog with the flag being false, hence allowing a different configuration to include it. There is no maximum corpus-wide record length; the longest one included is over two million nucleotides long, and the length threshold as well as all other filters is a query, not a corpus property. Reverse-complement augmentation is not performed due to the arbitrariness of contig-wise directionality in assemblies.

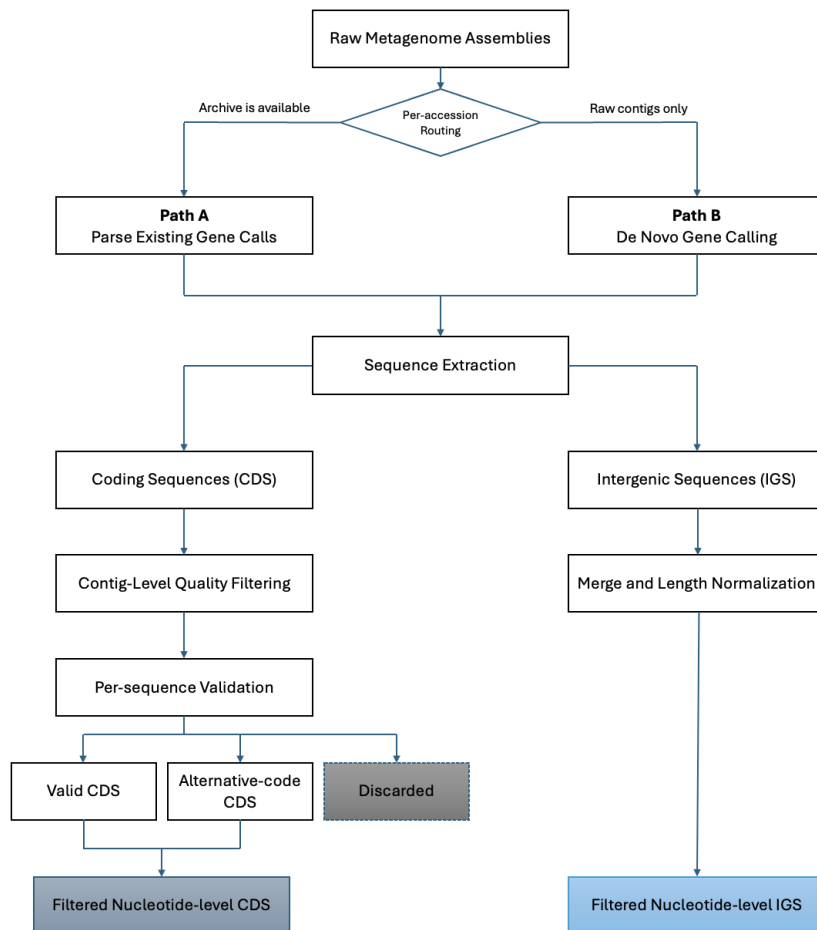


Figure 1: The metagenomic construction path

C.1 The Two Admission Paths

The sequence enters OIL using either of the two options. In most of the cases, the source provides the nucleotide sequence, which may either be coding or assembled contigs, and such sources are considered as observed. There are certain sources which do not provide the nucleotide sequence but protein sequence, such as antibody databases; for such sources, a record will be accepted if the nucleotide sequence of the record itself can be derived from the same physical entry, and if the derived nucleotide sequence translates to the amino acid sequence in the record. The translation of the sequence must be exact, and not a guess or alignment, and one whose nucleotide sequence is not recoverable for verification is discarded rather than rebuilt.

C.2 Metagenomic Coding-Sequence Validation

Each candidate sequence along the way of the metagenomic analysis process is subjected to seven different tests and kept only when it does not fail in any of them. Table 4 shows how many candidates violate each test. Two tests of structure, which discard many sequences, the requirement of a start codon and that of a stop codon, are more strict than necessary for a database designed simply for modeling of nucleotides. These rules arise due to the expected function; if a model positions the codon relative to the first triplet within a given entry, then the first triplet must itself be a start codon, and a coding sequence lacking a terminal stop codon is most likely a partial prediction than a complete gene.

Table 4: Per-criterion failure counts for metagenomic coding-sequence validation, over 1,918,832,613 candidates, as recorded by the r1.0 validation pass. The GC-content criterion was evaluated against a composition statistic we have since found to be computed by more than one formula across sources, so that row should be read as indicative rather than exact. A sequence may fail several criteria, so the column does not sum to the number discarded: 27,985,490 sequences (1.46%) fail at least one, while the per-criterion counts total 48,540,864.

Criterion	Requirement	Failures
Reading frame	length divisible by three	8,040,918
Minimum length	at least 60 nucleotides (20 codons)	242,991
Start codon	ATG, GTG, or TTG	15,816,081
Stop codon	terminal TAA, TAG, or TGA	16,139,309
Internal stop	none under the standard genetic code	7,080,012
GC content	between 0.15 and 0.85	247,360
Ambiguity	N-fraction below 0.05	974,193
Admitted	fails no criterion	1,890,847,123

C.3 Organisms Using Reassigned Genetic codes

Internal stop codons represent an additional challenge in that an internal stop codon within the framework of the standard genetic code might mean either a mistaken prediction or reassignment of that codon in the genetic code of an organism. There are several bacterial and ciliate groups where codons, which are considered as stops according to the standard code, are used for coding amino acids. Feeding such sequences into the training dataset of the standard code would not only be noisy but also fundamentally incorrect for a model which treats synonymous codons as belonging to one group, since the reassigned codon will be considered to be a synonym of the amino acid it codes for. Sequences coming from organisms using alternative code are consequently directed along an alternate path and not in the corpus, based on their taxonomic classification. In this release, the stream is empty, and this is due to three independent reasons. First, it was done intentionally, as mentioned earlier. Second, there is a gap in the ordering of the pipeline: the lookup based on which the stream works was created after validation, and therefore no sequence was sent to the stream at all. The third reason became obvious when we tested the previous two: the lookup table with the taxonomic information for these organisms was never populated from the multi-genus metagenomic archives where the organisms in question can be found, since the procedure of extracting the data in question does not operate on archives whose lineage variety exceeds a certain threshold, and most metagenomes fall under that description. The surviving lookup table has only a few hundred accessions in it, with no alternative-code genera included. So, running the router on it again gets us nothing, an assumption we have tested.

This result has a calculable upper limit. Of 7,080,012 discarded coding sequences that had an internal stop codon, 5,102,162 did not violate any other criterion, meaning they were the only ones that could theoretically be rescued by an alternate interpretation of the genetic code. In a sample of 200 metagenomic archives, 57 of which contain the annotation required to make an estimate, one would find on the order of a thousand sequences, or about 0.03% of the internal-stop-only group. There are at least two reasons why this number may be even lower. The taxonomic information pertains to the lineage of the most similar protein to the gene being examined, and does not specify anything about the source organism; thus, a sequence that is homologous to a protein translated with a ciliate genetic code doesn't have to use a ciliate genetic code itself. Also, most sampled libraries lacked taxonomic annotation. Against the total of two billion sequences, this fraction is insignificant. However, recovery still would require the analysis of each gene from the point of view of the taxonomic data rather than joining the databases, so the task will be put off for the time being.

D Deduplication

Deduplication narrows the 2,055,313,350 clustered records of coding and non-coding RNA to 300,971,252 clusters, of which 289,485,932 are coding; that is the figure against which the OIL framework compares itself to other codon corpora. The number of representative rows, 365 million, is

greater than the number of clusters, because clustered-assembled genomes increase the cluster number through various extracted rows: the metagenomics portion by 1.21 times, the snapshot of vertebrates by 1.18, MycoCosm by 1.13 and BV-BRC by 1.09 times. We state distinct-cluster numbers as the deduplicated number of sequences, with the representative rows stated when different. The intergenic stream demonstrates similar decrease: 1.77 billion fragments (388 Gbp) cluster to 253,528,927 representative (70.38 Gbp) that is 14.3% by count of surviving fragments. Only cross-source and joint clustering can remove that redundancy.

D.1 Residual Risk and Taxonomic Preservation

We state rather than round away this bounded residual risk of the operating threshold. The percentage of the metagenome-to-metagenome pairs that are part of all absorptions is 16.4%, and around one-third of these pairs at operating threshold fail to achieve nucleotide identity of 80%. Therefore, in the case where maximum diversity loss may be incurred, it is less than one-tenth of one percent of total representatives. There is taxonomic coverage; 100% at kingdom and genus level and 99.88% at species level at the pilot stage.

D.2 Anisotropy Correction

Semantic deduplication of the Evo 2 embeddings should be performed cautiously, since the space is extremely anisotropic, vectors lying on a very thin cone, an artifact already described for contextual language model embeddings [21, 22]. In the raw space, the mean cosine similarity of a random pair is 0.861, and 41.8% of such pairs have cosine greater than 0.90, thus the threshold determined by [20] for image embeddings collapses almost the entire corpus. Pair checks on sequence-level have shown that none of the absorbed pairs at any random raw threshold had similarity greater than 80%, meaning raw embedding cosine does not measure similarity at all. Thus, we restore isotropy via a whitening transformation [23], once per a million vectors, then the mean cosine of random pairs drops to 4×10^{-4} .

D.3 Relation to Prior Semantic Deduplication of Genomic Data

Semantically removing duplicate sequences is a common approach for cleaning up genome pretraining data, and the OMG authors used semantic deduplication for their dataset, removing 42% of the contig-level examples with an embedding distance cutoff of 10^{-3} [9]. However, the two approaches cannot be compared, since differences in reduction rates are due to different objectives, and the reduction rate is not lower in the OIL approach. The OMG’s pruning units are contig-level examples, while our pruning unit is a sequence representative; their embedding space was obtained through mean pooling from a genome language model trained on their corpus, while our embedding space is Whitened Evo 2. So the corresponding thresholds have no direct relation to each other. Moreover, while theirs aims at taxonomic balancing of an excessively large corpus, ours aims at reduction of redundancy with the clear diversity-preserving constraint. Most significantly, the stages are different; they use semantic deduplication as their main deduplication because they think sequence clustering would not be appropriate for variable-length genomic sequence, while OIL uses this deduplication at a later stage, after doing the identity clustering where the most duplicate sequences have been removed. Therefore, the semantics level of OIL is processing input in which redundancy is already removed. Thus, this small margin removal represents the ordering and not a less stringent criterion.

D.4 Non-coding Sequence and Taxonomic Capping

The semantic stage runs on coding representatives. Non-coding representatives, taken from the regulatory RNA and metagenomic viral datasets, are kept as is and are not subjected to duplication filtering based on embedding since the embedding is calculated and is defined on coding representatives only. Reference configurations do not impose the taxonomy limit and that is by design, not because of oversight. The pipeline supports genus-level capping stratification while that does not occur in reference release and is not even applicable on the metagenomic level, due to lack of taxonomy per record for virtually all representatives.

E Evaluation-Leakage Screening Procedure

All such sequences in the collected data are compared with test sequences in a set of downstream evaluation benchmark tasks. If a match exists between the sequence and the test sequence, then an eval leakage flag indicating the name of the associated benchmark is added. Flagged sequences stay in the collection so that when users are working with a different benchmark, then they can perform the screening process again for themselves without taking our opinion into account. We assess the codon and genome-level benchmarks most applicable to such type of data collection: iCodon [24] and Diverse Genomic Embedding Benchmark [25].

F Annotation Layer

The OIL framework has been engineered to transport the biological annotation layer with the sequences. Annotations are about functionality, structure, and classification for each record of the database which will allow to query the database with biological characteristics as well as sequence information.

The collected data includes the annotation scaffolding in its entirety, with a multi-level schema comprising three levels including cross-kingdom, vertebrate, and pathogen tiers. These levels are partitioned into five different classes (core, human, pathogen-extension, lookup-identifiers, and research-only) to make sure that the annotations carrying restrictions of their sources licenses do not co-exist with those which are freely redistributable. This is supplemented by another table containing the evaluation leakage class of the annotations as well as fourteen annotated drivers which include orthology propagation based annotations using eggNOG mapper [36], protein domain and RNAfamily based annotation (HMMER against Pfam and Infernal against Rfam), secondary structure annotation computing with ViennaRNA package, and cross-reference based annotations with other external functional databases.

Such an incremental model would be possible by virtue of using a mixed versioning approach, which splits sequence builds from annotation updates.

G Reproducibility

The OIL framework has been designed to be reproducible. All decisions made in the process of construction, sources acquired, parsers used, quality predicates applied, and the deduplication threshold, are captured in the methods ledger. The sources are snapshots, not versions, such that the rebuild uses the same sources despite database advancements.

Repeatability stems from our approach of starting with a catalog, and not from retaining a pipeline. Since our reference configurations are declarations in the form of predicates defined on the catalog itself, a repeated execution of the respective query gives us the repeatability of the configuration in question rather than having to rerun the ingestion process. Cluster assignments will persist per entry, which means that the unprocessed collection can be restored and other levels of identity can be formulated without re-clustering.

The pieces which are irreproducible by their nature are frozen. The transformation that serves for semantic duplicates removal, the whitening transform, is trained just once, its checksum is calculated and included into the repository so that the embedding geometry on the base of which the decision to remove semantic duplicate was made could be restored precisely, and there was no need to repeat fitting the transform. Evo 2 embedding storage is delivered with the corpora themselves so that it wasn't necessary to generate embeddings again for those analyses which depended on embeddings. Releases are versioned according to the hybrid versioning scheme of Appendix F, where majors refer to the sequence builds while minors to the annotations refresh within the sequence. The release manifest has documentation about the origin of the two layers independently. Hence, a result obtained from one particular release can be traced back to the actual sequence data and actual annotation data.

H Related Work

We position OIL relative to three groups of prior work, namely corpora for modeling at the codon level, large frame-independent and amino acid corpora, and mRNA therapeutic codon design systems.

H.1 Codon-resolved Corpora

The most relevant resources relative to the application of OIL to coding are those that present coding sequences as codons. The model CaLM trains on coding sequences from the European Nucleotide Archive (114M raw sequences, clustered by organism prior to training) and demonstrates that codon-based embeddings can surpass equivalent amino acid models in performance [11]. CodonBERT (10M mRNA coding sequences) predicts mRNA vaccine properties [12] and cdsBERT enhances protein language models with codon information [13]. Newer models have scaled up. SynCodonLM (43 million RefSeq CDS) limits predictions of codon-level signal by constraining output to synonymous codons [7], while CodonTranslator (around 62 million CDS) conditions on host codon use [37], and the two largest, mRNA-GPT (~165M CDS across the three domains of life) [14] and the EnCodon/CodonFM family (~131M CDS, ~22,000 species, MMseqs-deduplicated)[15] push codon modeling to foundation-model scale.

They share two commonalities, which limited their scope. Firstly, each relies exclusively on curated reference organisms (RefSeq reference assemblies or ENA assembled genomes) and are coding sequence only, devoid of any genomic context. In the case of CodonFM, the use of viral or pathogen sequences for human disease was excluded. Secondly, most of them were made available as model weights and scripts used to curate the corpus from the source database (CodonFM, mRNA-GPT, CaLM), or did not make available an already compiled corpus [38]. SynCodonLM's 43M sequence dataset is the main prior codon corpus released in an accessible way [7]. OIL has a unique position in both dimensions because OIL extends the codon resolution to uncultivated metagenomics as well as viruses, pathogens, therapeutics, and patented sequences, while it sends the entire assembled library that can not be constructed from RefSeq alone.

H.2 Frame-agnostic and Amino-acid Corpora at Scale

The second category consists of works which achieve scale and coverage on the biological side but are unable to preserve codons. The frame-agnostic DNA foundational models, the Nucleotide Transformer [1], and Evo 2 (based on the ~8.8 Tbp OpenGenome2 dataset [2]) operate with single nucleotide or k-mer representation of sequences and reveal no reading-frame structure. Larger on the aggregate, metagenomic datasets also tend to be either frame-agnostic or protein-based. OMG trains on 3.1 Tbp from the IMG JGI and MGnify databases [9], where coding sequences are represented through translated amino acids and intergenic nucleotides. METAGENE-1 is pretrained on the ~1.5 Tbp metagenome from wastewater with byte-pair tokenization [39]. GenomeOcean is trained on ~600 Gbp of metagenomics contigs [8], and FGeneBERT analyzes metagenomic genes in terms of their sequences on the amino acid level, specifically mentioning codon redundancy as an example to discard nucleotides [10]. The datasets provide the scope of metagenomics and genomics missing from codon corpora, but by design, they cannot solve problems about individual codons. The OIL is derived from the very same source upstream databases using independent ingestion and processing following the methodology described in OMG paper without loss of codons and using Evo 2 embeddings for the purpose of deduplication and thus can be considered complementary to the frameless models.

H.3 mRNA-Therapeutic Codon Design

The motivation of OIL comes from the practical issue of designing coding sequence in mRNA therapeutics with a focus on optimizing coding sequences with respect to structure to maximize expression, stability, and immunogenicity. Classical codon optimization utilizes the host frequency table. Principled approaches like LinearDesign optimize both the codon frequency table and the structure at once and show large in vivo improvement [3], but learning-based approaches play an increasingly important role in designing such as CodonTransformer for multi-species codon optimization [4] and deep generative models like RiboDecode [40]. This literature gives shape to the codon requirement, for models where synonymous codon selection is an engineering choice, the training data should contain codons that represent natural codon usage in their fullest biological complexity.

H.4 Positioning

OIL bridges the gap that remains open between the three bodies of literature: codon usage (unlike the frame-invariant and amino acid databases), across the scope of metagenomes and multiple provenance

sites with genomic context (unlike the codon databases), available in a usable artifact together with a queryable catalog (unlike the unreleased codon databases), and tied to a real therapeutic design requirement through the argument presented in Section 1.

I Corpus Characterization

Characterization of OIL does not involve the creation of any model trained on it. Our characterization is based on evidence that is primarily descriptive, including composition and diversity, deduplication-based characterization of redundancy, provenance-based accounting with respect to the inclusion policy, codon-usage based characterization calculated from OIL itself, estimation of evaluation leakage, and comparative characterization against currently available corpora.

I.1 Composition

The composition of each source individually is presented in Table 2. The metagenomic coding component is 1,890,800,696 of the 2,387,531,878 catalogue records, about 79%, with the remaining 497 million records, of which 253,528,927 are metagenomic intergenic, providing the domain coverage needed for this corpus. The catalogue accounts for 2.349 trillion bases; the intergenic stream is staged separately at 388 Gbp before deduplication and 70.38 Gbp after.

I.2 Deduplication Yield and The Structure of Redundancy

Deduplication brings the 2,055,313,350 clustered records down to 300,971,252 clusters, of which 289,485,932 are coding. Representative records stand at around 365 million, higher than the number of distinct clusters for the reasons given in Appendix D. Intergenic record stream is reduced from 1.77 billion records or 388 Gbp in size to 253,528,927 representatives, amounting to 70.38 Gbp in size. In terms of survival rate, only 14.3% records survive the process of deduplication. Deduplication yield will be reported here as a measurement, not a cleaning metric, as the survival fraction and cluster size distribution reflect the extent of biological redundancy represented by these samples, and they are quite different for coding versus intergenic and among the sources.

I.3 Diversity Preservation Under Semantic Deduplication

Since OIL's contribution to the dataset is coverage rather than scale, deduplication must be demonstrated to eliminate redundancy without compromising diversity. Pilot-scale deduplication has preserved 100% of kingdom and 99.88% of species diversity, and proof of threshold calibration is presented in Section 4.

I.4 Provenance Accounting

Inclusion Policy Implementation is very close to exclusivity. In total of 2,055,313,350 records, 2,054,982,118 (99.98%) are transcribed cDNA in native form, 331,211 records are optimized for codon usage, all of which are catalog-only and do not occur in any of reference configurations, and 21 records lack any provenance tag. None of the records are back-translated or generated. Provenance tags are applied at record-level rather than at source level.

I.5 Codon-usage Characterization

Since OIL retains the codons that have been observed, the codon usage statistics calculated based on OIL capture biological data and not an optimization heuristic. The coverage rate for codon-usage statistics is high for the part where the statistics are applicable. Sequences that do not have codon usage indices are predominantly non-coding. With coding regions of sequences alone, the level of coverage is virtually complete, achieving a minimum of 99% for all five metagenomic, fungal, plant, pathogen, and antibody parts as well as 98.9% for the traditional eukaryotic database. However, the lower number for OIL-Med is solely because of its increased regulatory RNA content.

The separation of the collection into metagenomic and reference components is consistent with the prediction of the coverage argument in terms of codon-usage space. The GC percentage at the third codon position varies from 0.27 to 0.88 in the case of metagenomic and from 0.36 to 0.65 in the case of reference sequences, thus metagenomics occupies codon-usage space that the references never reach.

I.6 Evaluation-Leakage Quantification

The comparison with the benchmarks detects 7,126 coding sequences and 39 intergenic sequences; of the 39 intergenic sequences, 37 are from iCodon and 2 from other targets. All these detected records are removed from the reference setup but kept in the catalog setup. This low overlap is in line with the makeup of the corpus since the benchmarks come from reference genomes and OIL from metagenomics data.

J Limitations

The limitations, as mentioned in Section 6, are outlined here in their entirety. OIL is a framework that one can create a dataset based on it, and this paper does not claim any modeling results. The hypothesis that natural codon diversity can serve as the proper dataset is inspired by previous examples in protein language modeling, not the result of this work, and whether this dataset performs well enough for therapeutic applications is beyond the scope of this paper. Composition is inconsistent with the relevant motivation, whereby there is a heavy presence of prokaryotes in the metagenomic portion, whereas mRNA construction is designed for eukaryotic gene expression. Taxonomic precision is constrained by the nature of metagenomics data. Almost all clusters lack any taxonomic assignment, since a gene identified within an assembled mixed community cannot have a single origin species, and thus methods that depend on taxonomic assignment for each record are not possible to perform on corpus scale. Coding sequences of metagenomic data are inferred, not observed. The whole process of assembly and gene calling are inference-based, and the certainty of the inclusion policy is that of the provenance. The policy also confines the coverage to situations where the upstream database contains amino-acid sequences only since there exist therapeutic antibodies, but in an infinitesimal quantity, and the occurrence of a domain in the upstream database does not imply the coverage of the same. The organisms which use reassignment of genetic codes are excluded, both because of the intent behind the policy and because of the deficiency in the sequence ordering process and the lack of taxonomic information for the multiple genus databases from which they occur. The upper limit of the recoverable pool is 5,102,162 internal-stop-only strings and is estimated using sampling up to order of one thousand (Appendix C). Deduplication requires certain threshold values determined using stratified sampling rather than exhaustive determination, and the leakage filtering process depends on how well the benchmark set is filtered.

Two more aspects to consider regarding the alternative genetic code exclusion are the following. In Evo 2, there are organisms that have different genetic codes, and it is stated that the model learns how to predict which genetic code should be used depending on the sequence context [2], so this exclusion was just an option for the model which could work without codon positioning. Also, the explanation of the mechanism of the alternative code stream is provided in Appendix C.