A machine learning framework for discovery and enrichment of metagenomics metadata from open access publications

Maaly Nassar

EMBL's European Bioinformatics Institute (EMBL-EBI) <https://orcid.org/0000-0003-3362-5202>

Alexander B. Rogers

EMBL's European Bioinformatics Institute (EMBL-EBI)

Francesco Talo'

EMBL's European Bioinformatics Institute (EMBL-EBI)

Santiago Sanchez

EMBL's European Bioinformatics Institute (EMBL-EBI)

Zunaira Shafique

EMBL's European Bioinformatics Institute (EMBL-EBI)

Robert D. Finn

EMBL-EBI <https://orcid.org/0000-0001-8626-2148>

Johanna McEntyre (✉ mcentyre@ebi.ac.uk)

EMBL's European Bioinformatics Institute (EMBL-EBI)

Article

Keywords:

Posted Date: February 25th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1396476/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at GigaScience on August 11th, 2022. See the published version at <https://doi.org/10.1093/gigascience/giac077>.

A machine learning framework for discovery and enrichment of metagenomics metadata from open access publications

Maaly Nassar, Alexander B. Rogers, Francesco Talo', Santiago Sanchez, Zunaira Shafique, Robert D. Finn, Johanna McEntyre*

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Corresponding. mcentyre@ebi.ac.uk

Abstract

Metagenomics is a culture-independent method for studying the microbes inhabiting a particular environment. Comparing the composition of samples (functionally/taxonomically), either from a longitudinal study or cross-sectional studies can provide clues into how the microbiota has adapted to the environment. However, a recurring challenge, especially when comparing results between independent studies, is that key metadata about the sample and molecular methods used to extract and sequence the genetic material are often missing from sequence records, making it difficult to account for confounding factors. Nevertheless, this missing metadata may be found in the narrative of publications describing the research. Here, we describe a machine learning framework that automatically extracts essential metadata for a wide range of metagenomics studies from the literature contained in Europe PMC. This framework has enabled the extraction of metadata from 114,099 publications in Europe PMC, including 19,900 publications describing metagenomics studies in ENA and MGnify.

Introduction

Whole genome shotgun DNA sequencing has emerged as a powerful tool for assessing the microbial content in a wide range of environmental samples. The advent of metagenomics analyses has enabled the phylogenetic and functional profiling of microbial communities found in a particular biome without the need of culturing. Meta-analysis between samples can reveal specific adaptations to environmental conditions, or candidate members of the community that may be responsible for health or disease¹⁻³. However, to establish causal relationships between observed phenotypes and compositional changes, potential confounding factors such as individual alcohol consumption or therapeutic interventions need to be taken into account^{4,5}. Thus, to undertake meaningful analyses spanning multiple metagenomics datasets, accurate contextual metadata about the environmental conditions from which the sample was taken and the experimental methods is required⁶. Despite the efforts to deposit, organise and analyse metagenomics data in public databases^{7,8}, major obstacles to their usability thereof remain, due to incomplete, missing and/or inaccurate contextual metadata.

Sequence databases, such as Sequence Read Archive (SRA)⁹ and European Nucleotide Archive (ENA)¹⁰, provide permanent storage for metagenome sequences, which offer the potential for knowledge databases, such as MGnify¹¹, MG-RAST¹² and Integrated Microbial Genomes with Microbiome Samples (IMG/M)¹³, to explore the taxonomic diversity and metabolic pathways of biome microbial communities. Although significant advances have been made in the computation and indexing of metagenomics data, the lack of detailed and structured microbiome metadata has dramatically hampered metagenomics cross-study comparisons. To address this issue, metagenomics databases, such as Terrestrial Metagenome DB¹⁴, Human Microbiome Project¹⁵, Human Metagenome DB¹⁶, Genomes OnLine Database¹⁷ (GOLD) and MGnify, have used controlled vocabularies from ontologies, such as ENVO (Environmental ontology)¹⁸, to manually annotate a broad range of metagenomics samples. However, many of these studies have incomplete detailed metadata about sample origins, due to the lack of metadata provided when depositing sequences into a database¹⁹ and the inflexibility of hierarchical ontology relationships for describing diverse and specific environments, for example ENVO lacks vocabularies that could describe samples from many engineered environments. A scalable way to overcome these shortcomings is the development of automated methods to extract sample and experiment metadata from research articles describing metagenomics studies and link them to metagenomics datasets.

Named Entity Recognition (NER) is one of the foundational text-mining methods used for retrieving information from the literature. In life sciences, traditional NER approaches have relied on annotating vocabularies onto texts^{20,21}. However, because of the limitation of their vocabularies, out-of-dictionary synonyms and restrictive hierarchical levels, current ontologies fail to capture essential data pertinent to diverse metagenomic studies; hence the urgent need for machine-guided approaches²². Consequently, hybrid machine and deep learning approaches, such as Bidirectional Long Short-Term Memory (BiLSTM) and Conditional Random Field (CRF) models, have been recently recognized to efficiently

identify diverse entities in biomedical texts. Principally, these NER classifiers have relied on generating context-independent word representations (embeddings) from unlabelled literature corpora using word2vec language models^{23,24}, followed by the fine-tuning of these embeddings with hand-labelled data and task-specific neural architectures, such as LSTM-^{25,26} or BiLSTM-CRF NER models²⁷⁻²⁹. But, whereas word2vec models aim to effectively capture the syntactic and semantic representations of words across diverse linguistic contexts, both context-independent and context-sensitive unidirectional representations³⁰ were found to be suboptimal for sentence- and token-classification tasks, where it is crucial to incorporate contexts from left and right directions (bidirectional representations)³¹. Hence, the emergence of the Bidirectional Encoder Representations from Transformers (BERT)-based models³². BERT alleviates both context-independent and context-sensitive unidirectional constraints with its unified architecture that can both pretrain and fine-tune bidirectional context-sensitive word representations for diverse Natural Language Processing (NLP) tasks. BERT-pretrained word representations from English Wikipedia and Books Corpus managed to outperform hybrid architectures - context-independent or context-sensitive pretraining and BiLSTM-CRF fine-tuning models - in sentence- and word-classification (NER) tasks. Recently, BioBERT, a BERT-based model pretrained on PubMed abstracts and PMC full-text articles, has significantly outperformed BiLSTM-CRF and the original BERT models on most of the biomedical NER tasks³³. Moreover, BioBERT has been recognized for its high performance in classifying medical entities after being fine-tuned with weakly supervised ontology-driven data³⁴.

Herein, we describe a machine-learning framework capable of enriching sample and experimental metadata of a wide range of metagenomics studies with terms extracted from open access publications in Europe PMC³⁵.

Results

Overview

A machine learning framework was developed and integrated for enriching a wide variety of metagenomics studies with essential and accurate metadata from open access research articles. This framework: (1) trained machine learning models on cross-referenced data to classify a broad range of biome publications (literature classification and triage); (2) defined and curated 16 novel bio-entities that describe vital metadata for diverse metagenomics studies (biocuration; [Table 1](#)); (3) trained and validated BioBERT-based NER models on curated datasets (BioBERT fine-tuning and internal validation); (4) deployed 16 high performance trained NER models in retrieving accurate metagenomics metadata from thousands of publications in Europe PMC (NER); (5) standardised BioBERT-model-predicted annotations with multiple domain-specific ontologies (normalisation); (6) validated the transfer of models accuracy and reliability to uncured datasets in Europe PMC and ENA (external validation); and lastly (7) developed an integrated and fully automated metagenomics annotations pipeline that regularly enriches up-to-date metagenomics studies

(19,209) with model-predicted metadata from research articles (114,099) (databases enrichments).

Literature classification

Predicting a wide variety of microbiome environments was an essential requirement to triage publications for NER tasks. This requirement was addressed by constructing supervised training datasets, where the GOLD annotations assigned to metagenomics studies were mapped to the corresponding MGnify cross-referenced publications. Subsequently, random forest models were trained on the hierarchical levels of GOLD ontology yielding diverse biome prediction models. Comparing models performance metrics revealed higher F1-scores of 91-97% for the top GOLD hierarchical levels (Engineered, Environmental and Host-associated). However, Doc2Vec models were outperformed by their TF-IDF alternatives ([Table 2](#), [Figure 1](#)). Hence, TF-IDF models were applied for the prediction of microbiome environments in new metagenomics publications (selected based on the publication being referenced from an ENA study record) and subsequently literature triage.

Classifier features	Class	Precision	Recall	F1-Score	Support
TF-IDF	Engineered	0.86	0.9	0.88*	20
	Environmental	0.86	0.9	0.88*	20
	Host-associated	0.94	0.85	0.89*	20
Doc2Vec ^{MGnify}	Engineered	0.77	0.85	0.81	20
	Environmental	0.89	0.8	0.84	20
	Host-associated	0.85	0.85	0.85	20

Table 2. Precision, Recall and F1-scores of the best performance random-forest biome classifiers. Classifiers features were either TF-IDF or Doc2Vec^{MGnify} (embeddings generated from MGnify cross-referenced publications). Support = number of publications in test datasets per class. *TF-IDF biome classifiers outperformed the ones trained on Doc2Vec^{MGnify}.

Figure 1. Receiver operating characteristic (ROC) curves of biome classifiers, using TF-IDF (A) or Doc2Vec^{MGnify} (B) as training features.

Named Entity Recognition (NER)

To identify metagenomics data in research articles, 16 novel metagenomics entities ([Table 1](#)) were defined, curated, and trained using NER models. Training and test datasets were constructed from the curated 140 ENA cross-referenced publications, which were not contained in the MGnify database to ensure unbiased and broader coverage of biome. Triage publications (see supplementary [Table 2](#)) were randomly selected and equally categorised using highest performance random forest classifiers ([Table 2](#), [Figure 1](#)). The manual curation of metagenomics entities in literature triage yielded 9567 annotations for 2496 sentences. As entities can be nested within each other, 16 individual datasets were constructed separately for fine-tuning 16 BioBERT models. Each entity dataset, comprising 2496 BERT-tokenized sentences with BIO-tagged entities, was partitioned for training (90%; 2246 sentences) and testing (10%; 250 sentences). During training, grid search over 5 learning rates and 7 epochs showed best performance models with diverse combinations of hyperparameters per entity. These varieties of combinations were expected to give the best performance, given the overlapping nature of the entities. Typically, the greater the contexts overlap, the greater the training the model required to achieve the best performance. [Table 3](#) shows the best token-wise precision (P), recall (R) and F1-score (F1) macro-averages for trained NER models per entity. Models achieved a precision of 80-100% and a F1-score between 71-98%.

Entity	Learning Rate	Epoch	Recall	Precision	F1-Score
Ecoregion	4e-5	50	0.95	1	0.98
Host	2e-5	90	0.89	0.93	0.9
Engineered	2e-5	10	0.65	0.93	0.75
Date	4e-5	90	0.78	0.91	0.83
Place	3e-5	90	0.78	0.86	0.82
Site	4e-5	10	0.71	0.85	0.77
Body-Site	4e-5	90	0.98	0.95	0.97
Sample-Material	5e-5	110	0.8	0.9	0.85
State	5e-5	110	0.65	0.8	0.71
Treatment	4e-5	30	0.66	0.8	0.73
Kit	2e-5	70	0.94	0.91	0.92
Primer	5e-5	70	0.94	0.97	0.96
Gene	1e-5	10	0.86	0.92	0.89
LS	5e-5	50	0.8	0.95	0.85
LCM	4e-5	50	0.86	1	0.92
Sequencing	5e-5	110	0.84	0.89	0.87

Table 3: Token-wise precision, recall and F1-score of the 16-best performance NER models.

To validate the accuracy of NER models predictions, 16 new NER models were trained from scratch on uncurated predicted annotations from 1500 new articles. The trained models were then tested in recognizing entities in the curated datasets (140 articles that were not included in the 1500 training articles). [Table 6](#) (supplementary) shows the best token-wise P, R and F1

macro-averages for the NER models trained on predicted entities, where F1-scores were 73-97%. These NER models outperformed those trained on the curated datasets for *date*, *place*, *state*, *treatment*, *kit*, *gene*, *LS*, *LCM* and *sequencing* entities.

Database enrichment

To enrich metagenomics studies with data from research articles, 114,099 publications in Europe PMC were processed and annotated using the highest NER performance models (Table 3). Those publications encompassed 19,900 publications linked to studies in the ENA and MGnify databases.

A new pipeline was developed to continuously provide metagenomics annotations for new open access publications. Users are now allowed to search, explore and retrieve metagenomics annotations programmatically from the Europe PMC website, using search queries (e.g. ANNOTATION_PROVIDER:"Metagenomics"), a SciLite application (Figure 2 : article view) and the annotations API (<https://europepmc.org/AnnotationsApi>), respectively.

4. Materials and Methods

4.1. Collection of Sponge Samples

Sponge samples were collected as part of a research cruise in the **equatorial Atlantic**; the **Tracing Ocean Processes** Using Corals and Sediments (TROPICS) expedition JC094, (13 October 2013–30 November 2013). Five locations were selected for sampling: from **east** to west, the **Carter and Knipovich seamounts** in the **eastern** basin, the **Vema fracture zone** at the **Mid-Atlantic Ridge** and the **Vayda and Gramberg seamounts** in the **western** basin. The **sponges** were subsampled in a **controlled-temperature (4 °C)** laboratory on board the ship before flash freezing them for storage at **−80 °C**. The **sponge** used for the current study was collected at a **depth of 971 m**, from the **Knipovich seamount** in the **eastern** basin of the **Atlantic Ocean** (**5° 37.5038' N, 26° 57.4780' W**). It was taxonomically assigned to the Class **Demospongiae** by microscopic identification of **sigma-C microscleres** during the research cruise (Supplementary Figure S1).

4.2. Sponge Processing and Isolation of Bacterial Strains

The **sponge sample** was prepared and homogenised as described by this group in Williams et al. 2020 [10]. The **sponge homogenate** was serially diluted (10^{-1} to 10^{-4}) with sterile **artificial sea water** (ASW; Crystal Sea Marine Mix, Marine Enterprise International, made to the manufacturer's instructions) and spread onto a variety of agar types [10], in duplicate. All agar was made with **ASW** and supplemented with **cycloheximide (10 µg/mL)** to inhibit fungal growth. To maximise recovery of environmental isolates, media included **sodium pyruvate (100 µg/mL)** [39], **trace metal solution (Trace Metal A5 with co, Sigma-Aldrich; 1 mL/L)** and **Basal Medium Eagle concentrated solution (BME, Sigma Aldrich; 1 mL/L)** [40]. **Nalidixic acid (30 mg/mL)** was also added to some media to inhibit the growth of Gram-negative bacteria. Duplicate plates were incubated at 4 and 28 °C for 4–10 weeks. A large number of **colonies** with a wide variety of morphologies were picked and streaked onto fresh agar of the same type to grow as a monoculture in the same isolation temperature. The axenic strains were then directly stocked from the agar plate into Microorganism Preservation System Protect Cryotubes (Technical Service Consultants Ltd., Heywood, UK) and stored at **−70 °C**.

4.3. 16S rRNA Gene Sequencing of Strain 28ISP2-46^T

Genomic DNA from strain 28ISP2-46^T was purified with a **GenElute Bacterial Genomic DNA Kit** (Sigma) following the manufacturer's protocol. A sample of the resulting DNA solution (1 µL) was mixed with 12.5 µL of CloneAmp HiFi PCR premix (ClonTech), 9.5 µL of ddH₂O, and 1 µL each of the primers **8F (AGAGTTTGATCCTGGCTCAG)** and **rP2 (ACGGCTACCTTGTACGACTT)** at 10 µM. The following thermocycler program was used: initial denaturation (1 cycle): 10 min at 95 °C. Amplification (35 cycles): 1 min at 95 °C, 1 min at 58 °C, 1.5 min at 72 °C. Final extension (1 cycle): 10 min at 72 °C.

Annotations Panel:

- Kit**
 - GenElute™ Bacterial Genomic DNA Kit (2) Find >
 - GenElute Bacterial Genomic DNA Kit (1) Find >
- Primer**
 - 8F (AGAGTTTGATCCTGGCTCAG) (1) Find >
 - rP2 (ACGGCTACCTTGTACGACTT) (1) Find >
- Gene**
 - 16S rRNA (1) Find >
 - kstA-D (1) Find >
 - kstRs (1) Find >
 - kstRg (1) Find >
 - kstU (1) Find >
- LS**
 - amplicon (1) Find >
 - TruSeq (1) Find >
 - shotgun (1) Find >
- LCM**
 - paired-end (1) Find >
- Sequencing**
 - Illumina (2) Find >
 - Illumina MiSeq (1) Find >
 - MinION (1) Find >

This site is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

Figure 2. Screenshot of Europe PMC article. Annotations panel with metagenomics entities and annotations (right). Highlighted annotations in full-text using SciLite tool (left) (SciLite article view).

To date, MGnify has enriched the metadata of 1600 studies with the metagenomics annotations from corresponding articles using the Europe PMC annotations API. Enriched metadata were made accessible in MGnify publications and sample web pages for researchers to explore in metagenomics analyses (Figure 3: study view).

Study MGYS00005760

EMG produced TPA metagenomics assembly of PRJNA343503 data set (fruitshake sequence reads using Oxford Nanopore MinION).

[Overview](#) [Analysis summary](#)

Last updated: Mon Jul 26 2021

Classification

root:Host-associated:Plants

Description

The Third Party Annotation (TPA) assembly was derived from the primary whole genome shotgun (WGS) data set PRJNA343503, and was assembled with Flye v2.8.3. This project includes samples from the following biomes: root:Host-associated:Plants.

Related studies

[MGYS00005761](#)

External links

[ENA website \(ERP128313\)](#)

Publications

EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies, *Nucleic Acids Res* (2018) doi:10.1093/nar/gkx967

Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, et al.

[Show metadata from Europe PMC Annotations](#)

Europe PMC Annotations

Powered by Europe PMC

Through automated text-mining, many publications may have additional metadata available in the form of Annotations from Europe PMC. Any Metagenomics-relevant annotations are shown here.

Sample processing

No Sample processing annotations

Other

Gene

Target gene(s) (e.g. hypervariable regions of 16S/18S rRNA gene)

16S

Mentioned in the Article section: [View on Europe PMC](#) [Definition of '16S'](#)

...The adoption of Infernal using a comprehensive library of ribosomal models means that the pipeline now identifies SSUs (or 16S and 18S rRNAs)....

16S rRNA/ssu

18S rRNA

ssu

ssus

Date

Analyses

Biome	Sample accession	Sample description	Run / Assembly accession	Pipeline version	Analysis accession
	SRS1720301	This sample has been submitted by pdaldkim136 on 2016-10-03; food metagenome	ERZ1949291	5.0	MGYA00585477
	SRS1720301	This sample has been submitted by pdaldkim136 on 2016-10-03; food metagenome	ERZ1949292	5.0	MGYA00585478

Figure 3. Composite screenshot showing annotations-enriched metadata for a MGnify study. Annotated Publications are highlighted within a Study (left) and annotations are shown in context (right).

To further evaluate our approach, the publication-derived metagenomics data were compared against author-submitted metadata for 19,209 studies. Both sets of metadata were mapped to the relevant MIxS checklist, allowing the annotations to be compared. Since both publication-derived metagenomics data and author-submitted metadata were provided as free text, comparing them was a non-trivial task; hence, the comparison of exact matches. This comparison revealed that for the vast majority of MIxS checklist terms, the numbers of publication-derived terms were greater than the numbers obtained from ENA author-submitted metadata, [Figure 4](#). In addition, metagenomics annotations demonstrated high quality coverage for many of the missing metadata in ENA, such as PCR primers (pcr_primers), phylogenetic marker genes (target_gene), nucleic acid extraction kit (nucl_acid_ext), environment phenomena (state), health or disease states (state) and treatment ([Table 4](#)). Moreover, overlaps were observed between metagenomics annotations and ENA metadata in numerous studies, regarding MIxS env_package (12,419 studies), lib_layout (8178 studies), source_uvig (7792 studies), env_medium (7316 studies), geo_loc_name (5571 studies), specific_host (2784 studies), host_spec_range (2784 studies), env_local_scale (2525 studies), seq_meth (1653 studies), collection_date (1002 studies), target_subfragment (241 studies), env_broad_scale (153 studies), health_disease_stat (137 studies), depth (122 studies), lat_lon (93 studies), target_gene (80 studies), pcr_primers (66 studies), elev (26

studies), alt (3 studies) and nucl_acid_ext (1 study). Yet, many studies have shown no or limited overlap, either due to missing essential metadata or having inaccurate, incorrect or inconsistent metadata (i.e. synonyms or unit formats) in ENA compared with their metagenomics annotations counterparts (see supplementary [Table 8](#)).

Figure 4. A clustered stacked bar plot showing the overlap of metadata terms between ENA (author submitted terms that are part of the study) and those derived from Europe PMC (EPMC) articles using our framework. For each study, we have evaluated twenty of the most populated MIxS fields in 19,209 studies, to establish if the metadata provided by both sources were identical (green), non identical (yellow), or provided by one of the metadata sources only (unique, red). For each MIxS field, each stacked bar shows the number of studies having identical, non identical and unique metadata from each source.

Study	PMC	MIxS	ENA Metadata	Metagenomics Annotations
PRJDB8863	PMC6941062	pcr_primers	-	RNA (rRNA) genes (forward: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGCCAGCMGCCGCGGTAA-3'; reverse: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGACTACHVGGGTWTCTAAT-3')
PRJDB9293	PMC8147061	target_gene	-	16S rRNA, V3-V4
PRJDB5614	PMC5745017	nucl_acid_ext	-	RNA PowerSoil total RNA isolation kit
PRJEB27411	PMC6072794	Env_package (state)	-	100-year drought, 75.6 mm of rainfall, mesotrophic, denitrification
PRJEB15392	PMC5405060	health_disease_stat- (state)	-	caries, gingivitis, medically healthy, oropharyngeal mucositis, poor oral hygiene, pulp diseases
PRJEB22207	PMC7044117	Env_package (treatment)	-	Antibiotic, benzylpenicillin, cefotaxime, gentamicin, meropenem, metronidazole, probiotic supplementation, vancomycin
PRJDB10581	PMC8151423	env_local_scale (body-site)	-	Oral, Rectal, cervical, posterior vaginal fornix

Table 4. This table provides some examples of how the metagenomics annotations extracted from the publications describing ENA records enrich the metadata for those studies. For example, in the last row of the table, the project PRJDB10581 is linked to PMCID PMC8151423, from which the terms “Oral, Rectal, cervical, posterior vaginal fornix” have been extracted to supplement the term “body-site” from the ENA record. Other examples include metadata about PCR primers (pcr_primers), marker genes (target_gene), nucleic acid extraction kit (nucl_acid_ext), body site (env_local_scale), environment phenomena (state), health or disease states (state) and treatment.

Discussion

In this work, we provide a machine learning (ML) framework that enriches a wide variety of microbiome studies (19,209) with essential and accurate metadata from open access research articles (114,099). This framework has been integrated into Europe PMC and further propagated to MGnify to regularly provide metadata for up-to-date diverse metagenomics studies, derived from host-associated (living organisms), ecoregion (natural environment) and engineered (man-made) environments, from research articles. To triage articles for curation, multiclass random forest models were trained to classify publications into these broad biome categories, which covered various and rich linguistic contexts for a wide range of organisms (e.g., mosquitoes, birds, chicken, ruminants and human), ecoregions (e.g., forests, oceans,

rivers, lakes) and engineered environments (e.g., bioreactors, wastewater treatment plants, food, and microbial fuel cells) (see supplementary [Table 2](#)). Subsequently, introducing 16 novel entities that flexibly accommodate the essential metadata of these environmental categories and training BioBERT NER models on their miscellaneous contexts have shown their utility for addressing the limitations of current databases and ontologies in describing extensive ranges of novel or hybrid microbiome environments.

Although metagenomics NER models demonstrated precisions of 80-100% ([Table 3](#)) for most of the entities, relatively lower F1-scores were observed with the categories engineered, site, state and treatment. Qualitative assessment of these entities revealed that their metadata can be represented by single (unigrams) or multiple tokens (ngrams). Curators seemed to be more comprehensive in annotating entity-related tokens when compared to models, which tend to generate few, yet precise annotations; hence, the relatively low token-wise recall (see supplementary [Table 9](#)). Accordingly, low F1-scores seem not necessarily reflecting poor model performance, but rather suggest that BERT-based models can perform better than curators in learning the contexts and patterns of weakly imprecise data³². Another limitation of the training dataset is that the metadata of a small subset of entities (e.g., primer, place and date) can be listed in publications tables, where rows and cells were merged and parsed as one sentence that might exceed the maximum sequence length (128) of model input. This resulted in the exclusion of tokens beyond maximum input length from being annotated by the model and subsequent low recall. On the other hand, training 16 individual NER models (one model per entity) on all curated sentences provided extensive negative dataset and considerable advantage for training overlapping entities, where each entity model has shown to require specific learning rate and training epochs to predict test datasets with higher precision. Alternatively, training one multiclass NER model on all entities was deemed to perform poorly due to the unbalanced BIO class distribution of single- (non-overlapping) versus multiclass (overlapping) entities, high number of trained entities (16) and small training dataset. Moreover, modifying BERT tokenizer to keep and unsegment words with long lengths (e.g., primers) or unknown characters (e.g., kits, chemical states) managed to improve NER model performance, via retaining their representations by NER models instead of excluding them as unknown tokens (UNK) from input data. It is worth noting that pretraining BioBERT models on new PMC publications would have enabled the recognition of those unknown tokens during fine-tuning. But, due to the computationally expensive generation of contextual word representations from millions of publications, neither BioBERT models or analogous neural architecture models, such as XLNet³⁶, RoBERTa³⁷ or ELECTRA³⁸, were pre-trained or compared. Overall, these training approaches and modifications have been shown to contribute to high prediction accuracy; hence, the implementation of metagenomics NER models in enriching the metadata of diverse microbiome studies in ENA and MGnify from Europe PMC publications.

We also demonstrated that the high accuracy of metagenomics NER models can extend to uncurated datasets. Training BioBERT-based NER models from scratch on large uncurated predicted datasets, generated from metagenomics NER models annotations, revealed the reversible transfer of their reliability and outperformance in predicting curated datasets with precision of 76-99% (see supplementary [Table 6](#)). Moreover, evaluating the accuracy of

model-derived annotations against ENA author-submitted metadata showed large overlaps between them in numerous studies regarding MIXS biome environmental packages, sequencing library strategy, material, location, host, sequencing methods, and collection dates ([Figure 4](#)). Conversely, although model-derived annotations showed greater MIXS coverage for more studies, limited overlaps were noticed due to incorrect, inaccurate, or inconsistent author-submitted data formats (see supplementary [Table 8](#)). The analysis is not without limitations, as it covered metadata only from ENA fields used in more than 50 studies. However, it still offers some interesting insights. For instance, metagenomics annotations provided metadata for nearly 2810 studies that had no sample or experimental metadata in ENA; hence the enrichment of ambiguous sequence datasets to a more meaningful and exploitable dataset. Examples of such cases include a study about the bacterial and fungal composition of soybean curd in tofu factories (PRJDB10470; PMC7563423) and the discovery of new bacterial candidate phylum, *Candidatus Kryptonia*, in Nevada hot springs (PRJEB11785; PMC4737851). In addition, exploring geographical metadata in ENA revealed that metagenomics annotations can enrich missing geographical locations for 7541 studies, including 336 studies in MGnify. Examples from those studies are a study about taxonomic composition of soil viruses along the Namib Desert (PRJEB11968; MGYS00000581; PMC5256219) and a study about the bacterial community identified from sequencing neanderthal genome from a 38,000-year-old fossil (PRJEB1198; MGYS00000318; PMC3643900). Thus, enriching ENA with normalised model-predicted metadata is worth pursuing.

As demonstrated in [Figure 4](#), this framework captures metadata terms that have a high degree of overlap to the corresponding metadata found in ENA, even when considering only exact terms. Moreover, the additional metadata identified through this framework will facilitate the more extensive use of existing datasets, which have been collected and sequenced at great expense and cannot simply be recreated. The expanded set of metadata will also simplify new meta-analyses, for example the analysis of amplicon sequence variants (ASVs)³⁹, a method that can compare taxonomic profiles between different studies. However, ASV analysis requires that only equivalent regions are compared. Thus, the expansion of metadata related to amplicon datasets, namely “target gene”, “target subfragment” and “PCR primer”, means that datasets can now be rapidly selected using a single API call to Europe PMC, without the labouring reading of papers that would otherwise be required. Note, that for very common terms like “rRNA”, it may be necessary to pair this with other terms, such as “library strategy” to ensure that the queries are restricted to amplicon-based studies. Furthermore, with the enriched metadata for shotgun metagenomics datasets, it will become easier to control for confounding factors by either producing closely paired datasets or statistically evaluating the significance of metadata differences. The wide range of metadata fields and biomes covered by the Europe PMC Submission System for Annotations means that these results are applicable to a broad range of researchers. This system is already linked to MGnify to enable the dynamic retrieval and display of additional metadata terms alongside users submitted annotations. Given that MGnify provides a uniform analysis across databases, this eliminates confounding factors from bioinformatics analysis. Combined, the enriched metadata and uniform analyses will help eliminate many of the confounding factors

that currently exist, making it easier to establish the link between compositional changes in the microbiota and phenotypic differences, and the identification of the underlying biological mechanism(s).

Methods

Literature classification

Training dataset. Due to the lack of public datasets for training metagenomics-specific literature classifiers, we constructed our own training dataset based on MGnify, which provides access to a broad range of metagenomics studies and associated analysis results, but importantly curates microbiome studies with standardised biome annotations and provides direct links to their relevant publications. Accordingly, a supervised training dataset was constructed by mapping the curated GOLD biome annotations assigned to the metagenomics studies in MGnify (<https://www.ebi.ac.uk/metagenomics/>) to the corresponding publications in Europe PMC describing these studies (674 articles accessed on 17 February 2020). First, metagenomics studies were collected from MGnify with both PMC identifiers and GOLD annotations. Then, the full-text XML of publications corresponding to the PMC identifiers were retrieved with corresponding metadata from Europe PMC. Following XML preprocessing, sentences were parsed from the XML tags corresponding to each article section (i.e. introduction, method, results, discussion) and merged into individual section texts. Each text was aligned to the GOLD annotation of the corresponding metagenomic study and TF-IDF (Term Frequency-Inverse Document Frequency) or a vector of precomputed 200-dimensional Doc2Vec embedding⁴⁰ was used as features for biome classifiers. To generate Doc2Vec embeddings, publications cited in MGnify records were trained using the document-embedding neural network model Doc2Vec with the following parameters: 15 training iterations, a window size of 5, worker threads of 8 and a dimensional size of 200 (Python genism package 3.8: Doc2Vec).

Training random forest biome classifiers on cross-referenced publications. A total of 5 GOLD multiclass hierarchical levels (supplementary [Table 1](#)) were used to train nearly 50 random forest models (Python Scikit-learn package 0.22.1: RandomForestClassifier). For each of the TF-IDF and Doc2Vec approaches, a separate multiclass model was trained using the selected features for every hierarchical level. Each model trained a random selection of balanced datasets that comprised 20-100 texts per class. To validate model performance, datasets were divided into training datasets (80%; 80 publications) and test datasets (20%; 20 publications). The hyperparameter settings of random forest models were determined using a grid search among {50, 100, 150, 200, 250, 300} decision trees in the forest (`n_estimators`) and {25, 50, 75, 100, 125, 150, None} maximum tree depths (`max_depth`). Features selection was assessed through 5-fold shuffle-split cross-validation stratified by classes labels and random state was fixed at model initialization for reproducibility. The precision (P), recall (R) and F1-score (F1) model performance metrics were compared and evaluated per hierarchical level.

Biome classifiers prediction and implementation. A query was constructed to retrieve all publications associated with all metagenomics studies from the ENA database (determined by

the descriptor “library source: METAGENOMIC” via <https://www.ebi.ac.uk/ena/browser/home> and XREF APIs (<https://www.ebi.ac.uk/ena/xref/rest/#!/JSON/getSourcesUsingGET>). This resulted in 4513 publications from 9949 ENA metagenomics studies accessed on 20 February 2020. For each publication, full texts were then retrieved and processed using the same workflow employed in constructing the classifiers training dataset. The highest performing model with the broadest biome coverage was then selected to classify ENA cross-referenced publications. Using the model prediction probabilities, publications with probabilities of $P \geq 0.4$ were selected to have higher publications coverage per category with 70-95% precision. Lastly, a random sample of ENA cross-referenced publications that represents the breadth of biome classes (44-50 publications per class) were manually verified to validate the predicted biomes and subsequently used for the literature triage (supplementary [Table 2](#)).

Named Entity Recognition (NER)

Training dataset. A literature triage comprising 140 ENA cross-referenced publications were categorised into environmental (N=50), engineered (N=44), and host-associated (N=46) publications (supplementary [Table 2](#)). A total of 16 novel metagenomics entities, covering biome and experimental data, were proposed, and defined for biocuration ([Table 1](#)). Words pertinent to the aforementioned entities were curated in the publications validated for biocuration by a team of experienced curators, using Hypothes.is (<https://web.hypothes.is/>) a web-based open-source annotation tool. Every manuscript was assigned to a single curator to annotate the entities under the supervision of another curator, followed by revisions from 2 different curators (see supplementary [curation guidelines](#)). After evaluation, annotations were retrieved via the Hypothesis API (<https://h.readthedocs.io/en/latest/api-reference/>) and mapped to their corresponding sentences. Those sentences were then tokenized with BERT WordPiece tokenizer⁴¹, which was modified to retain words that are unknown to the BERT model (e.g., long primers, kits or chemical states with unknown characters), to be represented and trained by NER models instead of being excluded as unknown tokens (UNK). As entities can be nested within each other (overlapping entities), 16 individual training datasets were constructed to train each entity, separately. Each training dataset consisted of all curated sentences with their entity-related tokens tagged, by the standard NER classification scheme BIO⁴², as the beginning (B), inside (I) or outside (O) of an entity.

Entity	Definition
Ecoregion	Microbiome natural environment
Host	Microbiome living organism or host
Engineered	Microbiome man-made environment
Date	Microbiome sample collection date
Place	The place of microbiome environment or host
Site	The site of microbiome sample within place
Body-Site	The organ or tissue of microbiome sample
Sample-Material	The material of the microbiome sample (e.g., water, mucus, soil)
State	The state of microbiome environment or host (e.g., disease)
Treatment	Any treatment performed on the host (e.g., medicine) or the environment (e.g., fertiliser) from which the sample was collected
Kit	DNA extraction kit
Primer	PCR primers
Gene	Microbiome target genes, e.g., rRNA subunit and amplified region(s)
LS	Library source or library strategy (e.g., amplicon, whole genome)
LCM	Library construction method or layout (e.g., paired-end, single-end)
Sequencing	Sequencing platform

Table 1: Metagenomics entity types

Training NER models on curated datasets. A total of 16 BioBERT models were trained to recognize metagenomics entities in publications, yielding 16 new metagenomics NER models. BioBERT pretrained weights generated from pre-training BERT on PubMed abstracts and PMC full texts were fine-tuned with each of the 16 training datasets. To validate the performance of each model, each dataset was divided into a training dataset (90%) and test dataset (10%). Fine-tuning was performed with a batch size of 32, maximum sequence length of 128 and fixed random state for reproducibility. A grid search was performed over 5 learning rates {1e-5, 2e-5, 3e-5, 4e-5} and 7 epochs {10,30,50,70,90,110,130} to select the highest performance hyperparameters for each model; thus, different hyperparameters for each training model. The token-wise macro-average precision (P), recall (R) and F1-score (F1) were compared and evaluated per each learning rate and epoch combination, separately.

NER models prediction and implementation. A total of 16 high-performance NER models (BioModels⁴³ accessions: [MODEL2202160002](#), [MODEL2202170001](#)-[MODEL2202170015](#)) were used to annotate the methods sections of 114, 099 metagenomics publications. Since the best model performance metrics were observed when biome classifiers were trained on method texts (see supplementary [Table 3](#)), method sections were identified as the most representative article section for metagenomics data. The 114, 099 metagenomics publications were identified as follows: 1) cross-referencing metagenomics studies in genome and biosamples databases (ENA, NCBI Taxonomy, BioProject and MGnify) against literature repositories and 2) querying literature repositories (Europe PMC, PMC and PubMed) using metagenomics- and microbiome-related search and MeSH terms (for queries, see supplementary [Table 4](#)). Then, methods sections were extracted from full-text XML via parsing the sentences wrapped within methods XML tags. Because metagenomics publications sometimes lack scientific methods sections, a multiclass random forest model

was trained to classify full-text into abstract-, introduction-, methods-, results- and discussion-related texts. This section classifier (see supplementary [Table 5](#) and [Figure 1](#)) was then used to extract methods-related text from publications lacking methods sections.

Training NER models on predicted datasets. To further demonstrate the accuracy of predicted annotations, a random subset of new metagenomics publications (1500 articles) was retrieved with their NER-models-predicted annotations to train new 16 NER BioBERT models from scratch. To construct this subset, trained biome classifiers were used to classify new metagenomics publications into 3 broad biome environments. Then, a random subset of 500 articles, with predicted probabilities of $P \geq 0.5$, was selected per biome class (i.e., engineered, environmental and host-associated classes), followed by the retrieval of their NER-models-predicted annotations. In this approach, those new publications with their NER-predicted annotations were considered as the training dataset, whereas the curated dataset was considered as the test dataset. This approach is like the student self-distillation process used in improving ImageNet classification⁴⁴ and protein structure prediction (AlphaFold)⁴⁵ models, where student models were trained on datasets predicted by teacher models (models trained on curated dataset). Though in our approach, it was hypothesised that teacher NER models - models trained on curated datasets - can generalise to novel biome publications, if students NER models - models trained on datasets predicted by teacher models - predicted curated datasets with token-wise macro-average F1-score $\geq 70\%$.

Database enrichment

Europe PMC enrichment. Metagenomics annotations of 114,099 publications in Europe PMC were represented in the annotations data model provided by the Europe PMC annotations submission system (<https://europepmc.org/AnnotationsSubmission>). To make it easier to categorise and search publications with metagenomics annotations, specific metagenomics ontologies were required to standardise the predicted annotations. However, with the surge in different environments being sampled, it became clear that metagenomics ontologies, including GOLD and ENVO, are insufficient for standardising most of the predicted annotations. Thus, ZOOMA (<https://www.ebi.ac.uk/spot/zooma/>), a semantic ontology mapping tool was used to map each of the predicted non-standardized annotations to a wide range of ontologies in the Ontology Lookup Service (OLS; <https://www.ebi.ac.uk/ols/index>). Processed annotations were then uploaded into the Europe PMC annotations API (<https://europepmc.org/AnnotationsApi>) for users to search, explore and retrieve programmatically.

ENA and MGnify enrichment. Metagenomics annotations of publications describing 19,209 studies in ENA were amassed from the Europe PMC annotations API. To assess the enrichment of experimental and sample metadata, metagenomics annotations (NER-models-predicted annotations) were compared with the corresponding author-submitted metadata for 19,209 metagenomics studies in ENA. However, no sample or experimental metadata were associated with nearly 2810 studies in ENA. For each study, sample, experiment and run records identifiers - accessions - were parsed from ENA-SAMPLE, ENA-EXPERIMENT and ENA-RUN tags in the study XML, respectively, followed by the retrieval of run and

experimental metadata from the identifier-associated records. Then, the entities of metagenomics annotations and a subset of ENA fields used in more than 50 studies were mapped to a corresponding subset of the minimum information about any (x) sequence (MIxS)³³ checklist (see supplementary [Table 7](#)). For each MIxS term, data were pooled from mapped ENA fields (author-submitted metadata) and metagenomics entities (metagenomics annotations) and compared per study. Studies with any identical or non identical metadata in both sources, or having metadata from one source only were then counted per each MIxS term, with the results from the twenty most common MIxS fields shown in Figure 4.

Data Availability

A simple version of metagenomics annotation pipeline with the training datasets and models are freely available from https://gitlab.com/maaly7/emerald_metagenomics_annotations.

Acknowledgement and Funding

The authors would like to thank the curators from Molecular Connections (<https://www.molecularconnections.com/>) for their biocuration efforts, Zhan Huang and Lynne Faulk (EMBL-EBI) for user interface design and implementation, Audrey Hamelers (EMBL-EBI) for proofreading the manuscript, and Awais Athar and Santosh Tirunagari (EMBL-EBI) for testing and validating the Gitlab repository. Francesco Talo' and Zunaira Shafique are funded by the Europe PMC grant, provided by 33 funders of life science research (<https://europepmc.org/Funders/>) under Wellcome Trust Grant 221523. Maaly Nassar and Santiago Sanchez were funded by Biotechnology and Biological Sciences Research Council (BB/S009043/1).

Additional Information

Competing Interests: The authors declare no competing interests

Authors Contributions: Maaly Nassar contributed to the study design and curation of training datasets, trained and validated the machine and deep learning models, developed the metagenomics annotations pipeline, analysed the data and wrote the manuscript. Alexander Rogers integrated metagenomics annotations into the MGnify database. Francesco Talo' and Zunaira Shafique integrated metagenomics annotations into the Europe PMC annotations platform. Santiago Sanchez contributed to the curation and revision of training datasets. Robert Finn and Johanna McEntyre conceived the project, contributed to the study design and contributed to, and reviewed, the manuscript.

References

1. Reitmeier, S. *et al.* Arrhythmic Gut Microbiome Signatures Predict Risk of Type 2 Diabetes. *Cell Host & Microbe* **28**, 258-272.e6 (2020).
2. Oh, T. G. *et al.* A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. *Cell Metabolism* **32**, 878-888.e6 (2020).

3. Weber, L. *et al.* Microbial signatures of protected and impacted Northern Caribbean reefs: changes from Cuba to the Florida Keys. *Environmental Microbiology* **22**, 499 (2020).
4. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**:7581 **528**, 262–266 (2015).
5. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**:7834 **587**, 448–454 (2020).
6. Nayfach, S. & Pollard, K. S. Toward Accurate and Quantitative Comparative Metagenomics. *Cell* **166**, 1103–1116 (2016).
7. Mirzayi, C. *et al.* Reporting guidelines for human microbiome research: the STORMS checklist. *Nature Medicine* **27**:11 **27**, 1885–1892 (2021).
8. Yilmaz, P. *et al.* The genomic standards consortium: bringing standards to life for microbial ecology. *The ISME Journal* **5**, 1565–1567 (2011).
9. Kodama, Y., Shumway, M. & Leinonen, R. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Research* **40**, D54 (2012).
10. Harrison, P. W. *et al.* The European Nucleotide Archive in 2020. *Nucleic Acids Research* **49**, D82 (2021).
11. Mitchell, A. L. *et al.* EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Research* **46**, D726 (2018).
12. Wilke, A. *et al.* The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Research* **44**, D590 (2016).
13. Chen, I.-M. A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research* **45**, D507 (2017).
14. Corrêa, F. B., Saraiva, J. P., Stadler, P. F. & da Rocha, U. N. TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic acids research* **48**, D626–D632 (2020).
15. Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).
16. Kasmanas, J. C. *et al.* HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Research* **49**, D743 (2021).
17. Mukherjee, S. *et al.* Genomes OnLine Database (GOLD) v.8: Overview and updates. *Nucleic Acids Research* vol. 49 D723–D733 (2021).
18. Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J. & Lewis, S. E. The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics* **4**, 43 (2013).
19. Bagheri, H., Severin, A. J. & Rajan, H. Detecting and correcting misclassified sequences in the large-scale public databases. *Bioinformatics* **36**, 4699–4705 (2020).
20. Pappas, K. M. *et al.* PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. *Microorganisms* **2022**, Vol. 10, Page 293 **10**, 293 (2022).
21. Chaix, E., Deléger, L., Bossy, R. & Nédellec, C. Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiology* **81**, 63 (2019).
22. Arighi, C. N. *et al.* BioCreative-IV virtual issue. *Database: The Journal of Biological Databases and Curation* **2014**, (2014).

23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* (2013).
24. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* (2013).
25. Habibi, M., Weber, L., Neves, M., Wiegandt, D. L. & Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **33**, i37 (2017).
26. Giorgi, J. M. & Bader, G. D. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics* **34**, 4087 (2018).
27. K, X., Z, Y., P, K., Q, W. & W, L. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. *Computers in biology and medicine* **108**, 122–132 (2019).
28. Luo, L. *et al.* An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **34**, 1381–1388 (2018).
29. Dang, T. H., Le, H.-Q., Nguyen, T. M. & Vu, S. T. D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. *Bioinformatics* **34**, 3539–3546 (2018).
30. Vaswani, A. *et al.* Attention is all you need. in *Advances in Neural Information Processing Systems* vols. 2017-December 5999–6009 (Neural information processing systems foundation, 2017).
31. Peters, M. E. *et al.* Deep Contextualized Word Representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* **1**, 2227–2237 (2018).
32. Devlin Google, J. & Language, A. I. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Bidirectional Encoder Representations from Transformers)*.
33. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz682.
34. Fries, J. A. *et al.* Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications* **2021 12:1** **12**, 1–11 (2021).
35. Ferguson, C. *et al.* Europe PMC in 2020. *Nucleic acids research* **49**, D1507–D1514 (2021).
36. Yang, Z. *et al.* XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv* (2019).
37. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019).
38. Clark, K., Luong, M.-T., Le, Q. v. & Manning, C. D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv* (2020).
39. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **2017 11:12** **11**, 2639–2643 (2017).
40. Le, Q. v. & Mikolov, T. Distributed Representations of Sentences and Documents. *31st International Conference on Machine Learning, ICML 2014* **4**, 2931–2939 (2014).
41. Wu, Y. *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. (2016).

42. Sang, E. F. T. K. & Buchholz, S. Introduction to the CoNLL-2000 Shared Task: Chunking. 127–132 (2000).
43. Malik-Sheriff, R. S. *et al.* BioModels—15 years of sharing computational models in life science. *Nucleic Acids Research* **48**, D407–D415 (2020).
44. Xie, Q., Luong, M. T., Hovy, E. & Le, Q. v. Self-training with Noisy Student improves ImageNet classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 10684–10695 (2019) doi:10.1109/CVPR42600.2020.01070.
45. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Nassar2022supplementary.pdf](#)