Data Accessibility Paper

Joanna Colovas Adena Collens Patrick D. Schloss

Sep 18, 2025

- Abstract
- Importance
  - Incentivize authors to publish/make available their original data
  - Publishing data helps get more use out of research
  - Helps eliminate file drawer effect as it shows negative data
- Keywords
  - Data accessibility
  - Data reproducibility

Introduction

Data availability (DA) is the practice of making raw experimental data and analyses publicly accessible, often via upload in maintained databases. DA is a newly emerging and yet deeply important component of the scientific process in the digital age. With the latest and greatest methodologies available across fields, increasing amounts of data are being generated each day, especially in the biological sciences. Availability of these large quantities of study data and metadata (data about data) is a necessary resource for appropriate use and re-use of data and protocols as well as the recreation of analyses. We believe that policies of data "available on request" are not sufficient to be considered available data. One example was published in *Microbiome*, that a reader

in search of data may email the corresponding author, with varying results ((1)). We believe that this is simply unacceptable, and set out to determine rates of published raw data in the biological sciences, specifically microbiology, by examining the American Society for Microbiology's (ASM) library of published primary research journals.

Scientific Data as a Public Good

The United States Government spent over two hundred million dollars (USD) in 2024 on research expenditures ((2)). The result of all of these investments are data, paid for in part by taxpayers. Therefore, data is a public good, and best used as a benefit to those who provided the funds for it. Once data has been generated, it can be used not only for initial analyses, but over and over again in future studies or meta analyses. Additionally, data can be used to eliminate possible solutions to a problem by the publishing of "negative data." Thinking of data as a public good, if negative data is published, it can help researchers avoid sinking time and financial resources into the investigation of non-viable hypotheses. This lack of publication of non-fruitful investigation is more commonly known as the "file drawer effect." Data as a public good also is subject to the tragedy of the commons. If no one contributes to the public "commons", data that is available for public use, how can we advance our understanding of our study systems?

Current DA Policies

Current data availability guidelines have been informed by a number of policies created by funding agencies, peer-review journals, conference and special task groups, as well as community interest groups. In 2011, after the Future of Research Communication (FoRC) conference in Germany to establish FORCE11, a community interest group which seeks to encourage and promote data availability standards. Also in 2011, the Genomic Standards Consortium (GSC) published the MIMARKS/MIxS standards in *Nature Biotechnology* to promote the publication of the "minimum information about a marker gene sequence" (MIMARKS) or "minimum information about x sequence"

(MIxS) ((3)). These standards are checklists usable by data generators and uploaders towards inclusion of relevant data with sequence uploads in the International Nucleotide Sequence Database Collaboration (INSDC).

In 2014 the group published the Joint Declaration of Data Citation Principles (JDDCP), a document towards the standardization of data citation and its future availability((**Data?**) Citation Synthesis Group 2014). The Findable, Accessible, Interoperable, and Reuseable (FAIR) data science guiding principles were put forth in 2016 by Wilkinson et al in *Nature Scientific Data* urges readers to "improve the infrastructure supporting the reuse of scholarly data" ((4)). Neither the JDDCP nor the FAIR principles are enforcable by any agency. Finally, the National Institutes of Health (NIH) began enforcing the "Policy for Data Management and Sharing" (NOT-OD-21-013) in January of 2023, requiring NIH funded studies to submit a data management and sharing plan (DMS) with their funding applications, and comply with their DMS plan after generation and publication of the funded work((**NIH2023?**)). Non-compliance with NOT-OD-21-013 is identified by funding agencies during annual Research Performance Progress Reports (RPPRs), and may impact future funding decisions ((**NIH2023?**)).

ASM Journals

With the advent of next generation sequencing, microbiology research has generated large amounts of sequencing data, and it is common to upload sequence data to a public repository as well as to include data in research publications. The American Society for Microbiology(ASM) is the major professional body recognized by microbiologists. They have eighteen journals, thirteen primary research journals, three review journals, and two archive journals. In addition, several journals have been folded into others or renamed over time. The ASM family of journals requires that authors "make data fully available, without restriction, except in rare circumstances" ((**ASM?**) open data policy). They have adapted this policy from journals *Microbial Genomics* and *PLOS*. In the ASM open data policy they describe the use of a "Data Availability Statement" which includes "data description, name(s) of the repositories, and digital object identifiers (DOIs) or accession numbers"

and encourages publishing data on relevant public repositories ((**ASM?**) website open data policy). Consequences of non-compliance to the ASM open data policy include contacting research article authors to inform of non-compliance, publication of an "Expression of Concern" for the author and their compliance issues, sanctions on publication in ASM journals, as well as contacting the affiliated research institution and/or funding agencies of the authors ((**ASMcompliance?**)). We endeavor to evaluate how well the microbiology community is using reproducible data practices as we believe that this group of researchers will be early adopters of the technologies available as a result of both the ASM and NIH policies towards data availability.

Nucleic Acid Sequencing Efforts

Beginning in 1996 with the International Strategy Meeting on Human Genome Sequencing in Bermuda, researchers have prioritized the release of all human genome sequencing information so that it may "maximize its benefit to society" (@bermuda1996). The meeting participants agreed that "primary sequence data should be rapidly released", with "sequence assemblies [to] be released as soon as possible, in some centres, assemblies of greater than 1 kb would be released automatically on a daily basis", and that "finished annotated sequence should be submitted immediately to public databases" (@bermuda1996). The "Bermuda Principles," as they became known, have been embraced by the large scale Human Genome Project (HGP) since 1998. In 2003, another meeting, held in Ft. Lauderdale, FL, re-affirmed the 1996 Bermuda Principles, expanded upon them to apply more broadly towards sequencing data, and called for further support of these practices (@ftlauderdale). These foundation agreements set the stage for both the HGP and the Human Microbiome Project (HMP) to generate and share massive amounts of data over the course of their studies (@HGP/HMP source needed).

Starting with projects such as the HGP and HMP, nucleic acid sequencing efforts have been commonly uploaded and released using public databases. There are three major databases worldwide to support sequencing and sharing efforts. The National Library of Medicine's (NLM) National Center for Biotechnology (NCBI) in the United States, the Research Organization of

Information Systems' (ROIS) National Institute of Genetics (NIG) in Japan, and the European Molecular Biology Lab's (EMBL) European Bioinformatics Institue (EBI) in Europe. These three databases are part of the International Nucleotide Sequence Database Collaboration (INSDC). These large databases make research by comparison possible. Genetic lineages of microbes are determined by creating phylogenetic trees which compare a new sequence to existing sequences. Phylogenetic trees show how closely related a new microbial genetic sequence is related to others studied before both in terms of evolution and mutation and in structure and function. An important tool for creating phylogenies is the NCBI Basic Local Alignment Search Tool (BLAST) ((**altschul1990?**)). The BLAST algorithm allows users to compare a nucleic acid or protein sequence to the NCBI database of over 1TB of data to find similar and related sequences. Without the upload of sequences to the NCBI database, the use and success of BLAST would not be possible, despite the effort required on part of the researcher to upload of sequences to one of the INSDC databases.

Examples of Data Avalability

A key tenet of the scientific method is the ability to replicate scientific findings to ensure that they are not due to error. One way that scientific findings can be replicated is by re-completing the same analyses by another researcher. This is only possible if the data used to complete the original analyses is available for use. The availability of datasets also allows new questions to be answered with existing data or the combination of multiple datasets, such as the use of the Human Microbiome Project's (HMP) sequencing data by researchers to create over 650 scientific publications ((**HMP?**)), and the completion of metadata studies. Availablity of data contributed to the rapid sequencing of the SARS-CoV-2 virus during the 2020 pandemic and subsequent expedition of vaccine development ((**needCOVIDref?**)).

With microbiologists commonly uploading nucleic acid sequences to public databases, the aim of this study was to determine the current state of data availability in twelve primary research journals from the ASM family of journals. Primary research articles were classified with using two machine

learning models to answer two questions; “Does this paper contain new sequence data?”, and “Is the data available?” Once these questions were answered, we moved to analyses to answer further questions, “How does making my data available impact my citation metrics over time?”

Results

General Description of the Experiment

We set out to determine the current state of data availability in twelve primary research journals from the ASM family of journals. This objective was completed by first acquiring all papers published in the journals of interest between 2000 and 2024 using the Crossref database and command line tools. We then trained random forest machine learning models to differentiate if each paper contained “New Sequencing Data” (NSD) and then if the paper had “Data Available” (DA). To avoid overfitting the model, we trained each model multiple times, performing validations on a subset of data after each iteration. This allowed us to have a greater number of papers in the training dataset, as well as to have great accuracy and precision within our models. Using our trained models, we were able to classify over 150,000 papers from the whole dataset to determine if they were NSD or DA. After this, we could perform statistical modeling to describe the data, and generate summary statistics. We were especially interested in the ways in which NSD and DA impact citation metrics.

ASM Journals

We used twelve of the ASM primary research journals in this study. Of note, several journals had changes to their publication goals during the 2000-2024 time period. The *Journal of Bacteriology* was the primary place to publish new genome announcements until 2013 when ASM announced journal *Genome Announcements* as a more permanent place for this type of data. *Genome Announcements* was active from 2013 until 2018, when it was re-branded to *Microbiology Resource Announcements*, which has been active from 2018 until present. These two journals appear separately in our analyses as a result of the Crossref database. New genome announcements

have a high percentage of NSD papers, and a high percentage of DA within those papers. As a result, the *Journal of Bacteriology* has had fewer NSD papers, and fewer papers with DA since 2013. Another journal of note is *Microbiology Spectrum* and its re-brand. From 2013 until the fall of 2021, *Microbiology Spectrum* was a review journal. After this point, *Microbiology Spectrum* became a primary research journal. Review journals are less likely to publish articles with NSD, and to have DA. Several journals, including *Microbiology Spectrum*, do not span the entire time period for the study. Journals *mBio* (b.2010), *Microbiology Spectrum* (b. 2013, re-brand 2021), *mSphere* (b. 2016), *mSystems* (b. 2016), and *Genome Announcements* (2013-2018). Not all journals are equally likely to contain NSD and have sequencing DA as a result of their field of interest. Journals *Antimicrobial Agents and Chemotherapy; Infection and Immunity;* and the *Journal of Microbiology and Biology Education;* are all less likely to contain NSD and have DA than the other journals in the dataset.

Descriptive Statistics:

Table 1: Summary Statistics of Model Training Dataset

container.title	n_total	n_fract	n_nsd	fract_nsd	n_da	fract_da
Antimicrobial Agents and Chemotherapy	82	7.846890	23	28.04878	10	43.47826
Applied and Environmental Microbiology	92	8.803828	38	41.30435	28	73.68421
Genome Announcements	30	2.870813	29	96.66667	29	100.00000
Infection and Immunity	76	7.272727	16	21.05263	9	56.25000
Journal of Bacteriology	75	7.177034	25	33.33333	15	60.00000
Journal of Clinical Microbiology	77	7.368421	23	29.87013	14	60.86957
Journal of Microbiology & Biology Education	60	5.741627	0	0.00000	0	NaN
Journal of Virology	87	8.325359	21	24.13793	6	28.57143
mBio	63	6.028708	17	26.98413	11	64.70588

container.title	n_total	n_fract	n_nsd	fract_nsd	n_da	fract_da
Microbiology Resource Announcements	155	14.832536	146	94.19355	145	99.31507
Microbiology Spectrum	111	10.622010	49	44.14414	38	77.55102
mSphere	57	5.454546	25	43.85965	20	80.00000
mSystems	80	7.655502	45	56.25000	45	100.00000

Table 2: Summary Statstics of All Data

container.title	n_total	n_fract	n_nsd	fract_nsd	n_da	fract_da
Antimicrobial Agents and Chemotherapy	20297	13.15501232765		13.6227029	897	32.44123
Applied and Environmental Microbiology	21853	14.16349637261		33.22655932546		35.06404
Genome Announcements	6714	4.3515176	6628	98.71909446546		98.76282
Infection and Immunity	14490	9.3913449	1287	8.8819876	264	20.51282
Journal of Bacteriology	16806	10.89240463734		22.21825541412		37.81468
Journal of Clinical Microbiology	18421	11.93912803713		20.1563433	499	13.43927
Journal of Microbiology & Biology Education	1305	0.8458044	4	0.3065134	1	25.00000
Journal of Virology	29761	19.28887623079		10.3457545	972	31.56869
mBio	7705	4.9938104	2283	29.63011031473		64.52037
Microbiology Resource Announcements	5878	3.8096843	5744	97.72031305734		99.82591
Microbiology Spectrum	6119	3.9658827	2787	45.54665802174		78.00502
mSphere	2615	1.6948493	971	37.1319312	751	77.34295
mSystems	2327	1.5081891	1430	61.45251401268		88.67133

Whole Dataset

Using the Crossref database, with validation from WOS, NCBI, and Scopus databases, we downloaded N = 154720 unique records of papers published in ASM journals between 2000 and 2024.

After downloading the HTML content of each paper, we cleaned the HTML content and readied it to apply our machine learning models to classify each paper. Overall, 26.9428645% of papers had NSD, and 58.8614883% of papers with NSD, had DA. See table 1 for percentages of NSD and DA for each journal. The journal with the highest rate of NSD was Genome Announcements at 98.7190944%, and the lowest was Journal of Microbiology & Biology Education at 0.3065134%. The journal with the highest rate of DA was Microbiology Resource Announcements at 99.8259053%, and the lowest was Journal of Clinical Microbiology at 13.4392674%. This was expected as Genome Announcements publishes mainly new genomic sequence data and makes the data available. On average, papers in the dataset had a median of 25 citations/article. This number varies by journal, see table YYYY for data by journal. The journal with the highest median rate of citations/article was Applied and Environmental Microbiology at 38%, and the lowest was Microbiology Resource Announcements at 1%. The journals in the dataset span years 2000-2024. See table YYYYY for the distribution of papers per year in the dataset.

Training Dataset (pretty much the same as the above paragraph, but with stats for the training dataset)

*XX - add italics for journal names even when they're variables, do this in not visual mode

We created a subset of the whole dataset to train our machine learning models. The training dataset initially had N = 500 papers, but was increased over time due to gaps in the dataset, and after subsequent validation of the trained models (see below), with a total of N = 1045. See table 2 for the distribution of papers per journal in the training dataset. The journals in the dataset also span years 2000-2024. See table 3 for the distribution of papers per year in the training dataset. Overall, 43.7320574% of papers had NSD, and 80.9628009% of papers with NSD, had DA. See table 2 for percentages of NSD and DA for each journal. The journal with the highest rate of NSD was Genome Announcements at 96.6666667%, and the lowest was Journal of Microbiology & Biology Education at 0%. The journals with the highest rate of DA in NSD papers were Genome Announcements and mSystems at 100%, and the lowest was Journal of Virology at 28.5714286%. On average, papers

in the dataset had median 10 citations/article. This number varies by journal, see table YYYY for data by journal. The journal with the highest rate of citations/article was Infection and Immunity at 40%, and the lowest was Microbiology Resource Announcements at 0%.

Year Published Distribution Table

Table 3: Distribution of Year Published for Whole and Training Dataset

year.published	n_whole_dataset	n_training_dataset
2000	6009	30
2001	5825	43
2002	5807	23
2003	6170	22
2004	6461	21
2005	6961	36
2006	5873	29
2007	5911	18
2008	5476	12
2009	5561	13
2010	5622	21
2011	6206	41
2012	6609	24
2013	6618	29
2014	6875	31
2015	6745	34
2016	6417	38
2017	5893	40
2018	5899	40

year.published	n_whole_dataset	n_training_dataset
2019	5964	63
2020	5886	65
2021	6268	119
2022	6824	70
2023	6199	57
2024	6212	116
NA	429	10

Descriptive Statisitcs about the Trained Models

Table 4: Trained Model Summary Statistics

key	da_model	nsd_model
mtry	300.0000000	200.0000000
logLoss	0.1864109	0.2847761
AUC	0.9878144	0.9645148
prAUC	0.9468117	0.9465242
Accuracy	0.9464109	0.9035551
Kappa	0.8826440	0.8028524
F1	0.9585917	0.9160102
Sensitivity	0.9609538	0.9350193
Specificity	0.9199189	0.8631264
Pos_Pred_Value	0.9564984	0.8983854
Neg_Pred_Value	0.9290263	0.9127760
Precision	0.9564984	0.8983854
Recall	0.9609538	0.9350193

key	da_model	nsd_model
Detection_Rate	0.6203836	0.5257273
Balanced_Accuracy	0.9404364	0.8990728
logLossSD	0.0152552	0.0205940
AUCSD	0.0051133	0.0107140
prAUCSD	0.0131174	0.0127266
AccuracySD	0.0144639	0.0188672
KappaSD	0.0316769	0.0387904
F1SD	0.0112009	0.0162334
SensitivitySD	0.0162858	0.0232715
SpecificitySD	0.0293899	0.0363671
Pos_Pred_ValueSD	0.0152433	0.0240272
Neg_Pred_ValueSD	0.0276805	0.0283175
PrecisionSD	0.0152433	0.0240272
RecallSD	0.0162858	0.0232715
Detection_RateSD	0.0104779	0.0132581
Balanced_AccuracySD	0.0164811	0.0199135

Figures for trained models - are these for supplement??

Two random forest models were trained to predict if published scientific papers "contained new sequence data" (NSD), and if the paper "had data available" (DA), one model for each variable. Other models such as generalized linear regression and boosted trees were explored, but were ultimately discarded in favor of the random forest model (data not shown). Random forest models were chosen to aid in this classification problem as the creation of many decision trees helps to improve accuracy and precision. This type of model has one hyperparameter, 'mtry' or the number of predictors to be sampled at each decision. During iterative model training, a subset of papers

were validated after each completed training and deployment of each model. Papers from each journal, and extras from certain journals were hand-validated against model predictions to generate confusion matrices. Confusion matrices for the final version of each trained model are available in YYYY table(supplement?). The NSD model used an mtry value of 200 had an Area Under the Curve(AUC) of 0.9645148 and an accuracy of 0.9035551. The sensitivity of the NSD model was 0.9350193, and the specificity of the model was 0.8631264. The DA model used an mtry value of 300 had an AUC of 0.9878144 and an accuracy of 0.9464109. The sensitivity of the DA model was 0.9609538, and the specificity of the model was 0.9199189 (See table 4 for more information on trained machine learning models). This shows that the models fit the data well, and can provide classifications on new data with an expected error rate of less than 10%. We deemed this as acceptable, accounting for variability in papers and data, as well as the large size of the dataset on which we deployed the models.

Regression Model using Negative Binomial Models

In this study we sought to investigate the effect of NSD and DA on the number of citations received by a given paper. We focused on NSD papers to determine the effect of having DA. This led us to the use of a negative binomial regression model to best describe our data. All regression data was NSD yes. We focused on the continuous outcome of “number of citations” with predictor variables journal (categorical), age in months (continuous), and DA status (dichotomous). Due to the number of citations being fairly bell shaped with a long right tail (very few papers at advanced age with many citations), the model that best described our data was the negative binomial regression model. A negative binomial model is appropriate for data that begins at zero and has a long ‘tail’ of data. This model also includes a dispersion parameter to describe the spread of the data. We applied a log transformation to our age variable (age.in.months) to better describe the relationship between time and number of citations received. ???HOW DO I DESCRIBE THE PRODUCED MODEL??? In general, we found that NSD papers that made DA received more citations over time than those that did not. See figure YYYY for trends in each major journal. ??HOW else do i describe the figure??

Discussion

- Percent of ASM papers that have new sequence data available (nsd Yes, da Yes)
  - trends by journal
  - trends by year
- Making data available
  - provides more citations per paper than not doing so by *XXX* number
  - Allows for replication of studies
- Why bother doing this?
  - What advantage does it give you as a data generator
  - Is it worth the work?
-

Materials and Methods

Creation of the Training set

To train our random forest machine learning model, we first created an appropriate training data set. Using the crossref database, we first downloaded all papers from the selected ASM family of journals from the time period beginning January 1st, 2000, and ending on December 31st, 2024. The data was updated as of February 10th, 2025 with all citation counts frozen at that date. For our initial training set, we chose N = 500 papers from across each journal and time period, adding special emphasis to include papers that were part of our desired set of interest (i.e. contained published data) to ensure that our two models could adequately characterize each paper as a new sequencing paper and if it published raw sequencing data or not. After creating our initial dataset, it was necessary to identify the status of both variables by hand and determine if each paper contained “new sequencing data” (NSD), and if each one had “data available” (DA). This

was completed by opening each paper in an internet browser window, and searching for a “data availability” or similar statement. See table XXX for specific cases and how each of these cases were identified for the purpose of this study.

Adding Additional Training Set Papers

After initial trainings of our random forest models, a random sampling of papers was collected for each journal to audit the efficacy of the models. To audit the efficacy of the models, we hand identified the status of both variables of interest, NSD and DA. We looked for weaknesses in the models, and updated methodology to reflect important areas of interest. For example, in 2023 the ASM journals changed their formatting to include the data availability statement of a paper in a sidebar of the webpage. We identified this by noticing that all papers from journal *Microbiology Resource Announcements* from 2023-2024 were incorrectly characterized by the model as DA = No. The sidebar of the webpage was not included in the text the model was considering, and code had to be updated to include all sidebar data for all papers. These improvements to the model created a larger and more comprehensive training set of N = 9XX. These validations allowed us to create confusion matricies for each model. Confusion matricies for the final version of each trained model are available in YYYY table(supplement?).

Descriptive Statistics about the Training Set

There were XXX papers in the training data set from 12 ASM journals. These papers came from journals *Applied and Envrionmental Microbiology; Antimicrobial Agents and Chemotherapy; Infection and Immunity; Journal of Clinical Biology; Journal of Virology; Journal of Bacteriology; Journal of Microbiology and Biology Education; Microbiology Resource Announcments* (formerly known as *Genome Announcements*); * mSystems; mSphere; mBio; and Microbiology Spectrum.* See table XXXX for the number of papers from each journal in the training dataset. The training dataset includes journal articles published between January 1st, 2000 and December 31st, 2024.

See table XXXXX for the number of papers included from each year from 2000-2024. XXX% of training set data was NSD = Yes, and XXX% of training set data was DA = Yes.

Creation of the Training Data from Training dataset

To perform the computational steps required for these experiments, we used the python tool Snakemake ((**snakeref?**)), and the University of Michigan's high performance computing cluster ((**arc?**) ref). Using our selected papers from the training dataset, we downloaded the entirety of each paper's source HTML using the command line tool wget. This allowed us to use the source HTML multiple times for updated analyses without the need to re-query the ASM webservers numerous times. Next, we performed cleaning of the HTML using R packages rvest ((**rvest?**) ref) and xml2 ((**xml2?**) ref) to get the desired portions of the paper from the HTML including the abstract, the body of paper, all tables and figureswith captions, as well as the side panels for all papers, but especially those containing the data availability statements in papers published after the 2023 change in webpage format (see above). Then we removed unnecessary text using R packages tm(text manipulation)((**tm?**) ref) and textstem ((**textstem?**) ref), as well as converting all text to lowercase, and the removal of digits and non-alphabetic characters such as whitespace. To have the fewest number of unique words, we lemmatized (sort words by grouping inflected or variant forms of the same word) words to trace them back to their root words and eliminate any possible issues with word tense. After this, we created and counted our 'tokens', phrases of up to 1-3 consecutive words from the text of the paper using R package tokeinziers ((**tokeinziers?**) ref). Towards the goal of the fewest meaningful number of words, we used the 'Snowball' ((**snowball?**) ref) dictionary of 'stop words' to remove non-meaningful words such as articles 'a', 'an', and 'the'. We removed the 'space' character with an underscore in multi-word tokens for ease of procesing, and created a count table for the tokens in each paper.

Once the tokens in each paper were counted, we transformed the data into a sparse matrix format useable by the R package mikropml ((**mikropml?**) ref), using R packages caret and dplyr ((**caret?**) ref, (**dplyr?**) ref). Tokens were filtered to those which appear in greater than one paper. This allows

comparison between papers by the model. We removed near zero variants (tokens with frequency very close to zero) as well as collapsing perfectly correlated tokens (tokens that always appear together) using R packages caret and mikropml to reduce model complexity. The data was then simplified to keep only the following variables; tokens, frequency, journal information, and hand identified NSD and DA variables. This simiplified sparse matrix data had the mean and standard deviation calculcated and saved for the frequency of each token to later apply a z-scoring method to future data to be predicted by the model.

Training of the DA and NSD Models

We trained two random forest machine learning models using mikropml's "run_ml" function, one to determine if a paper contained new sequence data (NSD), and another to determine if the paper had data availabile (DA). The mikropml "run_ml" function uses methodology described by Topcuoglu et al ((**topcuoglu2020?**)) to split data for model training. Random forest models have one hyperparameter to tune, the mtry value. We began with mtry values of 100, 200, 300, 400, 500, and 600, to find peak hyperparameter performance given *N tokens*. We trained the models multiple times in accordance with existing methologies, first to find the optimal Area Under the Receiver-Operator Curve (AUROC) value for each model with N=100 seeds. Then to find the best mtry performance for each model, with N=1 seed. Finally, with N=1 seed to train each final model for use on experimental data.

Preparation of the Experimental Dataset

To fully answer our research questions, we created a larger database with N = 155779 papers curated from reference datasets Crossref, NCBI, Scopus, and the Web of Science ((**crossref?**), (**ncbi?**), (**scopus?**), (**wos?**)). These papers span all twelve ASM journals of interest from the start of 2000 to the end of 2024. Once database was curated, we applied the same steps to ready papers for application of machine learning models as the model training datasets. See above

for descriptions of webscraping html, cleaning html, removing unnecessary text, and creation of token count table for application in each of the machine learning models to determine the NSD and DA statuses for each paper. Once the frequency count tables were prepared for each paper, a z-score was applied using the saved data from each model appropriately, using the formula XXX((observed_token_frequency - model_token_frequency_mean)/(model_token_frequency_sd)). This z-scoring formula was applied to standardize the frequency of each token. Only tokens included in the machine learning models were retained in experimental datasets. Finally, each model was deployed on each paper to determine its NSD and DA status.

- Statistical methodology
  - negative binomial modeling using r package MASS
    - fixed modeling
    - log transformation of age.in.months
    - 95% CI
- Supplemental Material file list (where applicable)
- Acknowledgments
- References
- Figures/Tables/stats to make/get
  - table of conditions to add to the methods of classification?

1. **Langille MGI**, **Ravel J**, **Fricke WF**. 2018. “Available upon request”: Not good enough for microbiome data! Microbiome **6**:8. doi:10.1186/s40168-017-0394-z.

2. Federal Research and Development (R&D) Funding: FY2024. legislation.

345 3. **Yilmaz P**, **Field D**, **Knight R**, **Cole JR**, **Amaral-Zettler L**, **Gilbert JA**, **Karsch-Mizrachi I**, **Johnston A**, **Cochrane G**, **Vaughan R**, **Hunter C**, **Park J**, **Morrison N**, **Rocca-Serra P**, **Sterk P**, **Arumugam M**, **Bailey M**, **Baumgartner L**, **Birren BW**, **Blaser MJ**, **Bonazzi V**, **Booth T**, **Bork P**, **Bushman FD**, **Buttigieg PL**, **Chain PSG**, **Charlson E**, **Costello EK**, **Huot-Creasy H**, **Dawyndt P**, **DeSantis T**, **Fierer N**, **Fuhrman JA**, **Gallery RE**, **Gevers D**, **Gibbs RA**, **Gil IS**, **Gonzalez A**, **Gordon JI**, **Guralnick R**, **Hankeln W**, **Highlander S**, **Hugenholtz P**, **Jansson J**, **Kau AL**, **Kelley ST**, **Kennedy J**, **Knights D**, **Koren O**, **Kuczynski J**, **Kyrpides N**, **Larsen R**, **Lauber CL**, **Legg T**, **Ley RE**, **Lozupone CA**, **Ludwig W**, **Lyons D**, **Maguire E**, **Methé BA**, **Meyer F**, **Muegge B**, **Nakielny S**, **Nelson KE**, **Nemergut D**, **Neufeld JD**, **Newbold LK**, **Oliver AE**, **Pace NR**, **Palanisamy G**, **Peplies J**, **Petrosino J**, **Proctor L**, **Pruesse E**, **Quast C**, **Raes J**, **Ratnasingham S**, **Ravel J**, **Relman DA**, **Assunta-Sansone S**, **Schloss PD**, **Schriml L**, **Sinha R**, **Smith MI**, **Sodergren E**, **Spor A**, **Stombaugh J**, **Tiedje JM**, **Ward DV**, **Weinstock GM**, **Wendel D**, **White O**, **Whiteley A**, **Wilke A**, **Wortman JR**, **Yatsunenko T**, **Glöckner FO**. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology **29**:415–420. doi:10.1038/nbt.1823.

4. **Wilkinson MD**, **Dumontier M**, **Aalbersberg IjJ**, **Appleton G**, **Axton M**, **Baak A**, **Blomberg N**, **Boiten J-W**, **Silva Santos LB da**, **Bourne PE**, **Bouwman J**, **Brookes AJ**, **Clark T**, **Crosas M**, **Dillo I**, **Dumon O**, **Edmunds S**, **Evelo CT**, **Finkers R**, **Gonzalez-Beltran A**, **Gray AJG**, **Groth P**, **Goble C**, **Grethe JS**, **Heringa J**, **Hoen PAC 't**, **Hooft R**, **Kuhn T**, **Kok R**, **Kok J**, **Lusher SJ**, **Martone ME**, **Mons A**, **Packer AL**, **Persson B**, **Rocca-Serra P**, **Roos M**, **Schaik R van**, **Sansone S-A**, **Schultes E**, **Sengstag T**, **Slater T**, **Strawn G**, **Swertz MA**, **Thompson M**, **Lei J van der**, **Mulligen E van**, **Velterop J**, **Waagmeester A**, **Wittenburg P**, **Wolstencroft K**, **Zhao J**, **Mons B**. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**:160018. doi:10.1038/sdata.2016.18.