Reviewer: 1

Comments to the Author

The metadata annotation of SRA is a major issue for large-scale meta-analysis and the study by Klie et. al provides a potential solution using deep learning. Though the capacity of the model is limited and only a few attributes were tested and achieved acceptable accuracy, it is a promising approach and future follow-up work is expected.

Comment #1:

The TITLE is a general description of the study and samples. Many times the samples include both diseased samples and healthy samples, which is one of the most crucial information in analysis. In Condition/Disease prediction, is the model able to predict diseased vs. healthy samples? If not, is it possible to do it? If this requires too much work or is out of the scope, please comment possible solutions and future directions in the discussion.

We agree that this disease/healthy/control status would be useful, and our metadata prediction framework could theoretically be extended to this problem. The model could be used to find n-grams labeled as "Condition/Disease" in sample TITLEs (or other free-text that may carry information as to whether the underlying sample is in a healthy or diseased state). Such "Condition/Disease" n-grams could then be used as features to predict whether a given sample comes from a healthy or diseased biospecimen in a separate classification task. However, in SRA TITLEs are not always specific to or representative of individual biological samples - sometimes the same title is used for all samples in an SRA BioProject which could confound predictions of the disease state for a given sample. Despite this, we agree that this is an important potential application and have added text to the discussion (page 13-14) to comment on this as a future direction:

"We also note the applicability of this work to feature extraction for the downstream classification of samples based on free-text. As an example, consider the underlying disease state of a sample, a critical piece of information often poorly annotated in large public data repositories. We have shown that our model is able to select disease related entities from sample TITLEs with high accuracy (Table 1) by labeling n-grams as "Condition/Disease" metadata. Predicted "Condition/Disease" n-grams could in turn be used as input features for a separate classifier to predict whether a given sample comes from a healthy or diseased biospecimen. However, since sample TITLEs are not always specific to or representative of the underlying biological sample they annotate, it is likely that a larger set of free-text sources, such as those described above, will need to be considered for such a task. Classification of samples based on entity recognition represents an intriguing avenue of future development of this work." (pages 13-14)

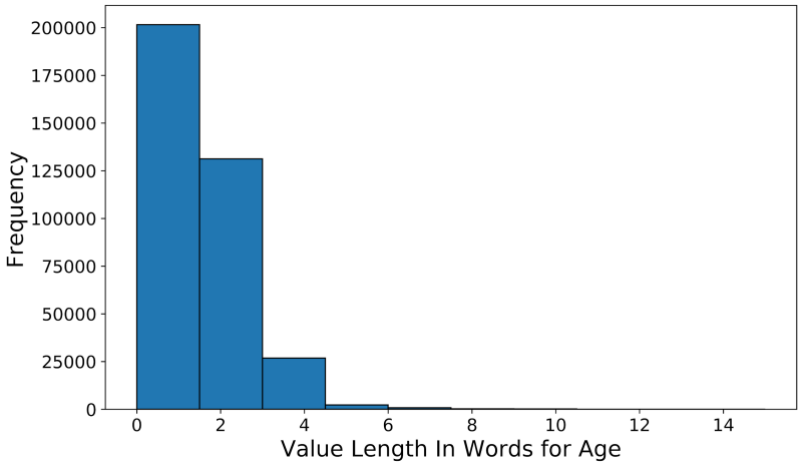
Comment #2:

How does the ground truth of age establish from TITLE, since age of the samples often varies and the TITLE cannot provide such detailed information?

As pointed out, the ground truth age of a BioSample is often not well established from its TITLE attribute due to the variability in Age ground truth labels, the variability of age related terms in TITLEs, and the fact that there is often a lack of correspondence between these two (examples from Supplementary Table 5 shown below). The reasons for these observed discrepancies are not always clear, but often it seems to be caused by a lack of sample level specificity of TITLEs.

Category Name	TITLE Sentence	Predicted Value	Actual Value
Age	2 week dark-adapted	2 week	7 days old
Age	Patient 2 Day 3	Day 3	21
Age	DMD sample Replicate 2 myotube day 2	DMD sample Replicate 2 myotube day 2	1 years old
Age	Model organism animal sample Dermacentor andersoni 2 days	2 days	adults
Age	4 hr FGF + LY294002 Rep2	4 hr	e13.5

We also observed that the prediction of Age was made more challenging by the inclusion of unigrams when establishing the ground truth of Age labels (e.g. rows 2, 4 and 5 of the table above). Our model often struggles to accurately capture these ground-truth unigrams due to the fact that we currently only consider n-grams of length 2 through 7 during metadata classification training and in metadata prediction. In fact, out of the 118 incorrect predictions on the Age metadata category made by our model, 84 (71.2%) of them were unigrams. Excluding unigrams prior to model prediction (after training on n-grams of length 2 through 7) does lead to a marked increase in prediction accuracy for Age (improves from 47.32% to 70.20%, now shown in Supplementary Figure 4D). However, unigrams make up a majority of the actual distribution of values for Age in the entire dataset (see below histogram and Supplementary Figure 4A and 4B) and are important in practical considerations of cleaning SRA metadata. Furthermore, including unigrams in the metadata classification training step led to a drop in test set accuracy across all metadata categories to 83.25% (Supplementary Figure 4C) and a significant drop in prediction performance in multiple metadata categories, including Age (from 47.32% to 36.09%, Supplementary Figure 4D).



Due to these observed challenges in establishing ground truth age metadata from TITLEs, it is likely that predictions for the Age metadata category will require multiple sources of free-text or a modification to our current prediction algorithm.

Comment #3:

How long does it take to train the model? What is the computing resource requirement?

Model training was completed in less than 5 minutes on a single machine with 32 Intel(R) Xeon(R) E5-2670 0 @ 2.60GHz CPUs and 64GB of total RAM. This information has been added to the "Bi-LSTM Training" section of the Methods in the main text (page 5, paragraph 4).

Comment #4:

How is the performance of this model compared to other available methods?

Multiple platforms with built-in user-interfaces have been developed to aid the standardized upload of metadata to large data repositories, including Qiita (Gonzalez et al., 2018) and SRA (Quiñones et al., 2020), but these webtools do not attempt to address the issues with existing, legacy metadata that were presented in this work. To the best of our knowledge, this is the first attempt to formulate the standardization of existing SRA metadata as a deep learning problem and to present a solution to this problem. It is our hope that more advanced deep learning methods will continue to be applied to this dataset in the future and yield even better results.

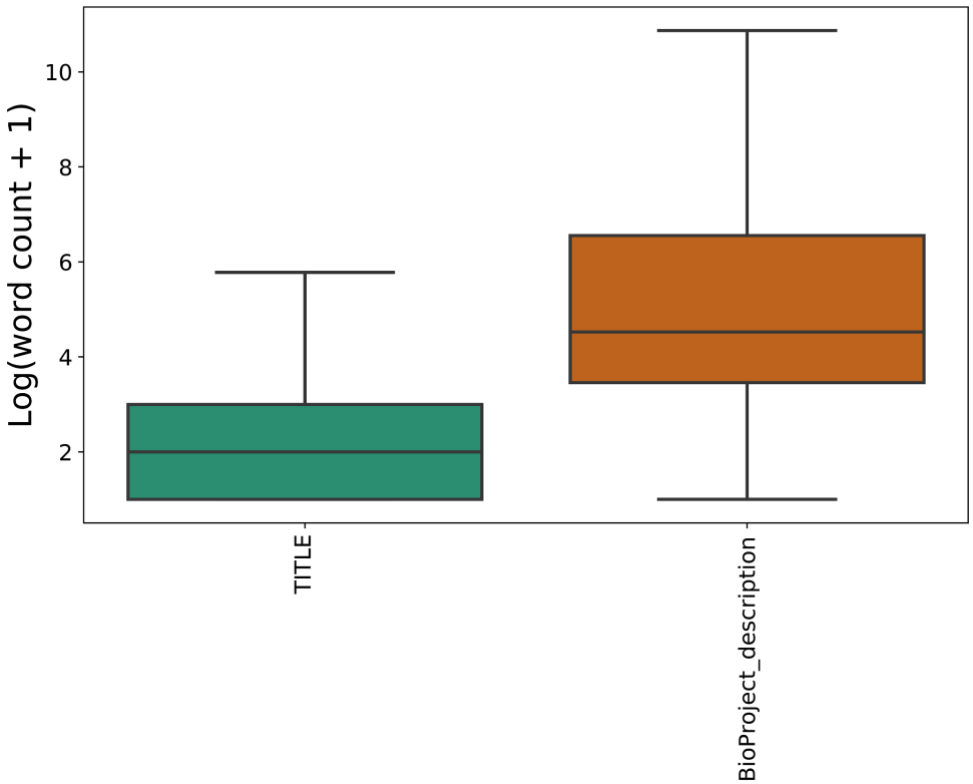
Comment #5:

The Abstract in the BioProject page associated with NCBI SRA contains more detailed information and is usually referred to in manual annotation. Could the model incorporate information from the abstract?

Our model could readily be applied to extract attributes from longer free-text sources such as BioProject descriptions, paper abstracts and full length articles (or sections of articles). As shown below, BioProject descriptions are significantly longer (in word count) than TITLEs (see boxplot

below). We also found that 45,629 BioProjects covering 1,312,806 total samples in our dataset have associated PubMed IDs with publications available. These sources of free-text likely contain many detectable metadata attributes not present in sample TITLEs, but are also more likely to be less specific to each individual sample. We believe that a combination of predicted metadata from sample TITLEs, BioProject descriptions and publications could increase the accuracy and plurality of our metadata prediction. We see incorporating predictions on these free-text sources as an exciting future direction and have included the following text pertaining to this in the discussion (page 13, paragraph 4):

"We limited the validation of our model to the free-text contained in the TITLE attribute of each BioSample due to the prevalence, length, and sample level specificity of this attribute. We note, however, that TITLEs may not be the lone source of usable free-text associated with SRA samples. BioProject descriptions and linked publications often represent multiple BioSamples and contain detailed descriptions of study designs and sample preparation protocols. Our model could readily be applied to such free-text sources to capture annotations not present in TITLEs, to verify predictions made by our model on TITLEs, or to refine current annotations in BioSample. However, the information contained in these types of free-text will often be less specific to each individual sample, and we hypothesize that an approach that considers multiple sources of free-text for metadata prediction will produce the most accurate annotation." (page 13, paragraph 4)



Comment #6:

Some critical information may be contained in the user-defined annotations. Is there a way to borrow information there to make more accurate predictions?

Indeed, user-defined attributes play a major role in building a training set for our model. We incorporate multiple user-defined attributes through the entity merging process described in the *Word Embeddings and Attribute Merging* section (pages 3-5). Supplementary Table 2A (shown below) highlights all the attributes used for model training. We have colored the attributes that are defined by NCBI (indicated in blue in the below table for clarity). The rest are user-defined attributes. Note that 93 out of the 105 attributes included in model training and prediction are user-defined and three of the metadata categories our model predicts are based on a user-defined attribute (SCIENTIFIC_NAME, platform, protocol). At prediction time, it is also possible to incorporate user-defined attributes as further sources of free-text, as described above.

Category Name	Attribute	Cosine Similarity
Genus/Species	**SCIENTIFIC_NAME**	1.000
	Organism	0.917
	host scientific name	0.911
	organism	0.903
	host_scientific_name	0.886
	host	0.884
	nat-host	0.859
	specific host	0.853
	host organism	0.851
	host species	0.831
	specific_host	0.805
	HostSpecies	0.801
Strain	**strain**	1.000
	strain background	0.932
	host_genotype	0.931
	background strain	0.927
	strain/background	0.925
	genetic background	0.912
	mouse strain	0.906
	Mouse_Strain	0.899
	StrainOrLine	0.894
	host strain	0.891
	stain	0.891
	host_strain	0.881
	strain or line	0.880

	strain name	0.879
	host infra-specific name	0.870
	host genotype	0.865
	background	0.864
	host_breed	0.849
	Strain	0.836
	maternal strain	0.835
	paternal strain	0.812
Cell type	**cell type**	1.000
	cell_type	0.897
	source_name	0.886
	source cell type	0.885
	CellType	0.869
	progenitor cell type	0.866
	cell types	0.865
	cell-type	0.853
	cell description	0.849
	CELL_TYPE	0.845
	biomaterial_type	0.841
	tissue/cell type	0.827
Genotype	**genotype**	1.000
	genotype/variation	0.940
	Genotype	0.895
	mutant	0.882
	mutation	0.847
	plant genotype	0.824
	genetic variation	0.812
	idh2.gene.mutation	0.801
	host genotype	0.801
Condition/Disease	disease	1.000
	disease status	0.833
	tumor type	0.828
	health state	0.828
	cancer type	0.826
	cell description	0.813
Tissue	**tissue**	1.000
	tissue_type	0.913
	organism part	0.848
	tissue-type	0.827

	source_name	0.800
Sex	**sex**	1.000
	host_sex	0.997
	Sex	0.996
	host ex	0.995
	sex_infant_1	0.995
	sex / reassigned sex	0.995
	Host_Gender	0.995
	babygender_m_f	0.994
	gender	0.994
	;	0.987
	breeding direction	0.981
	Gender	0.939
	host sex	0.930
	SEX	0.900
	sex_def_prob	0.865
Age	**age**	1.000
	Age	0.832
Data type	**molecular data type**	1.000
Platform	**platform**	1.000
	Platform	0.942
	instrument_model	0.935
	Sequencing_method	0.934
	sequencing method	0.932
	INSTRUMENT_MODEL	0.930
	SequencingTechnology	0.926
	sequencer	0.916
	illumina_technology	0.914
	labversion description	0.914
	illumina_technlogy	0.910
	Sequencer	0.910
	seq_meth	0.903
	sequencing_platform	0.892
	seq_methods	0.891
	sequencing_machine	0.890
	sequencing_method	0.874
	runchemistry	0.850
	seq_method	0.802
Protocol	**protocol**	1.000

	technology	0.887
	extract_protocol	0.860
	experiment type	0.858
	assay	0.850
	protocol description	0.844
	application	0.843
	library_type	0.800

Reviewer: 2

Comments to the Author

I welcome this manuscript from Klie and colleagues, addressing the serious problem of metadata quality in public repositories like SRA. Unfortunately, SRA proved too tough with limited features getting safely annotated. In any case I would welcome an integration of this system to SRA or as a server with an API for users to extract refined annotations for their samples of interest. Despite the limitations shown this is on the right path and I look forward to further improvements and utilization.

We thank the reviewer for their assessment. With a well put together API, we agree that the current iteration of our method could be used to help guide user searches for samples of interest and to help guide metadata annotation upon submission. We hope that more advanced deep learning methods will continue to be improved upon in this context and applied to this dataset. We see this work as a first step towards a set of harmonized and complete metadata in SRA.

Reviewer: 3

Comments to the Author

The manuscript presents a deep learning approach to extract metadata (i.e., attribute-value pairs) from titles in NCBI's Sequence Read Archive (SRA) BioSample. While there are standard attributes defined in SRA, the majority of commonly used attributes are user-defined and many samples do not have metadata. To address these issues, the authors use word embeddings and train a bi-LSTM (bidirectional long-short term memory) on existing attribute-value pairs in BioSample. The prediction is done by classifying attribute-value pairs based on 2- to 7-grams from BioSample titles. Among 11 metadata categories, the deep learning approach shows >80% accuracy in Condition/Disease. Genus/Species and Strain, however it shows less than 65% accuracy for other categories. The authors clearly explain the motivation of the manuscript and describe the pros and cons of their approach pretty well, but there are several major concerns as follows.

Comment #1:

1. The contribution of the manuscript is not clear. Using a deep learning approach is not new, and the performance on 11 metadata categories is not high enough to replace manual annotation (or it does not explain how one can save time and cost using the proposed approach). The reproducibility of the approach is also questionable, i.e., the authors did not share the data and scripts used for experiments.

To ensure reproducibility of our work, data and scripts used for all analyses can be found in the following code repository: https://github.com/cartercompbio/PredictMEE. We have now emphasized this link at the beginning of the text (see below). We have found that models trained with the same parameters and training set consistently reproduce the results we have presented in this work (Supplementary Tables 4A and 4B, shown below). Models can easily be trained and tested by following the instructions linked at the provided code repository.

"All data and code used for this analysis are available at the following Database URL: https://github.com/cartercompbio/PredictMEE" **(page 1)**

We have also added callouts to these tables in the text:

"We found these results to be reproducible, as 10 models trained separately showed similar performance (Supplementary Table 3A)." (page 11, paragraph 1)

"As with the metadata classification task, these results were reproducible across 10 independent trials (Supplementary Table 3B)." (page 12, paragraph 1)

Supplementary Tables 3A and 3B - Metadata category classification and prediction performance across 10 trials

Trial	Accuracy	Precision	Recall	F1
1	85.2%	0.851	0.852	0.850
2	84.7%	0.847	0.847	0.844
3	85.1%	0.851	0.851	0.849
4	84.9%	0.852	0.849	0.847
5	85.0%	0.848	0.850	0.848
6	84.9%	0.848	0.849	0.848
7	85.0%	0.853	0.850	0.847
8	85.2%	0.852	0.852	0.850
9	85.3%	0.855	0.853	0.851
10	84.6%	0.846	0.846	0.845
Average	**85.0%**	**0.850**	**0.850**	**0.848**
Standard Deviation	**0.2%**	**0.003**	**0.002**	**0.002**

			Trial											
Category	**TITLES**	**Predicted**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Average**	**Standard Deviation**
Age	1000	229.5	47.32%	45.42%	46.46%	45.19%	45.37%	47.11%	47.16%	47.11%	47.16%	46.32%	**46.46%**	0.85%
Cell type	702	230	47.22%	48.33%	48.74%	51.33%	45.71%	44.59%	45.56%	43.72%	50.25%	44.85%	**47.03%**	2.56%
Condition /Disease	122	25.6	95.65%	84.62%	85.19%	88.46%	84.62%	84.62%	92.00%	84.62%	88.00%	84.62%	**87.24%**	3.86%
Data type	78	12.7	83.33%	55.00%	81.82%	56.25%	84.62%	81.82%	77.78%	55.56%	100.0%	56.25%	**73.24%**	16.12%
Genotype	595	148	62.31%	58.96%	51.10%	50.89%	50.00%	50.55%	59.15%	57.04%	59.72%	62.89%	**56.26%**	5.12%
Platform	275	27.9	36.67%	48.84%	36.67%	52.38%	63.64%	61.11%	52.38%	48.84%	57.89%	52.38%	**51.08%**	9.04%
Sex	190	11.8	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	**100.00%**	0.00%
Genus/ Species	1000	765.4	94.85%	93.85%	93.35%	94.21%	93.43%	93.84%	92.38%	94.47%	94.12%	93.88%	**93.84%**	0.68%
Strain	1000	388.6	82.03%	86.45%	75.00%	82.87%	86.74%	64.97%	81.60%	71.05%	84.86%	80.24%	**79.58%**	7.11%
Tissue	1000	250.9	63.71%	62.13%	62.70%	63.35%	69.16%	72.14%	62.20%	67.42%	67.77%	64.14%	**65.47%**	3.43%

While new servers are coming online to improve metadata submission at the time of data upload, there are limited tools to help improve the annotation quality for existing data. To our knowledge, we are the first to apply a deep learning framework to address the missing metadata problem in the SRA. Moreover, while previous work has shown the utility of clustering metadata categories by word embeddings, to the best of our knowledge, we were the first to illustrate how this method could be applied to the SRA and to use it to increase the number of samples that could be used for training a deep neural network. Furthermore, whereas manually annotating or cleaning labels would require scanning large amounts of free-text manually, PredictMEE can easily and quickly handle millions of samples. It is also simple to re-train models on other datasets and models can be fine-tuned as new data is deposited. We agree that this model on its own is not sufficient to completely replace manual annotation without further improvement, but we believe that future coupling of our model with a well-designed user-interface could greatly facilitate the annotation of metadata.

We also see two other major avenues for expansion of our current framework: including more sources of free-text to pull metadata from and expanding the metadata categories that we make predictions for. Addressing the former, we believe that predicting metadata from an ensemble of free-text sources could lead to improved performance and have added a paragraph in the discussion highlighting this (page 13, paragraph 3, see below). As to the latter, we chose the 11 metadata categories presented in the manuscript because we felt that they are capable of describing most of the biological and technical variation in a sequencing sample. However, this framework can readily be extended to predict other metadata categories and attributes found in the SRA, or in other repositories, provided enough training examples are available.

"We limited the validation of our model to the free-text contained in the TITLE attribute of each BioSample due to the prevalence, length, and sample level specificity of this attribute. We note, however, that TITLEs may not be the lone source of usable free-text associated with SRA samples. BioProject descriptions and linked publications often represent multiple BioSamples and contain detailed descriptions of study designs and sample preparation protocols. Our model could readily be applied to such free-text sources to capture annotations not present in TITLEs, to verify predictions made by our model on TITLEs, or to refine current annotations in BioSample. However, the information contained in these types of free-text will often be less specific to each individual sample, and we hypothesize that an approach that considers multiple sources of free-text for metadata prediction will produce the most accurate annotation." (page 13, paragraph 4)

Comment #2:

2. The thresholds used for merging attributes (0.8 in Page 4) and evaluating prediction performance (0.7 in Page 6) seem arbitrary. It is not difficult to imagine those thresholds would play an important role in classification performance, but the manusciprt does not explain how and why the thresholds were determined.

We agree that these thresholds are important hyperparameters to examine in our presented framework. The threshold of 0.8 for merging attributes was chosen to balance increasing the training set size (and generality of the model) and merging categories that were too dissimilar. For evaluating prediction performance, 0.7 was chosen to balance too harsh a correctness criterion with one that is too lax. We have performed experiments to illustrate the choice of 0.8 and 0.7 for merging attributes and evaluating prediction performance, and have modified the text and included supplementary information (Supplementary Tables 4A and 4B, see below) to justify the threshold choice for these parameters:

"The threshold of 0.8 was chosen to balance increasing the training set size (and generality of the model) and merging categories that were too dissimilar." (page 8, paragraph 3)

"Furthermore, evaluation of performance with different cosine similarity thresholds for attribute merging validated the choice of 0.8 (Supplementary Table 4)." (page 12, paragraph 1)

We evaluated different cosine similarity thresholds for attribute merging to identify a value that resulted in the best performance in metadata prediction when applied to TITLEs. Cosine similarity thresholds of 0.7, 0.8 and 0.9 corresponded to 280, 105 and 53 merged attributes respectively. The larger the number of attributes merged, the more heterogeneous the values captured under a given metadata category become and the poorer we do on the metadata category classification task (Supplementary Table 4A). With fewer merged attributes, the metadata category classification task becomes easier, but the model does not generalize as well to the metadata prediction task using TITLEs, where we see performance decrease (Supplementary Table 4B).

Supplementary Tables 4A and 4B - Metadata category classification and prediction performance across merging thresholds

Merging Threshold	Accuracy	Precision	Recall	F1
0.7	72.0%	0.719	0.720	0.715
0.8	85.2%	0.851	0.852	0.850
0.9	**91.2%**	**0.917**	**0.912**	**0.913**

	Merging Threshold		
Category	**0.7**	**0.8**	**0.9**
Age	**49.29%**	47.32%	46.12%
Cell type	49.81%	47.22%	**53.55%**
Condition/Disease	91.67%	**95.65%**	91.67%
Data type	12.82%	**83.33%**	**83.33%**
Genotype	52.11%	**62.31%**	51.40%
Platform	25.00%	36.67%	**52.50%**
Sex	**100.00%**	**100.00%**	**100.00%**
Genus/Species	93.44%	**94.85%**	93.99%
Strain	77.84%	**82.03%**	78.90%
Tissue	**75.43%**	63.71%	66.25%

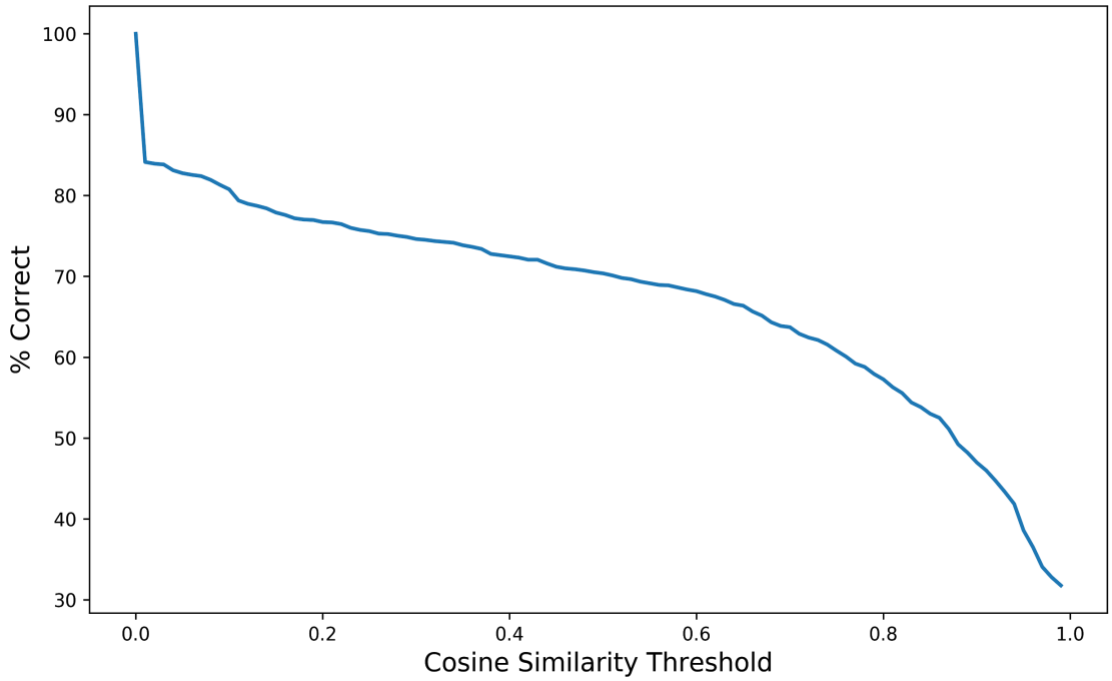
It is challenging to choose a cosine similarity threshold for evaluation of predictive performance due to the qualitative nature of assigning predictions as correct. When the threshold is set to 0.6, seemingly spurious predictions like the one shown below are counted as correct:

Predicted	Actual	Cosine similarity
Day 13	25 year	0.61

On the other hand, when the threshold is set to 0.7, small and (seemingly) insignificant differences can be counted as incorrect:

Predicted	Actual	Cosine similarity
Clostridium saccharogumia	[clostridium] saccharogumia	0.69

Wherever the threshold is set, cases like this can be found. Based on the plot shown below, we reasoned that a threshold of 0.7 approximated the best trade-off between avoiding the false labeling of predictions as true positives, while minimizing false negatives that are close to true negatives. Note that definitions of true positives and true negatives are inherently subjective in this case.



Even with a threshold of 0.7, many errors in prediction can be traced back to how predictions were evaluated for accuracy. Some examples from Supplementary Table 5 are shown below to illustrate this. Though the predicted annotations match the actual metadata well in text and in concept, they do not meet the cosine similarity threshold required for assigning a true positive. Refining the evaluation method of our framework represents another key aspect of developing a reliable metadata predictor.

Category Name	TITLE Sentence	Predicted	Actual
Genus/Species	SYNE2/ESR2A CALM1 loci Clupea pallisii PH3	Clupea pallisii	clupea pallasii
Strain	Pyronema omphalodes CBS 100304	CBS 100304	cbs100304
Genotype	Productive VB18 allele PP PNA low B cells VB18 passenger mice	VB18 allele	vb18 passenger mice
Tissue	Partial transcriptome abdominal tissue female Hypolimnas bolina	abdominal tissue	abdomen

The following text was also added to the manuscript:

"The threshold of 0.7 was chosen to avoid missing correct predictions with too harsh a correctness criterion and to avoid labeling spurious predictions as correct." (page 11, paragraph 2)

Comment #3:

3. The authors included samples with single words as ground truth for metadata prediction, but unigrams were not considered as candidates for prediction. If "ground truth annotations that

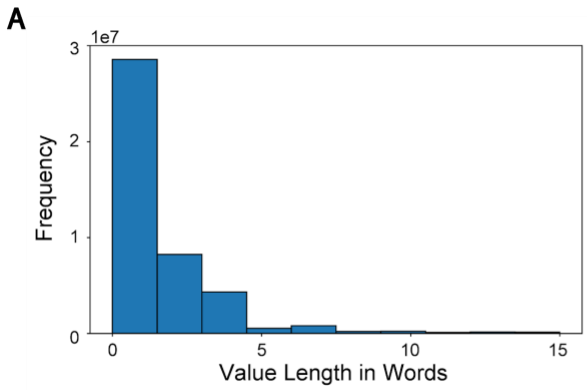
were a single word were challenging and often incorrect" (Page 12) is true, wouldn't it better to filter out such single word cases beforehand?

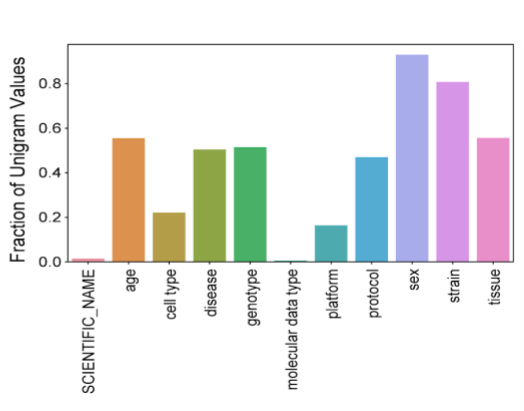
Excluding unigrams (single words) prior to prediction evaluation does lead to an increase in accuracy for multiple metadata categories. However, these unigrams were included in the ground truth labels to keep model validation consistent with the actual distribution of attribute value lengths (in words) in the entire dataset, of which a majority are unigrams (Supplementary Figure 4A and 4B, shown below).

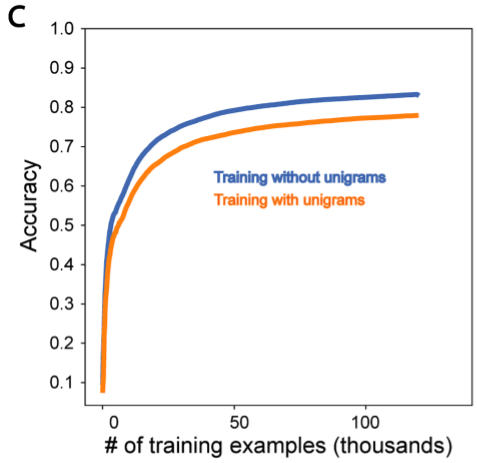
Including unigrams in model training *and* prediction led to a drop in overall test set accuracy (83.25%) and a significant drop in prediction performance across most metadata categories (Supplementary Figure 4C and 4D, below). It is likely that the prediction algorithm, model architecture, or free-text input sources will need to be modified to improve performance on unigrams.

The following text was added to clarify handling of unigrams:

"...Out of the 448 incorrect predictions made by our model, 217 (48.4%) of them were unigrams (Supplementary Table 5). While excluding unigrams prior to prediction does lead to a marked increase in prediction accuracy for most metadata categories, unigrams make up a majority of the actual distribution of annotation lengths across all attribute-value pairs in the dataset (Supplementary Figure 4A) and are a significant proportion of many of the attributes selected for classification (Supplementary Figure 4B). This problem was not solved by including unigrams in model training, as this led to an overall drop in training accuracy (Supplementary Figure 4C) and a significant drop in prediction performance in multiple metadata categories (Supplementary Figure 4D)..." (page 12, paragraph 2)







D

Category	Prediction with Unigrams	Prediction without Unigrams	Prediction after Training with Unigrams
Age	47.32%	**70.20%**	36.09%
Cell type	**47.22%**	45.45%	46.61%
Condition/Disease	**95.65%**	87.80%	52.17%
Data type	83.33%	**92.31%**	45.45%
Genotype	62.31%	**73.58%**	27.86%
Platform	36.67%	**51.22%**	44.68%
Sex	**100.00%**	**100.00%**	58.14%
Genus/Species	94.85%	**100.00%**	87.70%
Strain	82.03%	**94.34%**	62.58%
Tissue	63.71%	**75.28%**	66.82%

Supplementary Figure 4. Prevalence and influence of unigrams on training and prediction. **(A)** Value length distribution in terms of word count across all attribute value-pairs. **(B)** Fraction of values for each selected attribute that are unigrams across all attribute-value pairs for that attribute. **(C)** Metadata classification training accuracy versus iteration for models trained with and without unigrams included as training examples. **(D)** Metadata prediction performance of models with and without unigrams. The last column shows predictive performance after training on and predicting unigrams.

Minor comments

Comment #1:

- Page 3, "SCIENTIFIC_NAME" -> species (in Supplementary Table 1)

Since the value under the attribute SCIENTIFIC_NAME in the SRA does not always correspond to a species level annotation and most metadata in this category comes in the form "*genus species,*" we have changed the overall category name from Species to Genus/Species throughout the manuscript.

Comment #2:

- Page 6 (Figure 2B): It would be nice to indicate what are user-defined attributes or not.

We have added a legend to this panel indicating whether the attribute is user-defined or BioSample defined based on the color of the x-axis labels. This is also called out in the figure caption (see below).

A

SRS	Attribute	Value
SRS1024493	TITLE	Leaf sample from Camellia oleifera
SRS441497	platform	Illumina HiSeq 2000
SRS438070	SCIENTIFIC_NAME	Homo sapiens
SRS1860718	sex	female
SRS2222927	molecular data type	SNP Genotypes (NGS)
SRS1630076	age	2 years
SRS752582	disease	Acute Myeloid Leukemia
SRS1135004	genotype	Xrcc4-/- p53-/-
SRS1274093	protocol	48h dox induction

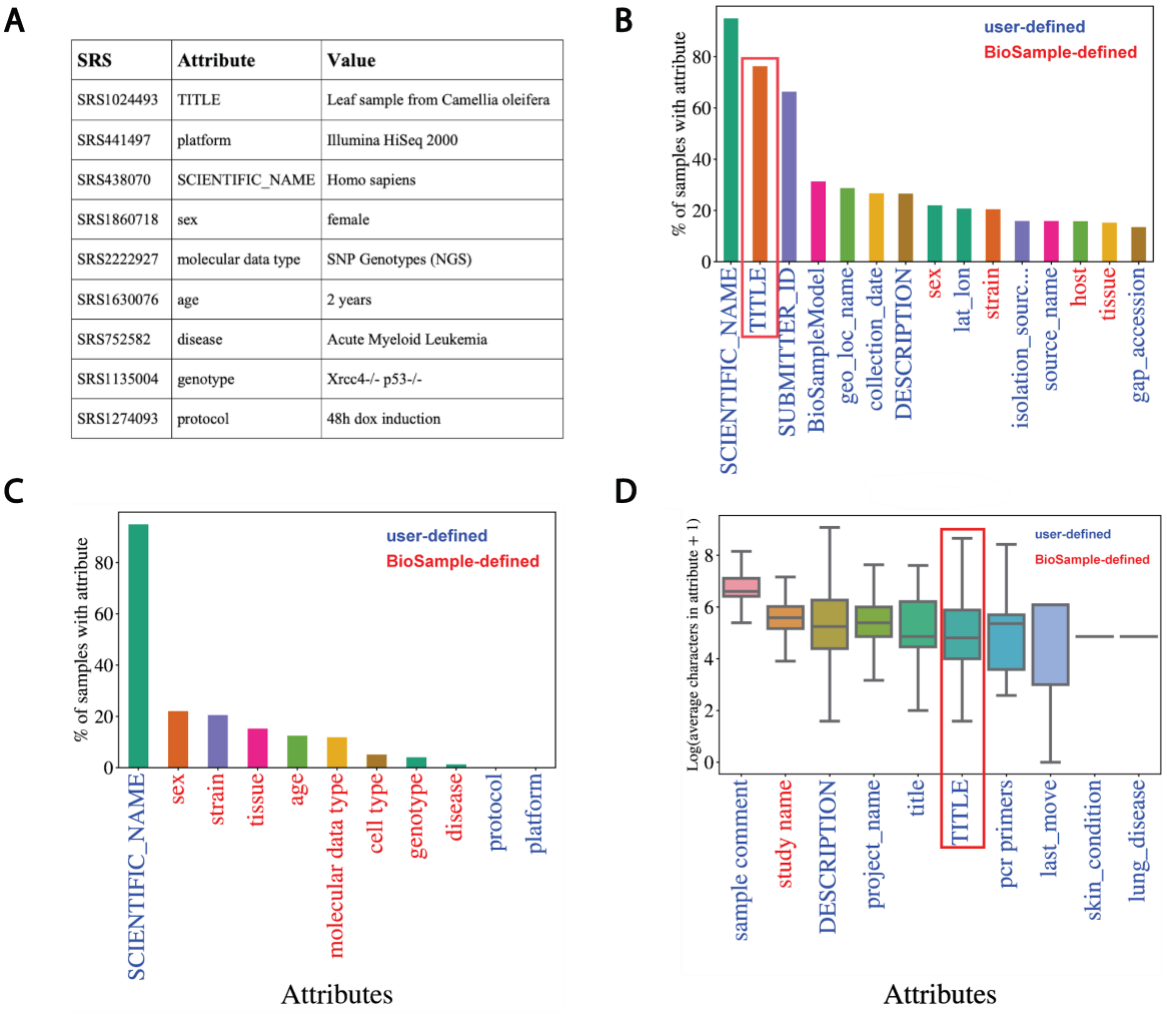


Figure 2. Missing metadata in SRA. **(A)** Examples of SRA attribute-value pairs. Percentage of all samples that contained annotations for the **(B)** top 15 most used attributes and **(C)** the 11 selected attributes. X-axis shows attribute type and y-axis shows the percentage of total samples that used the given attribute. **(D)** Distributions of the average number of characters for the 10 longest (by mean) attributes in BioSample annotations of SRA. X-axis shows attribute type and y-axis shows the Log_2(average characters) for a given attribute. Blue labels indicate a user defined attribute, red labels indicate a BioSample defined attribute. TITLE attribute in panels (B) and (D) is highlighted.

Comment #3:

- Pages 8-9 (Figure 4): Some figures are based on the training set and some are based on the test set, but it is not mentioned explicitly.

We have included panel titles and updated the caption for Figure 4 to reflect which panels represent metrics and performance on the test set and which represent the metrics and performance on the training set (see below).

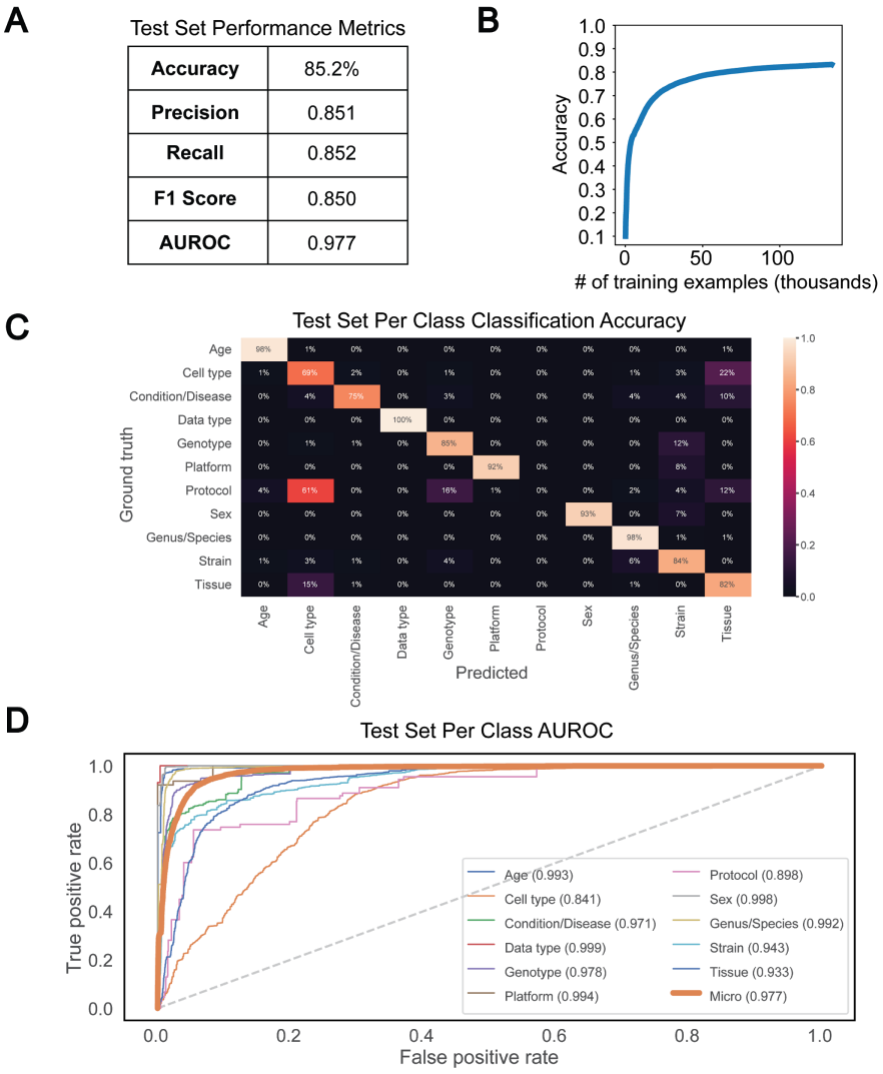


Figure 4. Performance of bi-LSTM in metadata category classification. **(A)** Accuracy, precision, recall, F1 score, and average AUROC calculated for all categories combined on held-out test set. **(B)** Accuracy of model classification on training set (y-axis) plotted against the number of training examples input (in thousands). **(C)** Percentage of each category correctly classified on held-out test set, shown as a heatmap, with predicted values on the x-axis and ground truth labels on the y-axis. **(D)** Receiver operating characteristic (ROC) curves for each category along with the average over all test set examples (micro average).

Comment #4:

- Page 13, "Using a similar framework with learned embeddings or with fine-tuned, pre-trained models like BERT may better delineate the classification categories and limit OOV words": BERT is different from typical word embeddings (e.g., word2vec), and it is a neural language model. The current approach in the manuscript cannot simply use BERT by replacing word2vec.

We agree that BERT would not simply replace word2vec in this setting. However, BERT and more recent language models use sub-word embeddings (e.g., "framework" may be divided into two subwords: "frame" and "work") so an OOV word may now be represented as a combination of subwords in our dataset. Utilizing learned embeddings or a pre-trained BERT neural language model could mitigate the OOV issue and may better delineate the classification categories.

References

Gonzalez, A., Navas-Molina, J.A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A.D., Orchanian, S.B., et al. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. Nat. Methods *15*, 796–798.

Quiñones, M., Liou, D.T., Shyu, C., Kim, W., Vujkovic-Cvijin, I., Belkaid, Y., and Hurt, D.E. (2020). “METAGENOTE: a simplified web platform for metadata annotation of genomic samples and streamlined submission to NCBI’s sequence read archive.” BMC Bioinformatics *21*, 378.