

# Annotating publicly-available samples and studies using interpretable modeling of unstructured metadata

Arjun Krishnan<sup>1,2</sup>, J. Christopher Love<sup>3</sup>, and J. Christopher Love<sup>4,\*</sup>

<sup>1</sup>Genetics and Genome Sciences Program, Michigan State University, East Lansing, MI 48823, United States

<sup>2</sup>Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, MI 48823, United States

<sup>3</sup>Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, United States

<sup>4</sup>Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, United States

\*Corresponding author. E-mail: arjun.krishnan@cuanschutz.edu

## Abstract

Publicly available omics data is often accompanied by unstructured metadata that is not easily machine-readable and unambiguously searchable. We present a method for annotating this metadata using interpretable modeling. The method involves training a transformer-based model on a large dataset of unstructured metadata to predict the most likely term from a set of candidate terms. The model is trained using a combination of supervised and unsupervised learning. The supervised learning component uses a dataset of manually annotated metadata, while the unsupervised learning component uses a dataset of unannotated metadata. The model is evaluated using a held-out dataset of manually annotated metadata. The results show that the model is able to accurately predict the most likely term for a given piece of metadata. The model is also able to generate a saliency map that highlights the words in the metadata that are most important for the prediction. This method provides a way to automatically annotate unstructured metadata, making it easier to search and analyze.

**Keywords:** metadata, annotation, transformer, interpretable modeling, unstructured data

## Introduction

Currently, there are millions of public omics samples available via resources like the Gene Expression Omnibus (GEO) [1], the Sequence Read Archive (SRA) [2], Proteomics Identification Database (PRIDE) [3], and MetaboLights [4]. In GEO alone, >7.1 million genomics samples from >224 thousand studies contribute to a vast collection of data from various biological contexts. This massive data collection can be incredibly valuable in revealing novel insights into the molecular basis of numerous tissues, phenotypes, diseases, and environments. However, although these data are available, finding datasets and samples relevant to a biological context of interest is still difficult because these data are described using unstructured, unstandardized, plain-text metadata, which is not easily machine-readable and unambiguously searchable [5].

To tackle this issue, significant efforts have been made to manually annotate datasets [6]. However, manual annotation is not feasible for the exponentially-growing volume of datasets, which now runs in the millions. To automate the annotation process using the metadata, natural language processing (NLP) have been employed to overcome these challenges. Rule-based NLP methods annotate metadata using text-matching or regular expressions [7–9]; however, these methods are vulnerable to misspellings or variations of a query term in the study or sample

descriptions, and cannot infer annotations based on biomedical concepts in the text that are different from but relevant to a query term.

The emergence of transformer architecture-based models has revolutionized the application of NLP in the biomedical domain [10, 11]. Utilizing the power of transformer models, previous methods have framed the annotation task as translation [12] or Q&A [13]. For example, GeMI [12] uses a fine-tuned GPT-2 model to annotate a wide array of term types related to a sample, including species, sequence type, tissue, and cell type, where the ‘question’ is metadata, and the ‘answer’ is term types and corresponding predicted terms. Further, to combat the black-box nature of transformer models, GeMI used the saliency map technique to highlight prediction-related text. However, GPT-based models require a restructuring of the input such that it follows a fixed template, which guides the model in generating coherent and meaningful responses. This constraint makes it difficult to adapt fine-tuned models for annotating biomedical text from sources that do not fit the template. Moreover, GeMI’s output labels are not assigned to controlled vocabularies from ontologies, leading to lingering ambiguity in the annotation terms, which in turn hinders data discovery. Additionally, the GPT model is not lightweight enough to efficiently predict samples at a large scale for millions of metadata [14, 15].

Received: 16, 2024. Revised: 31, 2024. Accepted: 13, 2024

© 2024 The Author(s). Published by Oxford University Press on behalf of the International Society for Bioinformatics and Biomathematics. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Traditional machine learning (ML)-based approaches have also been used to extract information from metadata [6, 16]. For example, Wang et al. [6] manually curated studies related to disease, drug, and gene perturbations from GEO. Then, they converted metadata into a term frequency-inverse document frequency (TF-IDF) matrix that considers the relative frequency of individual n-grams across all metadata documents. Combined with curated labels, these data were used to train ML classifiers to automatically extract signatures from GEO. However, since the features of the TF-IDF matrix are text, trained models cannot handle input text that includes unseen words, such as metadata from other databases, which limits the generalizability of the model.

Furthermore, the advancement of biomedical Large Language Models (LLMs) has allowed for the conversion of biomedical text to numerical representations, also called embeddings, which can effectively capture information in the text [17]. Leveraging this technique, txt2onto [18] represented sample descriptions as the average embeddings of words in metadata, then trained ML models to annotate tissue and cell types using these embeddings. Theoretically, the txt2onto framework can be adopted to annotate any kind of general biomedical text. However, one issue with the txt2onto approach is that averaging the embeddings of all words in a description can dampen the signal from informative biomedical terms. Furthermore, the trained model coefficients of the embedding features do not provide insight into which specific words in the sample descriptions contributed to the model's predictions, limiting interpretability.

To tackle the challenges mentioned above, we present txt2onto 2.0, a novel and lightweight approach that assigns standardized tissue, cell type, and disease annotations to unstructured sample- and study-level metadata. Txt2onto 2.0 combines the power of semantic relationships captured by LLMs (as in txt2onto 1.0 [18]) with the high interpretability offered by word-based modeling, leading to significant improvements in performance while providing transparent, easily understandable predictions. In the following sections, we demonstrate that our ML framework not only outperforms txt2onto 1.0 in sample-level tissue and study-level disease annotation, especially for highly specific tissues and disease terms with limited training instances, but also highlights relevant words in each metadata associated with the annotated tissues and diseases. Moreover, our model can differentiate between similar tissues and similar diseases, enhancing the specificity of the annotations. Furthermore, we demonstrate that our disease classification models, trained on descriptions of transcriptomics studies from GEO, are proficient at infer disease annotations for biomedical metadata from any source.

## Methods

### Overview of txt2onto 2.0 for disease and tissue annotation

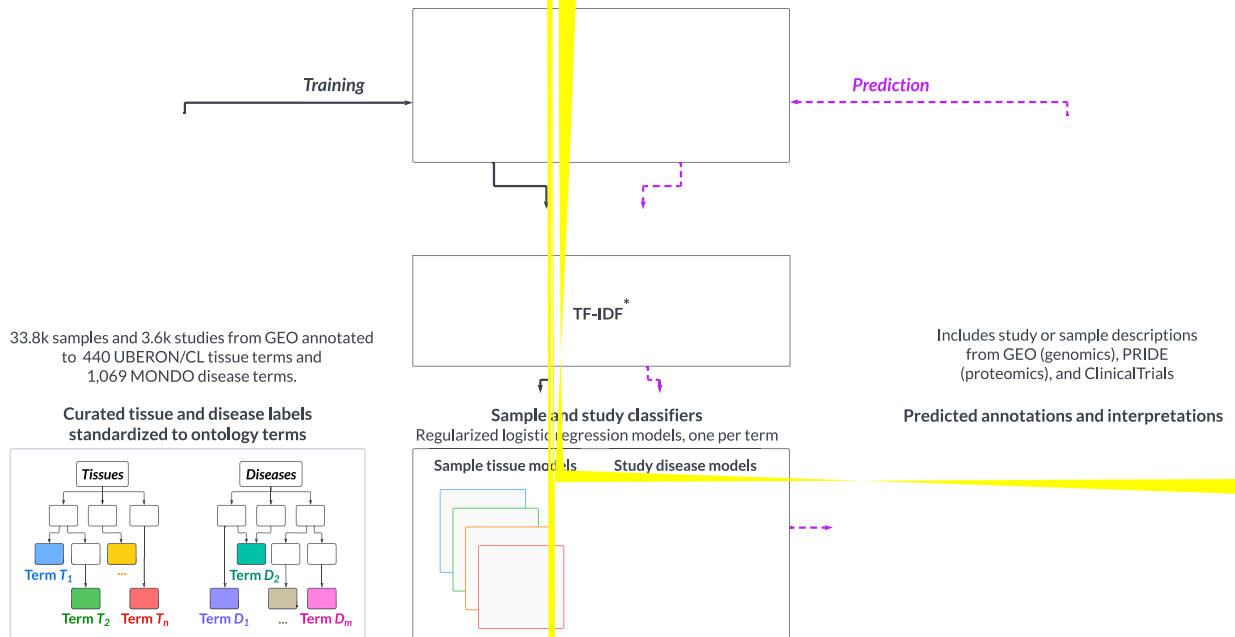
The primary goal of our work was to develop interpretable classifiers that annotate unstructured metadata of omics samples or studies to controlled tissue, cell type, and disease terms from ontologies. To increase the interpretability of our models compared to the previous state-of-the-art method, txt2onto 1.0 [18], txt2onto 2.0 introduces a key improvement: instead of using average word embeddings as features, it converts sample or study metadata into a TF-IDF vector, which serves as input to the ML classifier. During the prediction phase, this classifier accepts the TF-IDF vector of each new unlabelled metadata as input and leverages word embeddings from an LLM to map words in the

new metadata, including those unseen during training, to the training feature space. With text as features, our approach uses model coefficients to easily track the biomedical words/phrases that strongly influencing model predictions (Fig. 1). A detailed comparison of the key differences between txt2onto 1.0 and 2.0 is provided in Table S1.

### Collecting and processing training data

We began by collecting unstructured metadata from GEO to train the tissue and disease annotation classifiers. GEO metadata can be divided into two groups: study-level descriptions, which describe the study's aim and design, and sample-level descriptions, which detail the source and processing of individual samples within studies. These two metadata groups were qualitatively different in terms of the tissue and disease information they contained. Sample descriptions invariably contain information about the sample's tissue-of-origin while study descriptions often lack explicit tissue source information; further, a single study may include samples from multiple tissues. Hence, we decided to train the tissue classifiers at the sample-level. On the other hand, study descriptions contained notes about the disease being studied while sample descriptions mostly only containing modifiers such as 'yes/no' or 'wt/ctrl' without any mention of the disease. Therefore, we decided to train disease classifiers at the study-level.

To create the input text for these classifiers, we identified and extracted relevant fields from the original metadata. Some choices were clear: fields like source name and description contain information directly related to classification tasks while fields such as dates and contact information are irrelevant. Other fields like lab protocols and data processing methods were less clear because they may contain potentially useful information. We observed that the choice of which fields to include during training indeed impacts classifier performance (Fig. S1). For details about the fields used in the following analysis, see **Supplementary Methods, Preprocessing input**. After extracting the relevant fields from each metadata entry, we removed potentially uninformative elements (e.g. punctuation, URLs), converted the remaining words to lowercase, and concatenated the text from the selected fields. As we observed





might have even fewer positive instances in the training set (as low as two instances), which is insufficient for hyperparameter tuning. Our model has top performance even in these cases (Fig. S5a and b).

Among models using different feature types, word-based models slightly outperform phrase-based models (Fig. 2). One reason could be the high sparsity of phrase-based feature matrices compared to word-based matrices because multi-word phrases occur less frequently than individual words. Models based on clusters of semantically-similar word and phrase features show a decrease in performance for both tissue and disease classification (Fig. 2). Consequently, we selected the LR model with word features as the optimal approach for all subsequent analyses.

We also investigated the potential improvement of this approach by combining predictions with MetaSRA (Supplementary Methods, Combining txt2onto 2.0 and MetaSRA predictions). Previously, we showed that combining txt2onto and MetaSRA predictions leads to performance gains [18]. However, when using our novel word feature-based approach, there's minimal performance improvement when combining tissue predictions from both models (Fig. S6).

This choice was also confirmed by the comparison of the LR model with an even simpler baseline model that makes predictions by greedily aggregating the occurrence individual features one-by-one, weighted by each feature's informativeness (Supplementary Note 1). Although the performance of this baseline method is comparable to LR, it is oversensitive to uninformative words (Figs S7 and S8) and cannot output probabilities.

### Txt2onto 2.0 models learn features relevant to tissue and disease classification tasks

One of the primary goals in this new version of txt2onto was to develop interpretable models that capture words and phrases in sample and study metadata that are most predictive of annotations to specific ontology terms. To inspect the interpretability of the models, we summarized predictive word features from top-performing tissue and disease models as word clouds, where the size of each is proportional to its regression coefficient. For example, the keywords for *Glucagon-secreting cells* (CL:000170) (Fig. 3a top) are 'alpha', 'islets', 'pancreatic', 'developmental', 'fetal', 'adult', 'stage', and 'gestational'. These terms refer to the endocrine function of the pancreas, with 'alpha' referring to alpha cells that produce glucagon, a hormone that regulates blood sugar levels. 'Islets' are the islets of Langerhans, clusters of cells in the pancreas. The words 'developmental' and 'gestational' imply a context of fetal (early developmental) or adult stages where these cells are involved. The informative words for *Coronary artery disorder* (MONDO:0005010) (Fig. 3b top) include 'coronary', 'artery', 'disease', 'stenosis', 'atherosclerosis', 'myocardial', 'CAD', 'heart', 'blood', and 'ventricular'. These words are clearly linked to the cardiovascular system, specifically the arteries that supply blood to the heart. Coronary artery disease (CAD) involves the narrowing of the arteries due to atherosclerosis, potentially leading to myocardial infarction or heart attacks. Similarly, the appearance of key terms such as 'MECP2' for *Rett syndrome* (Fig. 3b, middle), 'merlin' for *Meningioma* (Fig. 3b, bottom), 'sickle' for *Erythrocyte* (Fig. 3a, middle), and 'bronchoalveolar' for *Pulmonary acinus* (Fig. 3a, bottom) demonstrate that our models successfully identify relevant features from the metadata text that significantly contribute to its the correct tissue and disease annotation predictions.

Worthy of note is that, by leveraging their ability to capture term-related words and phrases, our models are able to annotate

samples/studies even if the name of the tissue or disease term is absent in the metadata. *Pulmonary acinus* (UBERON:0008874) (Fig. 3a) is one among many good examples of terms to which where metadata are correctly annotated solely via the words and phrases associated with that term. To demonstrate this potential across models for both tissue and disease classification tasks, for each term, we compared the performance of our models before and after removing every mention of that term's name from the metadata corpus. Results show that removing the ontology term names leads to a negligible drop in performance, demonstrating our models' robustness to missing information in metadata (Fig. 3c).

### Txt2onto 2.0 models accurately predict instances related to specific tissues and diseases

To truly promote the discovery of existing data relevant to specific biomedical questions of interest, it is crucial that metadata annotation is accurate for specific tissue and disease terms (e.g. *Rett syndrome*) and not just general terms (e.g. *Neurological disorder*). Therefore, we evaluated the effect of term specificity on txt2onto 2.0's disease and tissue annotation performance. Using the structure of the appropriate ontology the tissue and disease term are a part of, we defined each term's specificity using its information content (IC) of a term. Higher IC indicates specific terms closer to the leaf nodes of the ontology while lower IC indicates general terms closer to the root of the ontology. We included all disease and tissue terms without redundancy filtering in the analysis to help examine model performance across a wide range of term specificities.

This analysis showed that there is a clear association between tissue/disease term specificity (IC) and model performance (Fig. 4). The smaller data points, representing terms with fewer positive samples in the training set, tend to cluster towards the higher end of the IC scale and the performance metric. When cast in the context of the number of positive examples available during training, these trends indicate that our models are capable of accurately annotating to very specific terms in both the tissue and disease ontologies even when the amount of training data is limited. This trend persists across models trained on various features, including phrases, word embeddings, phrase clusters, and word clusters (Fig. S9). When we examined four tissue and four disease outlier models which are outliers of this trend, i.e. specific (high IC) terms having low performance (Supplementary Note 2), we found that poor performance often resulted from mislabeled samples or studies, suggesting that tasks with very few positive instances in the training data are sensitive to mislabeling.

### Txt2onto 2.0 models can differentiate between similar tissues or diseases

Following the specificity analysis, we evaluated the ability of the txt2onto 2.0 models to correctly differentiate between similar or related diseases or tissues, such as distinguishing *Crohn's disease* from *Ulcerative colitis*, or *Skeletal muscle* from *Smooth muscle*. For this evaluation, we defined the semantic similarity between pairs of terms within an ontology as the cosine similarity between their corresponding text embedding vectors calculated using BioMed-BERT [29] (see Supplementary Methods, Differentiate similar terms). Then, we split all term pairs based on the cosine similarity into four equal-size quantile bins corresponding to term pairs with increasing similarity. Finally, within each bin, we quantified (using area under the receiver operating characteristic curve; auROC) each term model's ability to assign higher probabilities



**a**                      **Tissue classification**



**b**                      **Disease classification**

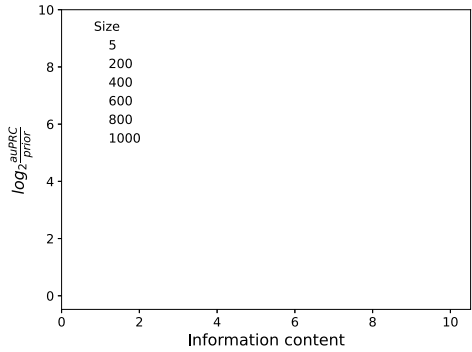


Table 1. Disease annotation performance for studies from PRIDE. This table displays the annotation performance of the top six disease classification models, evaluated based on their top ten and bottom ten predictions. Each row corresponds to a disease term, and each column represents a performance metric. The metrics included are as follows: Acc (Accuracy), calculated from unshuffled predictions;  $\log_2(\text{Acc}/\text{R Acc})$ , the accuracy normalized by the random accuracy calculated from shuffled predictions, then log base 2 transformed; and  $\log_2(\text{F1}/\text{R F1})$ , the F1 score normalized by the random F1 score calculated from shuffled predictions, then log base 2 transformed.

Term	Acc	F1	$\log_2(\text{Acc}/\text{R Acc})$	$\log_2(\text{F1}/\text{R F1})$
Invasive ductal breast carcinoma (MONDO:0004953)	0.83	0.67	1.30	-0.76
Cytomegalovirus infection (MONDO:0005132)	1.00	1.00	1.18	1.24
Pancreatic ductal adenocarcinoma (MONDO:0005184)	0.87	0.83	0.15	0.22
Uveal melanoma (MONDO:0006486)	0.77	0.68	0.89	4.26
Head and neck squamous cell carcinoma (MONDO:0010150)	1.00	1.00	1.29	1.29
Covid-19 (MONDO:0100096)	1.00	1.00	1.29	1.29

Table 2. Disease annotation performance for studies from ClinicalTrials. This table presents the annotation performance of the top six disease classification models, evaluated based on their top ten and bottom ten predictions. Each row corresponds to a disease term, and each column represents a performance metric. The metrics included are as follows: Acc (Accuracy), calculated from unshuffled predictions;  $\log_2(\text{Acc}/\text{R Acc})$ , the accuracy normalized by the random accuracy calculated from shuffled predictions, then log base 2 transformed; and  $\log_2(\text{F1}/\text{R F1})$ , the F1 score normalized by the random F1 score calculated from shuffled predictions, then log base 2 transformed.

Term	Acc	F1	$\log_2(\text{Acc}/\text{R Acc})$	$\log_2(\text{F1}/\text{R F1})$
Atrial fibrillation (MONDO:0004981)	0.90	0.86	0.92	0.90
Urinary bladder carcinoma (MONDO:0004986)	0.70	0.57	0.91	0.97
Gastric adenocarcinoma (MONDO:0005036)	0.80	0.75	1.10	1.58
Periodontitis (MONDO:0005076)	0.77	0.68	0.70	4.35
Pulmonary hypertension (MONDO:0005149)	0.97	0.96	1.13	1.39
Allergic rhinitis (MONDO:0011786)	0.93	0.93	0.20	0.19

contributing to the prediction, providing a way to easily evaluate and confirm the annotations.

## Discussion and conclusion

Reusing existing biomedical data is paramount to advancing scientific research and accelerating medical discoveries [31, 32]. However, the samples and studies stored in current vast biomedical data repositories are often described using unstandardized, unstructured, plain-text descriptions. This poor quality of metadata is a major hindrance for researchers in discovering the datasets most relevant to their context or question of interest. In this work, we propose txt2onto 2.0, a computational approach that combines NLP techniques and ML to annotate any biomedical descriptions to standardized tissue and disease ontology terms. Thus, by providing a way to index public metadata for easy search using controlled vocabularies, txt2onto 2.0 addresses the first goal of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [33].

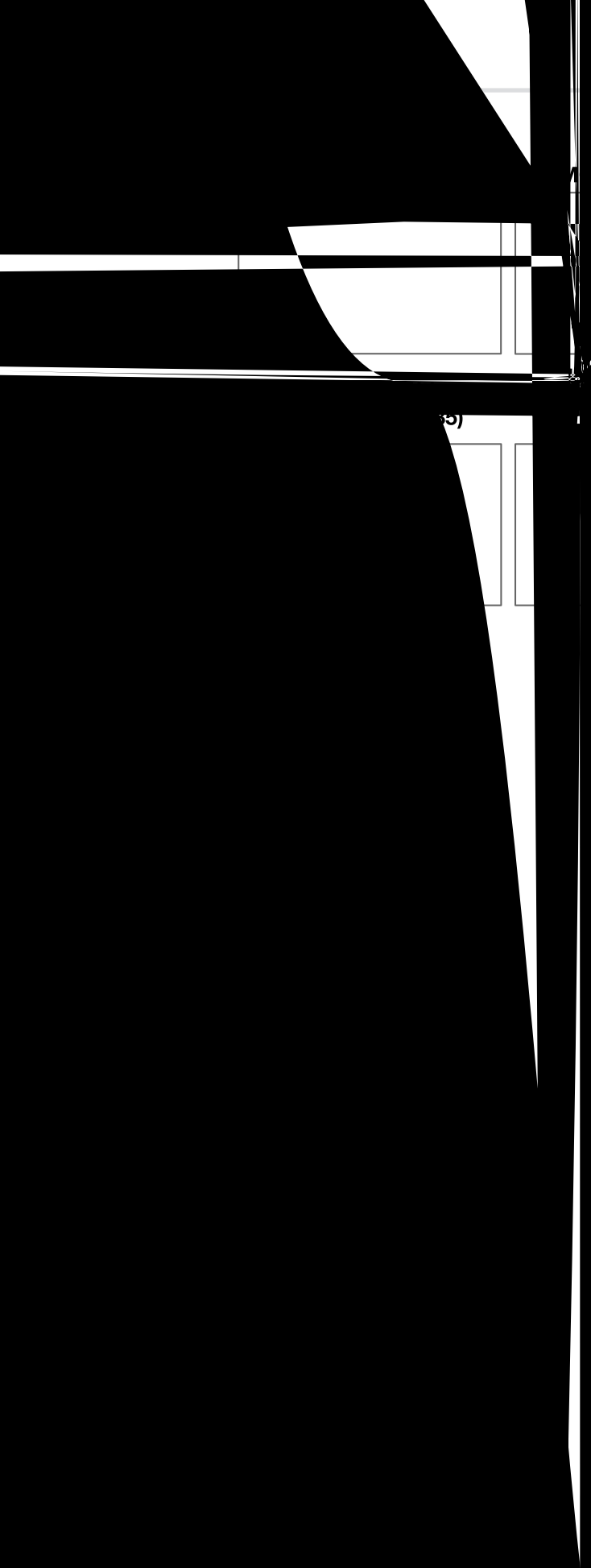
Through systematic and rigorous evaluations, we have demonstrated that our method outperforms the state-of-the-art method, txt2onto 1.0 [18], in disease and tissue classification. Especially in disease classification, our method markedly outperforms txt2onto 1.0, particularly when dealing with highly imbalanced cases. Disease classification poses greater challenges than tissue classification for the previous version due to the longer study descriptions used to infer diseases (Fig. S11). Long text inputs are not amenable to txt2onto 1.0's idea of representing metadata as the average of the embeddings of all the constituent words, which dampens the signal from informative words. Txt2onto 2.0 overcomes this limitation by utilizing a TF-IDF feature matrix to represent text, naturally avoiding the mixing of signals from predictive words with the rest of the text. This strategy enables txt2onto 2.0 to

predict a wider range of tasks than txt2onto 1.0, allowing for accurate predictions of understudied tissues and diseases.

In addition to inferring accurate tissue and disease annotations of metadata, it is crucial to address interpretability and explainability to increase trust in and verifiability of the annotation results. By using word-level features combined with a large language model, txt2onto 2.0 achieves both these desired qualities: (i) Each tissue or disease term model is highly interpretable in terms of the most informative text features learned during training, and (i)ii) Each predicted annotation (of an input metadata to a particular tissue or disease term) is explainable in terms of the specific text snippets in the new metadata that drove the prediction. As a contrast, using average word embeddings as features makes in txt2onto 1.0 makes both interpretability and explainability infeasible. Recent GPT-based models are promising for extracting disease and tissue labels from unstructured text due to their strong text comprehension powers. However, they operate as 'black boxes'. Efforts in explainable AI, such as GeMI's use of saliency maps to highlight words related to an annotation, have been made to explore the reasoning behind such models [13]. Nevertheless, these techniques have limitations, such as providing only post-hoc explanations without deeper insights into the model's internal logic and lacking systematic evaluation of interpretability. In the era of LLMs, we acknowledge that using better models like GPT-4 and state-of-the-art explainable AI methods could achieve better interpretability and explainability. However, it is critical to continue exploring simple and elegant methods like txt2onto 2.0 for building models with high annotation accuracy and inherent transparency that are also lightweight and cost-effective so that they can scale to exponentially growing number of samples and studies across biomedical data repositories.

Extremely imbalanced data is a common challenge in biomedical prediction studies [34]—an issue also present in ours





is important to realize this generalizability without having to (re-)training the underlying models based on curated metadata labels from every database. The models also need to be versatile in dealing with any generic biomedical metadata, regardless of the varying format or writing style. Our method meets this need by treating metadata as a bag-of-words, making it flexible enough to accept input from any source. Further, by using the latent embedding space that captures semantic similarities between words, txt2onto 2.0 is able to utilize words in new metadata to inform an annotation even if those words were never seen during training (Fig. 6a).

Our study opens up numerous avenues for refinement and exploration in the future. First, the informative keywords identified by our method occasionally include generic words because they are overrepresented among the few positive instances (compared to negative instances) available during training. For instance, words such as 'sample', 'week', and 'cells' are highlighted as predictive words for *Glucagon secreting cell* (CL:0000170) (Fig. 3a) despite not being specifically related to this cell type. This problem could be more prevalent in tissue annotation (of samples) compared to disease annotation (of studies), as the richer context in study summaries reduces the likelihood of selecting spurious words, while the limited context in sample descriptions makes the model more prone to latching on to generic words. In future work, we may employ causal inference techniques [38] or map extracted terms to biomedical knowledge graphs and perform graph-based reasoning to remove irrelevant words [39]. Second, we annotated disease labels at the study level but did not attempt sample-level disease annotation, i.e. annotating samples within a study as 'healthy' and 'disease'. This is because sample-level metadata in public databases is severely lacking in completeness and quality (e.g. sample metadata lacking any information about the disease being studied and containing, if at all available, only indicators like 'yes', 'wt', or 'ctrl'). In the future, sample-level disease annotation could be achieved by employing additional computational models that integrate information from samples across studies to train a single model to classify between healthy and disease samples. Third, future work can expand to other annotation categories beyond tissue and disease such as sex, age, phenotype, or environment. Finally, moving from one model per term towards a unified model for annotating multiple terms via a shared representation learning will likely lead to improve prediction performance across the board, and especially for terms with few positive instances.

Overall, txt2onto 2.0 is a novel, light-weight, interpretable, and explainable ML-based approach to annotate biomedical text from various sources with standardized tissue and disease labels, even when there are limited amounts of training instances. Biomedical researchers will be able to use the labels predicted by our method to drastically improve data organization and curation, and to effectively reuse existing data to make potentially novel scientific discoveries in downstream analyses. To ensure that our approach can benefit the research community for data reuse, we released txt2onto 2.0 on our GitHub repository. Users can either predict disease or tissue labels using provided models or build their own models from scratch.

#### Key Points

- We developed txt2onto 2.0, a computational method that combines language models and machine learning to annotate public samples and studies with standardized

tissue and disease terms, with a focus on interpretability and explainability.

- Txt2onto 2.0 uses word/phrase occurrence statistics to represent sample/study metadata, train machine learning models, and predict terms in controlled vocabularies to annotate each sample and study. This approach allows the model to keep track of predictive words related to model decisions and easily separate informative from uninformative words.
- Txt2onto 2.0 outperforms its predecessor, txt2onto 1.0, in tissue and disease annotation, especially when training data are limited.
- The predictive features learned by txt2onto 2.0 are highly interpretable. These features not only include explicit mentions of the actual disease or tissue terms but also related biomedical concepts, including words that are unseen by the model during training.
- Although trained on metadata of transcriptomes, txt2onto 2.0 is capable of annotating disease and tissue for any kind of biomedical metadata, making it a versatile tool for sample and study annotation.

## Acknowledgement

We would like to thank Keenan Manpearl for valuable suggestions on the organizations and documents of the Github repository, as well as code testing. We also would like to thank all members of the Krishnan Lab for valuable discussions and feedback on the project.

## Supplementary data

Supplementary data is available at Briefings in Bioinformatics online.

## Funding

This work was primarily supported by the National Science Foundation grant 2328140 to A.K.

## Code and data availability

We have made the trained disease and tissue classification models, a Python script for classification, and demo scripts, along with extensive documentation available at <https://github.com/krishnanlab/txt2onto2.0> (v1.0.0). The repository also includes utilities for training new models for a user-defined task.

## Author contributions

H.Y. and A.K. designed the study. H.Y. developed the software. H.Y., K.J., and L.V. curated annotations. H.Y., P.H., M.A., and K.J. performed the analyses. H.Y., P.H., M.A., and K.J. interpreted the results. H.Y., P.H., and M.A. wrote the final manuscript with feedback from A.K.

## References

1. Clough E, Barrett T. The gene expression omnibus database. *Stat Genom: Methods Protoc* 2016;**14**:93–110. [https://doi.org/10.1007/978-1-4939-3578-9\\_5](https://doi.org/10.1007/978-1-4939-3578-9_5).
2. Leinonen R, Sugawara H, Shumway M. et al. The sequence read archive. *Nucleic Acids Res* 2010;**39**:D19–21. <https://doi.org/10.1093/nar/gkq1019>.
3. Perez-Riverol Y, Bai J, Bandla C. et al. The pride database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022;**50**:D543–52. <https://doi.org/10.1093/nar/gkab1038>.
4. Kale NS, Haug K, Conesa P. et al. Metabolights: an open-access database repository for metabolomics data. *Curr Protoc Bioinform* 2016;**53**:14–3. <https://doi.org/10.1002/0471250953.bi1413s53>.
5. Wang Z, Lachmann A, Ma'ayan A. Mining data and metadata from the gene expression omnibus. *Biophys Rev* 2019;**11**:103–10. <https://doi.org/10.1007/s12551-018-0490-8>.
6. Wang Z, Monteiro CD, Jagodnik KM. et al. Extraction and analysis of signatures from the gene expression omnibus by the crowd. *Nat Commun* 2016;**7**:12846. <https://doi.org/10.1038/ncomms12846>.
7. Aronson AR, Lang F-M. An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**: 229–36. <https://doi.org/10.1136/jamia.2009.002733>.
8. Tanenblatt MA, Coden A, Sominsky IL. The conceptmapper approach to named entity recognition. *International Conference on Language Resources and Evaluation*, pp. 546–51, 2010.
9. Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the sequence read archive. *Bioinformatics* 2017;**33**:2914–23. <https://doi.org/10.1093/bioinformatics/btx334>.
10. Denecke K, May R, Rivera-Romero O. Transformer models in healthcare: a survey and thematic analysis of potentials, shortcomings and risks. *J Med Syst* 2024;**48**:23. <https://doi.org/10.1007/s10916-024-02043-5>.
11. Kalyan KS, Rajasekharan A, Sangeetha S. AMMU: a survey of transformer-based biomedical pretrained language models. *J Biomed Inform* 2022;**126**:103982. <https://doi.org/10.1016/j.jbi.2021.103982>.
12. Cannizzaro G, Leone M, Bernasconi A. et al. Automated integration of genomic metadata with sequence-to-sequence models. In: *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020*, 12461, pp. 187–203. Springer, 2021. [https://doi.org/10.1007/978-3-030-67670-4\\_12](https://doi.org/10.1007/978-3-030-67670-4_12).
13. Serna Garcia G, Leone M, Bernasconi A. et al. GEMI: interactive interface for transformer-based genomic metadata integration. *Database* 2022;**2022**:baac036. <https://doi.org/10.1093/database/baac036>.

38. Feder A, Keith KA, Manzoor E. *et al.* Causal inference in natural language processing: estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics* 2022;**10**:1138–58. [https://doi.org/10.1162/tacl\\_a\\_00511](https://doi.org/10.1162/tacl_a_00511).
39. Jaimini U, Sheth A. CausalKG: causal knowledge graph explainability using interventional and counterfactual reasoning. *IEEE Internet Comput* 2022;**26**:43–50. <https://doi.org/10.1109/MIC.2021.3133551>.