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

Hao Yuan[1,2], **Parker Hicks**[3], **Mansooreh Ahmadian**[4], **Kayla Johnson**[3], **Lydia Valtadoros**[3], and **Arjun Krishnan**[3]

[1]Michigan State University, Genetics and Genome Sciences Program, East Lansing, 48823, USA
[2]Michigan State University, Ecology, Evolution, and Behavior, East Lansing, 48823, USA
[3]University of Colorado Anschutz Medical Campus, Department of Biomedical Informatics, Aurora, 80045, USA
[4]University of Colorado Anschutz Medical Campus, Department of Biostatistics and Informatics, Aurora, 80045, USA
[*]Email: arjun.krishnan@cuanschutz.edu

Abstract

Reusing massive collections of publicly available biomedical data can significantly impact knowledge discovery. However, these public samples and studies are typically described using unstructured plain text, hindering the findability and further reuse of the data. To combat this problem, we propose *txt2onto 2.0*, a general-purpose method based on natural language processing and machine learning for annotating biomedical unstructured metadata to controlled vocabularies of diseases and tissues. Compared to the previous version (*txt2onto 1.0*), which uses numerical embeddings as features, this new version uses words as features, resulting in improved interpretability and performance, especially when few positive training instances are available. *Txt2onto 2.0* uses embeddings from a large language model during prediction to deal with unseen-yet-relevant words in the input text and to highlight biomedical concepts in the input text that are related to each disease and tissue term being predicted, thereby explaining the basis of every annotation. We demonstrate the generalizability of *txt2onto 2.0* by accurately predicting disease annotations for studies from independent datasets, using proteomics and clinical trials as examples. Overall, our approach can annotate biomedical text regardless of experimental types or sources. Code, data, and trained models are available at https://github.com/krishnanlab/txt2onto2.0.

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], 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, 8, 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].

Traditional machine learning (ML)-based approaches have also been used to extract information from metadata [16, 6]. For example, Wang *et al.* [6] manually curated studies related to disease, drug, and gene perturbations from GEO. Then, they converted metadata to 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 transcirptomics studies from GEO, are proficient at infer disease annotations for biomedical metadata from any source.

Results

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**).

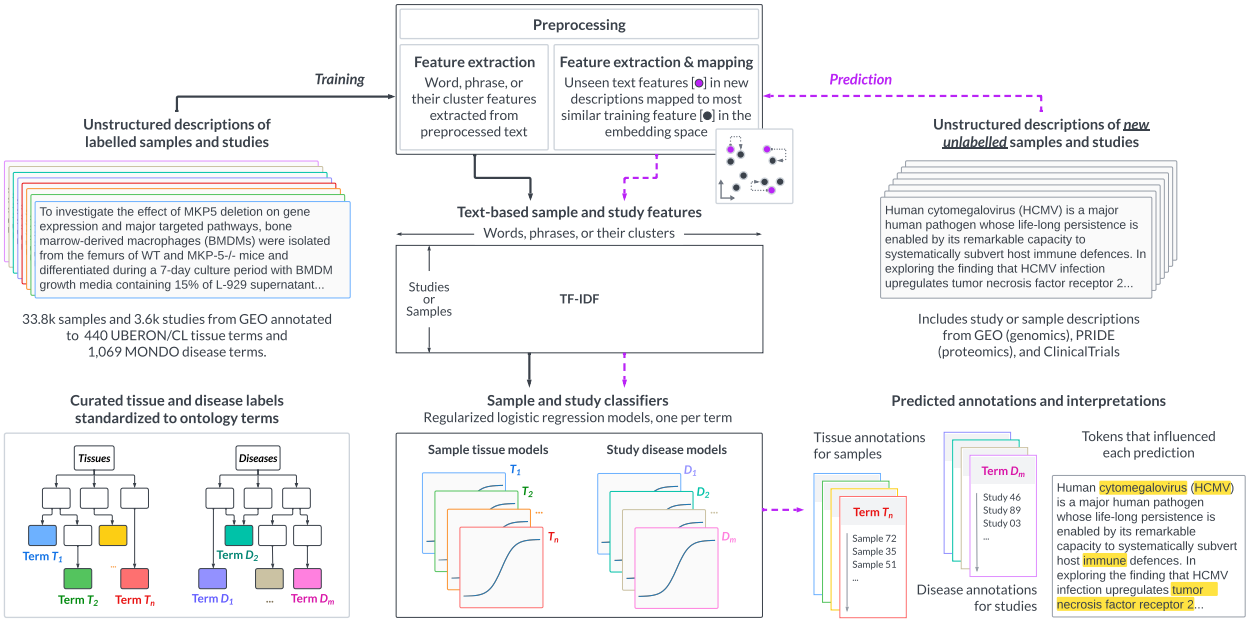


Figure 1: **Overview of *txt2onto 2.0***. Samples and studies from the GEO public database are annotated with cell type, tissue, and disease ontology terms to create a gold standard dataset with 33,849 samples and 3,676 studies. The text descriptions for these samples and studies are preprocessed to remove elements that do not contribute to identifying relevant tissue or disease terms. This step generates distinct corpora for samples and studies. TF-IDF features are then derived from these corpora for various entities, including individual words, word clusters, and phrase clusters in each sample or study. A supervised classification model is trained for each tissue (n=202) or disease (n=177) term using the extracted TF-IDF features. When predicting annotations for unseen samples or studies, unlabeled text descriptions undergo the same workflow: preprocessing, TF-IDF feature extraction, and then input into the trained models, which produce a probability of a given sample belonging to a specific tissue, cell type, or disease. However, during TF-IDF feature extraction for new metadata, words or phrases that were unseen during training are mapped to their most similar word or phrase in the training feature set calculated via the cosine similarity of their embeddings.

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 *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 that lemmatization (*i.e.*, reducing different word forms to a common root) had little to no effect on performance (**Fig. S2**), we kept all words in their original forms.

We then used the processed metadata to create a training corpus with a TF-IDF feature matrix. In this matrix, the rows represent samples or studies and the columns represent the text-based features. We extracted words and phrases (multi-word terms) as features. Further, to account for redundancy among semantically-related words and phrases, we grouped similar words or phrases into clusters, and treated each cluster as a new feature (see *Methods, Converting preprocessed text into TF-IDF matrix*). Overall, our approach considers four types of entities as features: words, phrases, word clusters, and phrase clusters.

To train classification models, we created a gold-standard dataset consisting of 33,849 samples and 3,676 studies from various sources [19, 20, 21, 22, 18, 23, 24] (including in-house curated data; **Table S1 and S2**) labeled with terms in the Uberon-CL tissue/cell ontology [25] and the MONDO disease ontology [26]. Upon filtering for redundancy and for availability of enough examples for hyperparameter tuning and independent testing (see *Methods, Selecting terms for model evaluation*), we were left with 202 specific tissue/cell-type terms and 177 disease terms. For each of these terms, using the gold-standard labels and TF-IDF feature matrix, we trained a one-vs-rest logistic regression (LR) classifier to predict whether a given sample or study is associated with that term or not. We selected LR model over other popular models such as support vector machine or random forest since LR outputs well-calibrated probability [27]. We used ElasticNet to regularize the classifier [28] and tuned the regularization strength and the mixing hyperparameters (see results in **Supplementary file 1**).

To make a prediction for a new sample or study, its metadata text was processed and transformed into a TF-IDF vector following the same procedure as above. Given that the new metadata might contain unseen but informative entities that do not exist among the features in the training set, we mapped unseen entities to their closest training features by comparing the cosine similarity between their text embeddings from a langauge model (see Methods, *Converting preprocessed text into TF-IDF matrix*). This approach allowed us to map the TF-IDF vector of the new metadata to the same feature space as the trained classifier. Once supplied with this input, the LR classifier for each term outputs a probability of that sample/study being annotated with that standardized tissue or disease term. Since *txt2onto 2.0* accepts just plain text as input, the bank of pre-trained LR classifiers can be used to make disease or tissue predictions from any type of biomedical text and are not limited to GEO metadata.

Txt2onto 2.0 outperforms *txt2onto 1.0* especially for terms with limited training data

We evaluated the performance of the new LR classifiers with various features (words, phrases, word clusters, and phrase clusters), comparing them to each other and to *txt2onto 1.0* to identify the best-performing

approach (**Fig. 2**). Overall, the new models outperform *txt2onto 1.0* in both disease and tissue prediction tasks (**Fig. 2a**). Among the new models, those using word-based features consistently provide the best performance compared to using any of the other text features. For tissue classification, word-based and phrase-based models deliver better results compared to *txt2onto 1.0*, while cluster-based models fall short. For the disease classification, all our proposed feature combinations demonstrate significant improvement over *txt2onto 1.0*.

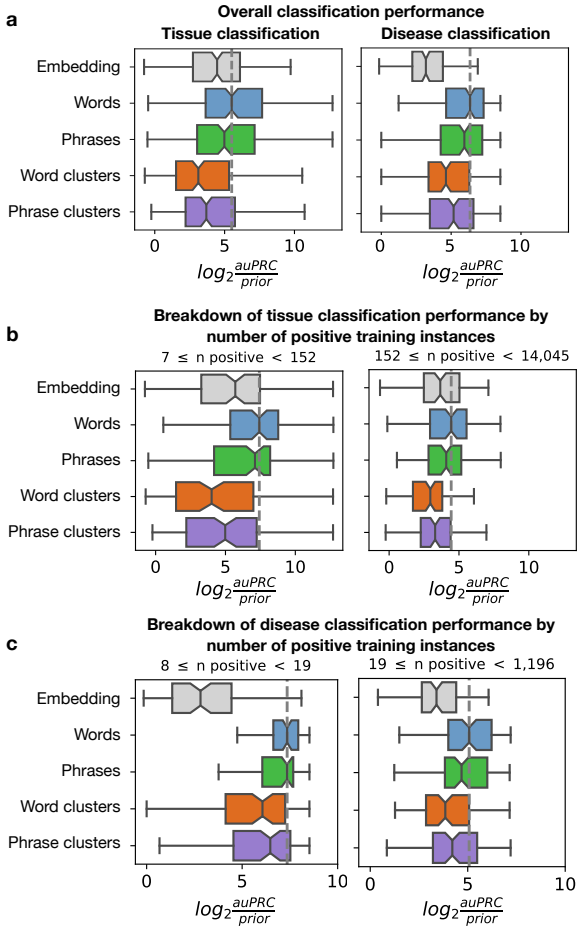


Figure 2: **Performance of disease and tissue predictions.** Boxplots display the distribution of $log_2(auPRC/prior)$ scores across 202 tissues and 177 diseases for each entities (words, phrases, word clusters, phrase clusters) using the LR model, along with *txt2onto 1.0* (embeddings). The auPRC is the area under the precision-recall curve. To account for varying class imbalances across terms, it is normalized by the prior, which is the ratio of positive instances to the total instances in the training set. Each point in the boxplots represents the performance of a disease or tissue model on the test set. Only non-redundant tasks, where the training set could be split into at least two folds, are included. The gray dashed lines indicate the median performance of the best-performing model (logistic regression with word features). (**a**) Boxplots showing overall performance for tissue (left) and disease (right) classification. (**b, c**) Boxplots in (b) and (c) show detailed performance for tissue and disease classifications, respectively, with tasks divided into two quantiles based on the number of positive instances in the training set.

Given that the number of positive instances (**Fig. S3**) can significantly impact model performance, we examined the performance within subsets of terms split by the numbers of positive instances. The results

show that our new models improve performance across terms with different training set sizes and are particularly suited for predicting terms when only a small number of positive examples are seen during training (**Fig. 2b-c**). This subset corresponds to terms that are either understudied in general or underrepresented in manually-curated labeled datasets. For instance, in our gold standard label set, 49 disease ontology terms contain only 8 to 11 positive instances each. We confirmed that our new model has the best performance for these terms with very few positives too (**Fig. S4a and b**). In extreme cases, some terms 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.

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 (**Fig. S6 and S7**) 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, in 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. In 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**).

Figure 3: **Predictive words for disease and tissue classification.** (**a, b**) Word clouds highlight the most predictive word features in the top-performing three tissue (a) and disease (b) classification models, which were trained using the best-performing model (logistic regression with word features). The size of each word is proportional to its regression coefficient in the logistic regression model, with larger words indicating higher coefficients and greater predictive importance. (**c**) The boxplots compare the change in the distribution of $log_2(auPRC/prior)$ scores for tissue (left) and disease (right) classification performance when task names are removed from the corpus.

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 information content 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. 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.

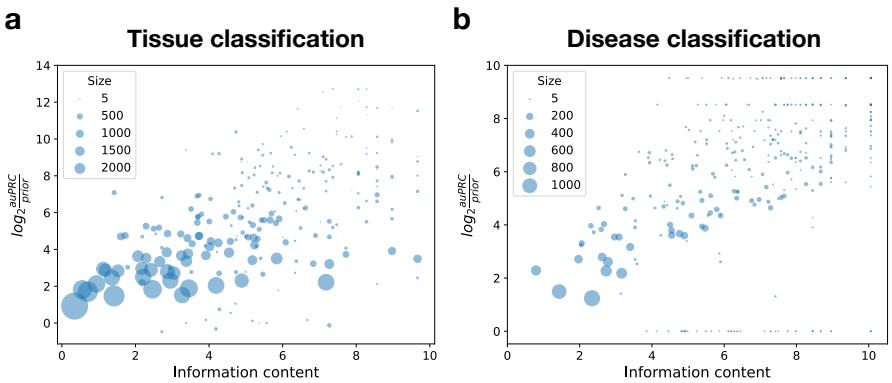


Figure 4: **Effect of term specificity on prediction performance.** Scatter plots show the prediction performance of the best model (logistic regression with word features) for tissue (**a**) and disease (**b**) classification tasks across various levels of term specificity. Each point represents the performance of a single tissue or disease model. The x-axis shows the specificity of the disease or tissue terms, defined by their information content (IC). The IC of a term is the negative logarithm of the fraction of terms in the ontology that are descendants of term of interest. The y-axis shows the model performance, measured by $log_2(auPRC/prior)$. The size of each data point is proportional to the number of positive instances in the training set for each task.

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 BioMedBERT [29] (see *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 to instances (sample or studies) labeled to that term compared to instances labeled to other terms in the same bin.

The auROC distribution within each bin being narrow and centered close to 1.0 confirms that our models can differentiate between terms irrespective of their semantic similarity (**Fig. 5**). Particularly, the *txt2onto*

2.0 classifiers can, in general, correctly prioritize instances annotated to the correct term over instances annotated to a closely-related term. However, as reflected in the increased number of outliers in the auROC distributions, there are indeed cases where highly semantically-similar terms prove challenging for our models. We examined four such challenging disease pairs, aiming to understand why they are difficult to distinguish (**Supplementary Note 3**). The prominent reason seems to be that predictive words extracted from the input text overlap between the disease pairs (**Fig. S8**). For instance, the words "lung", "cell", and "cancer" appeared frequently in the metadata of instances annotated to *Squamous cell lung carcinoma* and *Small cell lung carcinoma* (**Fig. S8e**). Although "squamous" is a predictive word for *Squamous cell lung carcinoma*, this word does not appear frequently among studies of this disease. Consequently, the signal from the word "squamous" is not strong enough to be useful in effectively distinguishing between these two textually-similar lung cancers. We observed that such dilution or absence of discerning predictive words in metadata might result from mislabeling. For example, GSE28582 and GSE21933, two studies annotated to *Squamous cell lung carcinoma* in our current gold-standard, should instead have been appropriately labeled as *Non-small cell lung carcinoma*, a more general disease term, which would explain the absence of the word "squamous" in the metadata of these two studies.

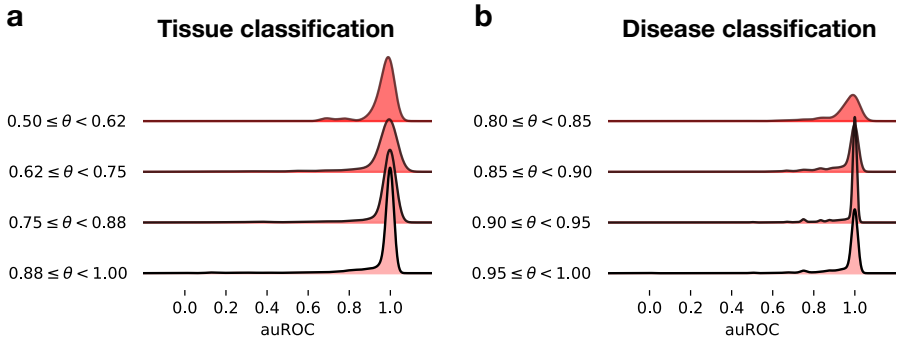


Figure 5: **Model's ability to annotate to the correct disease or tissue term over similar terms.** Ridgeline plots show the distribution of auROC scores for the best-performing model (logistic regression with word features) in distinguishing samples from pairs of (**a**) tissue terms and (**b**) disease terms with varying levels of semantic similarity (θ). Semantic similarities between terms are measured by the cosine similarity of text embeddings, where larger values indicate greater semantic similarity. The range of semantic similarities, from minimum to one, is divided into four equal-sized intervals. The x-axis represents the auROC score, while the height of each ridgeline plot represents the frequency of term pairs at each auROC level.

Txt2onto 2.0 models accurately predict disease labels for completely independent studies

Although all the models were only trained on metadata from GEO (principally from transcriptomics studies), we wanted to test the generalizability of the models by evaluating their ability to annotate the metadata of studies from completely independent data repositories. As study-level descriptions are more widely available than tissue-level descriptions, we focused on evaluating the disease annotation performance. First, we assessed if our models trained on transcriptomics metadata could annotate metadata of other omics studies. For this analysis, we used our models to annotate proteomics studies from the Proteomics Identification Database (PRIDE) [3]. Second, we tested whether our models could annotate biomedical studies beyond omics. We used the descriptions of clinical trials across various biomedical domains from ClinicalTrials [30] as a case study for this test. To quantify the prediction performance on these independent datasets, we conducted a double-blind test to manually assess the predictions from the top six models (see *Methods, Predicting on independent dataset*). Our models achieved high accuracy and F1 scores when predicting

disease labels for studies from PRIDE (**Table 1**) and ClinicalTrials (**Table 2**), demonstrating the potential generalizability of *txt2onto 2.0*.

Term	Acc	F1	log2(Acc/R Acc)	log2(F1/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 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: Acc (Accuracy), calculated from unshuffled predictions; log2(Acc/R Acc), the accuracy normalized by the random accuracy calculated from shuffled predictions, then log base two transformed; and log2(F1/R F1), the F1 score normalized by the random F1 score calculated from shuffled predictions, then log base two transformed.

Term	Acc	F1	log2(Acc/R Acc)	log2(F1/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

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: Acc (Accuracy), calculated from unshuffled predictions; log2(Acc/R Acc), the accuracy normalized by the random accuracy calculated from shuffled predictions, then log base two transformed; and log2(F1/R F1), the F1 score normalized by the random F1 score calculated from shuffled predictions, then log base two transformed.

Inspecting the predictive words in the input metadata confirmed that the text signals in these independent datasets that drive specific disease annotations are indeed related to the disease (**Fig. 6**). For example, the annotation of ClinicalTrials studies to *Allergic rhinitis* (MONDO:00117886) is based on words like "allergic", "rhinitis", "nasal", "allergen", and "asthma". These words align with typical symptoms and triggers for allergic rhinitis, which involves nasal inflammation due to allergens (**Fig. 6a**). Similarly, the annotation of PRIDE (proteomics) studies to *COVID-19* (MONDO:0100096) is based on keywords such as "sarscov2", "covid19", "infection", "severe", and "sarscov". These words reflect the main factors associated with COVID-19, including its causative virus, SARS-CoV-2, which leads to infection (**Fig. 6b**). Another noteworthy example is the prediction to the term *Gastric adenocarcinoma* (MONDO:0005036). Here, the word "gastrectomy" is present in metadata of relevant ClinicalTrials (**Fig. 6a**) studies influences their annotation, but is not one of training features for this disease. Nevertheless, our model aligns this unseen word to its closest training feature "gastric", enabling it to contribute to the prediction. Examples like these affirm our approach's capability of effectively using even previously unseen words if they are related to the annotation task under consideration.

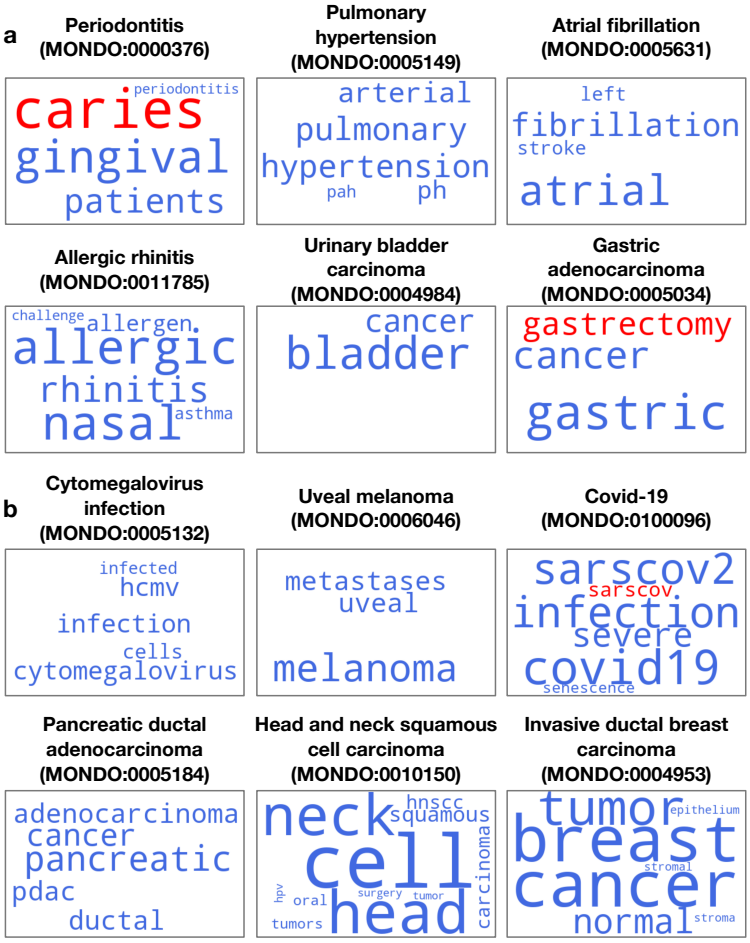


Figure 6: **Disease-related words derived from databases containing information beyond expression profiles.** Word clouds display disease-related words extracted from study titles and summaries of two independent datasets: (**a**) ClinicalTrials and (**b**) the PRIDE database. These predictions were generated by disease models trained on metadata from the GEO database, using the best-performing model (logistic regression with word features). The size of each word is proportional to its frequency among disease-related words in ClinicalTrials or PRIDE, with larger words being more prevalent. Blue words appear in both the predictive words from the independent datasets and the training features of the models, while red words are unique to the independent datasets.

More detailed results for the performance of our models in annotating text data from data sources other than GEO are given in the supplementary materials (see **Tables S3-S4**). Together, these analyses demonstrate the ability of *txt2onto 2.0* in annotating any biomedical text description to appropriate ontology terms. The interpretability of our models allows one to trace back the words contributing to the prediction, providing a way to easily evaluate and confirm the annotations.

Discussion

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. S9**). 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 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 (**Fig. S3**) due to most annotation tasks (tissue and disease terms) only having limited number of positive instances. Despite using a LR model, which typically outputs well-calibrated probabilities, the high class imbalance leads to underestimated probabilities for positive instances. We applied two strategies to identify likely positive predictions while dealing with underestimated probabilities: (1) Train a recalibration model to project underestimated probabilities to the 0–1 range with pivot point close to 0.5, or (2) compare predicted probabilities to expected random classification performance. We did not choose undersampling or oversampling strategies as they might not work for extremely imbalanced datasets, and correcting class imbalance may worsen the reliability of estimates [35]. Additionally, we found that models trained and tested on imbalanced datasets are extremely vulnerable to mislabeled instances because incorrect signals can easily dominate training or evaluation, leading to poor performance (see **Supplementary Note 2**). Possible approaches to combat this problem include using methods that allow for naturally updating the labels, such as active learning [36] and semi-supervised learning [37], or building a unified model for all tissue or disease classification tasks to borrow information from related tasks.

Given that inferring standardized annotations of metadata is a universal problem pertinent to discovering and reusing data in every biomedical repository, generalizability is a fundamental requirement of automatic

annotation inference methods. It 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.

Methods

Preparing gold standard

Collecting labeled examples and constructing gold standard

We obtained tissue labels for samples from GEMMA [21], Hawkins *et al.* [18], Gu *et al.* [22], Bernstein *et al.* [20], Sirota *et al.* [19], TissueNexus [23] and an in-house curated dataset, covering 33,849 samples and 440 tissue types. If samples overlapped among these datasets, labels were prioritized in the following order: Hawkins *et al.* > in-house > TissueNexus > Sirota *et al.* > Gu *et al.* > Bernstein *et al.*. We obtained disease labels for studies from multiple sources, including GEMMA (accessed on 7/9/2023) [21], Gu *et al.* [22], Sirota *et al.* [19], USRA-HD [24], and an in-house curated dataset, covering 3,676 studies and 1,069 disease types. If there were duplicated studies included, we prioritized the annotations in the following order: in-house > USRA-HD > Sirota *et al.* > Gu *et al.* > GEMMA. All tissue and disease annotations were mapped to MONDO and UBERON-CL, respectively.

Next, we used the annotations of samples and studies to build a gold standard following the procedure described in Lee *et al.* [40]. For every term, if a sample or study was directly annotated to the term or was a descendant of the term, it was labeled as positive. If a sample or study was annotated to an ancestor of a term, it was labeled as unknown since the annotation was ambiguous. The rest of the samples or studies were labeled as negative. Samples or studies from GEMMA and Bernstein *et al.* might be labeled to multiple terms. For those cases, if a sample or study could be labeled as positive based on at least one of the terms, it was labeled as positive. For samples or studies that could not be labeled as positive, they were labeled as negative if at least one of the terms was labeled as negative. The remaining cases were labeled as unknown.

Selecting terms for model evaluation

As described above, since we propagated labels along the ontology, positive labels of closely related parent-child terms might largely overlap. These redundant tasks could potentially bias the performance score distribution by inflating the scores for overlapping terms. To mitigate this issue, we found groups of redundant terms by comparing shared positive instances and then selecting representative terms from each group. The procedure is conducted as follows. First, we measured the similarity of positive instance sets between terms using the Jaccard index. Then, we constructed a network of terms in which two terms were connected if their Jaccard index exceeded the threshold, meaning the pair of terms shared a significant number of positive instances. We set 0.9 and 0.5 as the minimum Jaccard index threshold for tissue and disease, respectively. Next, we extracted all connected components from the term network, and each connected component contained redundant terms. To extract a set of representative terms from the connected components, we first ordered the terms in descending order based on the number of neighbors they have. For each term, we formed a cluster with the term and its neighbors. For each cluster, we selected a representative term that was annotated to the highest number of instances in the original sample or disease labels before propagation, then removed the entire cluster from the connected component. We repeated this process until all connected components were depleted. Notably, connected components could contain only one term, so we did not miss terms unrelated to any other terms.

To guarantee robust training and evaluation, we selected a subset of non-redundant terms with sufficient positive instances for training. For disease terms, we first split the dataset into training and testing sets in an 80/20 ratio. The training set was further split into at least two folds for hyperparameter tuning. To ensure unbiased model evaluation, the difference in the positive-to-negative ratio between the training and testing sets was controlled to be under 0.1. This same criterion was applied to the ratio difference between folds. The data splitting for tissue terms also required all the above criteria. Additionally, samples from the same study must be kept within the same set or fold to avoid data leakage.

Among all disease and tissue/cell terms in the ontology, 590 tissue and 1,166 disease terms were trainable. After removing redundant tasks, 237 non-redundant tissue/cell terms and 353 disease terms remained suitable for training. These tasks can be used for training but may not have enough positive instances for hyperparameter tuning. Of these trainable terms, 202 tissue/cell and 177 disease terms had sufficient instances in the training sets to be further split into at least two folds, enabling hyperparameter tuning through cross-validation.

Downloading metadata from GEO

GEO houses comprehensive study and sample descriptions, so we aimed to download all metadata from GEO [41]. Since samples from Gu *et al.* and Bernstein *et al.* were named with Sequence Read Archive (SRA) IDs, we used the efetch tool version 17.0 [42] to map SRA sample IDs to GEO sample IDs. Samples not included in GEO were removed from the analysis. Similarly, we used the efetch tool to obtain the GEO study ID for each sample. Then, we downloaded the soft format file for each study directly from NCBI and extracted all text related to the sample and study descriptions.

Txt2onto 2.0

Preprocessing input

Preprocessing has a crucial effect on the model's ability to make accurate predictions. In addition to basic preprocessing steps such as removing non-UTF-8 characters, URLs, punctuation, whole numbers, and converting text to lowercase, we tested several additional preprocessing settings that might impact prediction performance. One preprocessing step involved removing task-irrelevant fields, as information from those fields might confound model training and prediction. For tissue classification, we treated "Sample_title", "Sample_source_name_ch1", "Sample_description" and "Sample_characteristics_ch1" as task-relevant fields. For disease classification, we treated "Series_overall_design", "Series_summary", and "Series_title" as task-relevant fields. We found that removing task-irrelevant fields does affect model performance (**Fig. S1**). Lemmatization is another widely applied technique for preprocessing text, which aims to reduce the feature space by converting words to their original root form. Lemmatization has a limited effect on model performance (**Fig. S2**). However, ineffective lemmatization could yield meaningless strings, especially for biomedical words. Hence, we opted to keep the text in its original form. Additionally, we defined a pool of text snippets to be removed from the corpus, primarily consisting of names of biomolecular reagent companies and reagent kits.

Converting preprocessed text into TF-IDF matrix

After preprocessing, we converted the preprocessed text into a TF-IDF matrix. TF-IDF is a popular method for extracting information from free text, measuring how frequently a word is observed in the document (term frequency, TF) and how unique this term is across the document corpus (inverse document frequency, IDF). Since the training and testing sets may not share the same corpus, the TF-IDF matrix features are calculated separately for each set.

For the training set, the TF-IDF was calculated as follows:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log \frac{D}{df_t} \tag{1}$$

For a given metadata in the training set, $\text{TF}(t, d)$ was the term frequency of a word t within that metadata d. D was the total number of documents in the training set, and df_t was the frequency of this word t occurring across the entire training corpus D. For each feature in the testing set, we aligned it to the most similar feature in the training set by comparing its cosine similarity of text embedding to training features. If multiple testing features were mapped to the same training feature, their term frequencies were summed up, and the inverse document frequency was inherited from corresponding features in the training set. After normalizing the TF-IDF by instance, we obtained TF-IDF matrix of a testing set with the same features as the training set. The similarity between training and testing features was measured as the cosine similarity between their embeddings. We tested the effect of selecting different language models and found that the choice of embedding has a limited impact on performance (**Fig. S10**). We selected the base uncased version of BioMedBERT [29] pretrained on abstracts to generate embeddings using PyTorch 2.1.0 [43] and the Transformer 4.36.2 Python library [44].

Since individual words may not fully capture the meaning of biomedical terms, we also tried using phrases as features. Phrases were extracted from the text using spaCy 3.7.4 [45] conjunct with "en_core_sci_md" model, which is a spaCy model used to process biomedical text. In this case, TF was the phrase frequency, and IDF was the uniqueness of this phrase across the corpus. To reduce redundant entities, we built a K-nearest neighbor graph, where each node was an entity, and edges were the cosine similarity between entities ($K = 1$). Clusters were then found using Leiden community detection [46]. The TF of a cluster was calculated as the sum of TF of all entities in the cluster, and the IDF was the inverse frequency of at least one entity in the cluster being observed across the corpus.

Training the LR classifiers

We chose the logistic regression model as the predictive model due to its superior performance compared to other classifiers in *txt2onto 1.0* [18], and its output probability is well-calibrated. ElasticNet was selected as the penalty term since it can reduce the influence of features unrelated to the prediction while keeping features with similar meanings [28]. This is particularly useful when dealing with sparse matrices like the TF-IDF matrix. The model was trained using scikit-learn version 1.2.1 [47]. We conducted cross-fold validation training to determine the best hyperparameters. Based on the results of hyperparameter tuning (see **Supplementary file 1**), we built the model with the following parameters: $C = 1$, $alpha = 0.4$, $penalty = elasticnet$, and $solver = liblinear$. Notably, despite the model yielding the best performance when $C = 1.0$ and $alpha = 0.0$ in most cases, we raised alpha to 0.4, since we wanted the model to penalize features to zero while maintaining comparable performance when $C = 1.0$ and $alpha = 0.0$. By setting alpha to 0.4, the model can effectively regularize features to zero while maintaining a performance level similar to the optimal settings.

Evaluating classification results

We measured classification performance on the testing set using $\log_2$(auPRC/prior), the auPRC normalized by the prior, which is the ratio of positive instances to the total instances in the training set. This normalization addresses class imbalance, given that positive samples are often much fewer than negative ones for most tissue or disease terms. Normalization also allows for consistent comparison across different tissue and disease terms, despite variations in the distribution of positive samples.

Differentiating between similar terms

We compared the model's performance when differentiating similar and dissimilar terms. Semantic similarity between terms was measured by the cosine similarity between the text embeddings of each pair of terms' names, which was generated by BioMedBERT. For every pair of terms, we used the trained model from one of the terms to predict the probability of instances annotated to these two terms in the testing set. We treated the instances annotated to the same terms as the task of the model as positive instances, then treated the instances related to the other term in the pair as negative instances. We expected the probability of the positive instances to be higher than the negative instances, and measured the performance of differentiating positive and negative instances by auROC. We did not use $\log_2$(auPRC/prior) as in assessing the general performance of the models since the ratio between positive and negative instances is not proportional to the one in the training set, so the prior cannot inform random classification performance in this case. Additionally, since labels of instances have been propagated to general terms, an instance could be labeled to multiple terms. We removed instances that appeared in both positive and negative sets. We also did not include low-performance models with $\log_2$(auPRC/prior) lower than two in the analysis, as they cannot effectively distinguish among terms. Finally, we included 408 terms and 92,517 pairs of terms for disease classification, and 237 terms and 39,361 pairs of terms for tissue classification.

Predicting on independent datasets

Besides annotating metadata in GEO, we also wanted to test the performance of our model when predicting on independent datasets. We chose ClinicalTrials and PRIDE as test cases. We downloaded metadata for all studies from ClinicalTrials (accessed on 12/6/2023) and PRIDE (accessed on 1/22/2024). For efficient testing, we randomly selected 20,000 and 10,000 studies from ClinicalTrials and PRIDE, respectively, for disease annotation. Data were preprocessed in the same way as described for GEO metadata. We employed the LR model using words as features for prediction, which achieved the best performance over the rest of the entity types. Since the ratio of positive and negative instances was highly imbalanced in some tasks, the output probability from the model might be underestimated. We trained another Logistic Regression model using the probability before calibration and ground truth label from the testing set without penalty terms. Then, we applied it to recalibrate the prediction probability from independent datasets. Recalibrated

probabilities above 0.5 were considered positive predictions. Another useful metric for indicating good predictions was $log_2(probability/prior)$, which measured how much better the predicted probability was compared to the probability of random predictions. If this score exceeded 1, the prediction was considered a trustworthy positive prediction.

We also conducted a double-blind test to evaluate model performance on independent datasets. For each dataset, we selected six top-performing models with at least twenty positive predictions and at least three positive instances during training. Potential positive predictions were selected based on $log_2(probability/prior)$, considering instances as potential positives if the value exceeded 1. Then, we selected the top and bottom ten predictions and ordered them according to $log_2(probability/prior)$ in descending order, where the top ten were expected to be positive, and the bottom ten predictions were expected to be negative. We divided these predictions into two sets, each containing five positive and negative predictions. For the first set, everything remained the same, while for the second set, we kept the study order but shuffled the labels, purposely making some random predictions. Finally, we combined the two sets and shuffled the order of the studies with their labels. By comparing the performance of predictions from the first set, where the study and predicted label matched, to the second set, which represented random predictions, we can quantitatively evaluate the performance of predictions on independent datasets. The shuffled list was sent to three experts for manual evaluation. The averaged performance scores were used as the final score.

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.

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

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.

Ethics declarations

Competing interests

The authors declare no competing interests.

References

[1] Clough, E. & Barrett, T. The gene expression omnibus database. *Statistical Genomics: Methods and Protocols* 93–110 (2016).

[2] Leinonen, R., Sugawara, H., Shumway, M. & Collaboration, I. N. S. D. The sequence read archive. *Nucleic acids research* **39**, D19–D21 (2010).

[3] Perez-Riverol, Y. *et al.* The pride database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research* **50**, D543–D552 (2022).

[4] Kale, N. S. *et al.* Metabolights: an open-access database repository for metabolomics data. *Current protocols in bioinformatics* **53**, 14–13 (2016).

[5] Wang, Z., Lachmann, A. & Ma'ayan, A. Mining data and metadata from the gene expression omnibus. *Biophysical reviews* **11**, 103–110 (2019).

[6] Wang, Z. *et al.* Extraction and analysis of signatures from the gene expression omnibus by the crowd. *Nature communications* **7**, 12846 (2016).

[7] Aronson, A. R. & Lang, F.-M. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* **17**, 229–236 (2010).

[8] Tanenblatt, M. A., Coden, A. & Sominsky, I. L. The conceptmapper approach to named entity recognition. In *LREC*, 546–51 (2010).

[9] Bernstein, M. N., Doan, A. & Dewey, C. N. Metasra: normalized human sample-specific metadata for the sequence read archive. *Bioinformatics* **33**, 2914–2923 (2017).

[10] Denecke, K., May, R. & Rivera-Romero, O. Transformer models in healthcare: A survey and thematic analysis of potentials, shortcomings and risks. *Journal of Medical Systems* **48**, 23 (2024).

[11] Kalyan, K. S., Rajasekharan, A. & Sangeetha, S. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics* **126**, 103982 (2022).

[12] Cannizzaro, G., Leone, M., Bernasconi, A., Canakoglu, A. & Carman, M. J. 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, Proceedings, Part V*, 187–203 (Springer, 2021).

[13] Serna Garcia, G., Leone, M., Bernasconi, A. & Carman, M. J. Gemi: interactive interface for transformer-based genomic metadata integration. *Database* **2022**, baac036 (2022).

[14] Wu, Y. *et al.* Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[15] Tay, Y., Dehghani, M., Bahri, D. & Metzler, D. Efficient transformers: A survey. *ACM Computing Surveys* **55**, 1–28 (2022).

[16] Posch, L., Panahiazar, M., Dumontier, M. & Gevaert, O. Predicting structured metadata from unstructured metadata. *Database* **2016**, baw080 (2016).

[17] Wang, B. *et al.* Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys* **56**, 1–52 (2023).

[18] Hawkins, N. T., Maldaver, M., Yannakopoulos, A., Guare, L. A. & Krishnan, A. Systematic tissue annotations of genomics samples by modeling unstructured metadata. *Nature Communications* **13**, 6736 (2022).

[19] Sirota, M. *et al.* Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine* **3**, 96ra77–96ra77 (2011).

[20] Bernstein, M. N., Ma, Z., Gleicher, M. & Dewey, C. N. Cello: Comprehensive and hierarchical cell type classification of human cells with the cell ontology. *Iscience* **24** (2021).

[21] Lim, N. *et al.* Curation of over 10 000 transcriptomic studies to enable data reuse. *Database* **2021**, baab006 (2021).

[22] Gu, J., Dai, J., Lu, H. & Zhao, H. Comprehensive analysis of ubiquitously expressed genes in humans from a data-driven perspective. *Genomics, Proteomics and Bioinformatics* **21**, 164–176 (2023).

[23] Lin, C.-X., Li, H.-D., Deng, C., Guan, Y. & Wang, J. Tissuenexus: a database of human tissue functional gene networks built with a large compendium of curated rna-seq data. *Nucleic acids research* **50**, D710–D718 (2022).

[24] Lee, Y.-s. *et al.* A computational framework for genome-wide characterization of the human disease landscape. *Cell systems* **8**, 152–162 (2019).

[25] Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome biology* **13**, 1–20 (2012).

[26] Vasilevsky, N. A. *et al.* Mondo: Unifying diseases for the world, by the world. *medRxiv* 2022–04 (2022).

[27] Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, 625–632 (2005).

[28] Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301–320 (2005).

[29] Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing (2020). `arXiv:2007.15779`.

[30] U.S. National Library of Medicine. Clinicaltrials.gov. `https://clinicaltrials.gov/`. Accessed: 2023-12-06.

[31] Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics* **14**, 89–99 (2013).

[32] Sielemann, K., Hafner, A. & Pucker, B. The reuse of public datasets in the life sciences: potential risks and rewards. *PeerJ* **8**, e9954 (2020).

[33] Wilkinson, M. D. *et al.* The fair guiding principles for scientific data management and stewardship. *Scientific data* **3**, 1–9 (2016).

[34] Diz, J., Marreiros, G. & Freitas, A. Applying data mining techniques to improve breast cancer diagnosis. *Journal of medical systems* **40**, 1–7 (2016).

[35] van den Goorbergh, R., van Smeden, M., Timmerman, D. & Van Calster, B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association* **29**, 1525–1534 (2022).

[36] Zhang, Z., Strubell, E. & Hovy, E. A survey of active learning for natural language processing. *arXiv preprint arXiv:2210.10109* (2022).

[37] Shams, R. Semi-supervised classification for natural language processing. *arXiv preprint arXiv:1409.7612* (2014).

[38] Feder, A. *et al.* Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics* **10**, 1138–1158 (2022).

[39] Jaimini, U. & Sheth, A. Causalkg: Causal knowledge graph explainability using interventional and counterfactual reasoning. *IEEE Internet Computing* **26**, 43–50 (2022).

[40] Lee, Y.-s., Krishnan, A., Zhu, Q. & Troyanskaya, O. G. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics* **29**, 3036–3044 (2013).

[41] Edgar, R., Domrachev, M. & Lash, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207–210 (2002).

[42] National Center for Biotechnology Information (US). Entrez Programming Utilities (E-utilities). `https://www.ncbi.nlm.nih.gov/books/NBK25500/`. URL `https://www.ncbi.nlm.nih.gov/books/NBK25500/`. Accessed: 2024-04-10.

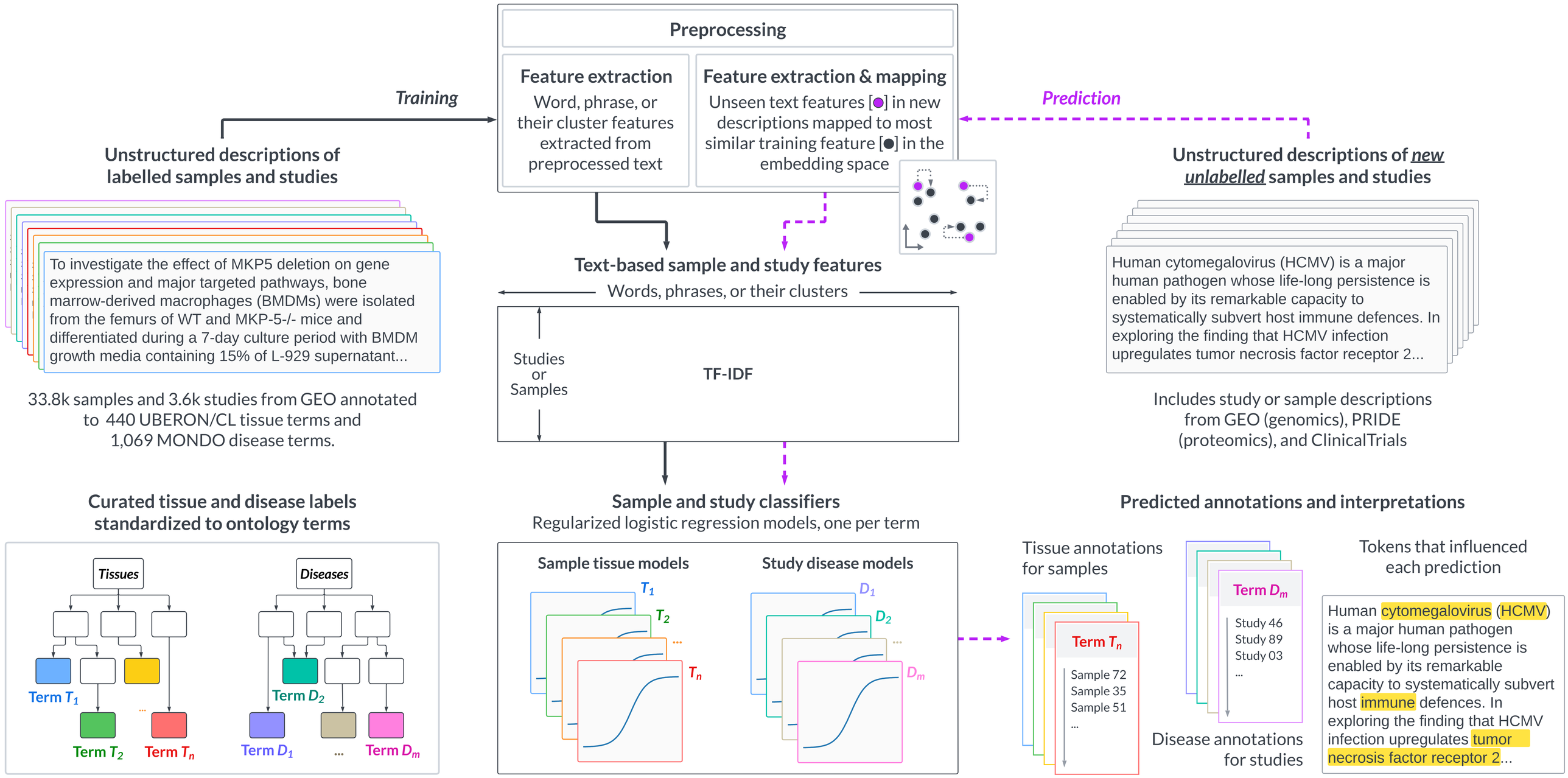
[43] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

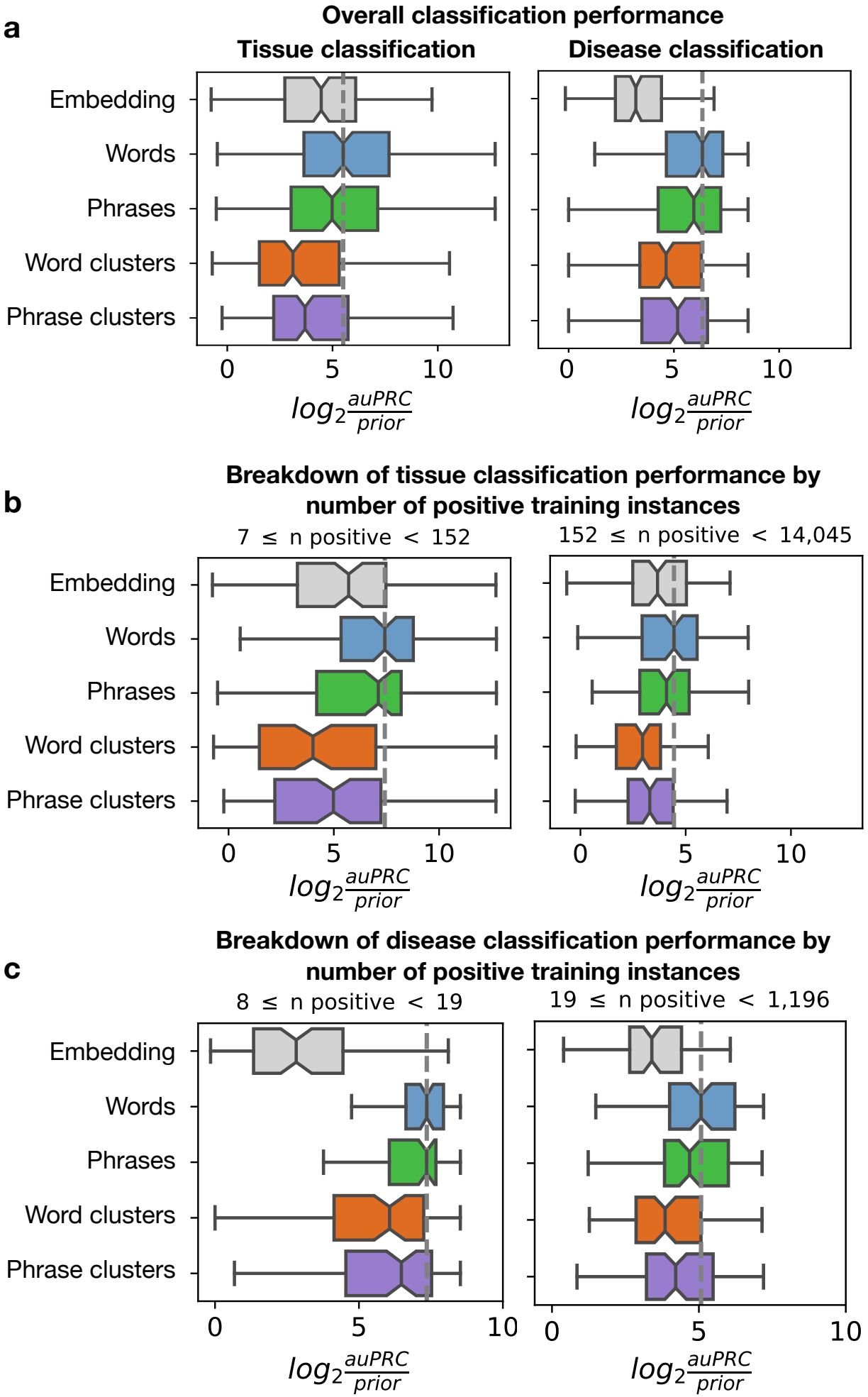
[44] Wolf, T. *et al.* Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[45] Honnibal, M. & Johnson, M. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1373–1378 (2015).

[46] Traag, V. A., Waltman, L. & Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 5233 (2019).

[47] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).


Preprocessing
Feature extraction
Word, phrase, or their cluster features extracted from preprocessed text
Feature extraction & mapping
Unseen text features [●] in new descriptions mapped to most similar training feature [●] in the embedding space
Training
Prediction
Unstructured descriptions of labelled samples and studies
To investigate the effect of MKP5 deletion on gene expression and major targeted pathways, bone marrow-derived macrophages (BMDMs) were isolated from the femurs of WT and MKP-5-/- mice and differentiated during a 7-day culture period with BMDM growth media containing 15% of L-929 supernatant...
33.8k samples and 3.6k studies from GEO annotated to 440 UBERON/CL tissue terms and 1,069 MONDO disease terms.
Text-based sample and study features
Words, phrases, or their clusters
Studies or Samples
TF-IDF
Unstructured descriptions of new unlabelled samples and studies
Human cytomegalovirus (HCMV) is a major human pathogen whose life-long persistence is enabled by its remarkable capacity to systematically subvert host immune defences. In exploring the finding that HCMV infection upregulates tumor necrosis factor receptor 2...
Includes study or sample descriptions from GEO (genomics), PRIDE (proteomics), and ClinicalTrials
Curated tissue and disease labels standardized to ontology terms
Tissues
Diseases
Term T_1
Term T_2
Term T_n
Term D_2
Term D_1
Term D_m
Sample and study classifiers
Regularized logistic regression models, one per term
Sample tissue models
Study disease models
T_1
T_2
T_n
D_1
D_2
D_m
Predicted annotations and interpretations
Tissue annotations for samples
Term T_n
Sample 72
Sample 35
Sample 51
...
Term D_m
Study 46
Study 89
Study 03
...
Disease annotations for studies
Tokens that influenced each prediction
Human cytomegalovirus (HCMV) is a major human pathogen whose life-long persistence is enabled by its remarkable capacity to systematically subvert host immune defences. In exploring the finding that HCMV infection upregulates tumor necrosis factor receptor 2...

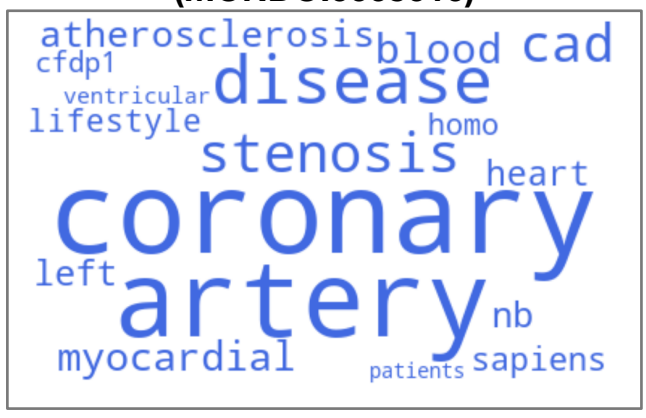

a
Overall classification performance
Tissue classification
Disease classification
Embedding
Words
Phrases
Word clusters
Phrase clusters
0
5
10
$log_2 \frac{auPRC}{prior}$
b
Breakdown of tissue classification performance by number of positive training instances
7 ≤ n positive < 152
152 ≤ n positive < 14,045
c
Breakdown of disease classification performance by number of positive training instances
8 ≤ n positive < 19
19 ≤ n positive < 1,196

a

Glucagon secreting cell (CL:0000170)

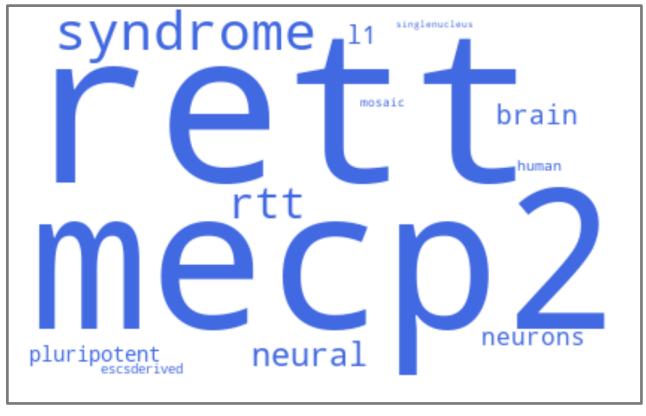
b

Coronary artery disorder (MONDO:0005010)

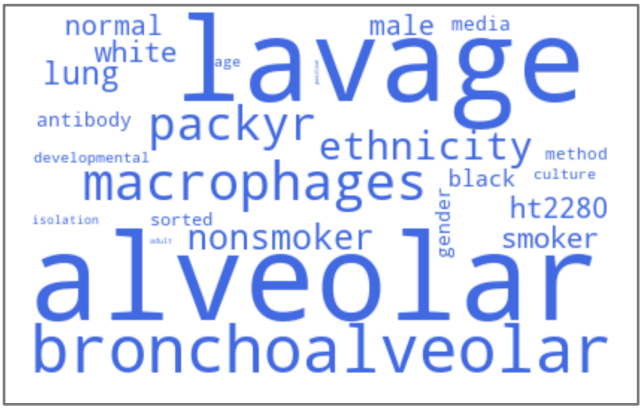


Erythrocyte (CL:0000232)

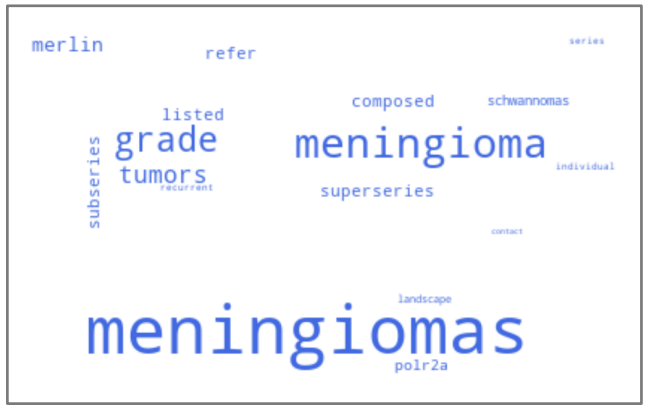
Rett syndrome (MONDO:0010726)



Pulmonary acinus (UBERON:0008874)



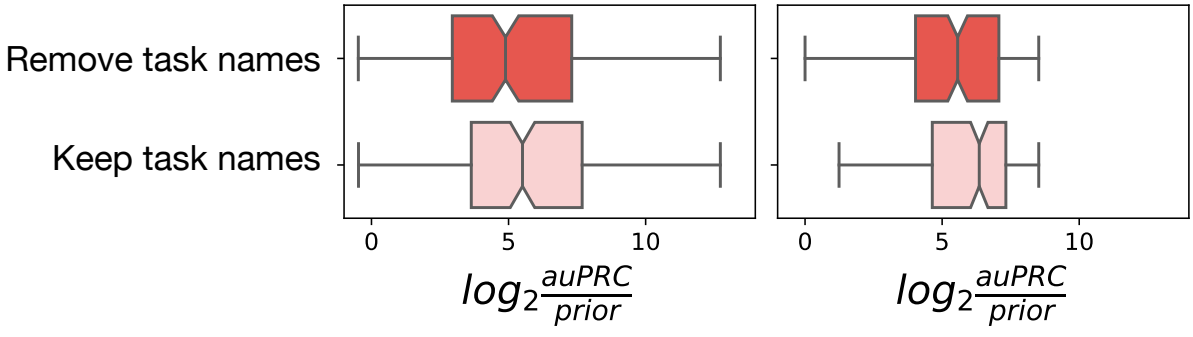
Meningioma (MONDO:0016642)

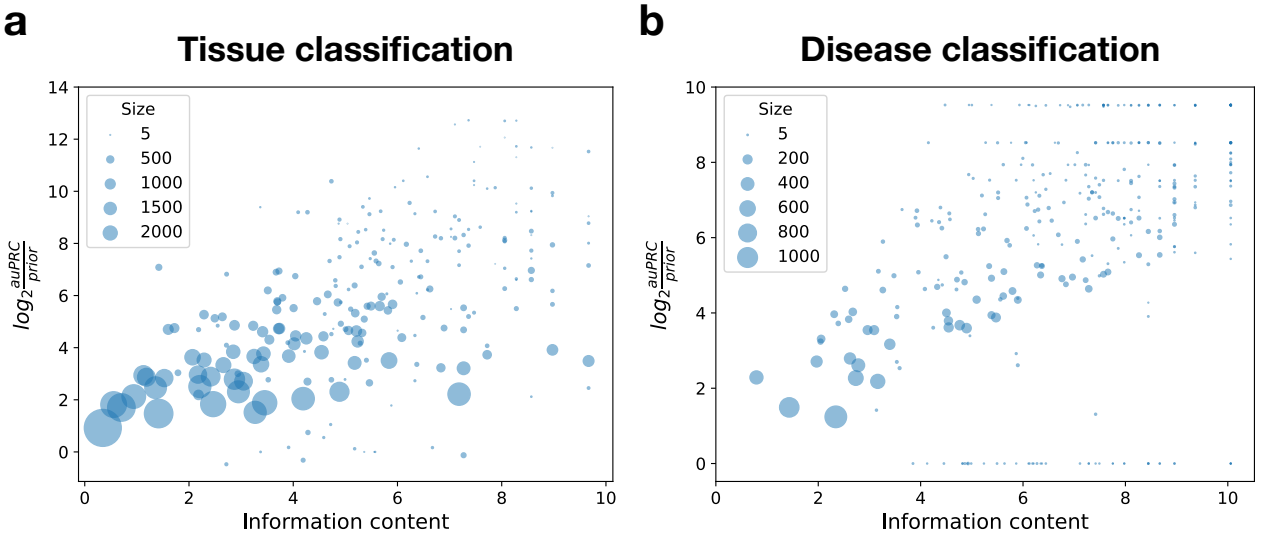


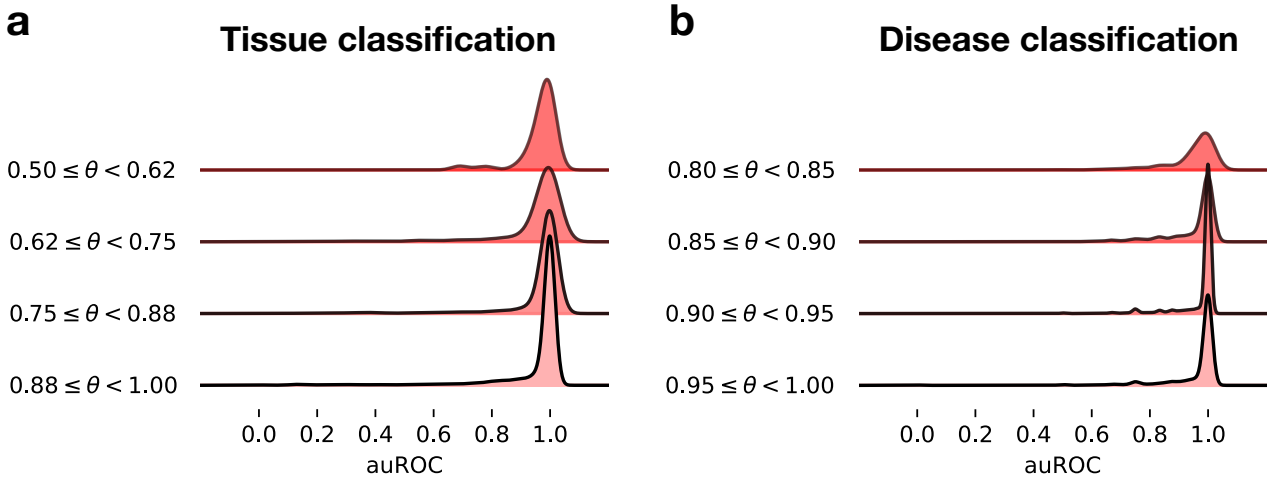
c

Overall performance of tissue classification

Overall performance of disease clasification



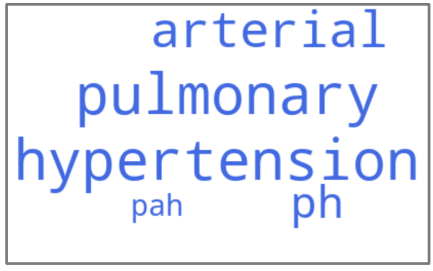

a
Tissue classification
Size
5
500
1000
1500
2000
$log_2 \frac{auPRC}{prior}$
Information content
b
Disease classification
Size
5
200
400
600
800
1000
$log_2 \frac{auPRC}{prior}$
Information content


a
Tissue classification
0.50 ≤ θ < 0.62
0.62 ≤ θ < 0.75
0.75 ≤ θ < 0.88
0.88 ≤ θ < 1.00
0.0
0.2
0.4
0.6
0.8
1.0
auROC
b
Disease classification
0.80 ≤ θ < 0.85
0.85 ≤ θ < 0.90
0.90 ≤ θ < 0.95
0.95 ≤ θ < 1.00
0.0
0.2
0.4
0.6
0.8
1.0
auROC

a

Periodontitis (MONDO:0000376)

Pulmonary hypertension (MONDO:0005149)



Atrial fibrillation (MONDO:0005631)

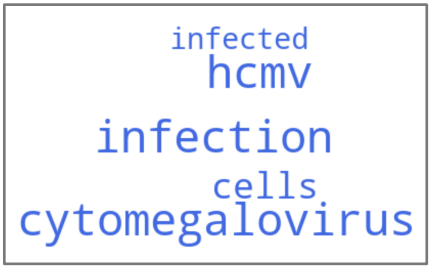
Allergic rhinitis (MONDO:0011785)

Urinary bladder carcinoma (MONDO:0004984)

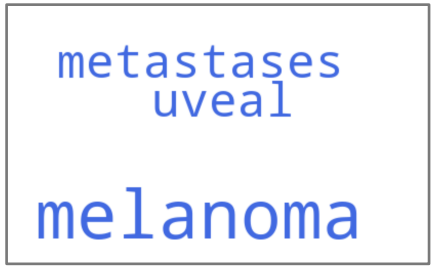
Gastric adenocarcinoma (MONDO:0005034)

b

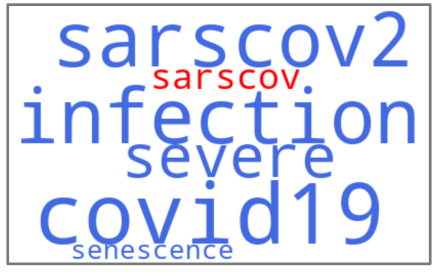
Cytomegalovirus infection (MONDO:0005132)



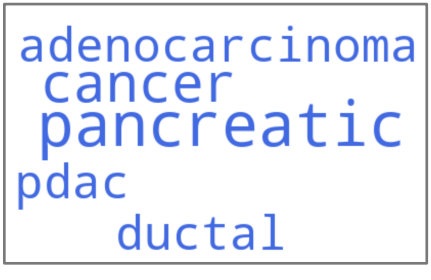
Uveal melanoma (MONDO:0006046)

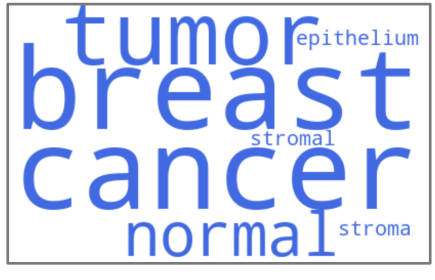


Covid-19 (MONDO:0100096)



Pancreatic ductal adenocarcinoma (MONDO:0005184)



Head and neck squamous cell carcinoma (MONDO:0010150)

Invasive ductal breast carcinoma (MONDO:0004953)