Systematic tissue annotations of genomic samples by modeling unstructured metadata

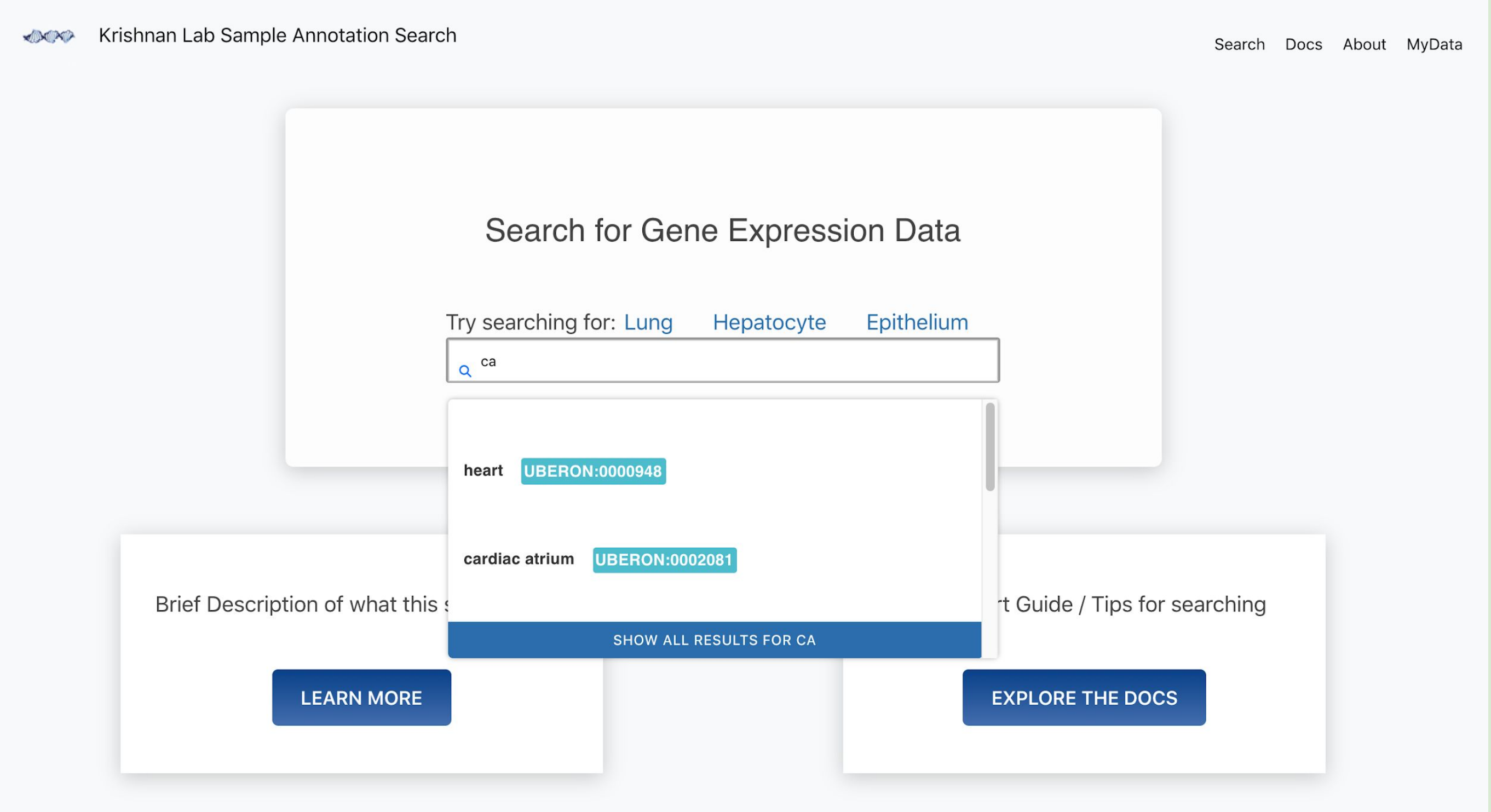
Nathaniel T. Hawkins[1], Marc Maldaver[1], Arjun Krishnan[1,2]
[1]Dept. of Computational Mathematics, Science and Engineering. Michigan State Univeristy. East Lansing, MI, United States.
[2]Dept. of Biochemistry and Molecular Biology. Michigan State University. East Lansing, MI, United States.

Overview

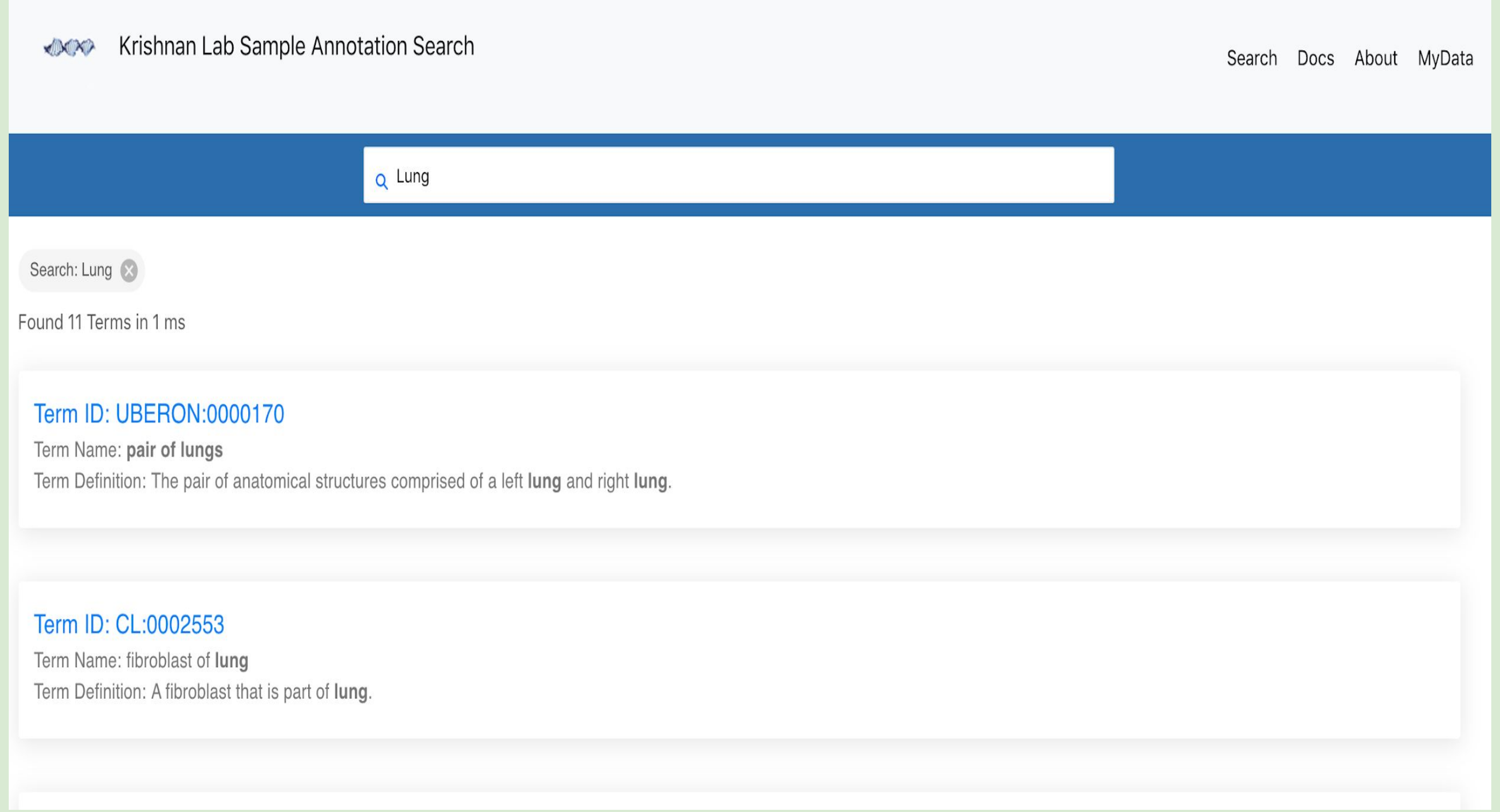
The reuse of publicly available omics data is limited by the lack of systematic annotations for tissues or cell types of origin. Previous work has used the underlying omics data to predict annotations for a given sample, but these tools are not widely usable due to the need to retrain for different data types. We propose a novel approach that leverages free-text, natural-language metadata to identify the tissue or cell-type of origin of omics samples. Using sample metadata files from ArrayExpress, we create a numerical representation for each sample using flair, a python NLP library. These representations are used train OVR logistic regression models, one for each tissue or cell type we have labels for, to predict a sample's annotations given its metadata. We compare our models' predictions to annotations predicted by models trained on gene expression data and those from two existing tools: TAGGER, an NER tool, and MetaSRA, a method that uses prior knowledge and ontologies.

WEB SERVER DEVELOPMENT

Example user query. Search text matched to available models using Elasticsearch


Krishnan Lab Sample Annotation Search
Search
Docs
About
MyData
Search for Gene Expression Data
Try searching for: Lung Hepatocyte Epithelium
ca
heart UBERON:0000948
cardiac atrium UBERON:0002081
SHOW ALL RESULTS FOR CA
Brief Description of what this s
LEARN MORE
rt Guide / Tips for searching
EXPLORE THE DOCS

Or user can search a term and select the most appropriate tissue from list of available models

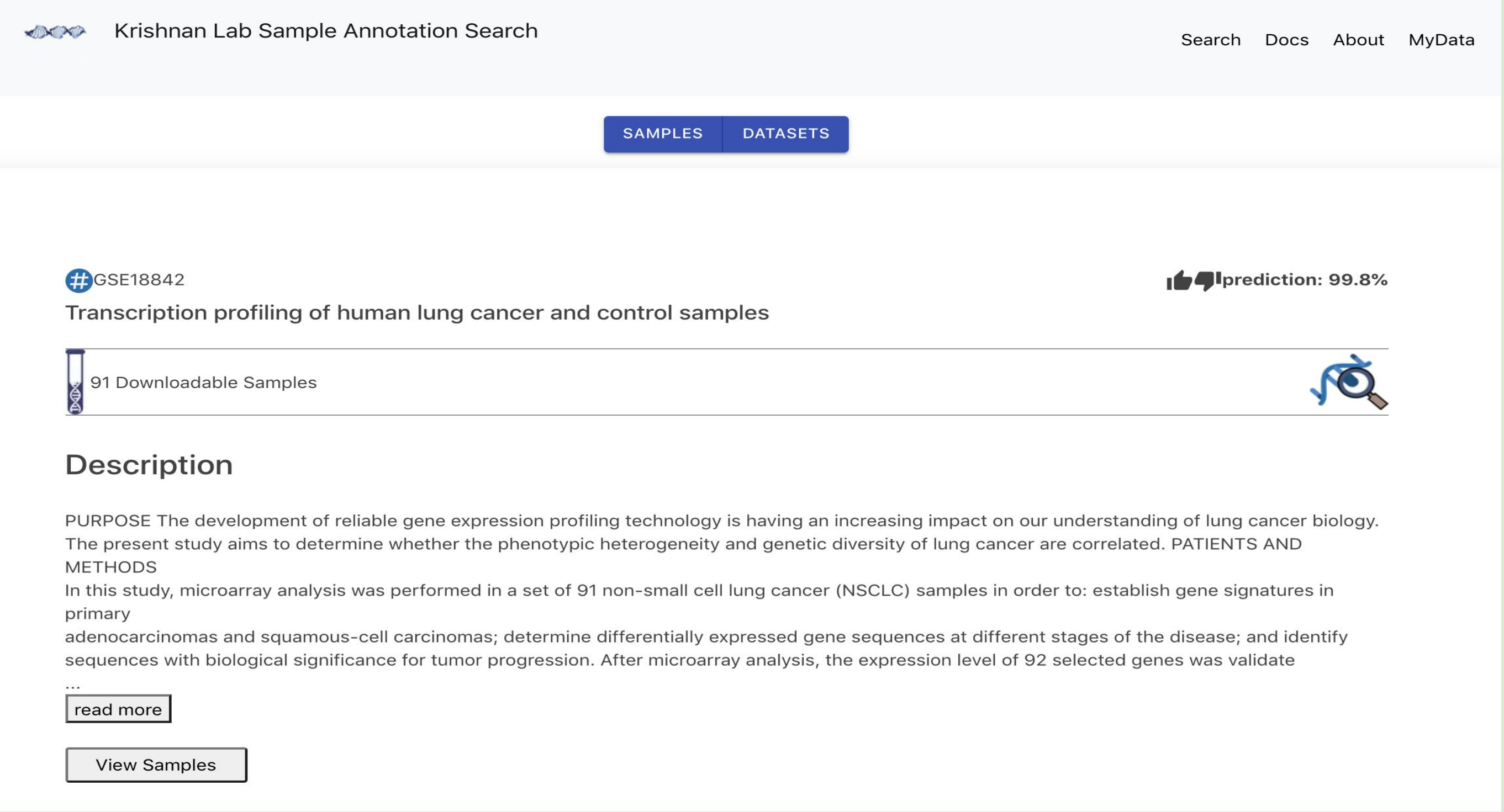


Found 11 Terms in 1 ms

Term ID: UBERON:0000170
Term Name: **pair of lungs**
Term Definition: The pair of anatomical structures comprised of a left **lung** and right **lung**.

Term ID: CL:0002553
Term Name: fibroblast of **lung**
Term Definition: A fibroblast that is part of **lung**.

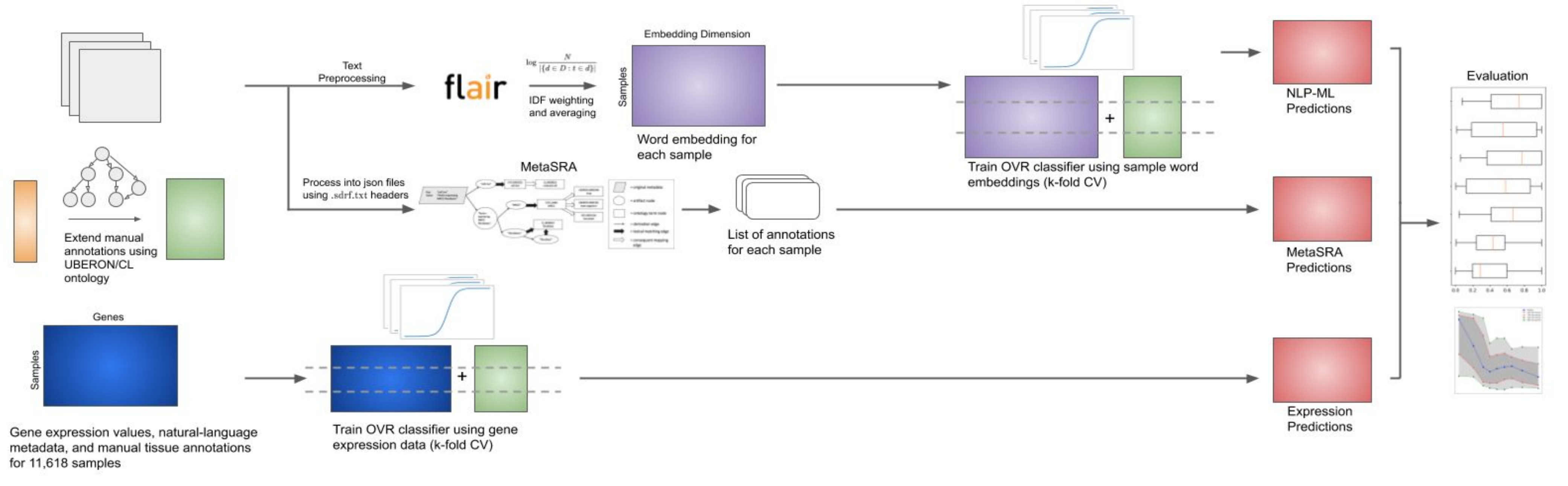
Query results. Series listed by predicted probability from NLP-ML models along with associated metadata. Individual samples in series can be selected



GSE18842 prediction: 99.8%
Transcription profiling of human lung cancer and control samples
91 Downloadable Samples

Description

PURPOSE The development of reliable gene expression profiling technology is having an increasing impact on our understanding of lung cancer biology. The present study aims to determine whether the phenotypic heterogeneity and genetic diversity of lung cancer are correlated. PATIENTS AND METHODS
In this study, microarray analysis was performed in a set of 91 non-small cell lung cancer (NSCLC) samples in order to: establish gene signatures in primary
adenocarcinomas and squamous-cell carcinomas; determine differentially expressed gene sequences at different stages of the disease; and identify sequences with biological significance for tumor progression. After microarray analysis, the expression level of 92 selected genes was validate
...


Text Preprocessing
flair
$\log \frac{N}{|\{d \in D : t \in d\}|}$
IDF weighting and averaging
Embedding Dimension
Samples
Word embedding for each sample
Train OVR classifier using sample word embeddings (k-fold CV)
NLP-ML Predictions
Extend manual annotations using UBERON/CL ontology
Process into json files using .sdrf.txt headers
MetaSRA
List of annotations for each sample
MetaSRA Predictions
Genes
Samples
Gene expression values, natural-language metadata, and manual tissue annotations for 11,618 samples
Train OVR classifier using gene expression data (k-fold CV)
Expression Predictions
Evaluation

PERFORMANCE COMPARISON

Combination of MetaSRA, NLP-ML, and gene expression models weighted by F1 score

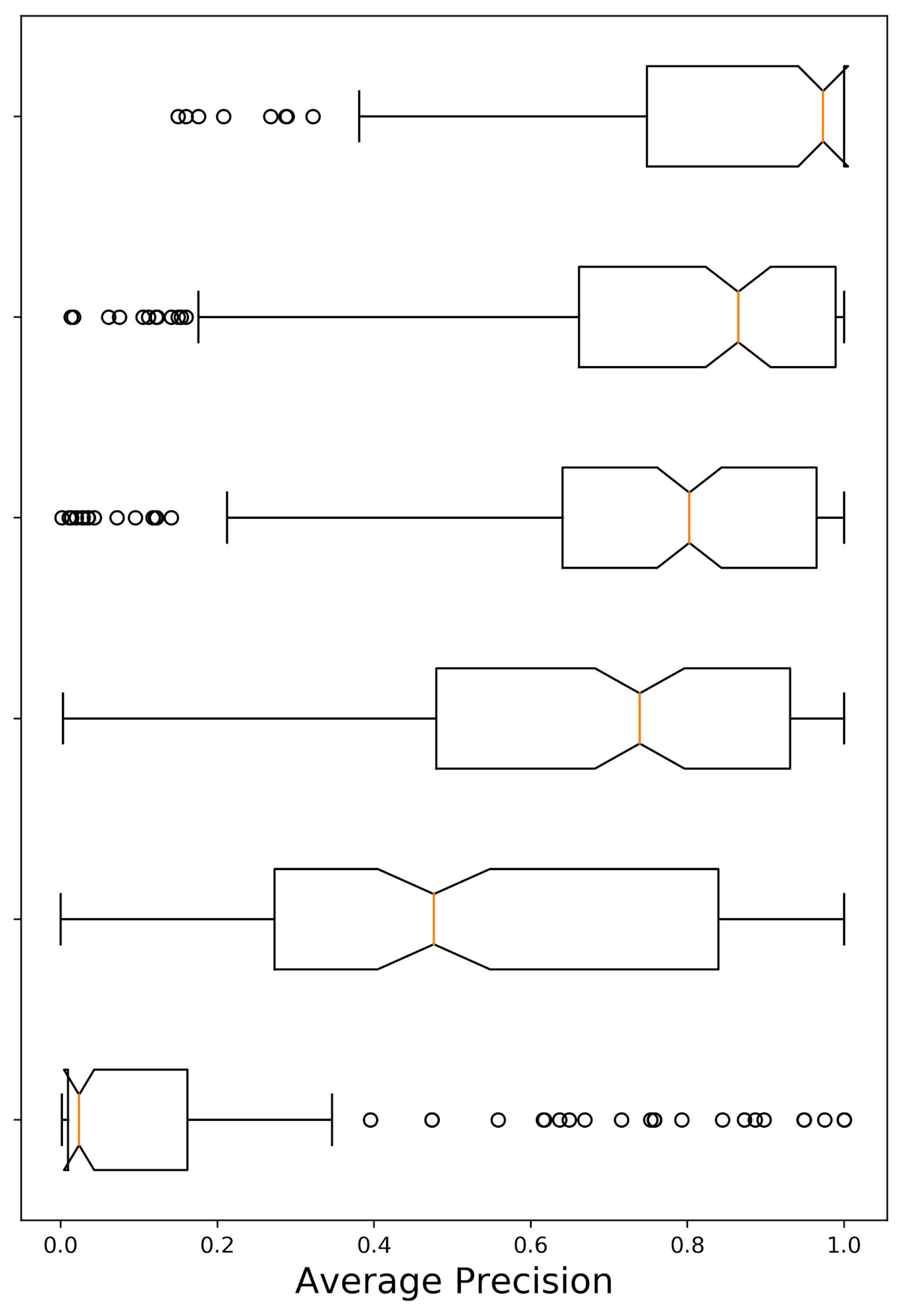
Combination of NLP-ML and gene expression models weighted by F1 score

Gene expression models: logistic regression models trained on normalized gene expression data

NLP-ML models: logistic regression models trained on word embedding representations of sample descriptions

MetaSRA: existing tool that constructs knowledge graph from text using ontologies and prior information

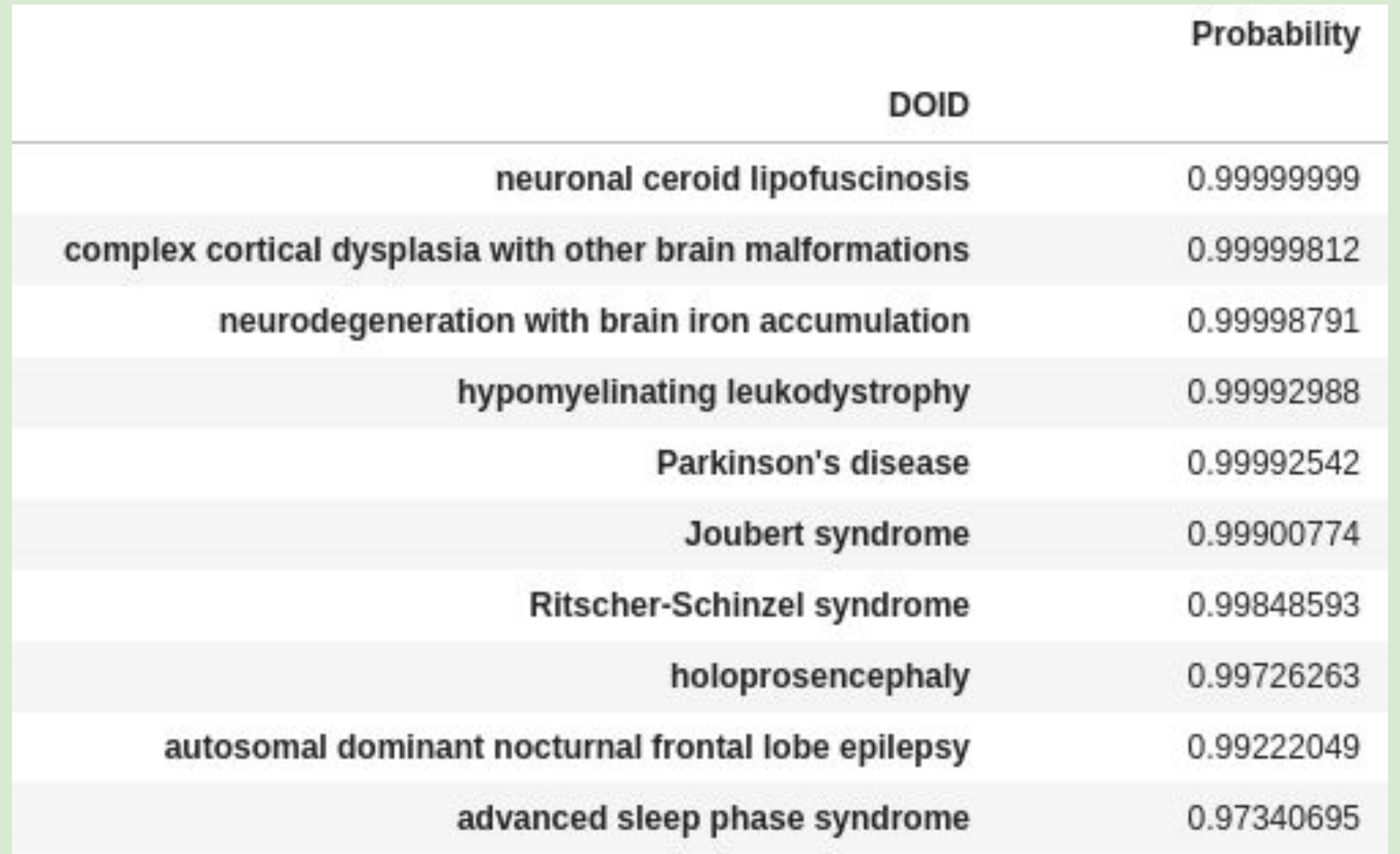
TAGGER: named entity recognition tool that directly matches text to a dictionary of ontology terms



Each point in the boxplot is a model to provide annotations for a single tissue/cell-type

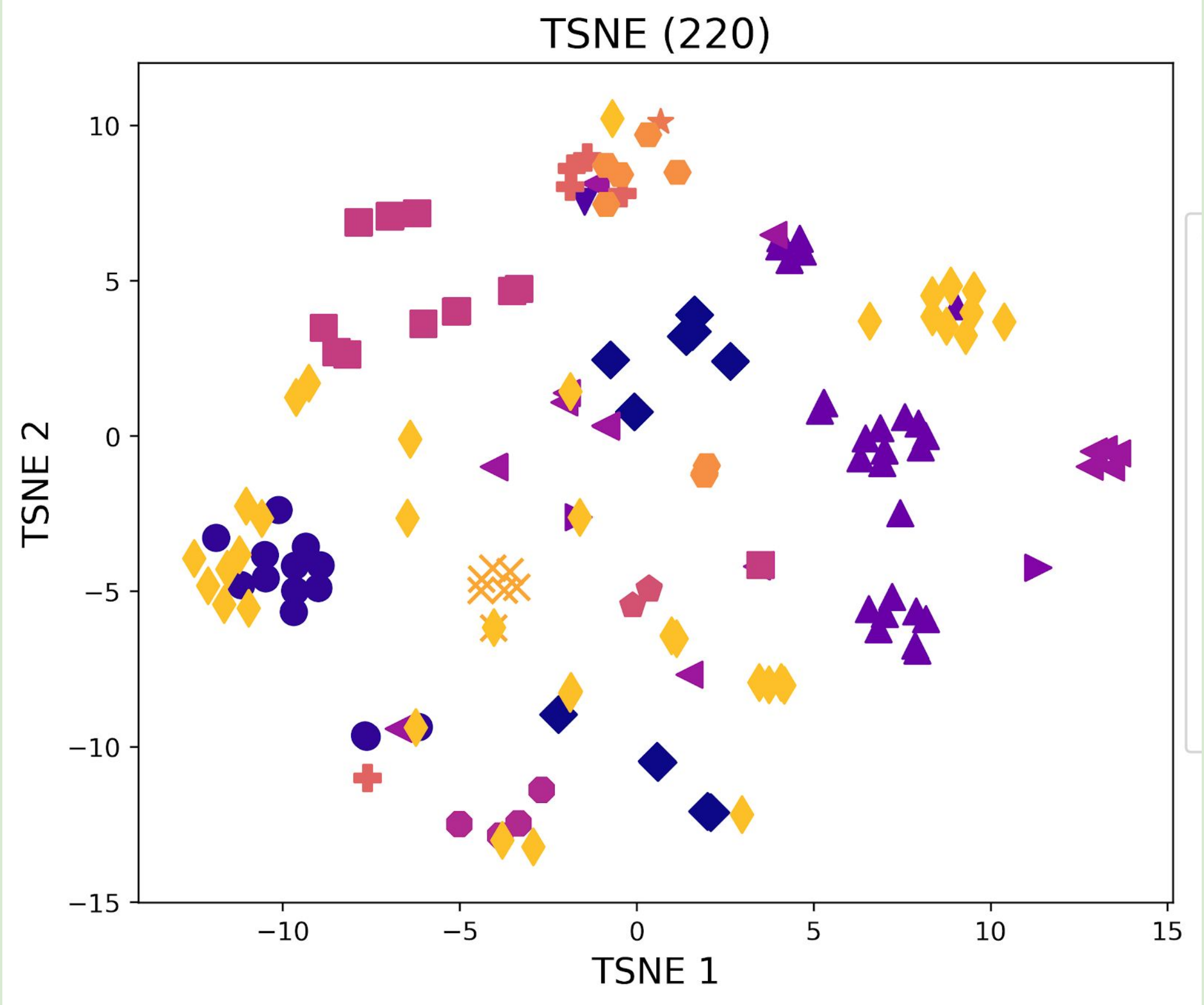
INTERPRETABILITY

Top predicted disease ontology terms for brain model using word embedding representations of ontology terms' description



	Probability
DOID	
neuronal ceroid lipofuscinosis	0.99999999
complex cortical dysplasia with other brain malformations	0.99999812
neurodegeneration with brain iron accumulation	0.99998791
hypomyelinating leukodystrophy	0.99992988
Parkinson's disease	0.99992542
Joubert syndrome	0.99900774
Ritscher-Schinzel syndrome	0.99848593
holoprosencephaly	0.99726263
autosomal dominant nocturnal frontal lobe epilepsy	0.99222049
advanced sleep phase syndrome	0.97340695

tSNE plot of normalized model coefficients colored by high-level anatomical system



1. Bernstein et al., 2017, *Bioinformatics*, 33, 18, 2914-2923.
2. Lee et al., 2013, *Bioinformatics*, 29, 23, 3036–3044.
3. Jensen, 2016, *bioRxiv*, doi.org/10.1101/067132.
4. Akbik et al., 2019, *NAACL Proceedings*, 724–728.

This work was primarily supported by US National Institutes of Health (NIH) grants R35 GM128765 to AK and in part by MSU start-up funds to AK and MSU Rasmussen Doctoral Recruitment Award and Engineering Distinguished Fellowship to NTH.

hawki2235@msu.edu