# Incorporating phenotypic similarity into trait description embeddings

**Soumyashree Kar[1], James P. Balhoff[1], Hilmar Lapp[2], Wasila Dahdul[3]**
**([1]RENCI, UNC at Chapel Hill, [2]Duke University, [3]UC Irvine)**

## Introduction

- Natural language descriptions of phenotypes are abundantly available.
- Developing computable traits or expressing phenotypes as logical statements amenable to machine reasoning, require considerable human effort.
- Phenoscape (https://phenoscape.org) curators annotate free-text phenotypic character state descriptions from morphological phylogenetic matrices, using the Entity–Quality semantic model. EQ associates an entity term from an anatomical ontology e.g., UBERON, with a quality term from the generic Phenotype and Trait Ontology (PATO).

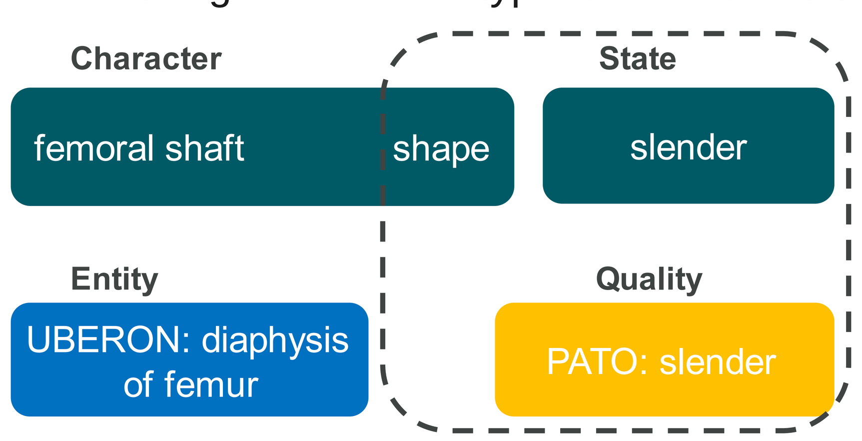


Figure 1. An example of EQ annotation

## Objective

Learning embeddings of trait descriptions that capture semantic similarity by incorporating background ontological knowledge.

Hypothesis: Ontology-based fine-tuning improves semantic textual similarity (STS) performance over just using free-text relationships.

- Develop a model to produce ontology-aligned text embeddings, without labor-intensive manual curation.
- Evaluate benchmarked models on trait description pairs, scored per their ontology-based semantic similarities.

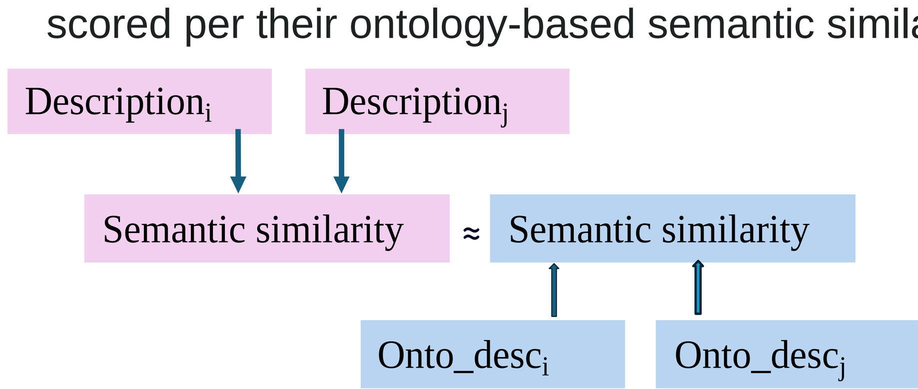


Figure 2. Schematic representation of the objective

## Data

The Phenoscape Knowledgebase (KB) contains ontology-annotated phenotypic data from 256 comparative studies.

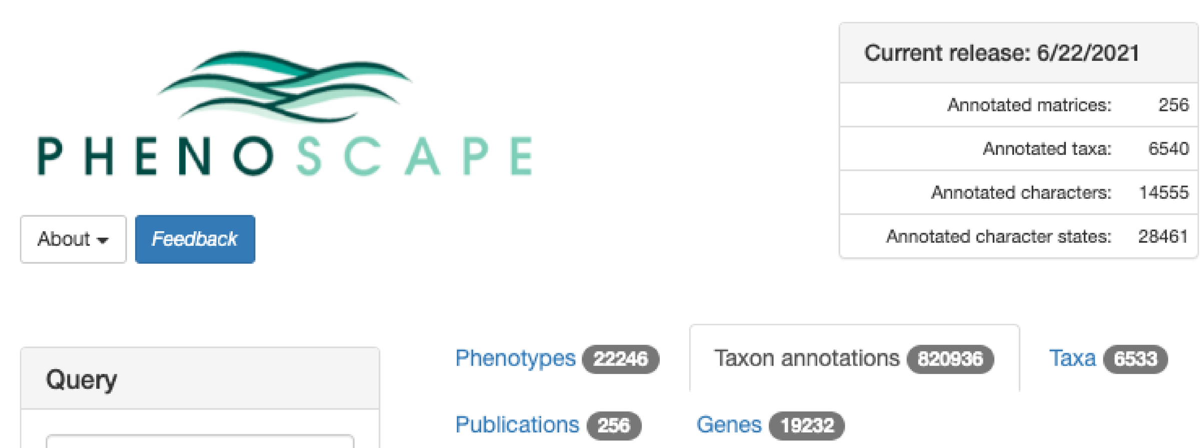


Figure 3. Snippet of the Phenoscape KB, the gold-standard data repository with annotations of phenotypes using ontologies.

Text input:

- 28461 individual character-state descriptions
- 405M unique pairs of character-state descriptions

Label: Ontology-based semantic similarities metrics:

- simGIC: proportion of subsumers in common, weighted by their information content.

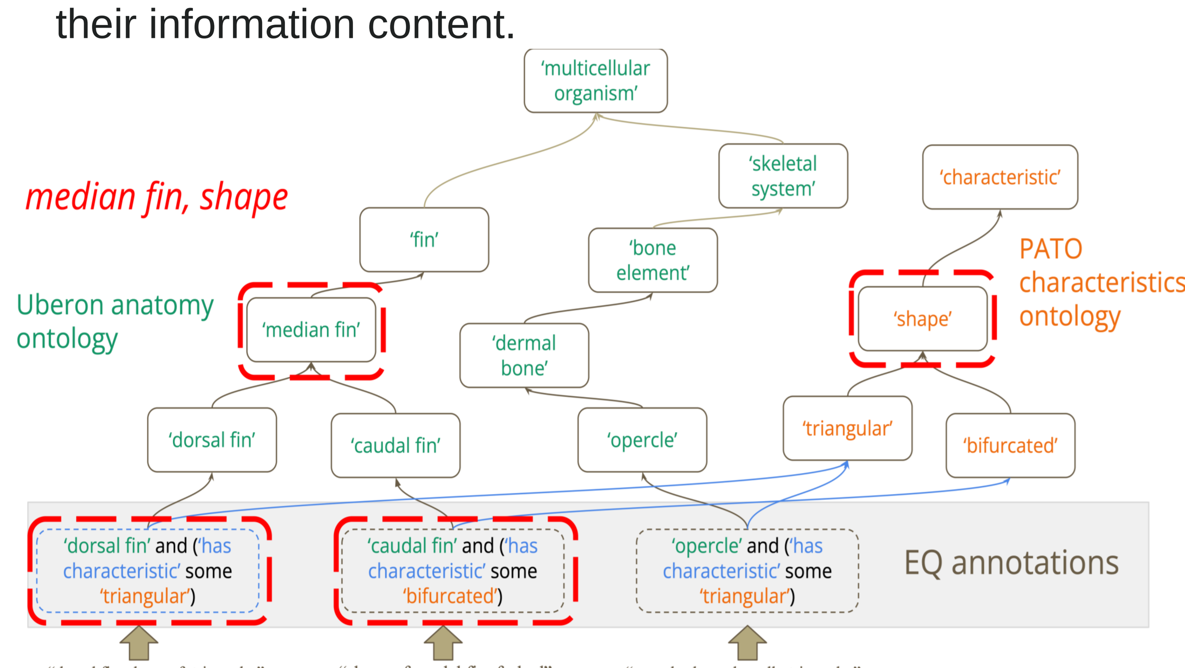


Figure 4. Illustration of EQ annotation using common ontology concepts.

- EQ annotations - logically connected to domain ontologies.
- Pairwise similarity of EQ annotations (and hence character state descriptions) are assessed using methods that consider the common ontology concepts connected via various relations (is_a, part_of, has_characteristic, etc.) as well as concept specificity.
- Concept specificity refers to the degree of detail or granularity of a concept within an ontology, quantified using information content.

## Methodology

- Inspect / filter (if any) duplicate trait-description pairs.
- Compute from KB ontology-based semantic similarities.
- Select the highest scoring metric (Fig. 5) as the label or target scores for the pairwise-similarities.
- Inspect for any noise in the dataset (non-English and coarse annotations) and generate filtered dataset.
- Perform semantic textual similarity analysis (Fig. 6).

1. Obtain raw-baseline performance of pre-trained and benchmarked sentence-transformer models.
2. Compare performance for different sequence lengths, embedding dimensions, and pooling methods.
3. Select the best (most accurate and efficient) model.
4. Finetune selected model on data: raw and filtered.
5. Compare performance metrics of both the models.

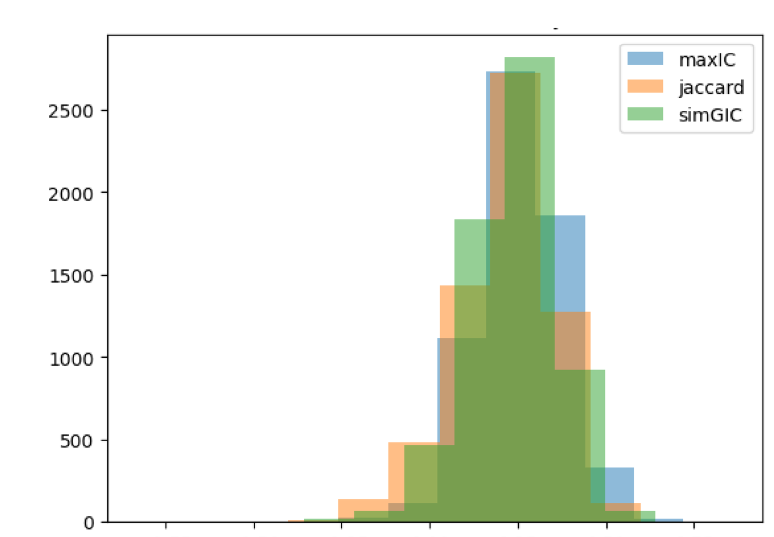


Figure 5. Distribution plot of maxIC, jaccard, and simGIC scores

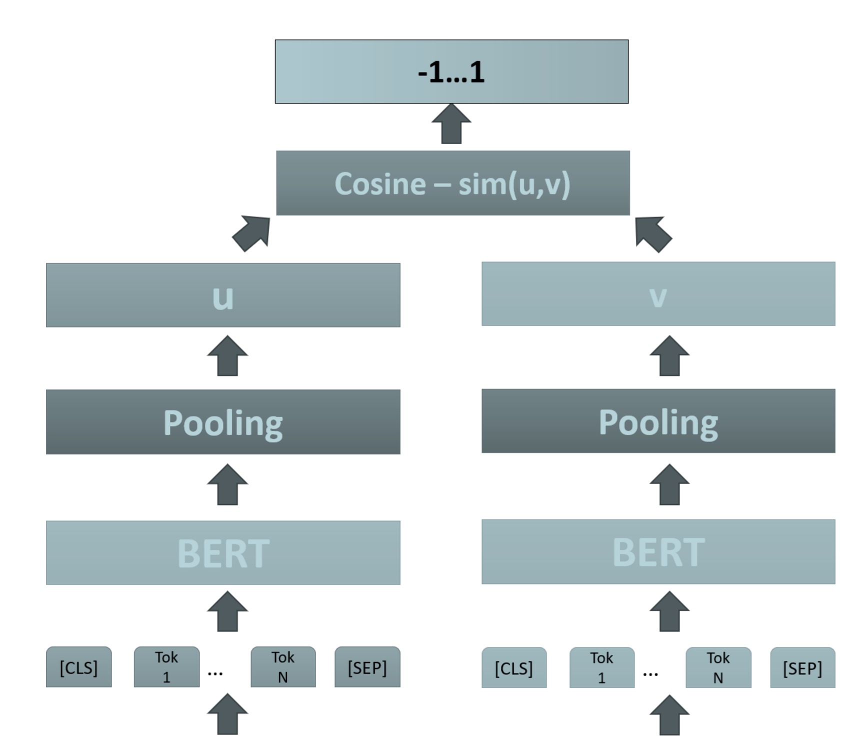


Figure 6. Siamese network of BERT-based models for semantic textual similarity (STS) analysis.

## Results

Table 1. Baseline performance of pretrained models for pooling methods (I - iii), embedding dim : sequence lengths.

i) CLS

| model | Pearson's r | | h:mm:ss | | Pearson's r | | h:mm:ss | |
|---|---|---|---|---|---|---|---|---|
| | 768 : 128 | 768 : 256 | 768 : 128 | 768 : 256 | 256 : 128 | 256 : 256 | 256 : 128 | 256 : 256 |
| all-distilroberta-v1 | 0.1618 | 0.1633 | 0:23:15 | 0:24:01 | 0.1484 | 0.1445 | 0:24:01 | 0:23:59 |
| all-MiniLM-L12-v2 | 0.1275 | 0.1277 | 0:32:02 | 0:35:14 | 0.1217 | 0.1203 | 0:33:28 | 0:33:47 |
| all-MiniLM-L6-v2 | 0.0941 | 0.0921 | 0:21:32 | 0:21:57 | 0.0900 | 0.0936 | 0:21:48 | 0:22:03 |
| all-mpnet-base-v2 | 0.2104 | 0.2109 | 0:25:24 | 0:27:04 | 0.1829 | 0.2158 | 0:26:33 | 0:26:48 |
| multi-qa-distilbert-cos-v1 | 0.1080 | 0.1063 | 0:21:37 | 0:22:40 | 0.1058 | 0.1052 | 0:22:20 | 0:22:42 |
| multi-qa-MiniLM-L6-cos-v1 | 0.0803 | 0.0794 | 0:21:29 | 1:14:57 | 0.0763 | 0.0807 | 0:21:59 | 0:22:08 |
| multi-qa-mpnet-base-dot-v1 | 0.1795 | 0.1753 | 0:34:57 | 0:36:39 | 0.1586 | 0.1683 | 0:36:18 | 0:36:31 |
| nli-distilbert-base | 0.0919 | 0.0910 | 0:21:32 | 0:21:50 | 0.0891 | 0.0913 | 0:21:58 | 0:22:04 |
| nli-distilbert-base-max-pooling | 0.0937 | 0.0921 | 0:21:15 | 0:21:49 | 0.0897 | 0.0909 | 0:21:53 | 0:22:06 |
| nli-distilroberta-base-v2 | 0.0895 | 0.0897 | 0:23:31 | 0:23:56 | 0.0871 | 0.0915 | 0:24:06 | 0:24:02 |
| nli-roberta-base-v2 | 0.0868 | 0.0876 | 0:34:30 | 0:35:42 | 0.0844 | 0.0901 | 0:35:33 | 0:35:39 |
| paraphrase-albert-small-v2 | 0.0761 | 0.0762 | 0:24:21 | 0:26:19 | 0.0739 | 0.0782 | 0:25:01 | 0:26:15 |
| paraphrase-MiniLM-L3-v2 | 0.0835 | 0.0842 | 0:59:33 | 0:16:05 | 0.0814 | 0.0866 | 0:16:03 | 0:29:03 |
| paraphrase-multilingual-MiniLM-L12-v2 | 0.0793 | 0.0794 | 0:31:51 | 0:33:48 | 0.0772 | 0.0820 | 0:33:09 | 0:46:24 |

ii) WEIGHTED MEAN

| model | Pearson's r | | h:mm:ss | | Pearson's r | | h:mm:ss | |
|---|---|---|---|---|---|---|---|---|
| | 768 : 128 | 768 : 256 | 768 : 128 | 768 : 256 | 256 : 128 | 256 : 256 | 256 : 128 | 256 : 256 |
| all-distilroberta-v1 | 0.1350 | 0.1254 | 0:25:42 | 0:26:25 | 0.1057 | 0.1131 | 0:26:10 | 0:33:04 |
| all-MiniLM-L12-v2 | 0.1330 | 0.1241 | 0:34:35 | 0:36:29 | 0.1073 | 0.1129 | 0:35:31 | 0:40:15 |
| all-MiniLM-L6-v2 | 0.1330 | 0.1250 | 0:54:05 | 0:23:54 | 0.1100 | 0.1193 | 0:23:41 | 0:25:51 |
| all-mpnet-base-v2 | 0.1826 | 0.1851 | 0:36:51 | 0:38:19 | 0.1539 | 0.1622 | 0:38:11 | 0:51:02 |
| multi-qa-distilbert-cos-v1 | 0.1327 | 0.1247 | 0:24:22 | 0:25:04 | 0.1088 | 0.1172 | 0:24:51 | 0:32:22 |
| multi-qa-MiniLM-L6-cos-v1 | 0.1351 | 0.1266 | 3:03:09 | 0:23:57 | 0.1121 | 0.1216 | 0:23:43 | 0:25:56 |
| multi-qa-mpnet-base-dot-v1 | 0.1617 | 0.1555 | 0:36:48 | 0:37:54 | 0.1279 | 0.1404 | 0:37:40 | 0:50:10 |
| nli-distilbert-base | 0.1323 | 0.1277 | 0:24:08 | 0:24:35 | 0.1172 | 0.1230 | 0:23:47 | 0:24:40 |
| nli-distilbert-base-max-pooling | 0.1256 | 0.1222 | 0:56:13 | 0:24:24 | 0.1116 | 0.1187 | 0:25:20 | 0:24:51 |
| nli-distilroberta-base-v2 | 0.1371 | 0.1311 | 0:25:39 | 0:26:07 | 0.1206 | 0.1262 | 0:25:58 | 0:26:17 |
| nli-roberta-base-v2 | 0.1377 | 0.1313 | 0:35:59 | 0:37:01 | 0.1198 | 0.1256 | 0:37:13 | 0:49:02 |
| paraphrase-albert-small-v2 | 0.1360 | 0.1286 | 0:27:27 | 0:28:32 | 0.1137 | 0.1235 | 0:27:12 | 0:36:25 |
| paraphrase-MiniLM-L3-v2 | 0.1370 | 0.1310 | 0:17:19 | 0:17:29 | 0.1183 | 0.1250 | 0:17:29 | 0:18:49 |
| paraphrase-multilingual-MiniLM-L12-v2 | 0.1362 | 0.1301 | 0:33:55 | 0:36:08 | 0.1161 | 0.1249 | 0:35:19 | 0:40:36 |

iii) MEAN

| model | Pearson's r | | h:mm:ss | | Pearson's r | | h:mm:ss | |
|---|---|---|---|---|---|---|---|---|
| | 768 : 128 | 768 : 256 | 768 : 128 | 768 : 256 | 256 : 128 | 256 : 256 | 256 : 128 | 256 : 256 |
| all-distilroberta-v1 | 0.1289 | 0.1315 | 0:24:14 | 0:24:34 | 0.1059 | 0.1108 | 0:24:15 | 0:24:25 |
| all-MiniLM-L12-v2 | 0.1320 | 0.1286 | 0:34:49 | 0:35:19 | 0.1090 | 0.1094 | 0:34:07 | 0:34:39 |
| all-MiniLM-L6-v2 | 0.1342 | 0.1314 | 0:22:15 | 0:22:45 | 0.1129 | 0.1120 | 0:22:11 | 0:22:13 |
| all-mpnet-base-v2 | 0.1757 | 0.1803 | 0:39:43 | 0:39:53 | 0.1480 | 0.1587 | 0:36:39 | 0:37:09 |
| multi-qa-distilbert-cos-v1 | 0.1039 | 0.1297 | 0:22:50 | 0:24:50 | 0.1132 | 0.1115 | 0:21:58 | 0:22:48 |
| multi-qa-MiniLM-L6-cos-v1 | 0.1312 | 0.1336 | 0:22:28 | 0:22:48 | 0.1017 | 0.1124 | 1:41:22 | 1:42:32 |
| multi-qa-mpnet-base-dot-v1 | 0.1527 | 0.1596 | 0:37:05 | 0:37:55 | 0.1208 | 0.1392 | 0:36:44 | 0:36:47 |
| nli-distilbert-base | 0.1106 | 0.1186 | 0:22:31 | 0:23:31 | 0.1016 | 0.1098 | 0:22:26 | 0:23:26 |
| nli-distilbert-base-max-pooling | 0.1022 | 0.1042 | 0:22:42 | 0:24:42 | 0.1035 | 0.1051 | 0:21:28 | 0:22:48 |
| nli-distilroberta-base-v2 | 0.1133 | 0.1243 | 0:24:34 | 0:25:14 | 0.1012 | 0.1137 | 0:23:01 | 0:24:01 |
| nli-roberta-base-v2 | 0.1212 | 0.1203 | 0:36:16 | 0:36:27 | 0.1114 | 0.1105 | 0:34:30 | 0:35:30 |
| paraphrase-albert-small-v2 | 0.1126 | 0.1265 | 0:27:03 | 0:27:50 | 0.1072 | 0.1096 | 0:25:52 | 0:26:52 |
| paraphrase-MiniLM-L3-v2 | 0.1170 | 0.1287 | 0:16:18 | 0:17:08 | 0.1131 | 0.1134 | 0:14:20 | 0:15:15 |
| paraphrase-multilingual-MiniLM-L12-v2 | 0.1255 | 0.1276 | 0:34:46 | 0:35:16 | 0.1100 | 0.1116 | 0:34:26 | 0:34:36 |

Table 2. Finetuned performance of all-mpnet-base-v2 on raw and filtered datasets.

| | Training loss | Validation loss | Spearman_max (val) | Pearson_max (val) |
|---|---|---|---|---|
| **RAW** | 0.0017 | 0.0015 | **0.9076** | 0.9544 |
| **FILTERED** | 0.0027 | 0.0136 | **0.9388** | 0.9432 |

- Baseline evaluation – all-mpnet-base-v2 (109M params) performed best (~0.22 correlation without finetuning).
- Model finetuned on filtered dataset showed better and more consistent performance, with overall correlation of 0.94.
- Ontology-based finetuning improves semantic similarity between trait descriptions.
- Finetuned embeddings to be evaluated for multimodal learning.

**Funding and Acknowledgement**

This work was supported by the NSF OAC 2118240 Imageomics Institute award.