Trait2Vec: Ontology-aware embeddings for organismal trait descriptions

Juan J. Garcia
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27514
jjgarcia@cs.unc.edu

Jim Balhoff
Renaissance Computing Institute
Chapel Hill, NC 27517
balhoff@renci.org

Hilmar Lapp
Department of Biostatistics & Bioinformatics
Duke University
Durham, NC 27708
hilmar.lapp@duke.edu

Abstract

Trait descriptions characterize how an organism looks, behaves or interacts. These descriptions are typically represented as text, but may be manually mapped within an ontology for downstream analysis. Nonetheless, the cost of this manual mapping is not scalable. In this work we propose a method to finetune a transformer model and embed textual trait descriptions in a latent space that captures the notion of distance within an ontology. The resulting model, which we coin Trait2Vec, can then embed trait descriptions in a scalable and biologically meaningful computational representation.

1 Introduction

Understanding and representing the semantic meaning of trait descriptions is a foundational challenge in building computational systems that reason over biological data. Trait descriptions, often expressed in natural language, characterize the phenotypic features of organisms. However, their syntactic and semantic variability poses a challenge for a consistent computational representation, and thus they are manually transformed into an ontological representation for further inference. Unfortunately, this manual transformation is not scalable, as it requires the expert curation of the trait description in terms of a meaningful ontology.

Recent advances in transformer-based language models [1, 2, 3] have been proposed to embed the ontology into a latent space [4]. Unfortunately, this is also not scalable, as it assumes access to an expensive and manual ontology representation of the organism. Instead, in this work, we propose to estimate a transformer-based language model to embed raw trait descriptions of the organism and thus bypass the need to annotate trait descriptions with a manual ontological representation. To do so, we propose to embed trait descriptions in a way that correlates with prior manually created ontological annotations of a set of trait descriptions.

To align these embeddings with the underlying ontological structure, we consider similarity measures (e.g. SimGC [5]) between trait descriptors that have been annotated with ontology-based expressions. These graph-based similarity measures inform how well the learned embeddings preserve rank. That is, how well similarity between representations in latent space correlates with similarity measures

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

of their corresponding ontology representations. We conjecture this alignment reflect the semantics provided by the ontology. This in turn induces the corresponding biological meaning in the similarity between embeddings. We test this conjecture using the Phenoscape knowledge base [6] to extract a collection of trait descriptions along with the corresponding pairwise similarity metrics.

By integrating language model embeddings with ontology-aware similarity metrics, we bypass the need to manually represent trait descriptors within an ontology before it is available for further computational inference. Accordingly, we obtain a biologically meaningful and scalable computational representation of a trait description that can be readily integrated in downstream tasks (e.g. extending imageomics pipelines [7]). More concretely, our contributions are:

1. A novel machine learning method to estimate a latent representation of trait descriptions from any collection of ontology annotations.
2. A dataset (Character similarity dataset [8]) of pairwise trait descriptors along with their corresponding SimGIC, maxIC and Jaccard similarity metrics.
3. A pre-trained model (Trait2Vec [9]) to embed textual trait descriptions.
4. Empirical evidence Trait2Vec embeddings capture ontology structure.

2 Related works

Embedding ontologies is a well explored idea. The work of [4] reviews current state-of-the-art strategies to capture the ontology structure in the embedding space. Closest to Trait2Vec are the sequence modeling approaches: Onto2Vec [1], OPA2Vec [2], OWL2Vec [3].

Onto2Vec uses a pretrained Word2Vec model to embed ontology structure axioms and entity term annotation axioms classes into a latent space. They then evaluate this embeddings on protein-to-protein interaction (PPI) by training a classifier on the embeddings. As an extension of Onto2Vec, the authors propose OPA2Vec and extend the ontology information with its corresponding meta data. They observe the embeddings from this sources improve the performance of PPI classification. Per OWL2Vec, Onto2Vec embed the axioms of the ontology and OPA2Vec complements this with lexical information. OWL2Vec complements their axiom corpus with a corpus generated by walking over RDF graphs that are transformed from the OWL ontology with its graph structure and logical constructors considered. In addition, to fully utilize the lexical information, OWL2Vec creates embeddings for not only the ontology entities as the current KG/ontology embedding methods but also for the words in the lexical information. The authors test the methods for membership prediction and subclass prediction. Nonetheless, unlike previous work, our method does not assume the future test input will have a corresponding representation in ontology space. Instead, our method aims to recover a serviceable embedding from a raw trait descriptor. Access to an ontology representation is expensive, as it requires expert knowledge to produce, thus limiting its scalability.

To the best of our knowledge, this is the first proposal of a trait embedding approach guided by aligning rank with ontology based metrics. Order preservation is important in numerous approaches that use semantic similarity for biological discovery. Accordingly, this work brings the biology community closer to a serviceable computational representation of trait descriptors that do not require manual annotation. This can potentially complement recently methodological developments that extract embeddings from images [10, 11] by directly encoding phenotypes that are more easily described via text (e.g. "spiny-rayed dorsal fin").

3 Methodology

3.1 Ontologies

Ontological representations are structured, formalized systems for categorizing and describing entities, their properties, and the relationships among them. In computational terms, an ontology provides a shared vocabulary and a logical framework for consistent annotation, integration, and reasoning across datasets. These representations typically consist of hierarchical structures (often directed acyclic graphs or trees), where entities are defined as classes or concepts and connected through relations such as "is-a" (subclass) or "part-of." In biology, the complexity and interdependence of

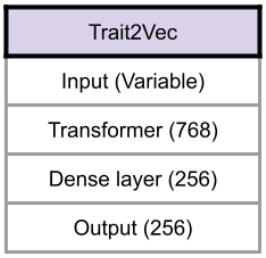

Figure 1: Layers in Trait2Vec model. Inside parenthesis is the output size of the given layer.

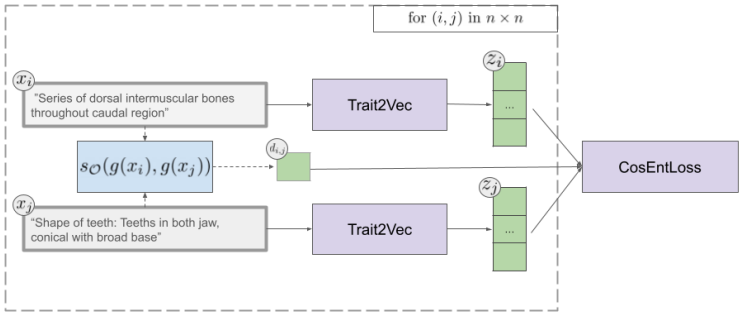

Figure 2: Loss computation for training Trait2Vec. Only Train2Vec parameters are estimated. Please refer to section 3.3 for more details.

organismal systems make ontologies valuable approaches for storing complex domain usage in a way that can be reused in a variety of applications. The hierarchical structure of ontologies enables computational methods (e.g., logical reasoners) bring together data at varying levels of granularity, as well as to make use of implied or indirect relationships between concepts, particularly through transitive relations such as part-of or develops-from. A useful application of ontology-annotated data is the ontology graph-based similarity metrics.

3.2 Ontological graph-based similarity metrics

Ontology-based similarity metrics (e.g. SimGIC, maxIC, Jaccard)[5, 12, 13] provide biologically meaningful metrics to compare trait descriptors and derive biological insight. SimGIC, in particular, computes similarity between two sets of ontology terms by considering the shared information content of their common ancestors relative to the total information content of all terms in both sets. This accounts not just for overlap but for how specific or informative the shared terms are within the ontology. Unfortunately, creating ontology annotations for trait descriptions has thus far been a primarily manual, expensive, task requiring expert knowledge. This cost in time and expertise motivates the question: Can we embed raw text descriptions of traits in a way that preserves the ontology structure induced by graph-based similarity measures?

3.3 Embedding trait descriptors

Assume access to a sample of n distinct textual trait descriptors and their corresponding ontological representation (i.e. $D_n = \{(x_i, g(x_i))\}_{i=1}^n$) where $g : \mathcal{X} \to \mathcal{O}$ is a manual mapping of trait descriptors to an ontology representation. We propose to optimize the following CosENT loss:

$$\mathcal{L}(\theta) = \log\Big(\sum_{((i,j),(h,k))} \{1 + \exp\{s(z_i, z_j) - s(z_h, z_k)\}\}\Big) \tag{1}$$

Where $s : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$ correspond to a similarity measure of the embeddings $z_i = f_\theta(x_i) \in \mathcal{Z}$ and $z_j = f_\theta(x_j) \in \mathcal{Z}$ for distinct traits x_i, x_j (e.g. $s(z_i, z_j) = \frac{1}{2}(\langle z_i, z_j \rangle / \|z_i\| \|z_j\| + 1)$ is proportional to cosine similarity). The function $f_\theta : \mathcal{X} \to \mathcal{Z}$ corresponds to the Trait2Vec model parametrized by θ (i.e. The parameters of the Dense and Transformer layers in figure 1). The indices (i, j) and (h, k), index pairs of descriptors such that the following order relation is satisfied: $d_{i,j} = s_{\mathcal{O}}(g(x_i), g(x_j)) \leq s_{\mathcal{O}}(g(x_h), g(x_k)) = d_{h,k}$. The function $s_{\mathcal{O}} : \mathcal{O} \times \mathcal{O} \to [0, 1]$ corresponds to the ontology-based distance metric (e.g., SimGIC). Intuitively, this loss penalizes embeddings whose cosine similarity does not respect the ranking (or order) induced by the similarity of the corresponding ontology representations. We visualize the loss computation pipeline in figure 2.

4 Experiments

Across all experiments we extract a collection of textual trait descriptors and corresponding ontology representation from the Phenoscape knowledgebase [6]. We collect over 500K trait descriptor pairs for training and leave out an extra 100K pairs for testing. No single trait descriptor is shared between train

and test sets. We initialize the transformer with the "all_mpnet_model_v2" and the Dense architure with the default initialization and hyperbolic tangent non-linearity from the sentence-transformer library [14]. We learn the Trait2Vec parameters (θ in (1)) using the Trainer class with hyperparameters (learning-rate, epochs, batch-size, warmup-ratio) estimated from a validation sample. We also perform early-stopping with a patience of 5. We measure embedding similarity with cosine similarity and measure ontology-based similarity with the SimGIC metric. All experiments are performed on an A100 gpu. Code for reproducing our experiments is available at https://github.com/Imageomics/char-sim.

4.1 Rank correlation

The goal of this experiment is to assess how well similarity between embeddings correlates with Sim-GIC on the test dataset. A positive correlation indicates the transformation preserves the ontological order of the traits. We measure the Spearman's rank correlation where the true ranking is given by the SimGIC metric. This statistic is useful because it is invariant to the scales of similarity metrics. Furthermore, we also measure training time in hours on a A100GPU, for a variety of architectural choices and sample sizes. Results in figure 3 suggest performance increases with more data albeit with diminishing returns. The best model is estimated with 400K training sample pairs. As a way to mitigate the sample size, the bronze and silver star models, perform a resampling of the validation data during training ten times (i.e. 10K total samples for validation). We observe it is possible to obtain similar performance with 60K samples as opposed to 500K. Lastly, we note fixing the transformer weights (i.e. "Frozen") diminishes performance substantially. We speculate this is because the difference between trait descriptors depends on the ontology structure rather than linguistic patterns.

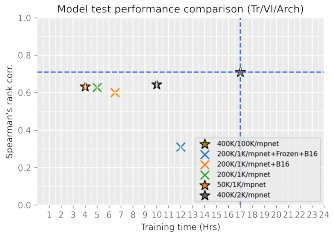

Figure 3: Performance and training time of the full Trait2Vec architecture (mpbnet), Trait2Vec with frozen transformer weights (mpbnet+Frozen), and Trait2Vec with lower precision weights (B16), for a variety of train/validation sample sizes.

4.2 Low-dimensional clustering performance

The goal of this experiment is to qualitatively assess how well similar trait descriptions cluster together in embedding space. We leverage the pre-trained Trait2Vec model (i.e. $f_{\hat{\theta}} : \mathcal{X} \to \mathcal{Z}$) estimated in section 4.1. For this experiment, we sample trait descriptions from 25 samples from each of two different groups (i.e. Jaw and Fin) which we denote as D_{jaw}, D_{fin}. We embed all samples and collect them into a matrix $Z \in \mathbb{R}^{n \times d}$ where $n = |D_{\text{jaw}} \cup D_{\text{fin}}|$, d corresponds to the embedding size ($d = 256$) and $Z_{i:} = f_{\hat{\theta}}(x_i) \in \mathbb{R}^d$ where $x_i \in D_{\text{jaw}} \cup D_{\text{fin}}$. We apply PCA to the matrix Z and plot the embeddings projected unto the two principal components in figure 4. Descriptors are color coded based on the category they belong to. Transparency indicates embeddings before (transparent) and after training (opaque). We observe that embeddings cluster into the expected groups after training. Furthermore, embedding distance is not entirely explained by syntactic similarity between traits. Finally, we measure the euclidean distance between trait-embeddings, before (Table 1) and after training (Table 2), and compute the 0%, 25%, 50%, 75% and 100% quantiles. In short, 0% would correspond to the closest trait descriptors and 100% to the farthest. We observe the post training ranking captures more semantics of the

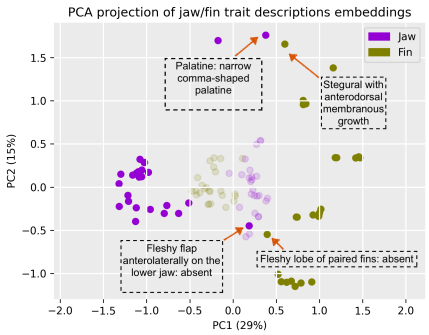

Figure 4: 2D-PCA projection of held-out Jaw and Fin descriptor embeddings. Transparent projections correspond to pre-training embeddings; opaque correspond to post-training embeddings. Some corresponding trait descriptors are indicated.

ontology structure. For instance, quantile 0.50 of Table 1 shows two Jaw trait descriptors that should be closer than the descriptors in the 0.25 quantile. This is not the case for post-training embeddings.

Quantile	Trait 1	Trait 2
0.00	(Jaw)Form of retroarticular: retroarticular elongate and cup-shaped, with its length in lateral view more than three times its height, with interopercularmandibular ligament attaching in cup-shaped depression near anterior margin of bone	(Jaw)Form of retroarticular: retroarticular elongate and cup-shaped, with its length in lateral view more than three times its height, with interopercular-mandibular ligament attaching in cup-shaped depression near anterior margin of bone
0.25	(Fin)Paired fins, pelvic girdle and scapulocoracoids absent in adults: Paired fins, pelvic girdle and scapulocoracoids absent in adults	(Jaw)Jaw teeth-shape and size gradation: strongly curved, moderate anteroposterior gradation in size
0.50	(Jaw)Fleshy flap anterolaterally on the lower jaw: absent	(Jaw)Metapterygoid: large, broad and in contact with quadrate and symplectic through cartilage
0.75	(Fin)Fin base articulation on scapulocoracoid: stenobasal	(Jaw)Vomer: arrow or T-shaped
1.00	(Fin)Stegural with anterodorsal membranous growth. Absence in Argentinoidei is secondary by parsimony optimization: Stegural with anterodorsal membranous growth	(Jaw)Vomer: arrow or T-shaped

Table 1: Quantiles of pairwise distances between descriptors before training.

Quantile	Trait 1	Trait 2
0.00	(Jaw)Jaw teeth-shape and size gradation: scarcely curved, moderate anteroposterior gradation in size	(Jaw)Jaw teeth-shape and size gradation: strongly curved, moderate anteroposterior gradation in size
0.25	(Fin)neural arch PU2: present	(Jaw)Form of retroarticular: retroarticular elongate and rod-shaped, with interopercular-mandibular ligament attaching at posterior margin of bone
0.50	(Fin)(H5.) Pectoral attachment rotated; primitive metapterygial axisfransverse or oblique to body axis: pectoral attachment rotated	(Jaw)Fleshy flap anterolaterally on the lower jaw: present
0.75	(Fin)Paired fins, pelvic girdle and scapulocoracoids absent in adults: Paired fins, pelvic girdle and scapulocoracoids absent in adults	(Jaw)Palatine: narrow comma-shaped palatine, lacking a posterior process
1.00	(Fin)Pectoral fin: rounded, horizontally placed, posteriormost tip reaching point midway between pectoral-fin origin and pelvic-fin origin when adpressed	(Jaw)Palatine: narrow comma-shaped palatine, lacking a posterior process

Table 2: Quantiles of pairwise distances between descriptors after training.

4.3 Taxon generalization

In this experiment we quantify the ranking capabilities to descriptors from different taxa. Generalizing to other taxa is important to reduce the need to collect taxon-specific ontological descriptors. We train the model on trait descriptions pairs within the fish order Characiformes and test on held-out trait description pairs of distinct fish orders Siluriformes, Cypriniformes, and Gymnotiformes. Figure Boxplot of the Spearman's rank correlation of 200 descriptors from Siluriformes (n=1327009), Cypriniformes (n=556447), Gymnotiformes (n=679). We sample the 200 descriptors pairs and measure the correlation coefficient. We repeat this experiment w/o replacement 30 times per taxon to produce the boxplot. The embedding is done by a model trained on the Characiformes taxon. Descriptors from other test taxa do not contain characters from Characiformes. The dashed red line indicates the Spearman's rank correlation of a Characiformes test samples. With the exception of "siluriformes" (i.e. catfish), the ordering in performance matches the evolutionary relationships between the taxa. We speculate this discrepancy stems from the specialized morphology exhibited by catfish traits.

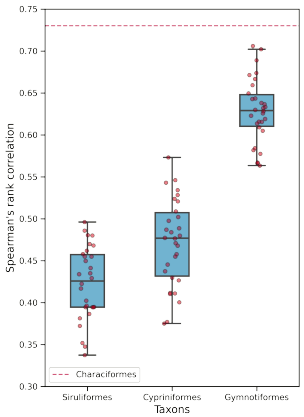

Figure 5: Boxplot of the Spearman's rank correlation accross different unseen taxa. The dashed red line indicates the average test performance on the observed taxon.

5 Conclusion

In this work we propose a novel machine learning method to learn a latent representation of trait descriptors where similarity metrics in latent space correlate with ontology-based similarity metrics. We evaluate this methodology on both quantitative and qualitative tasks, suggesting a these representations are capturing structure from the ontology. This is important to computational biologists that seek to derive biological insight from ontology based metrics (e.g. SimGIC) but do not have access to the ontology representation of a given collection of traits. Instead, we envision they can embed the collection of traits and aim to derive the same insight using the latent representations from Trait2Vec. It is important to highlight that the current model has the biases, coverage gaps, and evolving definitions of a single similarity metric and ontology. Biological conclusions may differ under alternative metrics (e.g., Jaccard) or other phenotype ontologies. Future work will consider alternative metrics, as well as embedding spaces with different geometric structure (e.g. Hyperbolic geometry). It would also be helpful to explore the complementary value Trait2Vec embeddings from text can provide embeddings from other sources (e.g. BioCLip2 [11] embeddings from Images).

Acknowledgments and Disclosure of Funding

The authors thank Soumyashree Kar for her contributions to this project. This work was supported by the NSF OAC 2118240 award: "HDR Institute: Imageomics: A New Frontier of Biological Information Powered by Knowledge-Guided Machine Learning"; by Phenoscape, supported by NSF grants DBI-1661456, DBI-1661529, DBI-1661516, and DBI-1661356 originally incubated and supported by the National Evolutionary Synthesis Center (NESCent), NSF EF-0905606.

References

[1] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, 34(13):i52–i60, July 2018.

[2] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics*, 35(12):2133–2140, June 2019.

[3] Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. OWL2Vec*: embedding of OWL ontologies. *Machine Learning*, 110(7):1813–1845, 2021. Publisher: Springer.

[4] Jiaoyan Chen, Olga Mashkova, Fernando Zhapa-Camacho, Robert Hoehndorf, Yuan He, and Ian Horrocks. Ontology Embedding: A Survey of Methods, Applications and Resources. *IEEE Transactions on Knowledge and Data Engineering*, 37(7):4193–4212, July 2025. arXiv:2406.10964 [cs].

[5] Catia Pesquita, Daniel Faria, Hugo Bastos, António Ferreira, André Falcão, and Francisco Couto. Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC bioinformatics*, 9 Suppl 5:S4, February 2008.

[6] Wasila Dahdul, Prashanti Manda, Hong Cui, James P Balhoff, T Alexander Dececchi, Nizar Ibrahim, Hilmar Lapp, Todd Vision, and Paula M Mabee. Annotation of phenotypes using ontologies: a gold standard for the training and evaluation of natural language processing systems. *Database*, 2018:bay110, January 2018.

[7] Meghan A. Balk, John Bradley, M. Maruf, Bahadir Altintaş, Yasin Bakiş, Henry L. Bart Jr, David Breen, Christopher R. Florian, Jane Greenberg, Anuj Karpatne, Kevin Karnani, Paula Mabee, Joel Pepper, Dom Jebbia, Thibault Tabarin, Xiaojun Wang, and Hilmar Lapp. A FAIR and modular image-based workflow for knowledge discovery in the emerging field of imageomics. *Methods in Ecology and Evolution*, 15(6):1129–1145, 2024. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14327.

[8] Jim Balhoff, Soumyashree Kar, Juan Garcia, and Hilmar Lapp. Character Similarity Dataset, 2025.

[9] Juan Garcia, Soumyashree Kar, Jim Balhoff, and Hilmar Lapp. Trait2Vec, 2025.

[10] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, and others. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19412–19424, 2024.

[11] Jianyang Gu, Samuel Stevens, Elizabeth G. Campolongo, Matthew J. Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, James Balhoff, Wasila Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP 2: Emergent Properties from Scaling Hierarchical Contrastive Learning, May 2025. arXiv:2505.23883 [cs].

[12] Catia Pesquita, Daniel Faria, André O. Falcão, Phillip Lord, and Francisco M. Couto. Semantic Similarity in Biomedical Ontologies. *PLOS Computational Biology*, 5(7):e1000443, July 2009. Publisher: Public Library of Science.

[13] Prashanti Manda and Todd Vision. An analysis and comparison of the statistical sensitivity of semantic similarity metrics, May 2018. Pages: 327833 Section: New Results.

[14] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.