Paper Review For ModalityChoices
by Anurag Banerjee

# Evaluation of input data modality choices on functional gene embeddings[1]

*Felix Brechtmann et. al.*

## 1 Jargon Glossary

**Allele** - or *Allelomorph*, is a variant of the sequence of nucleotides at some locus in a DNA molecule; a person inherits one allele from each parent for an autosomal gene

**Assays** - Some kind of test to determine content, function or qualities

**Autosomes** - Any of the numbered chromosomes other than the sex determining chromosomes

**Co-variates** - A co-variate affects the outcome of a response variable in a statistical trial along with the explanatory variables under consideration. A co-variate itself is not of direct interest

**Elastic regularisation** - Elastic Net is a regularized regression method that linearly combines the $L_1$ and $L_2$ penalities of LASSO and Ridge methods

**Exome** - A naïve understanding of genome could be that it is made of alternating *introns* and *exons*. The set of all exons is the Exome. When a genome is transcribed, a particular cell type is formed which has exons drawn from the Exome

**Genotype** - The genotype of an organism is its complete set of genetic material

**Modality** - May refer either to different sources or types of input data

**Multi-collinearity** - A statistical concept where several independent variables in a model are correlated

**Phenotype** - The set of observable characteristics of an organism including morphology, developmental processes, biochemical, physiological as well as behavioral properties - generated from genotype and environmental factors

**Physiology** - branch of biology that studies how a living organism operates (functions and mechanisms)

**SNPs** - *Single Nucleotide Polymorphisms*, is a germline ( *population of reproductive cells*) substitution of a single nucleotide at a locus in the genome

**Stratified** - When a population is heterogeneous with sub-groups of different sizes, instead of random sampling, the sub-groups are arranged in *strata* and sampling is done proportionally to mitigate sampling bias

**Traits** - Various traits form the Phenotype set. Traits can be quantitative such as *blood pressure*, *height* or qualitative such as *eye colour*. It characterises specific aspects of an organism

**Trait-gene** - May be a reference to a *dominant allele*, gene responsible for the expression of a trait


Figure 1: Mindmap

> **Z-score** - or, *standard score*, is the number of standard deviations that a data point is above or below the mean of a measured quantity

## 2 Problem Description

Whereas, previous work in this field has focused on either the embedding generation algorithm or on a particular modality that such an algorithm can accept, in this research, the authors attempt to compare the effect of **different data modalities** for generating **functional gene embeddings** on a set of downstream tasks such as '*predict disease-gene list*', '*predict cancer drivers*', '*predict phenotype-gene associations*' and '*predict scores from genome-wide studies*'.

The two overall objectives were:

1. generate embeddings from various modalities
2. ensure that the embeddings are free from bias

The various modalities utilized are - '**quantitative OMICS**', '**protein-protein interaction networks**' and '**literature**'. The authors have observed that there is no clear winner for the data source/type and in general there is a dependance on the particular downstream task. Further, they have also observed the problem of undercoverage bias in gene embeddings generated from literature. Figure 1 provides an overview of this paper.

## 3 Problem Relevance

Machine Learning based downstream tasks require that the genes, diseases etc. are represented numerically. **Representation Learning** converts complex data structures (free text, ontology based annotations) into numeric vectors (or, matrices and tensors) so that ML techniques may consume it. The complex data structures are built from various data sources, such as,

| | | |
|---|---|---|
| *occurences in scientific literature* | *CRISPR screens* | *protein sequence* |
| *protein-protein interaction network* | *gene expressions* | *gene ontology annotation* |

Sampling Bias, a type of *selection bias*, pertains to over-representation of certain members of a population in some experiment. In research literature, some genes/diseases are more explored than others. Functional gene embeddings generated from such a literature corpus will contain this type of bias. This results in **under-performance** of the embeddings on downstream tasks for less studied genes/diseases. The authors have referred to this as **annotation inequality bias** caused by

street-light-effect.

In this study, the authors *have not attempted to create any novel solution to the downstream tasks*, but rather, focused on studying the effect of various data sources on such tasks using off-the-shelf ML algorithms. They have, however, proposed an embedding generation scheme that they claim will remove the aforementioned bias.

# 4 Proposed Solution

> **Disclaimer**: *The following may not be accurate! Reader discretion is advised.*

In order to compare the **modality choices** for the generation of functional embeddings for *gene-products*, the authors have proceeded in two phases:

1. prepare embeddings
2. compare on downstream tasks

In the first part various embeddings from different modalities are devised, then, in the second part a comparison of these embeddings are performed on some benchmark downstream tasks.

## 4.1 Embedding preparations

Three embeddings' groups are devised, namely:

1. OMICS
2. STRINGS
3. PoPS

For the **OMICS embeddings**, a Variational Deep Tensor Factorization Model is setup (Figure 2), wherein, three data sources, namely, **DepMap**, **GTEx** and **ProtT5** are used. ProtT5 [2] is a *language-model*, that generates embeddings utilising protein sequences information from the previous two sources. So, the gene-sample matrices available from DepMap (essentiality screen) and GTEx (expression profile) are used directly for their part, and their processed version is output from ProtT5. The VDTF Model then attempts to learn gene embeddings and sample embeddings by reconstructing the three aforementioned matrices. A gradient reversal layer [3] ensures that the learnt gene embeddings ()$a_j$) are **unable** to distinguish between the three datasources, as part of bias mitigation scheme.

The **STRING embeddings** were prepared in two versions, the first utilised the *complete* gene-gene interaction graph and in the second the edges that were proven by manual curation were removed. On each GGI graphs, two algorithms were applied, *viz.*, **node2vec** and **VERSE** and the resulting embeddings per gene were concatenated.

For the **PoPS embeddings**, the *polygenic priority score* features were downloaded as feature matrices, and two versions were created. In the first all info was retained, but, in the second, manually curated information was excluded. These feature matrices form the embeddings.

The data sources are mentioned in detail in **Table 2** in the paper. The second versions for the STRING and PoPS embeddings are referred to as *experimental*, and they aim at removing the street light effect.

## 4.2 Comparison on downstream tasks

A total of 6 downstream tasks were setup, which can be grouped into three categories, namely, **human curated gene-lists/annotations** which had evidence in published literature, **statistical association scores**, where the task was to predict the known scores and **annotation inequality assessment** which attempted to identify the bias. The performance of the embeddings are *mixed* and the authors failed to conclude if one embedding was better than rest, in general.

### 4.2.1 Human curated

**1. Disease-Gene Prediction:** The dataset is obtained from 6 sources mentioned in **Table 1** in the paper. XGBoost model is trained and tested for each embedding group, the task being, for all 6 lists, whether or not a gene is part of it.

**2. HPO Annotation Prediction:** The dataset is based on NCBI, Ensembl and STRING; the data and baseline results are obtained from HPOFiller (GCN) based. A 2-layer multi-class classifier predicts the phenotype annotation for genes.

**3. Cancer Gene Prediction:** The dataset is based on STRING and TCGA; the data and baseline results are obtained from EMOGI (GCN based). Again, XGBoost model was used to predict cancerous genes.

### 4.2.2 Association Studies

**4. Trait-Gene Association Prediction:** The data is obtained from UK Biobank's Genebass study and the target scores are computed using MAGMA. This task considered the rare variants. A logistic regression setup was used to match the MAGMA scores for traits per gene.

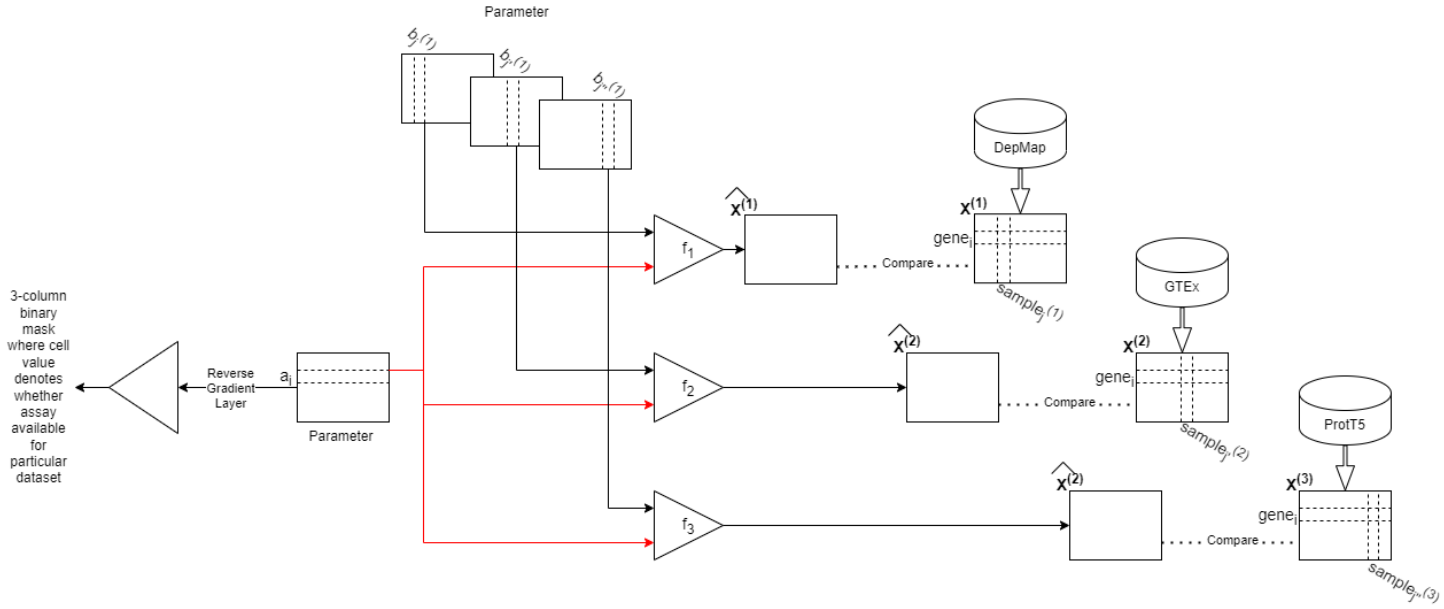


Figure 2: OMICS Embedding Generation

**5. Gene level Association:** The data is obtained from UK Biobank's GWAS study for 22 uncorrelated blood biomarkers and the target scores are computed using MAGMA. This task considered the common variants. The model was similar to that used in original PoPS study.

#### 4.2.3 Annotation Inequality

**6. Publication-count Bin Prediction:** The publications' list for each gene covered in previous benchmarks was obtained from PubMed, and some method was used to classify each gene into a publication-count bin. The objective was to test if embeddings that did not have street light effect leaked into them could still predict counts reliably.

## 5 Positive Points

- This study gives a good hint that modality for embeddings' generation is very much task dependent

## 6 Negative Points

- Too many experimental setups have been crammed into a single paper
- The description of the paper could have followed a more structured approach
- The publicly released source code [4] is not well structured and/or documented

## 7 Questions

1. The authors have observed that their study was inconclusive in terms of identifying a single robust modality for embedding generation or even embedding generation method. In light of that, what was even the point of publishing this paper?

## References

[1] F. Brechtmann, T. Bechtler, S. Londhe, C. Mertes, and J. Gagneur, "Evaluation of input data modality choices on functional gene embeddings," *NAR Genomics and Bioinformatics*, vol. 5, p. lqad095, 11 2023.

[2] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost, "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 7112–7127, oct 2022.

[3] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 11801189, JMLR.org, 2015.

[4] F. Brechtmann, "A genome-wide experiment-based functional gene embedding." Available: `https://github.com/gagneurlab/gene-embedding/tree/v1.0.0/`, Apr 2023. Accessed: 09 August 2024.