Benchmarking gene embeddings from sequence, expression, network, and text models for functional prediction tasks

Jeffrey Zhong[1], Lechuan Li[1], Ruth Dannenfelser[1], Vicky Yao[1,2,3*]

[1] Department of Computer Science, Rice University
[2] Ken Kennedy Institute, Rice University
[3] Rice Synthetic Biology Institute, Rice University
* *Correspondence to:* {vy}@rice.edu

ABSTRACT

Gene embeddings have emerged as transformative tools in computational biology, enabling the efficient translation of complex biological datasets into compact vector representations. This study presents a comprehensive benchmark by evaluating 38 classic and state-of-the-art gene embedding methods across a spectrum of functional prediction tasks. These embeddings, derived from data sources such as amino acid sequences, gene expression profiles, protein-protein interaction networks, and biomedical literature, are assessed for their performance in predicting individual gene attributes, paired gene interactions, and gene set relationships. Our analysis reveals that biomedical literature-based embeddings consistently excel in general predictive tasks, amino acid sequence embeddings outperform in functional and genetic interaction predictions, gene expression embeddings are particularly well-suited for disease-related tasks, and protein-protein interaction embeddings perform well in pairwise tasks. Importantly, we find that the type of training data has a greater influence on performance than the specific embedding construction method, with embedding dimensionality having only minimal impact. By elucidating the strengths and limitations of various gene embeddings, this work provides guidance for selecting and successfully leveraging gene embeddings for downstream biological prediction tasks. All associated code is available at https://github.com/ylaboratory/gene-embedding-benchmarks.

1 Introduction

Gene embeddings have been increasingly recognized as powerful tools in computational biology, providing a framework to transform complex, high-dimensional biological data into compact, more manageable numerical vector representations. These embeddings serve as a bridge between raw biological datasets and machine learning models, allowing researchers to better extract meaningful patterns that may be otherwise obscured by noise and heterogeneity in the data. By amplifying critical biological signals, embeddings can significantly improve the performance of downstream predictive tasks.

Embeddings have been shown to increase performance in predicting human disease associations [1, 2], gene function [3–5], and genetic interactions [6], among many other tasks in just the biomedical realm. In recent years, a variety of gene embeddings have emerged [7], expanding on classic biological vector representations [8–13], encoding gene-level information in a more compact form or over larger portions of integrated data. The latest methods employ a variety of techniques, such as: transformers, recurrent neural networks, multilayer perceptrons, skip-gram, and matrix factorization across different data inputs, such as: amino acid sequence, biomedical literature, gene expression, mutational profiles, and protein-protein interactions, potentially capturing different aspects of complex mammalian biology (Table 1). For example, one recent state-of-the-art method, Geneformer [14], uses a transformer-based architecture trained on large-

scale single-cell RNA sequencing data to produce gene representations. By encoding gene co-expression patterns within individual cells, Geneformer embeddings should effectively capture gene relationships across cellular states. In contrast, GenePT [15] uses a natural language processing approach, leveraging GPT-3.5 to generate embeddings from text descriptions of genes sourced from NCBI. This method encodes semantic information about gene functions and relationships as described in scientific literature, providing a complementary perspective to expression-based models.

These specific examples highlight an important question—if each method potentially captures different functional aspects of genes, do they capture shared or more specialized biological signals? Furthermore, are specific algorithmic or data source choices more or less suited for common downstream tasks? Unsal et al. [7] began to answer these questions by systematically evaluating representations across a variety of protein-centric tasks, highlighting the potential of embeddings and importance of training data type and model design on their performance. Existing gene-focused benchmarking studies typically consider only a single data type or narrowly defined task, such as drug response [16] or biological function [17] prediction, leaving a critical gap in understanding how embedding methods perform across a broader range of gene-associated applications.

In this study, we aim to address these questions by conducting a comprehensive benchmarking effort, systematically evaluating 38 state-of-the-art gene embedding methods (Table 1) across 3 categories of benchmarking tasks. Our benchmarks assess each method's performance in predicting: (1) individual gene level attributes and disease genes, (2) paired gene interactions and pathway edges, and (3) relationships between functionally meaningful collections of genes, referred to as gene sets. By systematically assessing embedding performance across these tasks, we aim to clarify their comparative advantages and limitations and provide a clearer roadmap for leveraging gene embeddings in future applications. Interestingly, we find that the underlying data type used in the creation of the embeddings is the most critical factor influencing performance, opposed to algorithm or the overall embedding dimension. Text-based models demonstrate strong generalizability across most tasks, whereas other models perform well in specific scenarios. More specifically, amino acid–based models excel in function and genetic interaction prediction, gene expression–based models perform particularly well in disease-related tasks, and methods that encode pairwise relationships such as protein-protein interaction networks or embeddings over large collections of integrated samples perform well on pairwise gene interaction and, to a lesser degree, gene set tasks.

2 Methods

2.1 Gene embedding acquisition and standardization

We performed a literature search for gene embedding methods and selected those that provided gene embeddings along with their publication. Each embedding was then processed through a standard pipeline, where all gene identifiers were converted to Entrez IDs using the mygene python library [36]. For genes with multiple identifiers corresponding to the same Entrez ID, the mean embedding was calculated to generate a single vector. Any genes that could not be mapped to Entrez IDs were removed. Only embeddings with at least 15,000 Entrez ID genes were kept to ensure a comparable intersection set of genes between embedding methods.

The embeddings AAC [11], ALBERT [23], APAAC [12], BERT-BFD [23], BERT-PFAM [37], BLAST [8], CPC-PROT [38], ESMB1 [26], GENE2VEC [30], HMMER [9], KSEP [13], LEARNED-VEC [21], MUT2VEC [32], PFAM [9], PROTVEC [22], SEQVEC [19], T5 [23], TCGA-EMBEDDING [29], UNIREP [20], and XLNET [23] were sourced from the Protein RepresentatiOn BEnchmark study [7]. The embeddings BioConceptVec [27], FRoGS [28], Geneformer [14], GenePT [15], Mashup [33], and scGPT [31] were retrieved from their respective publicly available repositories. In addition to these previously published embeddings, we created 3 additional embeddings (ESM2, Node2Vec, PPI-RAW). For Node2Vec and PPI-RAW, we used a consensus PPI

Table 1. Embedding Methods Tested. A summary of the embedding methods evaluated in this study, including their training input data types, underlying algorithm class, and original dimensionalities. For each method, the number of genes (before any preprocessing) and dimensions in the embedding are shown in parentheses.

Embedding Name	Training Input Data	Algorithm	Original Dimensions
AAC [7, 11]	Amino acid sequence	Non-ML	(18,910, 20)
APAAC [7, 12]	Amino acid sequence	Non-ML	(18,908, 80)
BLAST [7, 8]	Amino acid sequence	Non-ML	(18,910, 20,421)
HMMER [7, 9]	Amino acid sequence	Non-ML	(18,910, 20,421)
KSEP [7, 13]	Amino acid sequence	Non-ML	(18,835, 400)
PFAM [7, 9]	Amino acid sequence	Non-ML	(18,910, 6,227)
CPC-PROT [7, 18]	Amino acid sequence	RNN	(18,493, 512)
SEQVEC [7, 19]	Amino acid sequence	RNN	(18,910, 1,024)
UNIREP [7, 20]	Amino acid sequence	RNN	(18,910, 5,700)
LEARNED-VEC [7, 21]	Amino acid sequence	skip-gram/CBOW	(18,910, 64)
PROTVEC [7, 22]	Amino acid sequence	skip-gram/CBOW	(18,910, 100)
ALBERT [7, 23]	Amino acid sequence	Transformer	(18,493, 4,096)
BERT-BFD [7, 23]	Amino acid sequence	Transformer	(18,493, 1,024)
BERT-PFAM [7, 24]	Amino acid sequence	Transformer	(18,494, 768)
ESM2 [25]	Amino acid sequence	Transformer	(18,902, 1,280)
ESMB1 [7, 26]	Amino acid sequence	Transformer	(18,493, 1,280)
T5 [7, 23]	Amino acid sequence	Transformer	(18,493, 1,024)
XLNET [7, 23]	Amino acid sequence	Transformer	(18,493, 1,024)
BIOCONCEPTVEC-GLOVE [27]	Biomedical literature	Matrix factorization	(20,917, 100)
BIOCONCEPTVEC-CBOW [27]	Biomedical literature	skip-gram/CBOW	(20,917, 100)
BIOCONCEPTVEC-FASTTEXT [27]	Biomedical literature	skip-gram/CBOW	(20,917, 100)
BIOCONCEPTVEC-SKIP-GRAM [27]	Biomedical literature	skip-gram/CBOW	(20,917, 100)
GENEPT-ADA [15]	Biomedical literature	Transformer	(33,837, 1,536)
GENEPT-MODEL3 [15]	Biomedical literature	Transformer	(33,837, 3,072)
FROGS-ARCHS4 [28]	Gene expression (bulk)	MLP	(22,970, 256)
TCGA-EMBEDDING [7, 29]	Gene expression (bulk)	MLP	(18,432, 50)
GENE2VEC [7, 30]	Gene expression (bulk)	skip-gram/CBOW	(17,798, 200)
GF-12L30M [14]	Gene expression (single cell)	Transformer	(21,228, 512)
GF-12L95M [14]	Gene expression (single cell)	Transformer	(19,523, 512)
GF-12L95MCANCER [14]	Gene expression (single cell)	Transformer	(19,523, 512)
GF-20L95M [14]	Gene expression (single cell)	Transformer	(19,523, 896)
GF-6L30M [14]	Gene expression (single cell)	Transformer	(21,228, 256)
SCGPT-HUMAN [31]	Gene expression (single cell)	Transformer	(38,714, 512)
SCGPT-PANCANCER [31]	Gene expression (single cell)	Transformer	(38,714, 512)
MUT2VEC [7, 32]	Mutation profile, Literature, PPI	skip-gram/CBOW	(17,500, 300)
MASHUP [33]	PPI	Matrix factorization	(21,503, 800)
PPI-RAW [34]	PPI	Non-ML	(20,363, 20,363)
NODE2VEC [35]	PPI	skip-gram/CBOW	(20,362, 128)

network that combines experimentally-derived protein physical interactions from eight databases assembled previously [34]. Node2Vec generates random walks from the PPI network to create a gene emebdding with the skip-gram model. PPI-RAW uses the raw adjacency matrix of the consensus PPI as a simple "embedding," where each row of the matrix represents the interactions with every other gene. We also generated gene embeddings using a recent pre-trained protein language model (esm2_t33_650M_UR50D) [25], feeding it human protein sequences from Uniprot [39]. We then used the recommended method on the ESM2 GitHub for creating gene-level embeddings via mean pooling aggregation of corresponding amino acid embeddings. To prevent performance from being driven by gene inclusion and enable more direct comparability, all embeddings were filtered to a common set by taking the intersection of genes across all embeddings, resulting in a common set of 11,355 genes.

2.2 Embedding comparison

To compare across different embeddings, we transformed each embedding first into a gene-gene similarity matrix. Specifically, for each $N \times M$ embedding matrix, where N= # genes and M= # embedding dimensions, we computed the $N \times N$ Pearson correlation matrix of gene-similarities within each embedding. The

resultant gene-gene embedding similarity matrix thus eliminates differences in embedding dimensions across methods. Each gene-gene embedding correlation value was scaled to the range $[0, 1]$ using the transformation: $v' = \frac{v+1}{2}$, where v is the original Pearson correlation. Scaled values were then used to calculate the weighted Jaccard similarity index between pairs of embedding methods. For two scaled correlation vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, derived from two embedding methods, the weighted Jaccard similarity J_w was computed as:

$$J_w(\mathbf{v}_1, \mathbf{v}_2) = \frac{\sum_{k=1}^{n} \min(v_{1k}, v_{2k})}{\sum_{k=1}^{n} \max(v_{1k}, v_{2k})}.$$

2.3 Gene-level attribute benchmark

To benchmark gene-level attributes, we evaluated predictive performance on the Online Mendelian Inheritance in Man (OMIM) [40] and Gene Ontology (GO) [41] gene sets. We used the biological process annotation from Gene Ontology (2024-09-08). To reduce possible circularity, we only used GO annotations with experimental evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, HTP, HDA, HGI, and HEP). Leveraging the directed acyclic graph structure of the ontology, we propagated gene annotations from child to parent terms through “is a" and “part of" relationships. As a result, parent terms also contain genes annotated to their child term(s). Disease genes were obtained from OMIM (October 2023). We mapped OMIM ids to disease annotations through the Disease Ontology (DOID) [42] and followed a similar propagation process to propagate child annotations to parent terms.

We next filtered both OMIM and GO gene sets based on number of annotations, keeping only terms with more than 20 propagated genes. We also only kept terms that were children of terms in the Alliance of Genome Resources (AGR) slim (for both GO and OMIM). This filtering resulted in a total of 103 OMIM diseases and 385 GO terms. The GO terms were further reduced to a smaller size for faster benchmarking by selecting three random GO terms per AGR slim term, resulting in a total of 56 GO terms.

We trained individual Support Vector Machine (SVM) classifiers for each DOID and GO term, using the scikit-learn package [43]. For a given gene set, positive examples were defined as genes annotated to the gene set. To construct the negative set, we used the AGR slim to first exclude any genes annotated to a parent slim term for the given set from consideration. Next, we randomly sampled 10 times the number of positive examples across all non-parent slim terms, ensuring balanced representation from each slim term. We completely excluded 20% of the resultant gold standard as a holdout set. The remaining genes were used for three-fold cross-validation to select the SVM C parameter within [0.1, 1, 10, 100, 1000] that resulted in the best mean AUROC across folds for prediction on the holdout set. For all models, a balanced class weight and the radial basis function kernel was used, and all other hyperparameters were set to their default values.

2.4 Paired gene interaction benchmark

To evaluate the ability of gene embeddings to predict genetic interactions, we benchmarked their performance in predicting synthetic lethality (SL), negative genetic (NG) interactions, and transcription factor (TF) target relationships. Both the SL and NG genetic interaction data was obtained from BioGRID [44] (4.4.240), and transcription target information from the Transcription Factor Target Gene Database [45].

For the SL and NG tasks, positives consist of reported gene pairs for the respective genetic interaction. Negative examples were generated by permuting the genes present in the positive set across all gene pairs and selecting pairs that are not reported to show a genetic interaction. Each negative set was sampled to be 10 times larger than the number of positive examples. The sum of paired embeddings for each gene pair were used as input features. Similar to the previous benchmark, we used SVM with three-fold cross-validation for parameter tuning. 30% of the gene pairs were reserved as a holdout set for testing. To avoid information

leakage, each cross-validation fold and holdout set were split by genes instead of pairs (i.e., if gene a is in the cross-validation set, no gene pairs including a would appear in the holdout set).

For TF target prediction, we use TFs that have between 500 and 1,000 targets, yielding 116 TFs. To make the gold standard size more computationally manageable, we randomly downsampled the list of TF-target pairs to include only 5,000 positive pairs. Similar to the genetic interaction tasks, each cross-validation fold and holdout set are split by genes, and we trained SVMs for each TF using the same setup.

2.5 Gene set comparison benchmark

To evaluate the interpretability of gene embeddings in comparing gene sets, we used our previously published ANDES [46] tool to assess relationships between: (1) matched pathways from GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) [47] and (2) disease and tissues using disease-associated genes (OMIM) and tissue-specific genes obtained from the Brenda Tissue Ontology [48].

For the matched GO-KEGG comparison, the GO gene sets were the same as in the gene-level prediction task and the KEGG gene sets were obtained from ConsensusPathDB [49]. We then obtained KEGG-GO mappings using the external database cross-reference to GO biological processes from the KEGG web portal for 52 KEGG pathways. To evaluate the capacity to capture functional similarity beyond counting overlapping genes, we exclude all overlapping genes between the KEGG- and GO-matched gene sets from the GO gene sets during the evaluation of each method.

In the disease-tissue analysis, we compared disease-associated gene sets with tissue-specific gene annotations. The OMIM gene sets were the same as the previous gene-level prediction task. The BRENDA annotations were obtained from TISSUE [50], and we included only annotations with experimentally determined annotations with a confidence z-score$>$1.64. To ensure tissue-specificity, we define the “indirectly related set" for each term as tissue terms that cannot be reached without traversing through the root. Gene annotations were retained only if they were associated with fewer than 5% of the terms within this indirectly related set. Disease-tissue associations signify disease phenotypes observed in corresponding tissues and were made through manual curation of disease descriptions, resulting in 44 tissues terms associated with 79 disease terms. As the terms describe differing biology, unlike the GO-KEGG comparison, overlapping genes were retained when running ANDES.

3 Results

3.1 Gene embeddings capture distinct functional signals

We reviewed and categorized 38 gene embedding methods based on their input data types, algorithms, and embedding dimensionalities (Table 1). These embeddings span six broad data types, including gene expression (single-cell and bulk RNA-seq), mutation profiles, protein-protein interactions (PPI), amino acid sequences, and biomedical literature. We further categorized based on algorithm, into six categories: skip-gram/CBOW, transformers, multi-layer perceptrons (MLPs), non-machine learning methods (non-ML), recurrent neural networks (RNNs), and matrix factorization. Non-machine learning methods are a catch-all category for classic vector representations of genes with minimal further processing, such as protein sequence information from BLAST, PFAM, HMMER, or the raw PPI adjacency matrix. Dimensionalities of the embeddings range from compact vectors with as low as 20 dimensions to highly detailed representations with up to 20,421 dimensions.

Before standardizing all embeddings to a common intersecting gene set, we calculated raw gene coverage across embedding methods by calculating the Jaccard index (Supplementary Table). We found that most embeddings share a similar set of genes, with the notable exceptions of scGPT, GenePT, FRoGS-ARCHS4 and MASHUP. scGPT and GenePT had a broader gene coverage compared to other embeddings, thus lower Jaccard indices (< 0.6) when compared with most other embeddings, while compared with each other,

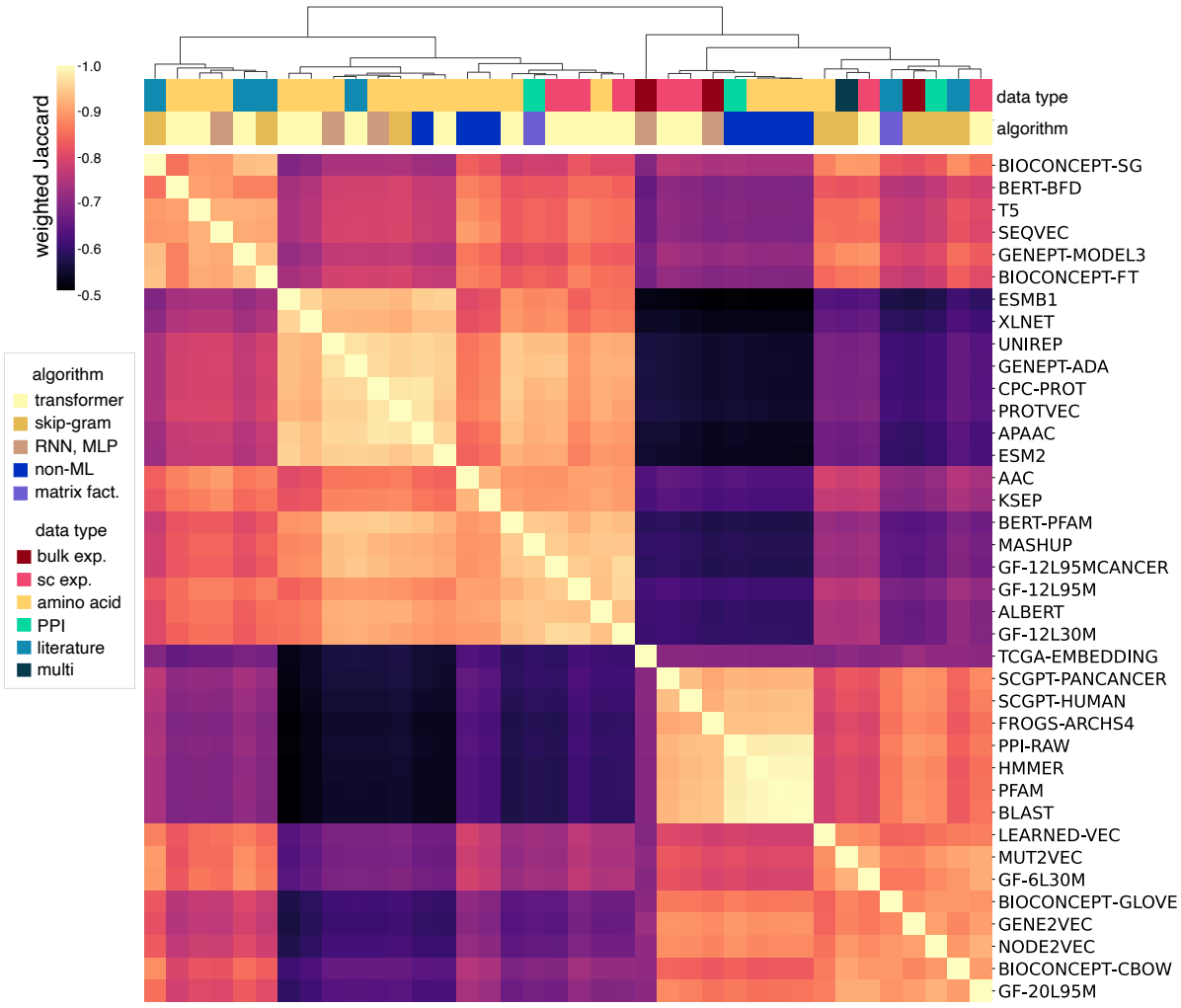

Figure 1. Embedding comparison using weighted Jaccard similarity. Higher weighted Jaccard scores indicate the encoding of similar gene-gene relationships.

they exhibited general concordance (Jaccard index ≈0.8). FRoGS and MASHUP, on the other hand, had Jaccard indices (≈0.6) with all other embeddings, suggesting they may cover a unique set of genes. After standardizing genes to a common intersecting set of 11,355 genes, we clustered the embeddings based on their weighted Jaccard similarities, where weights represent gene-gene similarity across embeddings (Figure 1). We observe distinct clustering patterns, influenced largely by the underlying data types and algorithms used. For example, amino acid sequence embeddings formed four smaller clusters, while gene expression embeddings grouped separately. Transformer-based embeddings tended to cluster together, distinct from classical non-machine learning methods such as BLAST, PPI-RAW, HMMER, and PFAM, which formed their own separate cluster.

3.2 Gene-level attribute benchmarks

We next evaluated how well embeddings predict individual gene-level attributes, focusing on two primary tasks: disease gene prediction (using OMIM annotations) and gene function prediction (using GO terms). Separate SVM classifiers were trained for each term using the set of different gene embeddings as features (Figures 2 and 3, Supplementary Table).

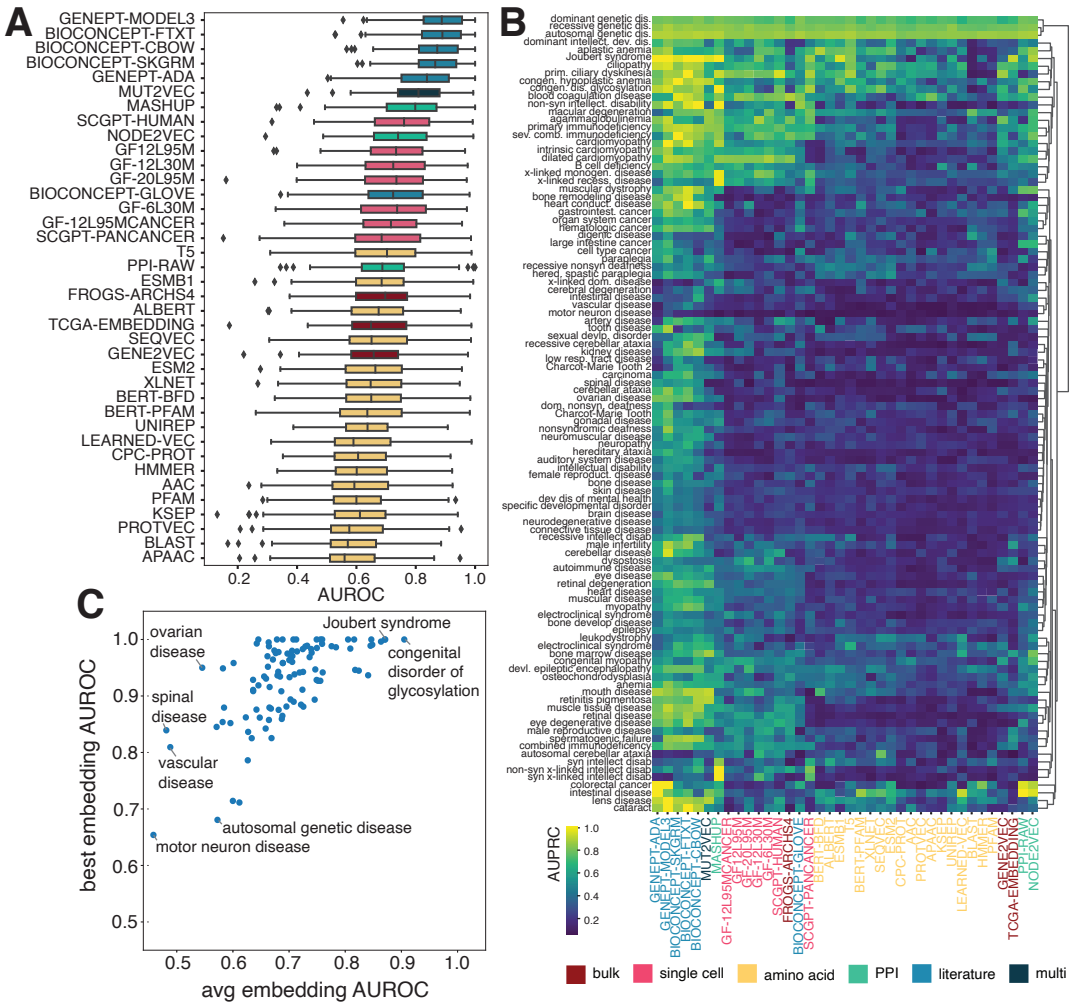

Figure 2. Disease gene prediction results. (A) Average AUROC scores across disease terms reported for each embedding and are colored by the underlying embedding data type. Biomedical text embeddings (blue-green bars) are the top performers, followed by PPI (green), gene expression (red), and amino acid embeddings (yellow). (B) AUPRC scores are shown for each disease term, with embedding names colored by embedding data type. Generally, literature-based models perform well, but some disease-gene predictions have strong performance regardless of embedding type. (C) Scatterplot highlighting the conservation of performance for specific diseases. Here, dots represent disease term prediction performance, comparing the average AUROC across embeddings vs the AUROC of the top-performing embedding.

For disease-gene prediction, embeddings derived from biomedical literature consistently outperformed other types (Figure 2A,B). BioConceptVec embeddings (FastText: mean AUROC=0.87; mean AUPRC=0.64) and GenePT (Model3: mean AUROC=0.88; mean AUPRC=0.68) emerged as the top-performing methods. Notably, BioConceptVec-FastText achieved comparable performance to GenePT-Model3 while using significantly fewer dimensions (Table 1), making BioConceptVec much more computationally efficient and approximately 16.5 times faster in our tests (Supplementary Table). Excluding biomedical text embeddings, those using PPI data, such as Mashup (mean AUROC=0.78; mean AUPRC=0.50) and Node2Vec (mean AUROC=0.74; mean AUPRC=0.40), and gene expression-based embeddings like scGPT (mean AUROC=0.74; mean AUPRC=0.41) and several Geneformer embeddings (mean AUROC=0.73; mean AUPRC=0.37) performed well. Embeddings derived from amino acid sequence data and non-ML approaches generally performed the worst overall (Figure 2A). Examining individual disease term predictions, we observed that a subset of diseases (dominant, recessive, and autosomal genetic disorders) were easy to predict across all

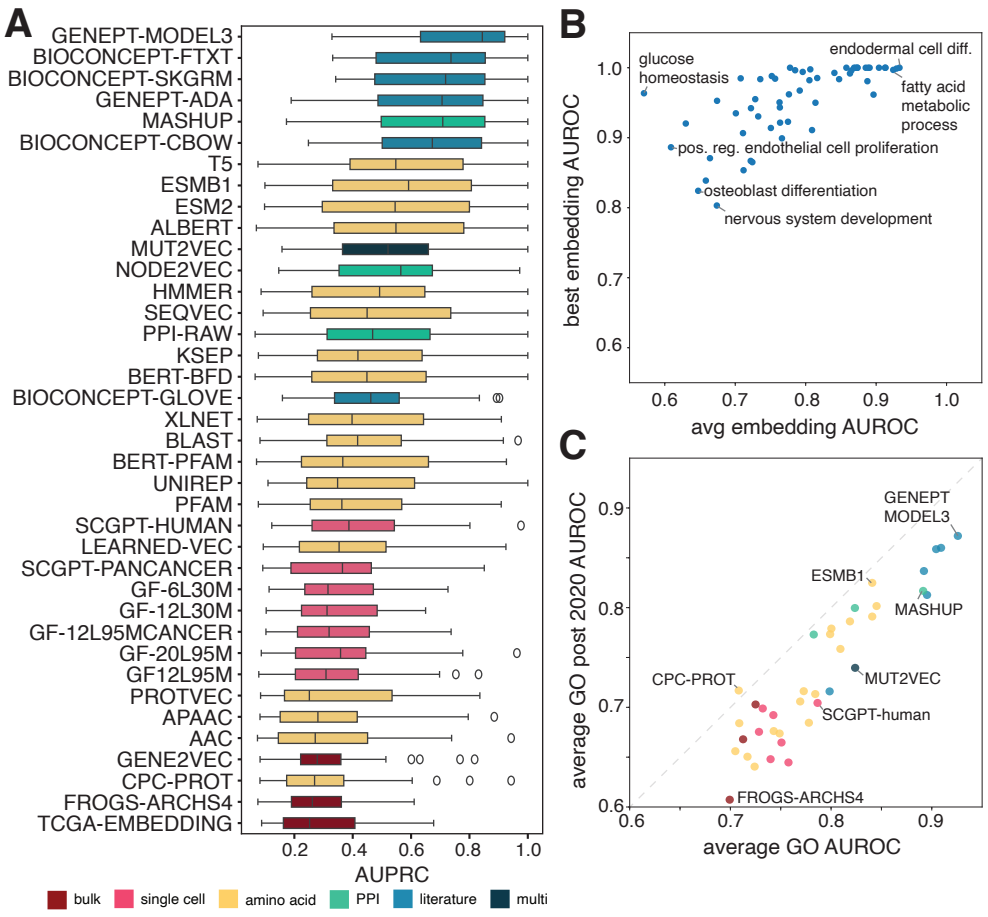

Figure 3. Gene function prediction results. (A) Boxplot showing the distribution of AUPRC scores across embeddings for all GO function prediction tasks. Boxes are colored by the underlying embedding data type. (B) Scatterplot highlighting the conservation of performance for specific GO terms. AUROCs are compared for each GO term between the average and best performing classifier. (C) Average performance of GO terms for each embedding method considering either all annotations (average GO) or only gene annotations to GO terms that have been added since 2020 (average GO post 2020). Methods that generalize well to newer GO data are above or near the dotted line.

embedding types (Figure 2B). Some disease genes had variable performance, particularly motor neuron disease and vascular disease, which were only predicted with high precision and recall by one literature-based model (GenePT-Model3 and BioConcept-FastText, respectively). In these cases, we speculate that the disease might be highly heterogenous, which can pose a problem for models that use noisy training data. To further investigate, we compared the average performance across all embeddings with the single best performing classifier (Figure 2C). Here, an interesting subset of diseases emerged, including ovarian disease and spinal disease, where again specific embeddings had much higher signal. Cross-referencing with Figure 2B, we see that these diseases were also predicted well only by the literature-based embedding models.

For GO gene function prediction, biomedical literature embeddings once again dominate in performance. GenePT-Model3 achieved the highest scores (mean AUROC=0.93; mean AUPRC=0.77), closely followed by BioConceptVec embeddings: FastText (mean AUROC=0.91; mean AUPRC=0.69), skip-gram (mean AUROC=0.90; mean AUPRC=0.68), and CBOW (mean AUROC=0.89; mean AUPRC=0.66). However, BioConceptVec-GloVe, a matrix factorization-based embedding, performed poorly (mean AUROC=0.78; mean AUPRC=0.46) compared to its counterparts. Interestingly, this was the case in the disease-gene prediction task as well. The amino acid sequence embeddings, such as T5, ESMB1 and ESM2, ranked well

(Figure 3A) for both AUROC and AUPRC (Supplementary Table). In contrast to the disease-gene prediction task, gene expression-based embeddings performed the worst for function prediction. Furthermore, much like the OMIM task, not all GO terms were easy to predict (Figure 3B). Nervous system development seems to be a challenging task for all the embeddings tested, while terms like glucose homeostasis only seem to do well with biomedical text or PPI-based embeddings.

One potential rationale for the strong performance of embeddings that use biomedical literature is data leakage—the literature-based embedding methods could use publications that report the GO term association or use those GO annotations as training data. While this cannot be fully ameliorated, we attempted to glean to what degree this may be an issue by not only restricting GO gene sets to experimentally derived annotations, but also conducting a temporal holdout using only GO annotations added after 2020 (Figure 3C). With the exception of GenePT (released in 2024), all text-based methods in this benchmark were created on training data prior to 2020. Given that GenePT uses a combination of NCBI gene summaries and ChatGPT for training, we are unable to ensure an independent test set, and we note that the strong performance of GenePT could be heavily driven by data leakage. We see that most methods perform worse using only post 2020 annotations, but mostly maintain their performance. The biggest drop in performance was in FROGS-ARCHS4 (Δmean AUC=-0.1). Interestingly, CPC-PROT was the only method that had a slight gain in performance on the 2020 subset (Δmean AUC=0.01).

3.3 Paired gene interaction benchmarks

To assess the ability of embeddings to predict paired gene interactions, we examined two broad categories of tasks: genetic interaction prediction (e.g., synthetic lethality (SL) and negative genetic interactions (NG)) and transcription factor target (TF) prediction (Figure 4, Supplementary Table). The TF prediction task was typically the easiest. Interestingly, though the overall top-performing methods consistently predicted SL next best, with NG having the worst performance, this was not consistently the case across all embedding methods, and for more than half of the embedding methods (21/38), the reverse was true.

Overall, GenePT-Model3 again emerged as the top performer across all tasks (AUROCs: SL=0.86; NG=0.76; TF=0.89). Other than GenePT-Model3, the PPI-based embeddings performed well overall. Notably, using the non-ML-based PPI adjacency matrix directly as an embedding performs second-best when considering the mean across all 3 tasks (AUROCs: SL=0.79; NG=0.73; TF=0.74), surpassing most embedding-based approaches, including those derived from PPI networks. The PPI-based embedding methods (Node2Vec and Mashup) in general outperformed BioConceptVec-based methods, but the performance improvement was heavily driven by the particularly poor performance of BioConceptVec methods in predicting NG relationships rather than dominating performance across all tasks. Overall, as with the gene-level attribute prediction tasks, biomedical literature-based embeddings still tended to perform strongly, though GenePT-ADA and BioConceptVec-CBOW performed much worse than their other counterparts. Unlike the gene-level prediction tasks, embeddings derived from amino acid sequence data performed very well, especially ESM2 (AUROCs: SL=0.82; NG=0.75; TF=0.75) and ESMB1 (AUROCs: SL=0.88; NG=0.70; TF=0.70).

For TF prediction, amino acid sequence-based embeddings as a group were the strongest performers, with 7 of the top 10 in holdout AUROC evaluations derived from this group. Among these, ALBERT (AUROC=0.87) stood out as a particularly strong performer, but this strong performance was not observed for predicting SL (AUROC=0.65) or NG (AUROC=0.67). Overall, gene expression-based embeddings performed poorly for TF prediction, comprising the bottom 9 worst-performing methods.

As noted earlier, the SL and NG prediction tasks tended to result in somewhat different prediction performance trends. PPI-based embeddings performed consistently well on both. However, biomedical literature generally outperformed amino acid and gene expression-based embeddings for SL prediction, while many of the amino acid-sequence based embedding performed very well on NG prediction.

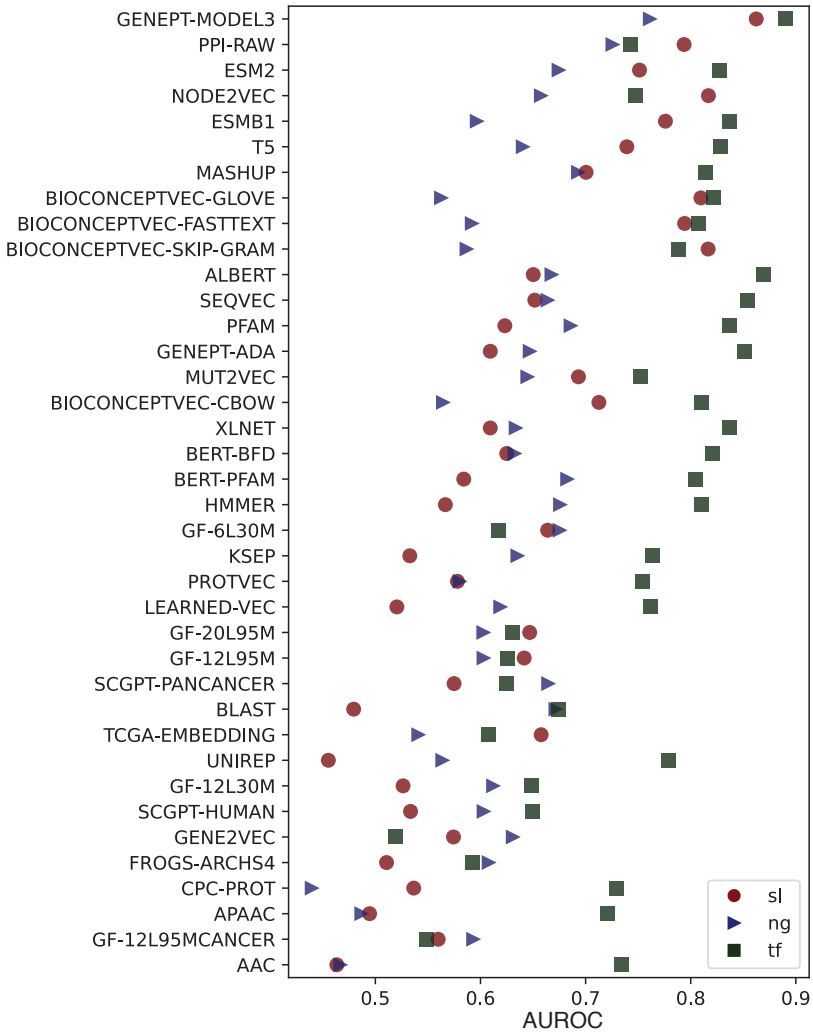

Figure 4. Paired gene relationship prediction results. Holdout AUROC scores for synthetic lethality (SL, red circle), negative genetic interaction (NG, green triangle), and transcription factor (TF, blue square) tasks, sorted by the mean AUROC across all tasks.

It is perhaps unsurprising that embeddings derived from PPI would perform well on the paired gene interaction benchmark, since physical interactions are known to be correlated with genetic interactions as well as TF-target binding, but it is interesting that using the PPI adjacency matrix directly as features on average outperforms corresponding denser embedding representations, suggesting that leveraging the direct interaction information was sufficient (rather than considering k-hop relationships). The fact that embeddings using amino acid representations performed so well at the TF tasks likely reflects that the corresponding embedding representations sufficiently capture sequence-based information underlying TF-target recognition (i.e., sequence motifs). Similar to the gene-level benchmarking tasks, we cannot rule out the possibility that the biomedical literature-based embeddings have an unfair advantage here due to data leakage, especially GenePT-Model3.

3.4 Gene set comparison benchmarks

Building upon our evaluations of individual genes and gene pairs, we extended our analysis to look at multiple genes within sets. To facilitate this, we used ANDES [46], a tool specifically designed for gene set

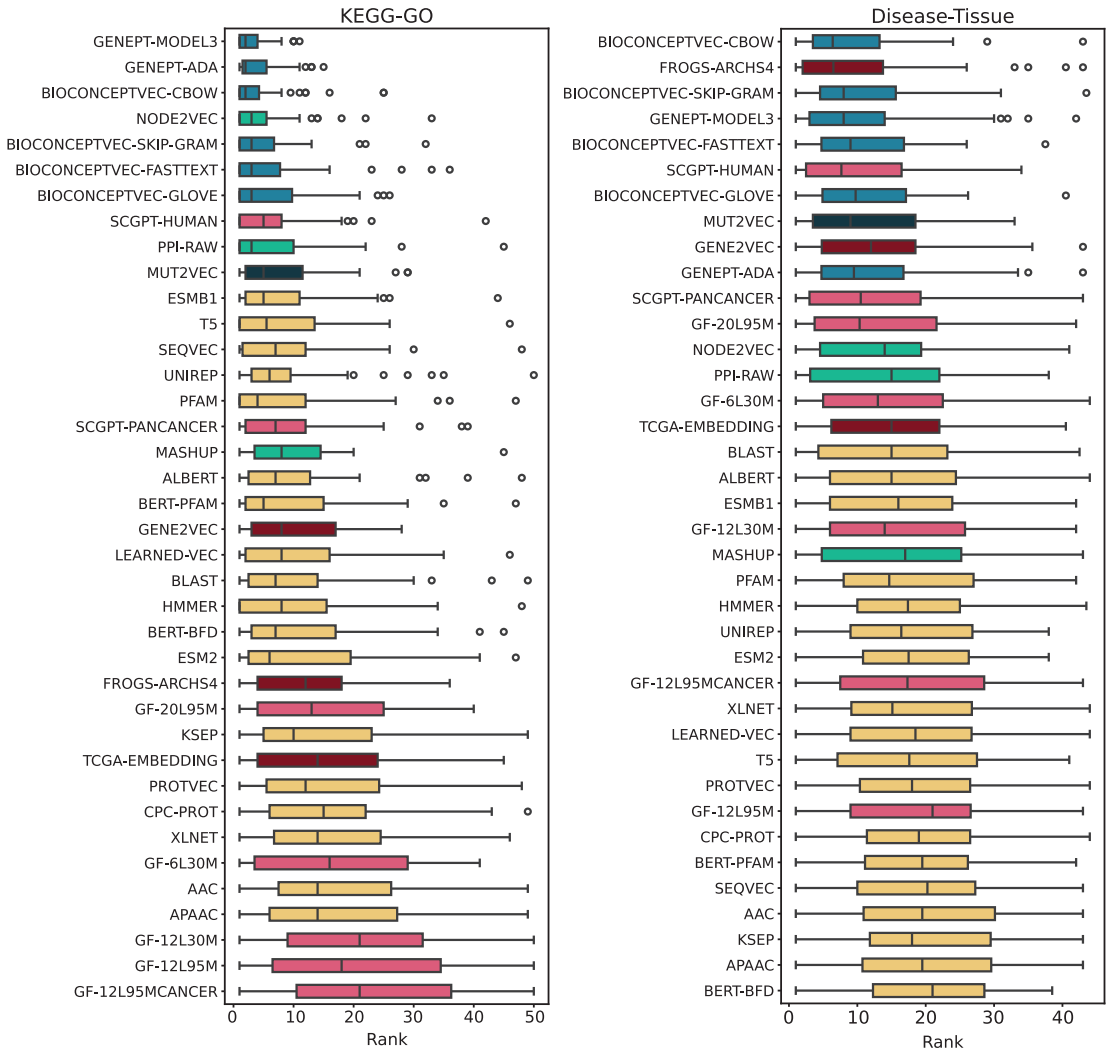

Figure 5. Comparing matching gene sets with ANDES. Boxplots of the ranking of the correct matching GO term for each KEGG term or tissue term for each disease term. Boxes are colored by the underlying input data type of each embedding.

analysis using embeddings. ANDES uses a best-match approach, calculating pairwise similarities between genes in two sets and identifying optimal matches in both directions. We applied this approach to assess the interpretability of embeddings by comparing gene sets in various biological contexts, first on GO and KEGG pathways with more directly corresponding functions and then on disease-tissue associations capturing more complex phenotype-associated signal. For GO and KEGG, we assessed how effectively each embedding could recover matched gene sets that describe the same biological processes across 52 matched term pairs (Figure 5, left). To minimize artificial inflation of matching performance due to overlapping genes, we removed genes shared between the paired KEGG and GO gene sets, retaining only the non-overlapping genes in GO. Among the embeddings, the 2 GenePT models performed best, followed by BioConceptVec and Node2vec, while Genformer embeddings ranked the lowest. In general, biomedical literature-based embeddings again outperformed other methods. Amino acid sequence embeddings also performed reasonably well, ranking generally higher than gene expression-based embeddings, where ESMB1, T5, and SeqVec performed the best. Of the gene expression-based embeddings, scGPT embeddings performed much better than its counterparts, beating all other gene expression-based embeddings.

We further evaluated the interpretability of embeddings using distinct but related gene sets, such as those linked through disease-to-tissue associations (Figure 5, right). In this task, we rank correctly matched tissue terms for each disease term for 44 tissue and 79 disease terms. Biomedical text based embeddings still performed the best, but interestingly, gene expression-based embeddings, which performed poorly in the GO-KEGG comparisons, demonstrated stronger performance in this task, while amino acid sequence embeddings demonstrated lower effectiveness. Notably, FROGS-ARCHS4, a gene expression embedding that struggled with most of the other benchmarks, was the second best-performing embedding overall, where BioConceptVec-CBOW was the best. scGPT also performed well, being the second and third best performing gene-expression method. When examining tissue-to-disease relationships, the overall trends and relative performance of embeddings remained largely consistent, with only minor shifts in their comparative outcomes.

3.5 Determining influential embedding attributes

To quantify the extent to which data type and other attributes of each embedding, such as algorithm and embedding dimensionality, affect performance, we fitted an ANOVA model per benchmark. Consistent with prior observations, data type emerged as the most influential factor across virtually all gene-level benchmarks. For example, data type accounted for 60.20% of the sum of squares for the OMIM task (p-value: 6×10^{-7}) and 53.44% for the GO task (p-value: 1.9×10^{-5}). In contrast, the choice of algorithm contributed 19.50% (p-value: 2.3×10^{-3}) and 20.21% (p-value: 9.6×10^{-3}) for OMIM and GO, respectively. Dimensionality, however, had negligible influence, with p-values > 0.75 for both tasks. This trend was consistent for most of the gene-pair benchmarks (Supplementary Table), further reinforcing that data type is the strongest determinant of performance, followed by the algorithm used, while dimensionality has little to no effect. It is worth noting that while less influential, algorithm choice still played an important role, with Transformer-based models consistently outperforming other algorithmic approaches within the same data type.

Interestingly, for gene set comparison tasks, some of these observed trends were slightly different. For the KEGG-GO matching task, data type remained significant (p-value: 1.1×10^{-4}), but this was not the case for the disease-tissue matching task (p-value: 0.16); in both cases, data type did still account for the largest proportion of the sum of squares (53% and 23% respectively). Though dimensionality had minimal effect on the gene-level and paired gene interaction benchmarks, it was somewhat influential for the KEGG-GO matching task (sum of squares: 12%, p-value: 0.02). We reason that for matching KEGG-GO terms, because the gene set comparison tasks do not rely on a separate classifier to draw a decision boundary, and as such, the information captured in the additional embedding dimensions can be more effectively leveraged. For the more challenging negative genetic interaction prediction and disease-tissue gene set matching tasks, we do not find any embedding attributes that are significantly associated with prediction performance.

4 Discussion

In this study, we undertook a comprehensive benchmarking effort to evaluate gene embeddings across a wide range of gene-centered tasks. One of the most notable findings from our benchmarks is the superior performance of text-based embeddings across all tasks. In particular, GenePT-Model3 was consistently a top performer, only dropping from the top position in the disease-tissue gene set matching task, which is the least well-characterized biological task. However, the impressive performance across all text-based embeddings raises concerns about potential data leakage. Since these embeddings are typically trained on a vast corpora of biomedical literature, there is a possibility that the information used in our functional prediction benchmarks may be implicitly contained within the embeddings themselves. This overlap could artificially inflate performance metrics over what may be encountered in novel settings. Nevertheless, our

post-2020 GO analysis (Figure 3C) did not reveal any major performance changes for text-based embeddings outside of what was seen for other embeddings.

Beyond text embeddings, newer representation learning-based methods using amino acid sequence generally outperformed classical counterparts, such as HMMER and BLAST. These modern embedding techniques are more compact, and require significantly less computational time while achieving superior performance (Supplementary Table). PPI-based embeddings performed reasonably well across all benchmarks, but particularly so on pairwise gene interaction benchmarks, and to some extent, the gene set comparisons. Mut2Vec stands out as a particularly interesting case, being the only method in our analysis that integrates multiple data modalities, including PPIs, text, and mutation profiles. Developed in 2018, Mut2Vec still demonstrated competitive performance, suggesting that integrating diverse data sources can enhance the robustness of gene embeddings for general purpose tasks. Given the advancements in language embedding and multimodal machine learning since Mut2Vec's inception, there is potential for modern embedding techniques to further improve by combining text-based data with gene expression or other biological datasets.

A limitation of our study is that some embedding methods offer broader gene coverage, enabling more comprehensive analyses across a wider range of genes. However, for fairness, we restricted our benchmarking analysis to the set of intersecting genes, potentially reducing the performance of some methods. Additionally, certain included embeddings were fine-tuned for targeted applications. For example, scGPT-cancer, which is optimized for cancer-related tasks, often underperformed more general models like scGPT-Human. This suggests that while specialized embeddings may excel within their target domains, their generalizability to diverse tasks may be constrained. Our study also predominantly focused on open-source embedding models with publicly available weights. By excluding proprietary, non-publicly or easily available embeddings, we may have missed some high-performing models.

Our benchmarking study revealed the general versatility and robustness of literature-based embeddings over amino acid sequence-, gene expression-, and network-based gene embeddings for enhancing functional prediction tasks. Together, these insights provide a guide post for selecting and leveraging gene embeddings tailored to specific downstream applications. Furthermore, they suggest new avenues for data integration to create hybrid embeddings that capture complementary biological signals. As the field continues to evolve, we advocate for more accessible embedding models that can help to further unravel the complexities of biological systems.

5 Acknowledgments

The authors thank Neel Mallipeddi and the members of ylaboratory for their feedback and suggestions. This work was supported by the Cancer Prevention & Research Institute of Texas (CPRIT RR190065) and the National Science Foundation (NSF DBI-2144534). VY is a CPRIT Scholar in Cancer Research.

References

1. Xiong, Y. *et al.* Heterogeneous Network Embedding Enabling Accurate Disease Association Predictions. *BMC Medical Genomics* **12,** 186. ISSN: 1755-8794. (2025) (Dec. 2019).
2. Yu, Z., Huang, F., Zhao, X., Xiao, W. & Zhang, W. Predicting Drug–Disease Associations through Layer Attention Graph Convolutional Network. *Briefings in Bioinformatics* **22,** bbaa243. ISSN: 1477-4054. (2025) (July 2021).
3. Kulmanov, M., Khan, M. A., Hoehndorf, R. & Wren, J. DeepGO: Predicting Protein Functions from Sequence and Interactions Using a Deep Ontology-Aware Classifier. *Bioinformatics (Oxford, England)* **34,** 660–668. ISSN: 1367-4811 (Feb. 2018).
4. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: Improved Protein Function Prediction from Sequence. *Bioinformatics (Oxford, England)* **36,** 422–429. ISSN: 1367-4811 (Jan. 2020).
5. Gligorijević, V. *et al.* Structure-Based Protein Function Prediction Using Graph Convolutional Networks. *Nature Communications* **12,** 3168. ISSN: 2041-1723. (2025) (May 2021).
6. Bryant, P., Pozzati, G. & Elofsson, A. Improved Prediction of Protein-Protein Interactions Using AlphaFold2. *Nature Communications* **13,** 1265. ISSN: 2041-1723. (2025) (Mar. 2022).
7. Unsal, S. *et al.* Learning Functional Properties of Proteins with Language Models. *Nature Machine Intelligence* **4,** 227–245. ISSN: 2522-5839. (2025) (Mar. 2022).
8. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND. *Nature Methods* **18,** 366–368. ISSN: 1548-7105. (2025) (Apr. 2021).
9. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure. *BMC Bioinformatics* **11,** 431. ISSN: 1471-2105. (2025) (Aug. 2010).
10. Mistry, J. *et al.* Pfam: The Protein Families Database in 2021. *Nucleic Acids Research* **49,** D412–D419. ISSN: 1362-4962 (Jan. 2021).
11. Gromiha, M. M. *Protein Bioinformatics: From Sequence to Function* ISBN: 978-0-12-388424-4 (Academic Press, Apr. 2011).
12. Chou, K.-C. Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics (Oxford, England)* **21,** 10–19. ISSN: 1367-4803 (Jan. 2005).
13. Wang, J. *et al.* POSSUM: A Bioinformatics Toolkit for Generating Numerical Sequence Feature Descriptors Based on PSSM Profiles. *Bioinformatics* **33,** 2756–2758. ISSN: 1367-4803. (2025) (Sept. 2017).
14. Theodoris, C. V. *et al.* Transfer Learning Enables Predictions in Network Biology. *Nature* **618,** 616–624. ISSN: 1476-4687. (2025) (June 2023).
15. Chen, Y. & Zou, J. *GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT* Mar. 2024. (2025).
16. Jia, P., Hu, R. & Zhao, Z. Benchmark of Embedding-Based Methods for Accurate and Transferable Prediction of Drug Response. *Briefings in Bioinformatics* **24,** bbad098. ISSN: 1477-4054. (2025) (May 2023).
17. West-Roberts, J., Kravitz, J., Jha, N., Cornman, A. & Hwang, Y. *Diverse Genomic Embedding Benchmark for Functional Evaluation across the Tree of Life* July 2024. (2025).
18. Lu, A. X., Zhang, H., Ghassemi, M. & Moses, A. Self-supervised contrastive learning of protein representations by mutual information maximization. *BioRxiv,* 2020–09 (2020).
19. Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics* **20,** 1–17 (2019).
20. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods* **16,** 1315–1322 (2019).
21. Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics* **34,** 2642–2648 (2018).

22. Asgari, E. & Mofrad, M. R. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one* **10,** e0141287 (2015).
23. Elnaggar, A. *et al.* Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **44,** 7112–7127 (2021).
24. Rao, R. *et al.* Evaluating protein transfer learning with TAPE. *Advances in neural information processing systems* **32** (2019).
25. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (2023).
26. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118,** e2016239118 (2021).
27. Chen, Q. *et al.* BioConceptVec: Creating and Evaluating Literature-Based Biomedical Concept Embeddings on a Large Scale. *PLOS Computational Biology* **16,** e1007617. ISSN: 1553-7358. (2025) (Apr. 2020).
28. Chen, H. *et al.* Drug Target Prediction through Deep Learning Functional Representation of Gene Signatures. *Nature Communications* **15,** 1853. ISSN: 2041-1723. (2025) (Feb. 2024).
29. Choy, C. T., Wong, C. H. & Chan, S. L. *Infer Related Genes from Large Scale Gene Expression Dataset with Embedding* July 2018. (2025).
30. Du, J. *et al.* Gene2vec: Distributed Representation of Genes Based on Co-Expression. *BMC genomics* **20,** 82. ISSN: 1471-2164 (Feb. 2019).
31. Cui, H. *et al.* scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI. *Nature Methods* **21,** 1470–1480. ISSN: 1548-7105. (2025) (Aug. 2024).
32. Kim, S., Lee, H., Kim, K. & Kang, J. Mut2Vec: Distributed Representation of Cancerous Mutations. *BMC Medical Genomics* **11,** 33. ISSN: 1755-8794. (2025) (Apr. 2018).
33. Cho, H., Berger, B. & Peng, J. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Systems* **3,** 540–548.e5. ISSN: 2405-4712, 2405-4720. (2025) (Dec. 2016).
34. Dannenfelser, R. & Yao, V. Splitpea: Quantifying Protein Interaction Network Rewiring Changes Due to Alternative Splicing in Cancer. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **29,** 579–593. ISSN: 2335-6936 (2024).
35. Grover, A. & Leskovec, J. *node2vec: Scalable feature learning for networks* in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (2016), 855–864.
36. Wu, C., MacLeod, I. & Su, A. I. BioGPS and MyGene.Info: Organizing Online, Gene-Centric Information. *Nucleic Acids Research* **41,** D561–D565. ISSN: 0305-1048. (2025) (Jan. 2013).
37. Rao, R. *et al.* Evaluating Protein Transfer Learning with TAPE. *Advances in neural information processing systems* **32,** 9689–9701. ISSN: 1049-5258. (2025) (Dec. 2019).
38. Lu, A. X., Zhang, H., Ghassemi, M. & Moses, A. *Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization* Sept. 2020. (2025).
39. Consortium, T. U. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51,** D523–D531. ISSN: 0305-1048. eprint: `https://academic.oup.com/nar/article-pdf/51/D1/D523/48441158/gkac1052.pdf`. `https://doi.org/10.1093/nar/gkac1052` (Nov. 2022).
40. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33,** D514–D517 (2005).
41. Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32,** D258–D261 (2004).
42. Schriml, L. M. *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research* **40,** D940–D946 (2012).

43. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12,** 2825–2830 (2011).
44. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **30,** 187–200 (2021).
45. Plaisier, C. L. *et al.* Causal mechanistic regulatory network for glioblastoma deciphered using systems genetics network analysis. *Cell systems* **3,** 172–186 (2016).
46. Li, L., Dannenfelser, R., Cruz, C. & Yao, V. A Best-Match Approach for Gene Set Analyses in Embedding Spaces. *Genome Research* **34,** 1421–1433. ISSN: 1549-5469 (Oct. 2024).
47. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28,** 27–30 (2000).
48. Gremse, M. *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research* **39,** D507–D513 (2010).
49. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic acids research* **41,** D793–D800 (2013).
50. Palasca, O., Santos, A., Stolte, C., Gorodkin, J. & Jensen, L. J. TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database* **2018,** bay003 (2018).