Enhancing gene set analysis in embedding spaces: a novel best-match approach

Lechuan Li[1], Ruth Dannenfelser[1], Charlie Cruz[1], Vicky Yao[1,*]

[1] Department of Computer Science, Rice University
* *Correspondence to:* {vy}@rice.edu

Abstract

Embedding methods have emerged as a valuable class of approaches for distilling essential information from complex high-dimensional data into more accessible lower-dimensional spaces. Applications of embedding methods to biological data have demonstrated that gene embeddings can effectively capture physical, structural, and functional relationships between genes. This utility has primarily been demonstrated by using gene embeddings for downstream machine learning tasks. Much less has been done to examine the embeddings directly, let alone embeddings for a set of genes beyond calculating simple mean embeddings between sets. Here, we propose a novel best-match approach that considers gene similarity while reconciling gene set diversity. This intuitive method has important downstream implications for improving the utility of embedding spaces for various tasks. Specifically, we show how our method, combined with different gene embeddings encoding protein-protein interactions, can be used as a novel overrepresentation-based and rank-based gene set enrichment analysis method that achieves state-of-the-art performance. Additionally, when used with a multi-organism joint gene embedding, our method can facilitate functional knowledge transfer across organisms, allowing for phenotype mapping across model systems. Our flexible, straightforward best-match methodology can be extended to other embedding spaces where there is a diverse community structure between set elements.

1 Introduction

Methods to build embedding representations have become ubiquitous in diverse fields spanning text-based [1, 2], image-based [3, 4], and domain-specific [5–10] applications. In addition to the computational benefits that lower dimension embedding representations provide, there is an implicit assumption that a quality lower dimension embedding amplifies the important signal in the data while reducing noise. In the biomedical realm, gene embeddings are gaining traction as a valuable approach for predicting function [11–13], disease associations [14, 15], among other applications[16–19].

Given the utility of gene embeddings for downstream machine learning tasks, it is intuitive that gene embeddings capture important gene-gene relationship information. However, surprisingly little attention is paid to exploring the resulting embedding spaces. In particular, in standard genomics analyses, gene set comparisons are very routine, as it is well understood that proteins work together to perform functional roles. Yet, there appears to be limited to no research considering the comparison of gene sets in embedding spaces. Existing methods outside of biological use cases use variations of averaging embedding information across genes, ultimately ignoring the functional diversity within sets.

One such averaging approach simply uses the centroid of all considered entities in the embedding space [20, 21]. However, gene sets, especially ones of interest, often contain a mixture of signals (e.g., a disease-associated gene set may include dysregulated genes from several pathways). The similarity calculated from

two gene set centroids would fail to consider different sub-functional groups in the gene set and instead obscure this signal, especially when the subgroups are distinct. Another relevant approach developed for interpreting gene set relationships in biological networks formulates the comparison as a t-test between the two gene sets, with a permutation-based background correction for the size of each gene set [22]. While the corrected t-test method does take into account the variability of the gene set, it is ultimately still anchored in a comparison of means. Another category of approaches tries to build individual embeddings for gene sets using network graph so that gene sets are directly comparable [23, 24]. However, these methods require the generation of new embeddings and cannot take advantage of existing embeddings, limiting their applicability. Furthermore, these methods are sub-class of graph representation methods and require a network backbone, which limits their ability to consider experimental information that cannot be easily transformed into a network graph format.

Here, we present an Algorithm for Network Data Embedding and Similarity analysis (ANDES), a best-match approach for gene set analysis that can be directly applied to existing embedding spaces (Figure 1). ANDES captures the diversity in sets by identifying the best-matching (most similar) gene in the other set and then taking a weighted sum between these best-matching similarities. Interestingly, this best-match concept has parallels to an early method proposed for biological text mining [25], but functional analyses in embedding spaces require adjustments for biases due to gene set cardinalities. ANDES incorporates a statistical significance estimation procedure that estimates the null distribution through Monte Carlo sampling to ensure comparable similarity estimations across different pairs of sets. The output of ANDES is an interpretable measure of similarity between two gene sets in the embedding space that considers gene set functional diversity.

We demonstrate that ANDES can better estimate gene set functional similarity in both single-species and cross-species settings compared to existing average-based methods, regardless of the underlying PPI gene embedding method. ANDES can be used directly for overrepresentation-based gene set enrichment analyses and, by extension, as a novel rank-based embedding-aware gene set enrichment analysis method. Leveraging our prior work on cross-organism joint embeddings[10], we also use ANDES with multi-organism gene embeddings for functional knowledge transfer. By matching phenotypes across human and mouse, we provide additional insight into opportunities for improved translational studies.

One of the fundamental limitations of both overrepresentation-based enrichment methods such as Fisher's exact test [26] and rank-based methods such as GSEA [27] is a complete reliance on gene annotations (e.g., functional annotations in the Gene Ontology); if a gene has no annotations, it cannot contribute to the enrichment analysis. There have been recent efforts to address this limitation. Network-Based Gene Set Enrichment Analysis (NGSEA) reranks the input gene list by incorporating the mean of its network neighbors [28]. Unfortunately, we are unable to test NGSEA because it is only available as a web portal and we are unable to change the underlying gene sets used. A later method, Gene set proximity analysis (GSPA), allows users to supplement gene set annotations by defining a radius in embedding space for each set [29]. ANDES does not require tuning any hyperparameters for gene set enrichment analysis and also outperforms Fisher's exact test, GSEA, as well as GSPA. The strong performance of ANDES across diverse tasks showcases the power of the best-match approach for gene set similarity comparisons.

2 Results

2.1 ANDES outperforms other set comparison approaches by effectively capturing substructure within gene sets

To highlight ANDES' ability to capture meaningful substructure, we first explore the extent to which set comparison methods can recover "functionally matched" gene sets in embedding spaces. Matched gene sets describe identical biological processes across different databases (we evaluate this for matching KEGG pathway and Gene Ontology (GO) biological process gene sets). Thus, we expect these gene sets

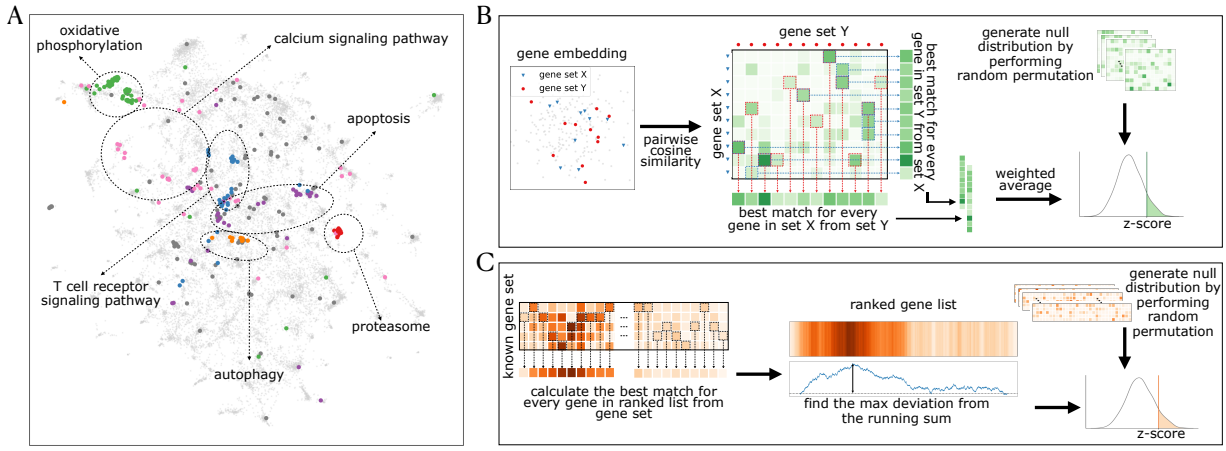

Figure 1. Overview of ANDES. (A) A UMAP plot of the node2vec gene embedding for a human PPI network with a set of Alzheimer's disease genes (hsa05010) highlighted. Within this set of disease genes, several sub-clustered biological processes scattered across the embedding space have clear biological function differences. ANDES is capable of considering this function diversity when matching gene sets. **(B)** Overview of the ANDES set similarity framework. Given two gene sets, ANDES first calculates the pairwise cosine similarity between every gene in each of the two sets. Based on the underlying similarity matrix, ANDES finds the best match for every gene and then calculates the weighted average to yield a single score. Statistical significance is estimated using a cardinality-aware null distribution. **(C)** Overview of ANDES rank-based gene set enrichment method. Given a ranked gene list based on an experimental result and a known gene set, ANDES calculates the best-match similarity for every gene in the ranked list. Walking down the ranked list, ANDES finds the max deviation from the running sum. The final enrichment score is also estimated using cardinality-aware null distribution.

to have similar patterns and substructures in an embedding space that represents gene relationships. The same process described in multiple gene sets may have overlapping genes, which can alter this matching problem into an easier one primarily driven by set overlap. To prevent this, we only keep one copy of overlapping genes and remove them from GO gene sets. We then compare ANDES against three methods: mean embedding, mean score, and the corrected t-score method. Mean embedding and mean score are two intuitive approaches for set comparisons in embedding spaces, and the corrected t-score method has been used for set comparisons in functional network representations [22]. To the best of our knowledge, these represent the current scope of embedding set comparison methods, highlighting the lack of method development for this problem.

All set comparison approaches are agnostic to the type of embedding method used. To see if the type of embedding impacts performance, we use three different methods to generate gene embeddings from a protein-protein interaction (PPI) network: node2vec, NetMF, and a structure-preserving autoencoder method (NN). ANDES consistently outperforms other methods and successfully identifies gene sets with similar functional roles, regardless of the underlying gene embedding method (Figure 2A). More specifically, ANDES significantly outperforms the mean score (node2vec: p=1.43e-2, NetMF: p=1.43e-2, NN: p=7.45e-3, Wilcoxon signed-rank test) and corrected t-score method (node2vec: p=5.50e-4, NetMF: p=5.59e-2, NN: p=5.54e-6, Wilcoxon signed-rank test). While ANDES' performance improvement over the mean embedding method is not statistically significant (node2vec: p=0.229, NetMF: p=0.413, NN: p=0.043, Wilcoxon signed-rank test), ANDES is much more stable. Practically, we observe that the mean embedding method can have more extreme failure cases compared to all other methods, likely due to the inherent limitation of collapsing information from all genes in the gene set into a single mean embedding before subsequent comparisons. As an example, we show one specific failure of mean embedding that occurs when matching KEGG term (hsa00071) with its corresponding GO term (GO:0009062) (Figure 2B). A direct inspection of the distribution of each gene set's genes in the embedding space (Figure 2B) quickly reveals that the correct KEGG-GO match has distinctly more similar embeddings, which ANDES can correctly identify and mean embedding cannot.

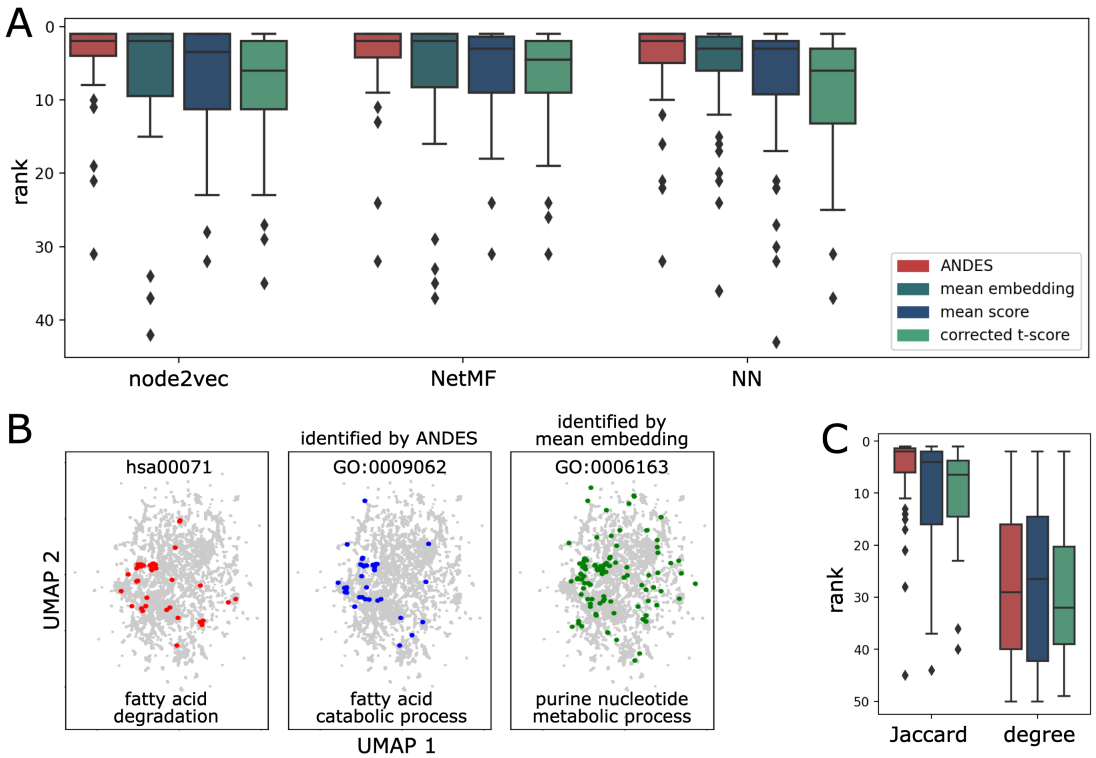

Figure 2. ANDES better matches gene sets that describe the same biological processes regardless of the underlying embedding or network structure. (A) Boxplots of the ranking of the correct matching GO term for 50 KEGG terms demonstrate that ANDES outperforms the mean embedding, corrected t-score, and mean score methods across three network embedding approaches (node2vec, NetMF, Neural Network (NN)). NN is a structure-aware autoencoder method (Methods). We also note that the handful of KEGG-GO pairs where ANDES performs poorly have consistently poor performance across methods (e.g., none of the 5 ANDES outlier KEGG terms in node2vec achieve a better ranking in any other methods). **(B)** UMAP of the node2vec PPI network embedding of genes in the KEGG fatty acid degradation gene set highlights a failure of the mean embedding method to capture meaningful substructure. Inspection of the embedding space reveals a similar substructure between the correct KEGG-GO term match prioritized by ANDES that is not seen in the top match for the mean embedding method. **(C)** Baseline approach for gene set matching in PPI networks. Matched KEGG-GO terms are ranked using pairwise similarity based on gene neighbor Jaccard similarity (Jaccard), or more naively, by the sum of node degrees (degree). Since embeddings are not used in this baseline, we cannot calculate mean embedding and instead compare ANDES to only the corrected t-score and mean score.

Indeed, both the mean score and corrected t-score methods rank the correct term higher than the mean embedding method. This highlights the importance of capturing diversity within gene sets to prioritize the correct matches.

As a baseline, we also examine the extent to which matching gene sets can be identified using naive non-embedding network-only approaches, such as shared neighbor profiles and node degree in the original PPI network. Interestingly, using the Jaccard index of shared neighbors between two genes also captures sufficient functional signals to identify several correct KEGG-GO matches. In this setup, ANDES still significantly outperforms the mean score (p=2.26e-2, Wilcoxon signed-rank test) and corrected t-score (p=6.31e-4, Wilcoxon signed-rank test) methods (Figure 2C). Since this baseline does not involve embeddings, it is impossible to calculate mean embeddings for this comparison. Though the shared neighbor profiles clearly capture some functional signals, we note that using embedding approaches such as node2vec as input still leads to better performance (ANDES: p=0.092, Wilcoxon signed-rank test). Using the simple sum of node degrees in the PPI network to measure gene similarity results in nearly random performance for the 3 methods explored in this comparison. Altogether, this demonstrates that the application of ANDES is

not limited to only embedding spaces, but its performance does improve with the informativeness of the similarity matrix that is used as input.

2.2 ANDES as a novel overrepresentation-based and rank-based gene set enrichment method

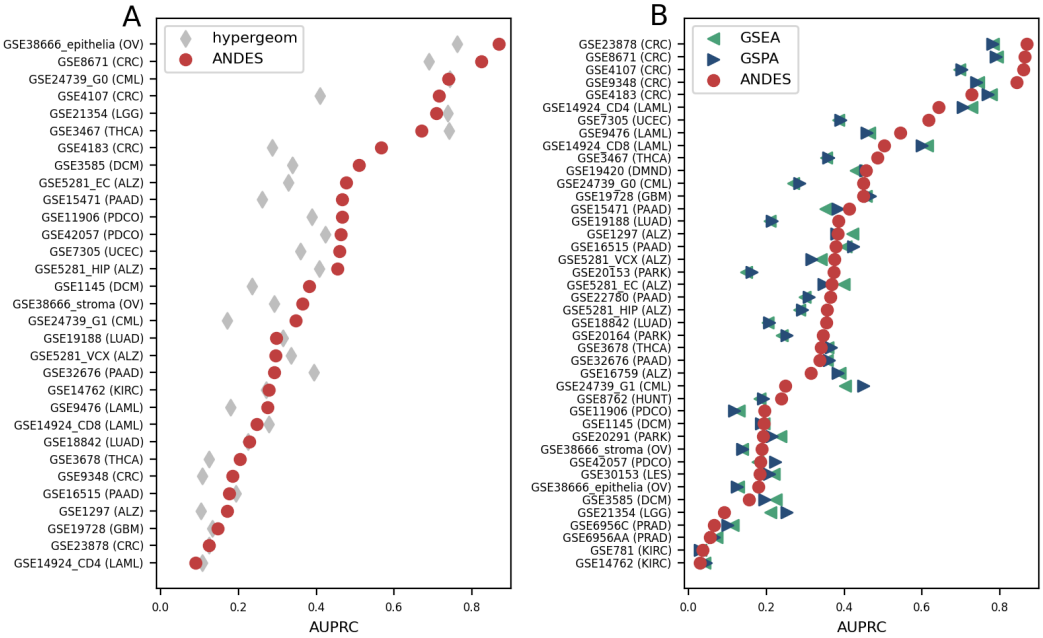

Figure 3. ANDES achieves state-of-the-art performance in overrepresentation-based and rank-based gene set enrichment for the GEO2KEGG [30] gene set enrichment benchmark. (A) Performance comparison between ANDES and hypergeometric test in retrieving annotated KEGG terms using genes that have FDR$\leq$0.05 in each dataset (where there are at least 10 genes that are significantly differentially expressed). **(B)** Performance comparison between ANDES, GSEA [27], and GSPA [29] in retrieving annotated KEGG terms using the full list of genes (no FDR cutoff), ranked by $\log_2$(fold change). In both cases, ANDES statistically outperforms other methods, demonstrating the advantage of incorporating gene embedding information using best-match into the gene set enrichment setting.

Gene set enrichment is a natural extension of the set matching abilities of ANDES. Gene set enrichment methods fall into two main classes: overrepresentation-based approaches [26, 31] and ranked-based approaches [27], of which hypergeometric test and Gene Set Enrichment Analysis (GSEA) are respectively representative methods [27]. We have already shown that when paired with gene embeddings, ANDES can work effectively as an improved set overlap method, and we now further solidify this finding on a well-established gene set enrichment benchmark (GEO2KEGG [30]). We also extend the ANDES framework to a rank-based embedding-aware enrichment approach, allowing users to calculate gene set enrichment given any ranked list of genes.

GEO2KEGG is a widely adopted gene set enrichment benchmarking dataset that annotates differential expression results (including FDR as well as log-fold-change per gene) with enriched KEGG pathways. The complete benchmark contains 42 microarray studies covering 200 KEGG pathways, and provides a unique opportunity to test both overrepresentation-based enrichment (through the use of significance or fold change cutoffs) and rank-based gene set enrichment (by ranking all genes based on their fold changes). Here, we calculate AUPRC based on the recovery of the corresponding KEGG pathways for each dataset. For the overrepresentation case, we compare ANDES performance with the hypergeometric test [26], testing gene sets by taking differentially expressed genes per dataset (FDR≤ 0.05), keeping gene sets that are larger than 10 genes, as is standard practice for overrepresentation analysis. In 22 out of 31 cases (71% datasets), ANDES set enrichment outperforms the hypergeometric test for KEGG pathway identification (p<7e-4, Wilcoxon

rank sum test, Figure 3A). For the rank-based version, we compare ANDES with GSEA [27] as well as a new embedding-based gene set enrichment method, GSPA [29]. Aggregating performance across all 42 datasets in the benchmark, ANDES' rank-based gene set enrichment significantly improves over both GSEA (p=0.041; Wilcoxon rank sum test) and GSPA (p=0.028; Wilcoxon rank sum test) (Figure 3B). Together, these results highlight ANDES' utility as a novel state-of-the-art method for overrepresentation-based and rank-based gene set enrichment. Furthermore, since ANDES utilizes gene embeddings, enrichments can be performed in cases without gene overlap, giving it a unique advantage, especially for the overrepresentation case.

2.3 ANDES can be used to recapitulate known relationships between drugs and prioritize new candidates for drug repurposing

To highlight how ANDES can be used to discover new biology, we use ANDES to match disease gene sets from OMIM [32] with drug target genes from DrugBank [33]. This is a use case of practical importance, since the drug designing process for new compounds is quite laborious, involving many layers of development to ensure compound safety, delivery, efficacy, and stability. Thus, a computational effort that can potentially be used to repurpose a vetted compound can greatly help accelerate the development of new therapies. We first calculate a disease similarity profile for drugs using their ANDES scores. After applying dimensionality reduction using principal components analysis (PCA) on these profiles, we find that drugs cluster by their therapeutic classes (Figure 4A). The clustering is particularly noticeable for both "nervous system" drugs and "antineoplastic and immunomodulating agents," which each target similar pathways, giving us confidence in the ability of ANDES to find biologically meaningful signal.

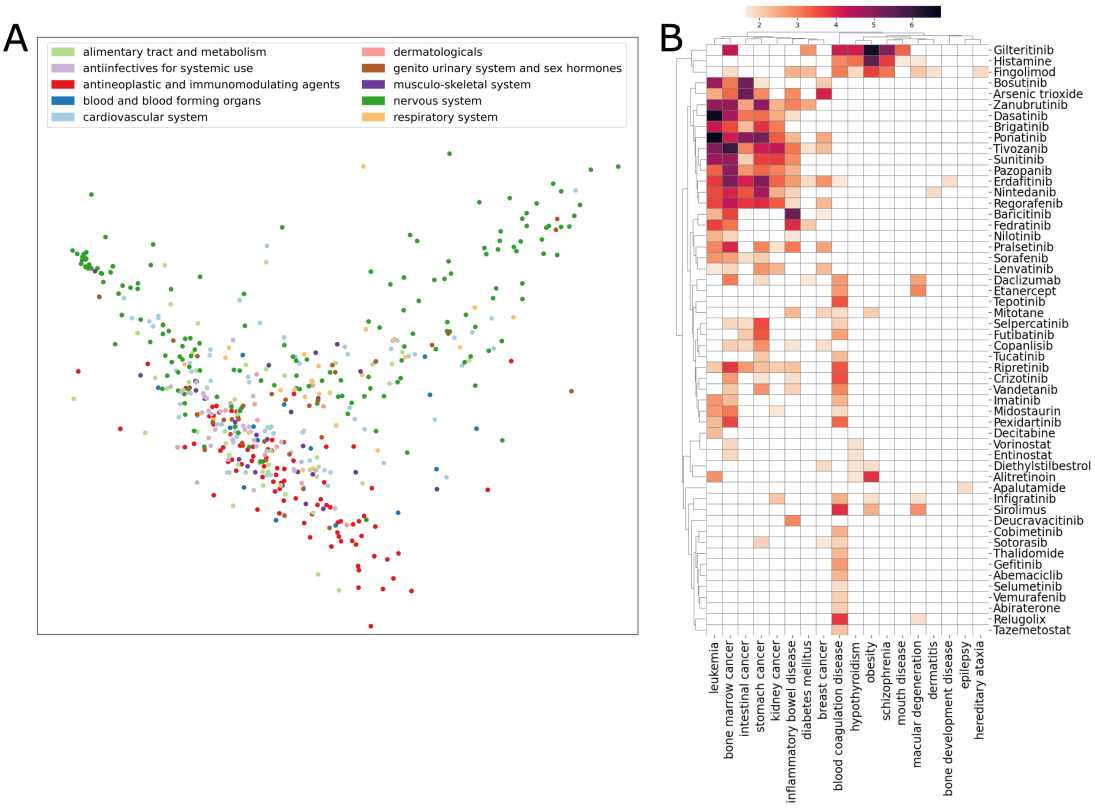

Figure 4. Analysis of drug-disease relationships using ANDES. (A) PCA plot showing the relationship between drugs based on their association with diseases. Colors are based on the 1st level of ATC groups. **(B)** Heatmap of ANDES gene set similarity z-scores (darker color indicates higher z-score) between diseases and drugs in the "antineoplastic and immunomodulating agents" therapeutic class. Diseases and drugs that have at least one association (z-score>1.64) are retained, yielding a heatmap of 18 diseases and 54 drugs.

As a proof of concept for highlighting avenues for drug repurposing, we take a closer look at the "antineoplastic and immunomodulating agents," examining the ANDES gene set similarity scores for all drugs in this class (Figure 4B). The ANDES similarity for other drug groups are in the supplement (figure S26-37). Only diseases and drugs that have at least one significant match ($z > 1.64$) are kept. Besides the most obvious cluster of cancers, ANDES also captures potentially novel indications or potential side effects. For example, ANDES predicts a strong association between obesity and two drugs, histamine and Gilteritinib, which are well-documented associations. Specifically, Histamine can decrease hunger by affecting the appetite control center in the brain [34], and weight gain is a listed side effect of Gilteritinib [35]. While these are all known relationships, ANDES also predicts less-documented, potentially novel drug-disease relationships such as a link between Fingolimod and schizophrenia. As recently as 2023, Fingolimod has been examined in rats for its potential to reverse schizophrenia phenotypes [36]. We observe a similar association between Sirolimus and macular degeneration. Sirolimus is an immunosuppressive agent used to prevent transplant rejection, but as of 2021 has been found to have emerging promise as a therapeutic for age-related macular degeneration [37]. These two seemingly disparate indications likely share an inflammatory pathological pathway, which can be picked up with ANDES.

2.4 ANDES can be effectively used with other types of gene embeddings, such as cross-organism embeddings for increasingly complex biological insights

We have already shown that ANDES performance is agnostic to the underlying embedding, making it a modular framework. To make it even more powerful, we swap the human PPI-based gene embeddings for joint cross-organism gene embeddings using our previously published network embedding alignment method, ETNA [10]. This analysis highlights two general principles: (1) ANDES can still prioritize relevant signal when the underlying gene embedding is structurally more complex and (2) the scope of new biological insights can be expanded with the use of different embeddings.

Beyond showcasing the power of ANDES, being able to successfully map coordinated gene sets such as pathways and processes between model organisms and humans is an important problem. Model organisms are critical for studying aspects of human biology that are technically infeasible or unethical to study directly. Thus, improving functional knowledge transfer increases the potential impact of model system study.

To determine if ANDES can aid in functional knowledge transfer, we use ETNA to build three pairwise joint gene embeddings between humans and three model organisms: *M. musculus*, *D. melanogaster*, and *C. elegans*. ETNA's joint embedding space enables the calculation of a similarity matrix for genes across species. Since genes across organisms can be annotated to the same GO terms, we can also evaluate to what extent the same GO term is prioritized in human when using the species-specific annotations in model organisms. For all three model organisms, ANDES consistently outperforms the mean embedding (*M. musculus*: p=2.71e-7, *D. melanogaster*: p=1.44e-1, *C. elegans*: p=9.29e-4), mean score (*M. musculus*: p=9.92e-19, *D. melanogaster*: p=4.25e-3, *C. elegans*: p=9.80e-4), and corrected t-score (*M. musculus*: p=3.43e-17, *D. melanogaster*: p=2.90e-2, *C. elegans*: p=7.19e-4) (Figure 5A).

When we group the GO terms shared between human and *M. musculus* by size, ANDES shows better performance, both when gene set is large (mean embedding: p=2.68e-2, mean score: 5.15e-14, corrected t-score: 1.70e-8) and small (mean embedding: p=8.14e-6, mean score: 2.06e-7, corrected t-score: 5.39e-12). We notice that larger gene sets are easier to match across organisms (Figure 5B). We speculate that having more genes can result in more distinct patterns in the embedding space, leading to better mappings, especially for ANDES and the mean embedding method. Meanwhile, the mean score and corrected t-score methods are not able to take advantage of the additional information in larger gene sets and perform similarly. Interestingly, the mean embedding method performs poorly when the gene sets are small; this is likely because "outlier" genes can more easily skew the mean embedding, especially when there are distinct subprocesses within a gene set. Overall, ANDES is the only method that is a strong performer, consistently robust to gene set size.

So far, we have simplified the relationship between a pair of gene sets to simply "matched" or "unmatched," but we can also evaluate unmatched terms based on how close they are to the correct target term in the ontology tree. To this end, we use Resnik measure[38], a semantic similarity measure leveraging the GO hierarchy, to quantify the similarity between two gene sets. Two gene sets that are close in the Gene Ontology are more likely to describe functionally similar biological processes and therefore have a higher Resnik score. We evaluate all gene set predictions using the Resnik score. Comparing predicted similarities for all gene sets between human and mouse, we calculate the average Resnik score of gene set pairs ranked by their similarity score in each set comparison method. Across all methods, the Resnik score is higher for the top-ranked pairs and gradually converges to randomness at around 30% of the ranked list of all pairs (Figure 5C). The trend also holds for *D. melanogaster* and *C. elegans* (Figure S2). Overall, ANDES consistently has the highest Resnik scores of all methods, demonstrating that it both identifies the exact match as well as other functionally related gene sets.

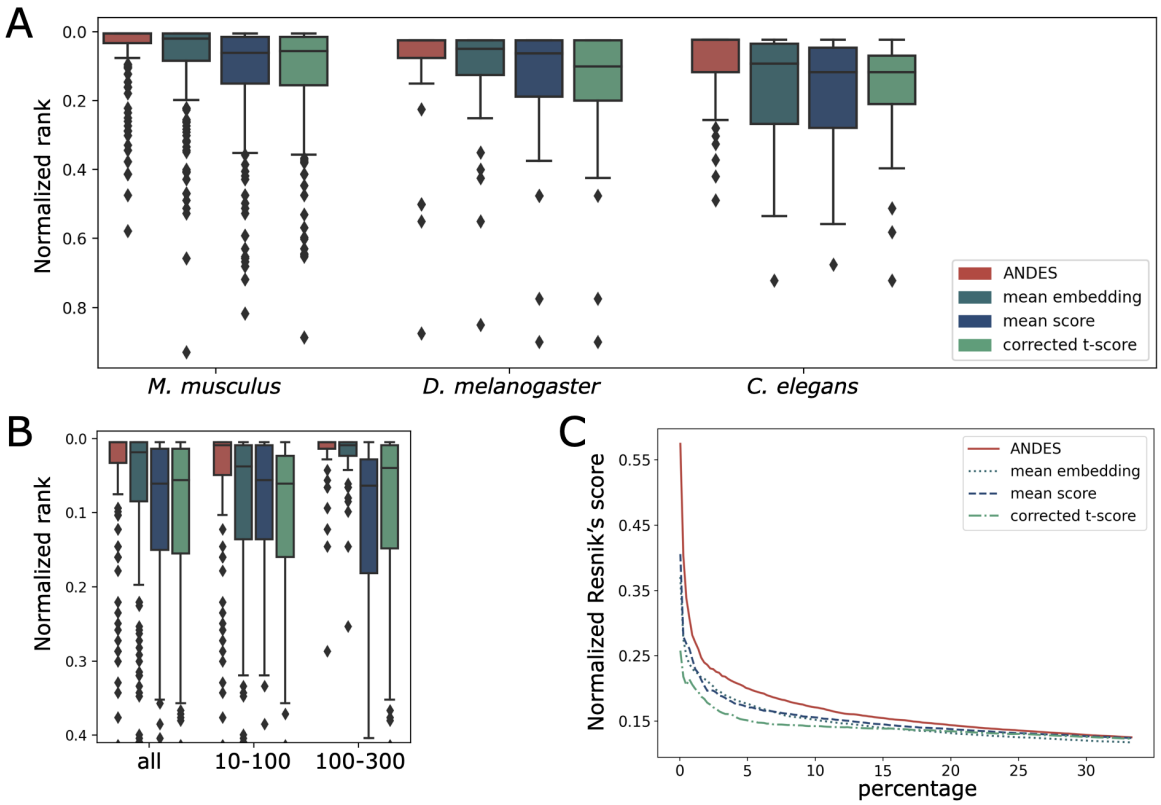

Figure 5. ANDES estimates gene set functional similarity across organisms better than existing methods. (A) Boxplot of the ranking of matched GO terms between human and three model organisms: *M. musculus*, *D. melanogaster*, and *C. elegans*, with 213, 40, and 43 shared GO slim terms, respectively. To facilitate comparison between organism pairs, the ranking is normalized by the number of shared GO terms. For each of the three organisms, ANDES consistently outperforms the mean embedding, corrected t-score, and mean score methods. **(B)** Boxplot of the ranking of matched GO terms for *H. sapiens* and *M. musculus*. Gene set pairs are grouped into two categories according to the sum of the number of genes in both gene sets (small [10-100] and large [101-300]). ANDES again consistently outperforms other methods regardless of gene set size. **(C)** Comparison of the cumulative average of the Resnik score walking down the ranked list for *H. sapiens* and *M. musculus*. ANDES consistently outperforms other methods until the score converges at ∼30% of all pairs.

2.5 Prioritizing mouse phenotypes for modeling human diseases with ANDES

After verifying that ANDES can recover conserved cross-organism functional signal (Figure 5), we further explore the potential of ANDES for cross-organism knowledge transfer. Phenotype prioritization is an

important aspect of effective knowledge transfer as some small phenotypic changes in the model organism (e.g., "decreased cervical vertebrae") can be an important marker of the presence or extent of a human disease phenotype. Good matches here can be potential candidates for phenotypic screens. Towards this end, we systematically test associations between human disease gene sets from OMIM [32] and mouse phenotype gene sets from MGI [39]. We identify a range of significantly related disease-phenotype pairs, many of which merit further exploration (Figure S3-25). As a proof of concept, we show a smaller slim set of 13 human diseases that span a wide range of organ systems and pathological mechanisms, along with the top 5 associated mouse phenotypes (Figure 6).

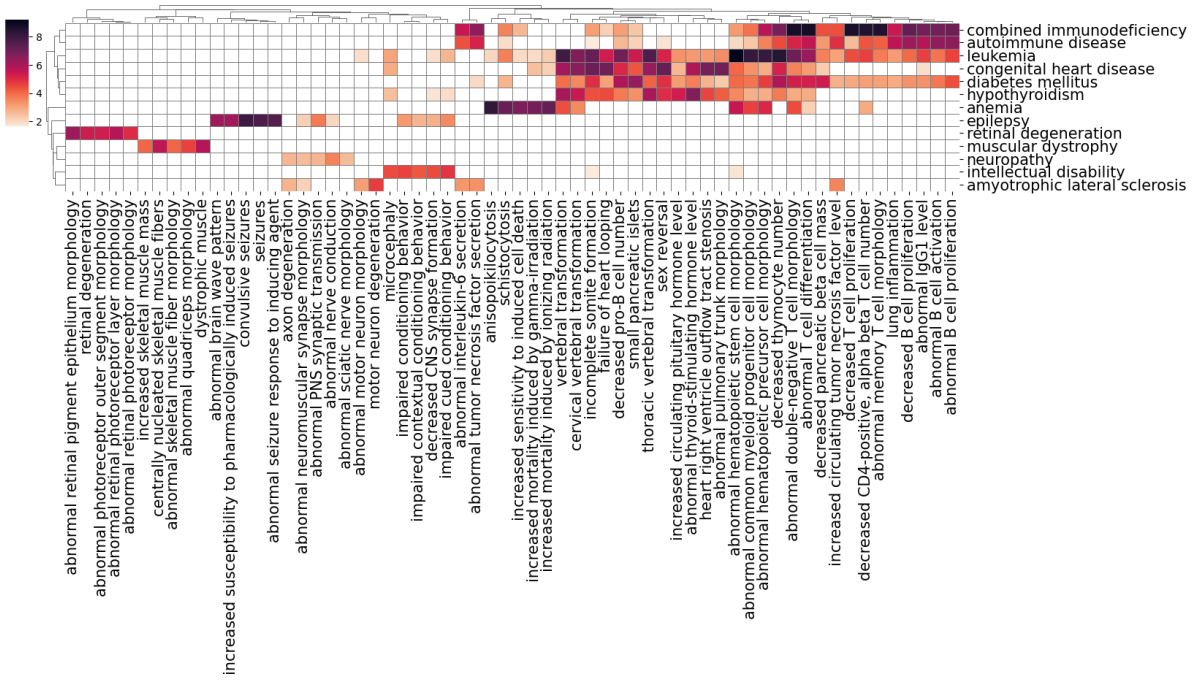

Figure 6. Heatmap of ANDES similarity scores for human disease and mouse phenotypes gene sets. Gene set similarity z-scores generated by ANDES are shown for 13 selected human diseases across various organ systems and pathological pathways. For each human disease, the top 5 mouse phenotypes predicted to be functionally similar are selected (62 mouse phenotypes). The intensity of color indicates the extent to which disease-phenotype associations exceed z-score=1.64 (corresponding to p-value=0.05).

While we do not have clear gold standards to evaluate mouse phenotype-human disease predictions, many of the disease-phenotype pairs we find make intuitive sense. Specifically, we find that diseases tend to cluster with ones that involve similar organ systems (e.g., combined immunodeficiency and autoimmune disease, Figure 6). Furthermore, phenotypes related to lymphocytes are associated with both immune diseases and leukemia, while "abnormal neuromuscular synapse morphology" is shared between neuropathy, epilepsy, and amyotrophic lateral sclerosis. Moreover, phenotype associations can also reflect differences between diseases related to the same organ system. Both epilepsy and intellectual disability's top related phenotypes, "impaired conditioning behavior" [40], but seizure phenotypes are specific to epilepsy.

We also believe that ANDES can capture both direct and secondary associations between human diseases and mouse phenotypes. For example, diabetes mellitus is related to the mouse phenotype "small pancreatic islets," capturing the fact that these cells produce insulin (Figure 6). Furthermore, anemia, a disease with many variants that affect red blood cells, is enriched for the mouse phenotype, "anisopoikilocytosis," which is a disorder where red blood cells have irregular sizes and shapes. ANDES can also identify secondary disease phenotypes not directly caused by the disease itself. For example, hypothyroidism is enriched in mouse phenotypes related to the brain and nervous system, which is a phenotype known to be associated

with thyroid disease in humans [41]. Together, these results highlight the exciting potential of ANDES to not only model existing human-mouse phenotype mappings, but also identify new translational opportunities to aid in the development of new model organism screens for specific human disease phenotypes.

3 Discussion

In this study, we introduce ANDES as a general-purpose method for comparing sets in embedding spaces. Applying ANDES to gene embeddings, we show how it can be used to prioritize functionally similar gene sets within a single organism, across organisms with more sophisticated joint embeddings, and how an extension of the framework can be used as novel gene set enrichment methods. Unlike current embedding set similarity methods that rely on averaging, ANDES identifies the best matches between individual elements in a set, thus considering the diversity within a set to better capture inherent substructure that may be otherwise lost.

Our novel ANDES framework has a myriad of downstream applications, especially when paired with different embeddings. Here, we simply scratch the surface by showing the potential of ANDES for function prediction and drug repurposing when paired with a human gene embedding space, and cross-organism translation tasks when paired with a joint gene embedding. We anticipate more interesting use cases with different gene embeddings or even embeddings of an entirely different modality. Furthermore, while we have analyzed several methods for generating PPI-based gene embeddings, integrating more gene information beyond PPIs may yield further improved gene set matching.

Beyond enrichment and set matching tasks, ANDES can also potentially be used to evaluate the quality of different embeddings when some set relationships are known a priori. In the gene embedding case, embedding spaces constructed in different ways might highlight different gene attributes. Analyzing known similarities (e.g., pathways, gene function, etc.) and how gene set matching changes might yield more unique insights into the information encoded in the latent space.

One limitation of our best-match approach is that it can sometimes be sensitive to outliers. We try to reduce this effect by taking the weighted average between best-matching pairs and estimating statistical significance using cardinality-aware null distributions. Computation speed is another reason why we have taken a best-match approach here. Co-clustering approaches may yield more accuracy, but the number of pairwise computations while performing gene set analysis quickly becomes intractable.

In conclusion, here we have presented a novel algorithm, ANDES, for set comparisons in embedding spaces. We hope that our best-match framework paired with various embeddings will be widely adopted and further adapted for additional novel use cases. Our implementation of ANDES and code for the analysis described herein is available on GitHub at `https://github.com/ylaboratory/ANDES`.

4 Methods

As depicted in Figure 1A, a single gene set can comprise a mixture of different biological processes scattered throughout the embedding space. ANDES considers similarity while reconciling gene set diversity when measuring the relationship between two gene sets. Specifically, for each gene in a gene set, the method focuses on the "best-matching" gene (i.e., closest gene) in the other set, allowing ANDES to quantify the presence of functionally similar genes between the sets. The mean of these best-match scores represents the similarity between two gene sets. Finally, to enable systematic comparisons across different gene set pairs, background correction with an estimated null distribution is applied to standardize the scores.

4.1 Calculation of the ANDES gene set similarity score

Given a low dimensional gene representation $E \in \mathbb{R}^{n \times d}$, where n is the number of genes and d is the embedding dimension, ANDES computes a pairwise similarity score between every pair of genes in the

embedding. Here, we use cosine similarity, as it measures the angle between two vectors rather than only the distance and is scale-invariant, which means that it is not affected by the magnitude of the initial vector. Formally, $S = cos(E, E) \in \mathbb{R}^{n \times n}$, where each entry $S_{i,j}$ of the matrix represents the similarity between two genes i and j. For two gene sets $X = x_1, x_2, \ldots, x_m$ and $Y = y_1, y_2, \ldots, y_k$, where $|X| = m$ and $|Y| = k$, we define a similarity matrix for X and Y as $A \in \mathbb{R}^{m \times k}$, the sub-matrix of S with the corresponding entries matching genes from X and Y as rows and columns, respectively. The gene set similarity score (GS) is defined as:

$$GS = \frac{\sum_{i=1}^{m} \max_{1 \leq j \leq k} A_{ij} + \sum_{j=1}^{k} \max_{1 \leq i \leq m} A_{ij}}{m + k}. \tag{1}$$

ANDES thus finds the best match for every gene in set X from set Y and vice versa. A large GS means most of the genes from X and Y can find similar genes to each other in the embedding space and are more likely to involve similar processes.

4.2 Estimation of null distribution and statistical significance

ANDES uses an asymptotic approximation of Monte Carlo sampling to calculate a statistical significance score for every pair of gene sets. This procedure facilitates the comparison between different gene set pairs with varying numbers of genes (cardinalities). Since ANDES uses the max operator, this step is particularly crucial. In contrast to the mean operation, which has the same expected value for different random samples drawn from a Gaussian distribution, the expected maximum value will increase as set cardinality grows. Therefore, an appropriate cardinality-aware null distribution is essential for ANDES to eliminate bias resulting from varied gene set sizes.

The null distribution of the ANDES score between a pair of gene sets is approximated by 1,000 Monte Carlo samples, where each of the 2 sets has the same cardinality as the original pair. A restricted background gene list helps prevent the statistical significance of gene set similarities from becoming artificially inflated. While in most cases, the background can be all genes in the embedding, E, for systematic comparisons with a target annotation database (e.g., Gene Ontology (GO)), we use a more conservative background gene list that includes only genes that are in both the embedding and the target annotation database (e.g., genes with at least one GO annotation). To balance power and computational cost, we use a normal asymptotic approximation to estimate a z-score: $z_{(GS)} = \frac{GS - \mu_0}{\sigma_0}$, where μ_0 and σ_0 are the mean and standard deviation of the Monte Carlo approximations of the null distribution.

Since GS and $z_{(GS)}$ can be used directly to quantify embedding-aware gene set similarities, they can be applied directly to gene set enrichment via overrepresentation analyses. In comparison with the standard Fisher's exact test that is typically used for such comparisons, the immediate advantage of ANDES is that the overrepresentation analyses can identify significantly "related" gene sets even in the scenario where two gene sets of interest have completely no overlap.

4.3 ANDES as a rank-based gene set enrichment method

In addition to overrepresentation analyses, we apply the best-match concept to develop a novel rank-based gene set enrichment method that considers distances between sets in gene embedding spaces.

Given a ranked gene list $L = g_1, g_2, \cdots, g_l$ and a gene set of interest $Y = y_1, y_2, \ldots, y_k$, we define the similarity matrix for L and Y as $A \in \mathbb{R}^{l \times k}$, a sub-matrix of similarity matrix S with the corresponding entries matching genes from L and Y as rows and columns, respectively.

The best-match gene set enrichment score is:

$$ES = \underset{1 \leq i \leq l}{\text{maxdev}} \left(\sum_{1 \leq t \leq i} \max_{1 \leq j \leq k} A_{tj} - \frac{1}{l} \sum_{r=1}^{l} \max_{1 \leq j \leq k} A_{rj} \right), \tag{2}$$

where maxdev is the maximum deviation from 0 and $\frac{1}{l}\sum_{r=1}^{l}\max_{1\leq j\leq k} A_{rj}$ is the mean of all best-match scores from L to Y. At each position i in L, ANDES is thus calculating the cumulative sum of mean-corrected best-match scores from genes $g_1, g_2, \ldots, g_i$ to genes in Y. As such, ES reflects the extent to which genes in Y are close (in embedding space) near the extremes (top or bottom) of the ranked gene list L. In this way, there are some similarities to the running-sum calculation used in the GSEA method [27], which updates an enrichment score based on the fraction of gene "hits" and "misses" as L is traversed. Similar to the limitation with Fisher's exact test, GSEA only considers genes in L that have direct annotations in Y. The best-match approach that ANDES takes is able to consider the extent to which each gene in L is close to genes in Y, even if it does not have a direct annotation.

To assess the significance of ES for a given gene set Y, ANDES uses an approach similar to that described above (Section 4.2) to calculate a normalized enrichment score (NES) through Monte Carlo sampling and asymptotic approximation, ensuring that the random gene sets have the same cardinality as Y. Systematic comparisons against a target annotation database, such as GO, also use the more conservative background gene list (e.g., genes with at least one GO annotation).

4.4 Gene embeddings using a protein-protein interaction network

While our proposed framework is agnostic to embedding method and data type, here we focus on gene embeddings generated from protein-protein interactions (PPIs). We use our previously assembled consensus PPI network [42], which considers physical interaction information from 8 different data sources, resulting in an unweighted and undirected network of 20,363 genes (vertices) and 822,311 interactions (edges). We compare three different embedding approaches: node2vec [43], NetMF [44], and a structure-preserving autoencoder method based on the architecture in [10], which we abbreviate as the neural network (NN) approach. Each method takes a different approach to embed the gene relationships characterized by the PPI network. Node2vec employs random walks on the graph, followed by the skip-gram model to embed node relationships. NetMF generates latent representations by solving a closed-form matrix factorization problem. Lastly, NN uses an autoencoder backbone with the following objective function:

$$L = \text{BCE}(\text{sigmoid}(\hat{M}), M) + \text{BCE}(\cos(Z, Z), A) \odot A + \lambda_1 L_2 + \lambda_2 L_{\text{norm}}, \tag{3}$$

where A is adjacency matrix, Z is gene embedding matrix, M is the NetMF matrix, and BCE is binary cross entropy. The L_2 norm on the autoencoder parameters is included to avoid overfitting and $L_{\text{norm}} = \sum_{i=1}^{n} \|z_i\|_2^2$ to avoid exploding norms. By optimizing this objective function, the neural network preserves the global and local structural information simultaneously. We fixed the embedding dimension sizes from all 3 methods to 128.

4.5 Gene set processing

For benchmarking, we leverage curated gene sets describing pathways, function, tissues, diseases, phenotypes, and drugs. We describe the gene sets and processing in the sections below.

Gene Ontology (GO): To assess gene function, we use the biological process (BP) annotations from GO [45] (16 July 2020). To ensure high-quality annotations, we only keep terms with low-throughput experimental evidence codes (EXP, IDA, IMP, IGI, IEP). Furthermore, to avoid any circularity with the underlying PPIs used to construct embeddings, we exclude terms with evidence code IPI (Inferred from Physical Interaction). We further restrict the total number of GO terms using an expert-curated set of slim terms designed to emphasize key biological processes [22]. Leveraging the directed acyclic graph structure of the ontology, we propagate annotations for the "is a" and "part of" relationships, such that parent terms also contain genes that participate in more specific (child term) processes. After propagation, we apply a final filter to preserve terms with more than 10 and fewer than 300 annotated genes.

Kyoto Encyclopedia of Genes and Genomes (KEGG): We obtain pathway gene sets from KEGG [46] using ConsensusPathDB [47]. In total, we obtain 333 unique human pathway gene sets.

Online Mendelian Inheritance in Man (OMIM): We collect disease gene sets from OMIM (October 2023) [32]. These gene sets are then mapped to Disease Ontology [48] and propagated through the ontology structure, resulting in 284 unique disease gene sets.

Mouse Genome Informatics Phenotypes (MGI): We assemble mouse phenotype gene sets from MGI (March 2022) [39]. After propagating genes from children to parents, we obtain 3,738 mouse phenotype gene sets.

DrugBank: Drug target information for 725 drugs, as well as other descriptions, such as ATC codes, were parsed from the academic licensed version of DrugBank [33] (Jan 2023).

4.6 Benchmarking gene set similarity metrics in embedding spaces

The most straightforward to compare sets in embedding space is to summarize their similarity through averaging. We compare ANDES with two variants of averaging (mean score and mean embedding) as a benchmark along with a corrected t-score approach [22].

Given two gene set embeddings $X \in \mathbb{R}^{m \times d}$ and $Y \in \mathbb{R}^{k \times d}$, where m and k are the number of genes in the set and d is the embedding dimension, the mean score method first calculates the pairwise cosine similarity S where $S_{ij} = \cos(\vec{x}_i, \vec{y}_j)$, where $\vec{x}_i$ is the i-th row vector of X and $\vec{y}_j$ is the j-th row vector of Y. The mean score method then simply computes the mean of these similarities, $\frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} S_{ij}$.

The mean embedding approach instead first takes the average within a gene set, resulting in two gene set-level pooled embeddings $\vec{p}$ and $\vec{q}$, where

$$\vec{p} = \frac{1}{d} \begin{bmatrix} \sum_{j=1}^{d} X_{1j} \\ \sum_{j=1}^{d} X_{2j} \\ \dots \\ \sum_{j=1}^{d} X_{mj} \end{bmatrix} \qquad \vec{q} = \frac{1}{d} \begin{bmatrix} \sum_{j=1}^{d} Y_{1j} \\ \sum_{j=1}^{d} Y_{2j} \\ \dots \\ \sum_{j=1}^{d} Y_{kj} \end{bmatrix}$$

The final mean embedding score is the cosine similarity between the gene set-level pooled embeddings, $\cos(\vec{p}, \vec{q})$.

The corrected t-score method [22] calculates an unequal variance t-test on two score distributions: the pairwise similarity score between two gene sets (between) and the scores associated with cardinality-matched gene sets across the genome (background). The final score is determined by comparing the between scores against a null distribution of the background scores.

We explore two gene set matching evaluations: (1) matching paired functional annotation datasets (KEGG and GO) with each other, and (2) matching GO gene sets across different model organisms. For the KEGG and GO comparisons, annotations for 50 KEGG pathways to corresponding GO biological processes are obtained based on the external database annotations from the KEGG web portal. To assess the ability to capture functional similarity beyond overlapping genes, we remove all overlapping genes between the KEGG- and GO-matched gene sets from GO gene sets when evaluating each method.

For the cross-species evaluation, we look for exact matches between the same GO slim term in humans versus three model organisms: *M. musculus*, *S. cerevisiae*, *D. melanogaster*. While cross-species GO annotations can capture conserved biological processes, we need updated embedding spaces that jointly model genes from both species. To that end, we use our previously developed method, ETNA [10], to construct pairwise gene embedding spaces for human and each of the three model organisms. ETNA uses an autoencoder approach to generate within-species network embeddings based on PPI networks, then uses a cross-training approach with known orthologous genes as anchors to align the two embeddings into a joint embedding. This joint embedding enables cross-species comparisons of all genes represented in each PPI network.

4.7 Evaluating gene set enrichment methods

To evaluate ANDES' ability to identify functionally relevant pathways in an enrichment-analysis setting, we use a gold standard compendium of pathway-annotated gene expression data, GEO2KEGG, that has been routinely used for benchmarking gene set enrichment analyses [29, 30, 49]. GEO2KEGG consists of 42 human microarray profiles matched to various diseases, each of which has a set of curated KEGG pathway annotations. Using the corresponding annotated KEGG pathways as a gold standard, we can then calculate AUPRC for each of the 42 datasets. We compare the results of ANDES against three existing gene set analysis methods: hypergeometric test [26], GSEA [27], and GSPA [29]. We are unable to use NGSEA [28] and EnrichNet [50] for this benchmarking analysis because only web portals are available, with no ability to change the underlying gene sets for a fair comparison across methods. We note that GSPA does report that they outperform both NGSEA and EnrichNet. The comparison with hypergeometric test uses ANDES' gene set similarity score and statistical significance calculation (Sections 4.1 and 4.2), taking genes with FDR$\leq$0.05. Comparisons with rank-based gene set enrichment methods use ANDES' best-match-based enrichment method (Section 4.3), where the input gene list is ranked using $\log_2$(fold change).

References

1. Church, K. W. Word2Vec. *Natural Language Engineering* **23,** 155–162 (2017).
2. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
3. Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I. & Lempitsky, V. *Hyperbolic image embeddings* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 6418–6428.
4. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
5. Zhang, F., Yuan, N. J., Lian, D., Xie, X. & Ma, W.-Y. *Collaborative knowledge base embedding for recommender systems* in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), 353–362.
6. Stanojevic, S., Li, Y., Ristivojevic, A. & Garmire, L. X. Computational methods for single-cell multi-omics integration and alignment. *Genomics, Proteomics & Bioinformatics* **20,** 836–849 (2022).
7. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative methods and practical challenges for single-cell multi-omics. *Trends in biotechnology* **38,** 1007–1022 (2020).
8. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature,* 1–9 (2023).
9. Du, J. *et al.* Gene2vec: distributed representation of genes based on co-expression. *BMC genomics* **20,** 7–15 (2019).
10. Li, L. *et al.* Joint embedding of biological networks for cross-species functional alignment. *Bioinformatics* **39,** btad529 (2023).
11. Kulmanov, M., Khan, M. A. & Hoehndorf, R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34,** 660–668 (2018).
12. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36,** 422–429 (2020).
13. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nature communications* **12,** 3168 (2021).
14. Xiong, Y. *et al.* Heterogeneous network embedding enabling accurate disease association predictions. *BMC medical genomics* **12,** 1–17 (2019).
15. 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 (2021).
16. Mostavi, M., Chiu, Y.-C., Huang, Y. & Chen, Y. Convolutional neural network models for cancer type prediction based on gene expression. *BMC medical genomics* **13,** 1–13 (2020).
17. Kim, S., Lee, H., Kim, K. & Kang, J. Mut2Vec: distributed representation of cancerous mutations. *BMC medical genomics* **11,** 57–69 (2018).
18. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature communications* **13,** 1265 (2022).
19. Gao, K. Y. *et al. Interpretable drug target prediction using deep neural representation.* in *IJCAI* **2018** (2018), 3371–3377.
20. Wieting, J., Bansal, M., Gimpel, K. & Livescu, K. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198* (2015).
21. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (2023).
22. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* **47,** 569–576 (2015).
23. Wang, S., Flynn, E. R. & Altman, R. B. Gaussian embedding for large-scale gene set analysis. *Nature machine intelligence* **2,** 387–395 (2020).

24. Cavallari, S., Zheng, V. W., Cai, H., Chang, K. C.-C. & Cambria, E. *Learning community embedding with community detection and node embedding on graphs* in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017), 377–386.
25. Azuaje, F., Wang, H. & Bodenreider, O. *Ontology-driven similarity approaches to supporting gene functional assessment* in *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies* **2005** (2005), 9–10.
26. Hahne, F. *et al.* Hypergeometric testing used for gene set enrichment analysis. *Bioconductor case studies,* 207–220 (2008).
27. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550 (2005).
28. Han, H., Lee, S. & Lee, I. NGSEA: network-based gene set enrichment analysis for interpreting gene expression phenotypes with functional gene sets. *Molecules and cells* **42,** 579 (2019).
29. Cousins, H. *et al.* Gene set proximity analysis: expanding gene set enrichment analysis through learned geometric embeddings, with drug-repurposing applications in COVID-19. *Bioinformatics* **39,** btac735 (2023).
30. Tarca, A. L., Bhatti, G. & Romero, R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS one* **8,** e79217 (2013).
31. Kim, S.-Y. & Volsky, D. J. PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics* **6,** 1–12 (2005).
32. 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).
33. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* **46,** D1074–D1082 (2018).
34. Jørgensen, E. A., Knigge, U., Warberg, J. & Kjær, A. Histamine and the regulation of body weight. *Neuroendocrinology* **86,** 210–214 (2007).
35. Perl, A. E. *et al.* Follow-up of patients with R/R FLT3-mutation–positive AML treated with gilteritinib in the phase 3 ADMIRAL trial. *Blood, The Journal of the American Society of Hematology* **139,** 3366–3375 (2022).
36. Yu, X. *et al.* Fingolimod ameliorates schizophrenia-like cognitive impairments induced by phencyclidine in male rats. *British Journal of Pharmacology* **180,** 161–173 (2023).
37. Suri, R. *et al.* Sirolimus loaded chitosan functionalized poly (lactic-co-glycolic acid)(PLGA) nanoparticles for potential treatment of age-related macular degeneration. *International journal of biological macromolecules* **191,** 548–559 (2021).
38. Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research* **11,** 95–130 (1999).
39. Eppig, J. T. *et al.* Mouse Genome Informatics (MGI): resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. *Systems Genetics: Methods and Protocols,* 47–73 (2017).
40. Holley, A. J. & Lugo, J. N. Effects of an acute seizure on associative learning and memory. *Epilepsy & Behavior* **54,** 51–57 (2016).
41. Khaleghzadeh-Ahangar, H., Talebi, A. & Mohseni-Moghaddam, P. Thyroid disorders and development of cognitive impairment: A review study. *Neuroendocrinology* **112,** 835–844 (2022).
42. Dannenfelser, R. & Yao, V. Splitpea: quantifying protein interaction network rewiring changes due to alternative splicing in cancer. *bioRxiv,* 2023–09 (2023).
43. 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.

44. Qiu, J. *et al. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec* in *Proceedings of the eleventh ACM international conference on web search and data mining* (2018), 459–467.
45. Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32,** D258–D261 (2004).
46. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28,** 27–30 (2000).
47. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic acids research* **41,** D793–D800 (2013).
48. Schriml, L. M. *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research* **40,** D940–D946 (2012).
49. Tarca, A. L., Draghici, S., Bhatti, G. & Romero, R. Down-weighting overlapping genes improves gene set analysis. *BMC bioinformatics* **13,** 1–14 (2012).
50. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* **28,** i451–i457 (2012).