A best-match approach for gene set analyses in embedding spaces

Lechuan Li, Ruth Dannenfelser, Charlie Cruz, and Vicky Yao

Department of Computer Science, Rice University, Houston, Texas 77005, USA

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. However, this utility has been primarily realized by using gene embeddings for downstream machine-learning tasks. Much less has been done to examine the embeddings directly, especially analyses of gene sets in embedding spaces. Here, we propose an Algorithm for Network Data Embedding and Similarity (ANDES), a novel best-match approach that can be used with existing gene embeddings to compare gene sets 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 ANDES, when applied to different gene embeddings encoding protein–protein interactions, can be used as a novel overrepresentation- and rank-based gene set enrichment analysis method that achieves state-of-the-art performance. Additionally, ANDES can use multiorganism joint gene embeddings to 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 with diverse community structures between set elements.

[Supplemental material is available for this article.]

Methods to build embedding representations have become ubiquitous in diverse fields spanning text-based (Church 2017; Devlin et al. 2019), image-based (Dosovitskiy et al. 2020; Khrulkov et al. 2020), and domain-specific (Zhang et al. 2016; Du et al. 2019; Ma et al. 2020; Stanojevic et al. 2022; Li et al. 2023; Theodoris et al. 2023) applications. In addition to the computational benefits that lower-dimension embedding representations provide, there is an implicit assumption that a quality 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 (Kulmanov et al. 2018; Kulmanov and Hoehndorf 2020; Gligorijević et al. 2021), disease associations (Xiong et al. 2019; Yu et al. 2021), expanding gene set representations (Chen et al. 2018; Wang et al. 2020), among other applications (Gao et al. 2018; Kim et al. 2018; Mostavi et al. 2020; Bryant et al. 2022).

Given the utility of gene embeddings for downstream machine-learning tasks, it is intuitive that gene embeddings capture important gene–gene relationship information (Fig. 1A). However, little attention is paid to exploring the resulting embedding spaces, especially for the analyses of gene sets. In standard genomics analyses, gene sets (e.g., a set of differentially expressed genes, reported genes from genome-wide association studies (GWAS), or even a group of genes annotated to a particular pathway) are often a fundamental "functional unit." Comparisons between sets are very routine, including gene set enrichment analysis (GSEA) (Subramanian et al. 2005; Hahne et al. 2008), disease-gene associations (Wang et al. 2011; Yao et al. 2018), and drug repurposing (Peyvandipour et al. 2018; Reay and Cairns 2021). Yet, there appears to be limited to no research considering gene set comparisons in the context of embedding spaces.

Here, we present an Algorithm for Network Data Embedding and Similarity (ANDES) analysis (Fig. 1). The goal of ANDES is to calculate an interpretable measure of gene set similarity that accounts for the presence of functional diversity. Toward this end, ANDES identifies best-matching (most similar) genes between two sets reciprocally and calculates a score based on the embedding distances between these best-match similarities (Fig. 1B). This best-match concept has parallels to an early method proposed for biological text mining (Azuaje et al. 2005), but functional analyses in embedding spaces require adjustments for biases due to gene set cardinalities. ANDES thus further 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.

Outside of biological use cases, previous methods to summarize set relationships in embedding spaces typically consider variations of averaging embedding information across set members, ultimately ignoring the potential diversity within the set. One such averaging approach simply uses the centroid of all considered entities in the embedding space (Wieting et al. 2015; Lin et al. 2023). 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 subfunctional groups in the gene set and instead obscure this signal, especially when the subgroups are distinct. Though network-based gene set comparison methods have typically not been applied to embedding spaces, they can sometimes transfer to this domain. One such network-based method formulates the gene set comparison problem as a *t*-test between the two gene

Corresponding author: vy@rice.edu

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279141.124. Freely available online through the *Genome Research* Open Access option.

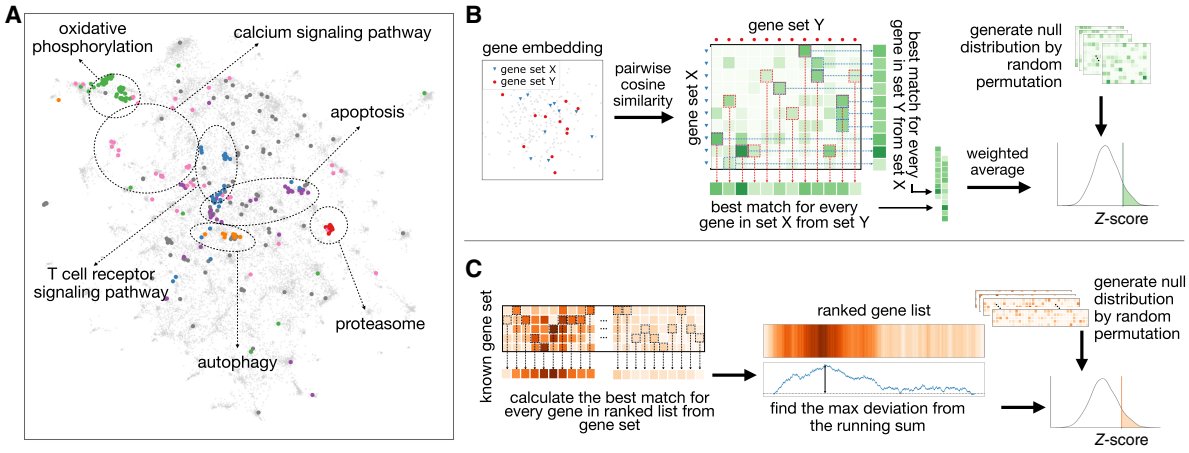

Figure 1. Overview of ANDES. (*A*) A Uniform Manifold Approximation and Projection (UMAP) plot (McInnes et al. 2018) of the node2vec gene embedding for a human protein–protein interaction network with a set of Alzheimer's disease genes (hsa05010) highlighted. Within this set of disease genes, several subclustered biological processes representing diverse biological functions are scattered across the embedding space. ANDES is capable of considering this functional 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 (in both directions), and then calculates the weighted average (taking into account gene set size) to yield a single score. Statistical significance is estimated using a cardinality-aware null distribution. (*C*) Overview of the 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 maximum deviation from the running sum. The final enrichment score is also estimated using a cardinality-aware null distribution.

sets, with a permutation-based background correction for the size of each gene set (Greene et al. 2015). However, although 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.

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 (Hahne et al. 2008) and ranked-based approaches (Kim and Volsky 2005; Subramanian et al. 2005), of which hypergeometric test and GSEA are, respectively, representative methods (Subramanian et al. 2005). One of the fundamental limitations of both categories of methods is a complete reliance on gene annotations (e.g., functional annotations in the Gene Ontology [GO]); if a gene has no annotations, it cannot contribute to the enrichment analysis. Considering genes in functionally meaningful embedding spaces is one way to circumvent this limitation. ANDES can be used directly as an overrepresentation-based approach, and we also extend ANDES' best-match approach for rank-based gene set enrichment (Fig. 1C).

Other recent methods that attempt to lessen the dependency of gene set enrichment on existing annotations include Network-based Gene Set Enrichment Analysis (NGSEA), which reranks the input gene list by incorporating the mean of its network neighbors (Han et al. 2019), and Gene Set Proximity Analysis (GSPA), which allows users to supplement gene set annotations using a radius in an embedding space (Cousins et al. 2023). Unfortunately, we are unable to systematically evaluate NGSEA because it is only available as a web portal, and we cannot change the underlying gene sets used.

Through a series of evaluations, we demonstrate that ANDES can better estimate gene set functional similarity in gene embedding spaces compared to existing average-based methods. Furthermore, ANDES outperforms previous methods for gene set enrichment, and we also showcase its ability to prioritize new candidates for drug repurposing. Overall, we find that leveraging the intuition of best-match comparisons is an effective, generalizable approach that has important implications for interpretable analyses of embedding spaces.

Results

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

We first explore the extent to which ANDES and other set comparison methods can recover "functionally matched" gene sets in embedding spaces. "Matched" gene sets describe the same biological processes across different databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and GO biological processes. When examining these gene sets in an embedding space that captures gene relationships, we expect a good set comparison method to identify these matches. We note that the same process described in KEGG and GO can naturally have overlapping genes, which would alter this set matching problem into an easier one primarily driven by set overlap. To prevent this, we only keep overlapping genes in the KEGG gene sets and remove them from GO gene sets. We compare ANDES against the mean embedding, mean score, and corrected *t*-score methods. 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 (Greene et al. 2015). 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 (Grover and Leskovec 2016), NetMF (Qiu et al. 2018), and a structure-preserving autoencoder method based on the architecture in Li et al. (2023). ANDES consistently outperforms other methods and successfully identifies gene sets with similar

functional roles, regardless of the underlying gene embedding method (Fig. 2A). More specifically, ANDES significantly outperforms the mean score (node2vec: $P = 1.84 \times 10^{-6}$, NetMF: $P = 2.75 \times 10^{-3}$, neural network [NN]: $P = 3.14 \times 10^{-5}$, Wilcoxon signed-rank test) and corrected *t*-score method (node2vec: $P = 7.56 \times 10^{-6}$, NetMF: $P = 2.06 \times 10^{-4}$, NN: $P = 2.23 \times 10^{-7}$, Wilcoxon signed-rank test). ANDES also significantly outperforms the mean embedding method for the node2vec and autoencoder embedding methods (node2vec: $P = 0.039$, NetMF: $P = 0.117$, NN: $P = 0.015$, Wilcoxon signed-rank test), and in general, ANDES is more stable. We note that if we do not remove the overlapping genes between matching KEGG and GO terms, ANDES shows an even larger performance advantage compared to existing methods (Supplemental Fig. S1).

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 with a matched pair of KEGG and GO terms (KEGG: hsa00071-fatty acid degradation, GO: GO:0009062-fatty acid catabolic process) (Fig. 2B). A direct inspection of the distribution of each gene set's genes in the embedding space (Fig. 2B) quickly reveals that the correct KEGG–GO match has distinctly more similar embeddings, which ANDES can correctly identify and mean embedding cannot. Both the mean score and corrected *t*-score methods also rank the correct term higher than the mean embedding method.

Because ANDES' best-match framework is not limited to using only embeddings as input, we also examine the extent to which matching gene sets can be identified using naive nonembedding network-only approaches, such as shared neighbor profiles, graph diffusion, and node degrees, using the original PPI network. Because these approaches do not use embeddings, it is impossible to calculate mean embeddings for the following comparisons. Using the Jaccard index of shared neighbors between two genes also captures sufficient functional signal to identify several correct KEGG–GO matches. In this setup, ANDES still significantly outperforms the mean score ($P = 1.22 \times 10^{-5}$, Wilcoxon signed-rank test) and corrected *t*-score ($P = 2.19 \times 10^{-4}$, Wilcoxon signed-rank test) methods (Fig. 2C). We observe similar performance trends using a heat diffusion kernel on the PPI network, though the difference between ANDES and mean score is not significant (mean score: $P = 0.158$, corrected *t*-score: $P = 4.12 \times 10^{-5}$) (Supplemental Fig. S2). Although both shared neighbor profiles and heat diffusion clearly capture functional signal, we note that using embedding approaches such as node2vec as input still leads to better performance (Jaccard: $P = 6.52 \times 10^{-3}$, heat diffusion: $P = 7.82 \times 10^{-3}$, Wilcoxon signed-rank test). Using a more naive exponential diffusion kernel and the simple sum of node degrees in the PPI network to measure gene similarity results in nearly random performance for all three methods explored in this comparison. In general, we observe that although ANDES can be successfully applied to these nonembedding approaches, the sparsity in the resultant PPI similarity matrix may be a limiting factor. Unlike gene similarity matrices based on embedding spaces, gene pairs with weak relationships have scores of zero using nonembedding approaches, which decreases the stability of the results. Altogether, this set of analyses demonstrates that the application of ANDES is not limited to only embedding spaces, but ANDES' performance improves with the informativeness of the similarity matrix that is used as input.

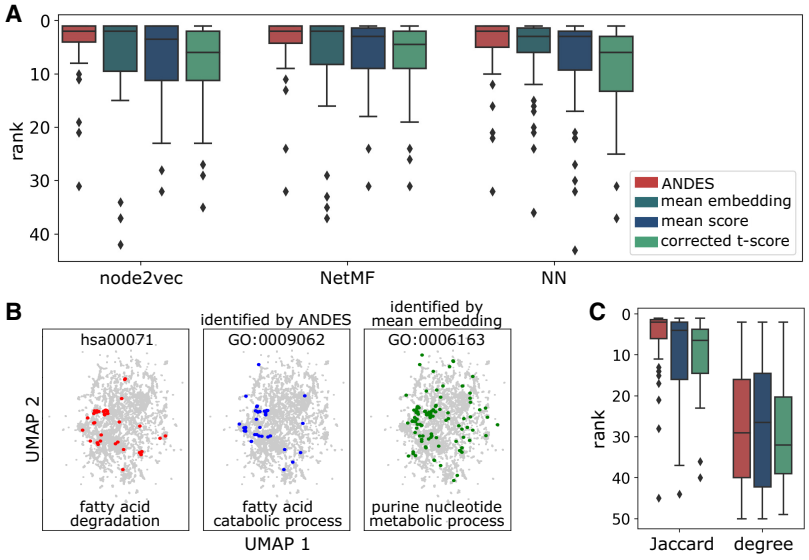

Figure 2. ANDES better matches gene sets that describe the same biological processes regardless of the underlying embedding or network structure. (*A*) Boxplots of the ranking of the correct matching GO term for 50 KEGG terms demonstrate that ANDES outperforms the mean embedding, corrected *t*-score, and mean score methods across three network embedding approaches (node2vec, NetMF, 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 five 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). Because these pairwise similarity matrices are directly calculated from network properties without using embeddings, we cannot calculate the mean embedding method and instead compare ANDES to only the corrected *t*-score and mean score.

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

Because ANDES is a general framework to compare sets, comparing a single gene set of interest against a collection of annotated gene sets (e.g., KEGG) is directly comparable to existing overrepresentation-based approaches. We also further extend ANDES to handle comparisons of a ranked list of genes to gene sets, which allows ANDES to be used as a rank-based gene set enrichment approach (Fig. 1C).

Here, our evaluations use a well-established gene set enrichment benchmark (GEO2KEGG [Tarca et al. 2013]). GEO2KEGG annotates differential

expression results (including FDR and log-fold-change per gene) to related pathways for 42 microarray studies, covering over 200 KEGG pathways. The rich data in GEO2KEGG enables us to test the recovery of corresponding KEGG pathways for each data set through overrepresentation-based enrichment (using significance or fold change cutoffs) as well as rank-based gene set enrichment (by ranking all genes based on their fold changes). For the overrepresentation case, we compare ANDES' performance with the hypergeometric test (Hahne et al. 2008), considering gene sets of interest as differentially expressed genes per data set (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% data sets), ANDES set enrichment outperforms the hypergeometric test for KEGG pathway identification ($P = 6.78 \times 10^{-4}$, Wilcoxon signed-rank test) (Fig. 3A). For the rank-based version, we compare ANDES with GSEA (Subramanian et al. 2005) and the embedding-based gene set enrichment method, GSPA (Cousins et al. 2023). Aggregating performance across all 42 data sets in the benchmark, ANDES' rank-based gene set enrichment significantly improves over both GSEA ($P = 0.041$; Wilcoxon signed-rank test) and GSPA ($P = 0.028$; Wilcoxon signed-rank test) (Fig. 3B). In general, we observe that ANDES consistently performs better than other methods, but especially when the original data set has more samples (Supplemental Fig. S3). We thus also wanted to more carefully perform the enrichment analysis while ameliorating potential biases due to differing differential expression signals in the original GEO2KEGG data sets. Thus, for data sets with sufficient samples, we also further compute empirical *P*-values for enrichment scores by comparing against an estimated null distribution of the same data set generated using 100 different condition label permutations. ANDES still outperforms GSEA ($P = 0.060$; Wilcoxon signed-rank test) and GSPA ($P = 0.041$; Wilcoxon signed-rank test) (Fig. 3C, Supplemental Fig. S4), demonstrating ANDES' ability of better prioritizing relevant functions based on true differential signals.

Together, these results highlight ANDES' utility as a novel state-of-the-art method for overrepresentation-based and rank-based gene set enrichment. Furthermore, because ANDES uses gene embeddings, enrichment analyses can be performed in cases where existing annotations have low or even no overlap with the genes of interest, making it particularly valuable for the overrepresentation case.

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 Online Mendelian Inheritance in Man (OMIM) (Hamosh et al. 2005) with drug target genes from DrugBank (Wishart et al. 2018). This is a use case of practical importance, because 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

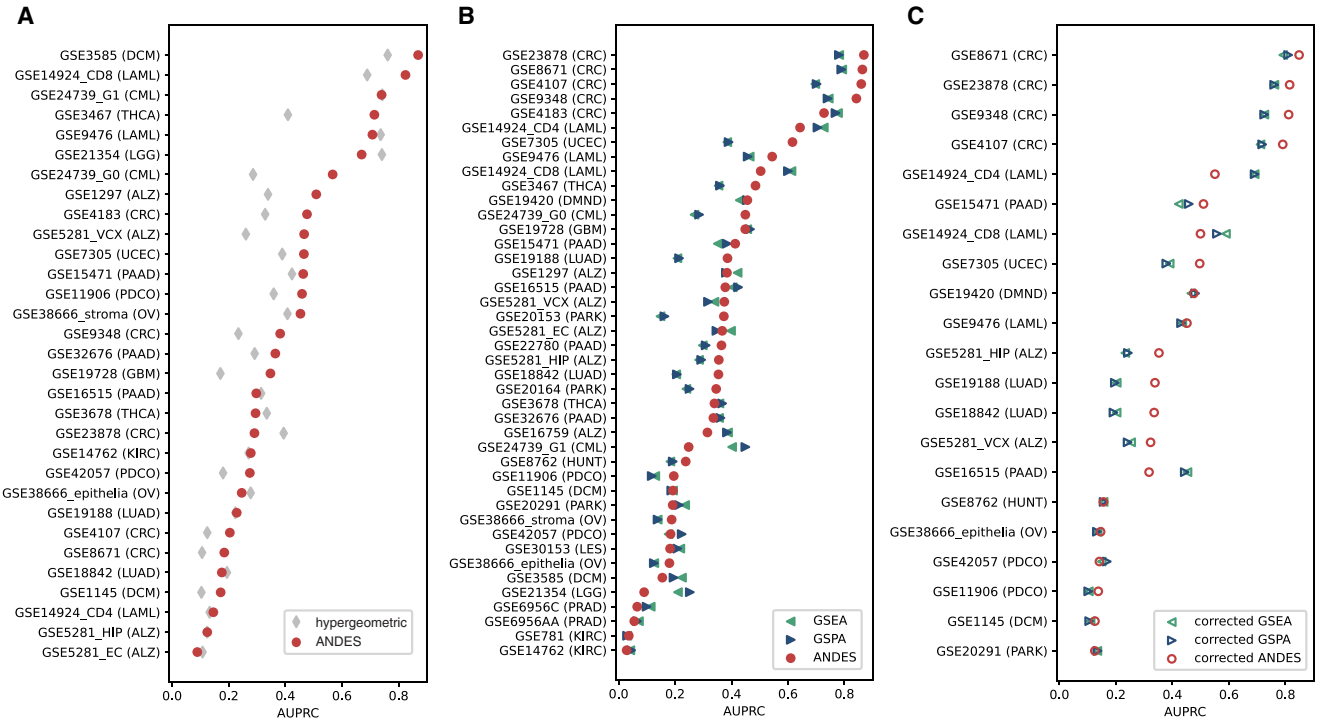

Figure 3. ANDES achieves state-of-the-art performance in overrepresentation-based and rank-based gene set enrichment for the GEO2KEGG (Tarca et al. 2013) gene set enrichment benchmark. (*A*) Performance comparison between ANDES and hypergeometric test in retrieving annotated KEGG terms using genes that have FDR ≤ 0.05 in each data set (where there are at least 10 genes that are significantly differentially expressed). (*B*) Performance comparisons between ANDES, GSEA (Subramanian et al. 2005), and GSPA (Cousins et al. 2023) 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 the best-match principle into the gene set enrichment setting. (*C*) Performance comparisons between ANDES, GSEA, and GSPA with empirically estimated *P*-values in retrieving annotated KEGG terms. Only expression data sets with at least 10 samples in both normal and diseased conditions (21 data sets) are included for sufficient variability in label permutations. Corrected-ANDES still outperforms corrected-GSEA and corrected-GSPA.

development of new therapies. We first calculate a disease similarity profile for drugs using their ANDES scores, then apply dimensionality reduction using principal component analysis (PCA) on these profiles (Fig. 4A; Supplemental Fig. S5). We see that there are several "nervous system" drugs that are more distinct than others, as well as a subgroup that has overlap with other classes, including drugs that are "antineoplastic and immunomodulating agents."

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 (Fig. 4B). Results for the other drug classes can be found in Supplemental Figures S6–S17. Only diseases and drugs with at least one significant match ($z > 1.64$) are kept. Besides the most apparent 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 (Jørgensen et al. 2007), and weight gain is a listed side effect of Gilteritinib (Perl et al. 2022). Although 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 (Yu et al. 2023). We observe a similar association between Sirolimus and macular degeneration. Sirolimus is an immunosuppressive

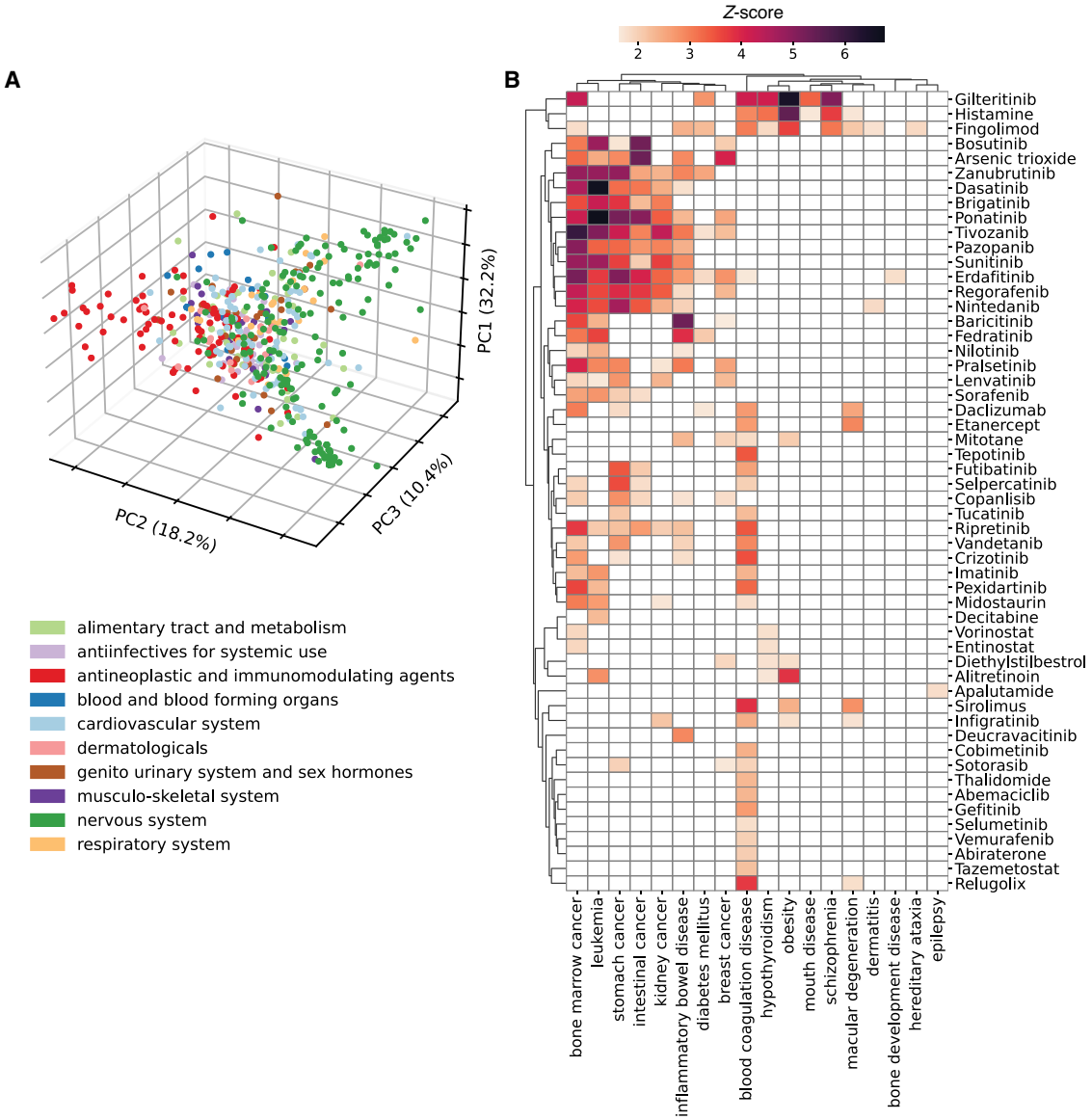

Figure 4. Analysis of drug–disease relationships using ANDES. (*A*) PCA plot of the first three principal components (PCs) showing the relationship between drugs based on their association with diseases. Colors are based on the first level of ATC groups. The first three PCs explain 32.2%, 18.2%, and 10.4% of the total variance, respectively. (*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.

agent used to prevent transplant rejection, but as of 2021, it has been found to have emerging promise as a therapeutic for age-related macular degeneration (Suri et al. 2021). These two seemingly disparate indications likely share an inflammatory pathological pathway, which can be picked up with ANDES.

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 used, 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, Embeddings to Network Alignment (ETNA) (Li et al. 2023). This analysis highlights two general principles: (1) ANDES can still prioritize relevant signals 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.

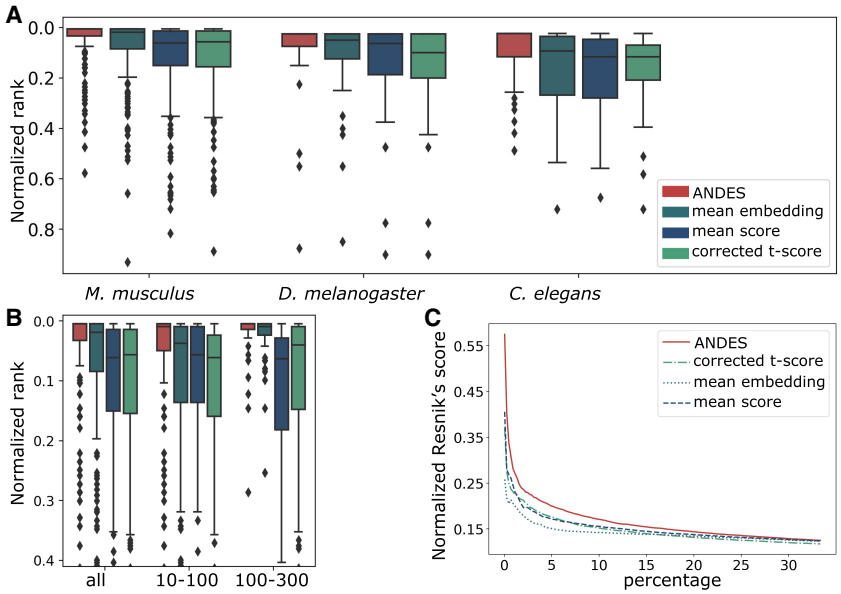

Figure 5. ANDES estimates gene set functional similarity across organisms better than existing methods. (*A*) Boxplot of the ranking of matched GO terms between human and three model organisms: *M. musculus*, *D. melanogaster*, and *C. elegans*, with 213, 40, and 43 shared GO slim terms, respectively. 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 *Homo 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's 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.

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: *Mus musculus*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. ETNA's joint embedding space enables the calculation of a similarity matrix for genes across species. Because 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*: $P = 4.30 \times 10^{-10}$, *D. melanogaster*: $P = 0.147$, *C. elegans*: $P = 2.28 \times 10^{-3}$, Wilcoxon signed-rank test), mean score (*M. musculus*: $P = 1.01 \times 10^{-25}$, *D. melanogaster*: $P = 1.19 \times 10^{-3}$, *C. elegans*: $P = 9.88 \times 10^{-4}$, Wilcoxon signed-rank test), and corrected *t*-score (*M. musculus*: $P = 3.66 \times 10^{-25}$, *D. melanogaster*: $P = 3.62 \times 10^{-3}$, *C. elegans*: $P = 2.27 \times 10^{-4}$, Wilcoxon signed-rank test) (Fig. 5A).

When we group the GO terms shared between human and *M. musculus* by size, ANDES shows better performance, both when the gene set is large (mean embedding: $P = 0.023$, mean score: 5.39×10^{-11}, corrected *t*-score: 1.96×10^{-12}, Wilcoxon signed-rank test) and small (mean embedding: $P = 5.91 \times 10^{-9}$, mean score: 8.18×10^{-17}, corrected *t*-score: 1.55×10^{-14}, Wilcoxon signed-rank test). We notice that larger gene sets are easier to match across organisms (Fig. 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. 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 the Resnik's measure (Resnik 1999), a semantic similarity measure leveraging the GO hierarchy, to quantify the similarity between two gene sets. Two gene sets that are close in the GO are more likely to describe functionally similar biological processes and, therefore, have a higher Resnik's score. Comparing predicted similarities for all gene sets between human and mouse, we calculate the cumulative average Resnik's score of gene set pairs ranked by their similarity score in each set comparison method. Across all methods, the Resnik's score is higher for the top-ranked pairs and gradually converges to randomness at around 30% of the ranked list of all pairs (Fig. 5C). The trend also holds for *D. melanogaster* and *C. elegans* (Supplemental Fig. S18). Overall, ANDES consistently has the highest Resnik's scores of all methods, demonstrating that it both identifies the exact match as well as other functionally related gene sets.

Prioritizing mouse phenotypes for modeling human diseases with ANDES

After verifying that ANDES can recover conserved cross-organism functional signal (Fig. 5), we further explore the potential of ANDES for cross-organism knowledge transfer. Phenotype prioritization is a vital 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. Toward this end, we systematically test associations between human disease gene sets from OMIM (Hamosh et al. 2005) and mouse phenotype gene sets from Mouse Genome Informatics Phenotypes (MGI) (Eppig et al. 2017). We identify a range of significantly related disease-phenotype pairs, many of which merit further exploration (Supplemental Figs. S19–S41). 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 (Fig. 6).

Although 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) (Fig. 6). Furthermore, phenotypes related to lymphocytes are associated with both immune diseases and leukemia, whereas "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" (Holley and Lugo 2016), but seizure phenotypes are specific to epilepsy.

We also find 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 (Fig. 6). Furthermore, anemia, a disease with many variants that affect red blood cells, is enriched for the mouse phenotype, "anisopoikilocytosis," 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 (Khaleghzadeh-Ahangar et al. 2022). 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 developing new model organism screens for specific human disease phenotypes.

Discussion

Here, we introduce ANDES as a general-purpose method for comparing sets by considering best-match elements. In exploring ANDES' application to gene embeddings, we have demonstrated how it can be used to prioritize functionally similar gene sets within a single organism or across organisms using more sophisticated joint embeddings. ANDES can leverage functional information from gene embeddings to avoid a complete reliance on gene

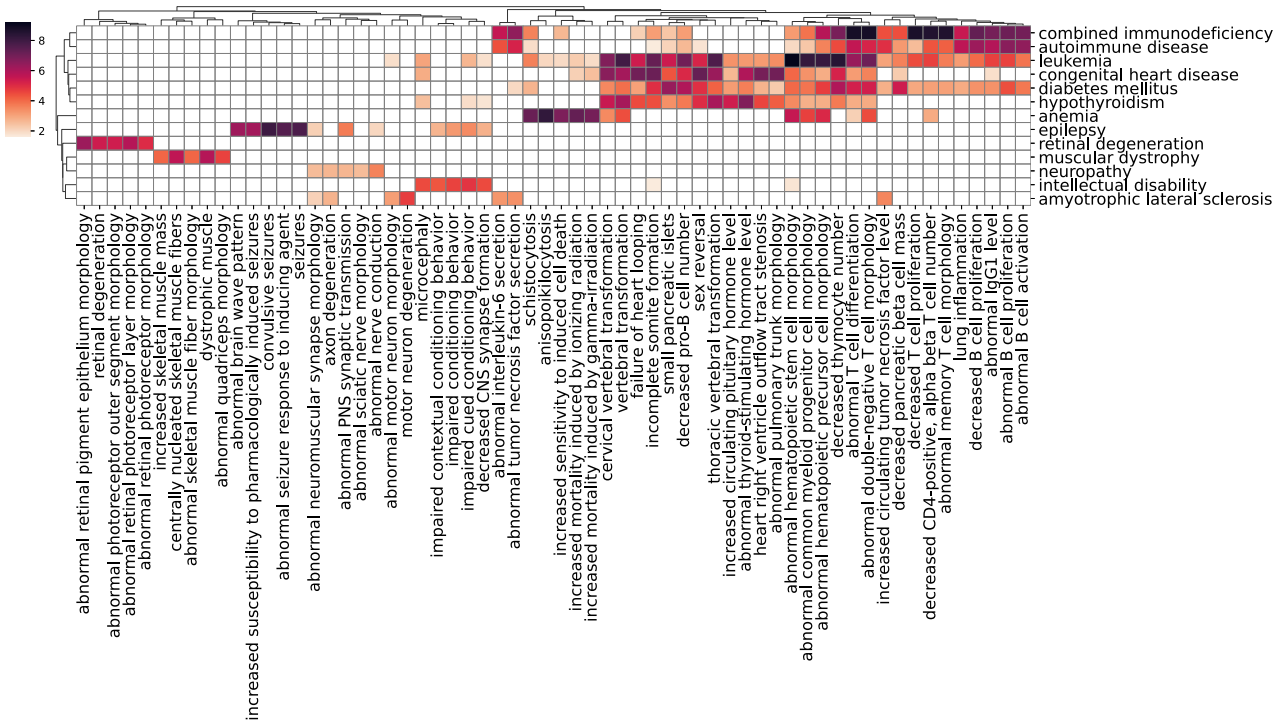

Figure 6. Heatmap of ANDES similarity scores for human disease and mouse phenotype gene sets. Gene set similarity *Z*-scores generated by ANDES are shown for 13 selected human diseases across various organ systems and pathological pathways. For each human disease, the top five mouse phenotypes predicted 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).

annotations in both overrepresentation- and rank-based GSEA, and by doing so, achieves state-of-the-art performance. 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.

One limitation of our best-match approach is that it could potentially be sensitive to outliers. ANDES currently addresses this by combining information from all best-matching pairs and estimating statistical significance using cardinality-aware null distributions. But beyond these strategies, we note that we can also expand the ANDES framework to consider the top k matches per gene instead. The argument for using more matching genes would be to diminish the effect of outliers in driving the assessment of set similarity. This framing would place the best-match and mean score approaches at two ends of the spectrum with respect to choosing k, as $k = 1$ (a single element) is by definition the best-match approach, and $k = 100\%$ (all elements in both gene sets) is equivalent to the mean score approach. A cursory exploration of the effect of varying k on ANDES' performance shows that the ability to identify functionally similar GO and KEGG gene sets using node2vec-embedded PPI networks decreases as k increases, eventually converging to the significantly lower mean score performance (Supplemental Fig. S42). Thus, at least with PPI embeddings, we find that using the best-match approach (i.e., $k = 1$) can avoid the introduction of an additional hyperparameter and performs well in practice. In other situations where outliers are of particular concern, it may be more worthwhile to examine the effect of varying k.

Another key aspect of ANDES is the similarity metric used to determine the best matches. ANDES currently uses cosine similarity, but there are of course several possible alternatives, two natural ones being Euclidean distance and Pearson's correlation. We find that using (inverse) Euclidean distance as the similarity metric results in the largest performance drop (Supplemental Fig. S43), likely due to the sensitivity of Euclidean distance to the magnitudes of the gene embeddings. Overall, the performance is similar when using the two scale-invariant methods (cosine similarity and Pearson's correlation), though cosine similarity performs slightly better and would be our generally recommended metric.

Our novel ANDES framework has a myriad of downstream applications, especially when paired with different embeddings. Here, we only scratch the surface by showing the potential of ANDES for function prediction and drug repurposing when paired with a human gene embedding space, as well as cross-organism functional knowledge translation tasks when paired with a joint gene embedding. By matching phenotypes across human and mouse, we provide additional insight into opportunities for improved translational studies. We anticipate more interesting use cases with different gene embeddings or even embeddings of an entirely different modality. Furthermore, although we have analyzed several methods for generating PPI-based gene embeddings, integrating more gene information beyond PPIs may yield further improved gene set matching.

For example, genetic interaction (Dixon et al. 2009), coexpression (Obayashi et al. 2023), and functional networks (Yao et al. 2018) can capture additional, complementary information with respect to gene functional similarity. Furthermore, because ubiquitously expressed genes can give rise to different disease-susceptibility and phenotypes in different tissues (Hekselman and Yeger-Lotem 2020), using tissue-specific network models as input to ANDES could bring additional depth to the analyses and lead to new biological insights. We explore the possibility of extending ANDES' enrichment analyses with tissue-specific functional networks (Greene et al. 2015), considering data sets from GEO2KEGG from five diseases: chronic obstructive pulmonary disease (PDCO), kidney renal clear cell carcinoma (KIRC), pancreatic adenocarcinoma (PAAD), Alzheimer's disease (ALZ), and acute myeloid leukemia (LAML), with their corresponding tissue-specific networks (lung, kidney, pancreas, brain, and blood, respectively). We find that using the tissue-specific networks consistently improves the enrichment for PDCO, KIRC, and PAAD. However, used directly, they do not result in much performance improvement on the ALZ and LAML data sets (Supplemental Fig. S44). For the LAML data sets, the CD8 samples in the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) GSE14924 data set seem to benefit slightly from using the blood tissue network whereas the CD4 samples from the same data set perform worse. These results suggest that the general blood tissue network may better capture and highlight CD8-related T cell signal rather than CD4 signal. In general, we reason that brain and blood may exhibit more regional and cell specificity with differing implications for disease, and thus using the general brain or blood tissue networks may not be specific enough to highlight the signal for ALZ and LAML. In these cases, more fine-grained tissue networks could be more beneficial. For example, Alzheimer's disease modeling could benefit from using an entorhinal cortex network (Roussarie et al. 2020) and likewise for leukemia.

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. A related extension for ANDES would be to further identify the best-match scores that drive different set matching results, thus providing gene-level insights for downstream analyses, similar to how network representations are used to provide functional insights for individual genes in Ietswaart et al. (2021).

Though we focus primarily on the utility of ANDES for gene embeddings, conceptually, ANDES' best-match approach can be applied to any set comparisons as long as a corresponding similarity matrix for set elements exists, regardless of whether the input is an embedding representation. For PPI networks, we have found that embedding representations are better able to prioritize functionally similar gene sets than using the nonembedded network or network properties as input (Fig. 2; Supplemental Fig. S2), but there could be alternative scenarios where the end-user is specifically interested in exploring set similarity using the nonembedded representation. In addition, by expanding the types of entities ANDES can be applied to, we foresee exciting applications for improved interpretability in other domains, such as protein language models (Rives et al. 2021; Villegas-Morcillo et al. 2022; Weissenow et al. 2022) and single-cell embeddings (Liu et al. 2019; Zhao et al. 2021; Chen et al. 2024).

In conclusion, here we have presented a novel algorithm for set comparisons, ANDES, that can improve the utility and interpretability of analyses using 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.

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.

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. This magnitude-invariance is an attractive feature because we find that the magnitude of embedding vectors is negatively correlated with the degree of the corresponding vertex in all three embedding methods used in our comparisons (Supplemental Fig. S45). Though degree has been shown to be a meaningful network property that relates to gene function, it can also capture study bias and is thus not preferable. 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 submatrix 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:

$$\mathrm{GS} = \frac{\sum_{i=1}^{m} \max_{1 \le j \le k} A_{ij} + \sum_{j=1}^{k} \max_{1 \le i \le m} A_{ij}}{m + k}. \quad (1)$$

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.

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). Because 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 1000 Monte Carlo samples, where each of the two 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. Whereas in most cases, the background can be all genes in the embedding, E, for systematic comparisons with a target annotation database (e.g., 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: $z_{(\mathrm{GS})} = \frac{\mathrm{GS} - \mu_0}{\sigma_0}$, where μ_0 and σ_0 are the mean and standard deviation of the Monte Carlo approximations of the null distribution.

Because GS and $z_{(\mathrm{GS})}$ 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.

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 $L = g_1, g_2, \ldots, g_l$ 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 submatrix 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

$$\mathrm{ES} = \max_{1 \le i \le l} \mathrm{dev}\left(\sum_{1 \le t \le i} \left(\max_{1 \le j \le k} A_{tj} - \frac{1}{l} \sum_{r=1}^{l} \max_{1 \le j \le k} A_{rj} \right) \right), \quad (2)$$

where maxdev is the maximum deviation from 0 and $\frac{1}{l} \sum_{r=1}^{l} \max_{1 \le j \le k} A_{rj}$ 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 (Subramanian et al. 2005), 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 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).

Gene embeddings using a protein–protein interaction network

Although our proposed framework is agnostic to embedding method and data type, here we focus on gene embeddings generated from PPIs. We use our previously assembled consensus PPI network (Dannenfelser and Yao 2024), which considers physical interaction information from eight 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 (Grover and Leskovec 2016),

NetMF (Qiu et al. 2018), and a structure-preserving autoencoder method based on the architecture in Li et al. (2023), which we abbreviate as the NN approach. Each method takes a different approach to embed the gene relationships characterized by the PPI network. Node2vec uses 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:

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

where A is the adjacency matrix, Z is the 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 $L_{\mathrm{norm}} = \sum_{i=1}^{n} \| z_i \|_2^2$ to avoid exploding norms. By optimizing this objective function, the NN preserves the global and local structural information simultaneously. We fix the embedding dimension sizes from all three methods to 128.

Gene embeddings using tissue-specific functional networks

We use five human tissue-specific functional networks (lung, kidney, pancreas, brain, and blood) from GIANT (Greene et al. 2015), downloaded from humanbase.io. Specifically, we use the "top edges" (edges with evidence supporting a tissue-specific functional interaction) as input to build the embeddings. To better use the weighted edge information in tissue-specific networks, we use PecanPy's (Liu and Krishnan 2021) efficient node2vec+ (Liu et al. 2023) implementation. Node2vec+ is an extension of the node2vec method that demonstrates consistently strong performance in our benchmarks; the node2vec+ extension considers input edge weights during the random walk sampling process and thus is able to more fully take advantage of the functional networks.

Gene set processing

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

GO: To assess gene function, we use the biological process annotations from GO (Gene Ontology Consortium 2004) (July 16, 2020). To ensure high-quality annotations, we only keep terms with low-throughput experimental evidence codes (EXP, IDA, IMP, IGI, and 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 (Greene et al. 2015). Leveraging the directed acyclic graph structure of the ontology, we propagate gene annotations from child terms to parent terms based on annotated "is a" and "part of" relationships; parent terms thus 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.

KEGG: We obtain pathway gene sets from KEGG (Kanehisa and Goto 2000) using ConsensusPathDB (Kamburov et al. 2013). In total, we obtain 333 unique human pathway gene sets.

OMIM: We collect disease gene sets from OMIM (October 2023) (Hamosh et al. 2005). These gene sets are then mapped to Disease Ontology (Schriml et al. 2012) and propagated through the ontology structure, resulting in 284 unique disease gene sets.

MGI: We assemble mouse phenotype gene sets from MGI (March 2022) (Eppig et al. 2017). After propagating genes from children to parents, we obtain 3738 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 (Wishart et al. 2018) (Jan 2023).

Benchmarking gene set similarity metrics in embedding spaces

The most straightforward way 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 (Greene et al. 2015).

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 ith row vector of X and $\vec{y}_j$ is the jth 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

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

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 (Greene et al. 2015) 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 data sets (KEGG and GO) with each other and (2) matching GO gene sets across different model organisms. For the matched 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. In addition to comparing the different embedding methods using this evaluation paradigm, we also evaluate several baseline approaches that capture gene similarity based directly on the original PPI network, including shared neighbor profiles, graph diffusion (heat diffusion as well as exponential diffusion), and node degree. For the shared neighbor profile baseline, we use the Jaccard index of shared neighbors between a pair of genes as a measure of functional similarity. The heat diffusion analysis uses a diffusion step of 0.1 as recommended by Vandin et al. (2012). Exponential diffusion and node degree constitute the most naive similarity approaches, where node degree is simply using the sum of a pair of genes' degrees in the PPI as a measure of similarity.

For the cross-species evaluation, we look for exact matches between the same GO slim term in humans versus three model

organisms: *M. musculus*, *Saccharomyces cerevisiae*, and *D. melanogaster*. Although 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 (Li et al. 2023), 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.

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 (Tarca et al. 2012, 2013; Cousins et al. 2023). 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 data sets. We compare the results of ANDES against three existing gene set analysis methods: hypergeometric test (Hahne et al. 2008), GSEA (Subramanian et al. 2005), and GSPA (Cousins et al. 2023). We are unable to use NGSEA (Han et al. 2019) and EnrichNet (Glaab et al. 2012) 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 the hypergeometric test uses ANDES' gene set similarity score and statistical significance calculation, taking genes with FDR ≤ 0.05. Comparisons with rank-based gene set enrichment methods use ANDES' best-match-based enrichment method, where the input gene list is ranked using $\log_2$(fold change).

To calculate empirical *P*-values to correct for potential biases in the amount of differential expression signal in the original data set, we generate null distributions for each data set with at least 10 samples in both normal and diseased conditions (21 data sets total) by permuting the sample labels 100 times. We include only data sets with at least 10 normal and 10 diseased samples to ensure more consistency across permutations. Using ANDES scores calculated on the data sets with permuted labels, we can then compute empirical *P*-values for each KEGG term and expression data set pair by comparing the ANDES score based on the true data set.

Software availability

Our implementation of ANDES and code for the analysis described herein is available on GitHub (https://github.com/ylaboratory/ANDES), released under the BSD 3-clause license for open source use, and also included as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

The authors would like to thank members of the ylaboratory research group for helpful discussions. This work was supported by the Cancer Prevention and Research Institute of Texas (CPRIT RR190065) and the National Science Foundation (NSF DBI-2144534). V.Y. is a CPRIT Scholar in Cancer Research.

Author contributions: L.L. and V.Y. developed the method; L.L. implemented the method, with assistance in computational efficiency improvements from C.C.; L.L. performed methods evaluations; L.L. and R.D. applied the method to different case studies; V.Y. supervised the study; L.L., R.D., and V.Y. wrote the manuscript. All authors read and approved the final manuscript.

References

Azuaje F, Wang H, Bodenreider O. 2005. Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings of the ISMB'2005 SIG Meeting on Bio-ontologies*, Detroit, Vol. 2005, pp. 9–10, Citeseer.

Bryant P, Pozzati G, Elofsson A. 2022. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* **13:** 1265. doi:10.1038/s41467-022-28865-w

Chen H-IH, Chiu YC, Zhang T, Zhang S, Huang Y, Chen Y. 2018. GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization. *BMC Syst Biol* **12:** 45–57. doi:10.1186/s12918-018-0642-2

Chen H, Ryu J, Vinyard ME, Lerer A, Pinello L. 2024. SIMBA: single-cell embedding along with features. *Nat Methods* **21:** 1003–1013. doi:10.1038/s41592-023-01899-8

Church KW. 2017. Word2Vec. *Nat Lang Eng* **23:** 155–162. doi:10.1017/S1351324916000334

Cousins H, Hall T, Guo Y, Tso L, Tzeng KTH, Cong L, Altman RB. 2023. Gene set proximity analysis: expanding gene set enrichment analysis through learned geometric embeddings, with drug-repurposing applications in COVID-19. *Bioinformatics* **39:** btac735. doi:10.1093/bioinformatics/btac735

Dannenfelser R, Yao V. 2024. Splitpea: quantifying protein interaction network rewiring changes due to alternative splicing in cancer. *Pac Symp Biocomput* **29:** 579–593. doi:10.1142/9789811286421_0044

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, Minneapolis, Vol. 1, pp. 4171–4186. Association for Computational Linguistics, Stroudsburg, PA.

Dixon SJ, Costanzo M, Baryshnikova A, Andrews B, Boone C. 2009. Systematic mapping of genetic interaction networks. *Annu Rev Genet* **43:** 601–625. doi:10.1146/annurev.genet.39.073003.114751

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. 2020. An image is worth 16×16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Du J, Jia P, Dai Y, Tao C, Zhao Z, Zhi D. 2019. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics* **20:** 7–15. doi:10.1186/s12864-018-5370-x

Eppig JT, Smith CL, Blake JA, Ringwald M, Kadin JA, Richardson JE, Bult CJ. 2017. Mouse genome informatics (MGI): resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. *Methods Mol Biol* **1488:** 47–73. doi:10.1007/978-1-4939-6427-7_3

Gao KY, Fokoue A, Luo H, Iyengar A, Dey S, Zhang P. 2018. Interpretable drug target prediction using deep neural representation. *IJCAI (US)* **2018:** 3371–3377. doi:10.24963/ijcai.2018/468

Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**(Database issue)**:** D258–D261. doi:10.1093/nar/gkh036

Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. 2012. Enrichnet: network-based gene set enrichment analysis. *Bioinformatics* **28:** i451–i457. doi:10.1093/bioinformatics/bts389

Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, et al. 2021. Structure-based protein function prediction using graph convolutional networks. *Nat Commun* **12:** 3168. doi:10.1038/s41467-021-23303-9

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, et al. 2015. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* **47:** 569–576. doi:10.1038/ng.3259

Grover A, Leskovec J. 2016. Node2vec: scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, pp. 855–864. Association for Computing Machinery, New York.

Hahne F, Huber W, Gentleman R, Falcon S, Falcon S, Gentleman R. 2008. Hypergeometric testing used for gene set enrichment analysis. In *Bioconductor case studies*, pp. 207–220. SpringerLink, New York.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33(suppl_1):** D514–D517. doi:10.1093/nar/gki033

Han H, Lee S, Lee I. 2019. NGSEA: network-based gene set enrichment analysis for interpreting gene expression phenotypes with functional gene sets. *Mol Cells* **42:** 579. doi:10.1101/636498

Hekselman I, Yeger-Lotem E. 2020. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat Rev Genet* **21:** 137–150. doi:10.1038/s41576-019-0200-9

Holley AJ, Lugo JN. 2016. Effects of an acute seizure on associative learning and memory. *Epilepsy Behav* **54:** 51–57. doi:10.1016/j.yebeh.2015.11.001

Ietswaart R, Gyori BM, Bachman JA, Sorger PK, Churchman LS. 2021. GeneWalk identifies relevant gene functions for a biological context using network representation learning. *Genome Biol* **22:** 55. doi:10.1186/s13059-021-02264-8

Jørgensen EA, Knigge U, Warberg J, Kjær A. 2007. Histamine and the regulation of body weight. *Neuroendocrinology* **86:** 210–214. doi:10.1159/000108341

Kamburov A, Stelzl U, Lehrach H, Herwig R. 2013. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* **41:** D793–D800. doi:10.1093/nar/gks1055

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27–30. doi:10.1093/nar/28.1.27

Khaleghzadeh-Ahangar H, Talebi A, Mohseni-Moghaddam P. 2022. Thyroid disorders and development of cognitive impairment: a review study. *Neuroendocrinology* **112:** 835–844. doi:10.1159/000521650

Khrulkov V, Mirvakhabova L, Ustinova E, Oseledets I, Lempitsky V. 2020. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, pp. 6418–6428. Institute of Electrical and Electronics Engineers, New York.

Kim S-Y, Volsky DJ. 2005. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* **6:** 144. doi:10.1186/1471-2105-6-144

Kim S, Lee H, Kim K, Kang J. 2018. Mut2Vec: distributed representation of cancerous mutations. *BMC Med Genomics* **11:** 57–69. doi:10.1186/s12920-018-0349-7

Kulmanov M, Hoehndorf R. 2020. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36:** 422–429. doi:10.1093/bioinformatics/btz595

Kulmanov M, Khan MA, Hoehndorf R. 2018. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34:** 660–668. doi:10.1093/bioinformatics/btx624

Li L, Dannenfelser R, Zhu Y, Hejduk N, Segarra S, Yao V. 2023. Joint embedding of biological networks for cross-species functional alignment. *Bioinformatics* **39:** btad529. doi:10.1093/bioinformatics/btad529

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379:** 1123–1130. doi:10.1126/science.ade2574

Liu R, Krishnan A. 2021. PecanPy: a fast, efficient and parallelized Python implementation of node2vec. *Bioinformatics* **37:** 3377–3379. doi:10.1093/bioinformatics/btab202

Liu J, Huang Y, Singh R, Vert J-P, Noble WS. 2019. Jointly embedding multiple single-cell omics measurements. In *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, Niagara Falls, NY, Vol. 143, NIH Public Access.

Liu R, Hirn M, Krishnan A. 2023. Accurately modeling biased random walks on weighted networks using node2vec+. *Bioinformatics* **39:** btad047. doi:10.1093/bioinformatics/btad047

Ma A, McDermaid A, Xu J, Chang Y, Ma Q. 2020. Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol* **38:** 1007–1022. doi:10.1016/j.tibtech.2020.02.013

McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw* **3:** 861. doi:10.21105/joss.00861

Mostavi M, Chiu Y-C, Huang Y, Chen Y. 2020. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med Genomics* **13:** 44. doi:10.1186/s12920-020-0677-2

Obayashi T, Kodate S, Hibara H, Kagaya Y, Kinoshita K. 2023. COXPRESdb v8: an animal gene coexpression database navigating from a global view to detailed investigations. *Nucleic Acids Res* **51:** D80–D87. doi:10.1093/nar/gkac983

Perl AE, Larson RA, Podoltsev NA, Strickland S, Wang ES, Atallah E, Schiller GJ, Martinelli G, Neubauer A, Sierra J, et al. 2022. Follow-up of patients with R/R FLT3-mutation–positive AML treated with gilteritinib in the phase 3 ADMIRAL trial. *Blood* **139:** 3366–3375. doi:10.1182/blood.2021011583

Peyvandipour A, Saberian N, Shafi A, Donato M, Draghici S. 2018. A novel computational approach for drug repurposing using systems biology. *Bioinformatics* **34:** 2817–2825. doi:10.1093/bioinformatics/bty133

Qiu J, Dong Y, Ma H, Li J, Wang K, Tang J. 2018. 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*, Marina Del Rey, CA, pp. 459–467. Association for Computing Machinery, New York.

Reay WR, Cairns MJ. 2021. Advancing the use of genome-wide association studies for drug repurposing. *Nat Rev Genet* **22:** 658–671. doi:10.1038/s41576-021-00387-z

Resnik P. 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J Artif Intell Res* **11:** 95–130. doi:10.1613/jair.514

Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* **118:** e2016239118. doi:10.1073/pnas.2016239118

Roussarie J-P, Yao V, Rodriguez-Rodriguez P, Oughtred R, Rust J, Plautz Z, Kasturia S, Albornoz C, Wang W, Schmidt EF, et al. 2020. Selective neuronal vulnerability in Alzheimer's disease: a network-based analysis. *Neuron* **107:** 821–835.e12. doi:10.1016/j.neuron.2020.06.010

Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA. 2012. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res* **40:** D940–D946. doi:10.1093/nar/gkr972

Stanojevic S, Li Y, Ristivojevic A, Garmire LX. 2022. Computational methods for single-cell multi-omics integration and alignment. *Genomics Proteomics Bioinformatics* **20:** 836–849. doi:10.1016/j.gpb.2022.11.013

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102:** 15545–15550. doi:10.1073/pnas.0506580102

Suri R, Neupane YR, Mehra N, Nematullah M, Khan F, Alam O, Iqubal A, Jain GK, Kohli K. 2021. Sirolimus loaded chitosan functionalized poly (lactic-co-glycolic acid) (PLGA) nanoparticles for potential treatment of age-related macular degeneration. *Int J Biol Macromol* **191:** 548–559. doi:10.1016/j.ijbiomac.2021.09.069

Tarca AL, Draghici S, Bhatti G, Romero R. 2012. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics* **13:** 136. doi:10.1186/1471-2105-13-136

Tarca AL, Bhatti G, Romero R. 2013. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One* **8:** e79217. doi:10.1371/journal.pone.0079217

Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng Z, Liu XS, et al. 2023. Transfer learning enables predictions in network biology. *Nature* **618:** 616–624. doi:10.1038/s41586-023-06139-9

Vandin F, Clay P, Upfal E, Raphael BJ. 2012. Discovery of mutated subnetworks associated with clinical data in cancer. *Biocomputing* **2012:** 55–66. doi:10.1142/9789814366496_0006

Villegas-Morcillo A, Gomez AM, Sanchez V. 2022. An analysis of protein language model embeddings for fold prediction. *Brief Bioinform* **23:** bbac142. doi:10.1093/bib/bbac142

Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z. 2011. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics* **98:** 1–8. doi:10.1016/j.ygeno.2011.04.006

Wang S, Flynn ER, Altman RB. 2020. Gaussian embedding for large-scale gene set analysis. *Nat Mach Intell* **2:** 387–395. doi:10.1038/s42256-020-0193-2

Weissenow K, Heinzinger M, Rost B. 2022. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure* **30:** 1169–1177.e4. doi:10.1016/j.str.2022.05.001

Wieting J, Bansal M, Gimpel K, Livescu K. 2015. Towards universal paraphrastic sentence embeddings. In *Proceedings of International Conference on Learning Representations*, San Juan, Puerto Rico.

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al. 2018. Drugbank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46:** D1074–D1082. doi:10.1093/nar/gkx1037

Xiong Y, Guo M, Ruan L, Kong X, Tang C, Zhu Y, Wang W. 2019. Heterogeneous network embedding enabling accurate disease association predictions. *BMC Med Genomics* **12:** 186. doi:10.1186/s12920-019-0623-3

Yao V, Wong AK, Troyanskaya OG. 2018. Enabling precision medicine through integrative network models. *J Mol Biol* **430:** 2913–2923. doi:10.1016/j.jmb.2018.07.004

Yu Z, Huang F, Zhao X, Xiao W, Zhang W. 2021. Predicting drug-disease associations through layer attention graph convolutional network. *Brief Bioinform* **22:** bbaa243. doi:10.1093/bib/bbaa243

Yu X, Qi X, Wei L, Zhao L, Deng W, Guo W, Wang Q, Ma X, Hu X, Ni P, et al. 2023. Fingolimod ameliorates schizophrenia-like cognitive impairments induced by phencyclidine in male rats. *Br J Pharmacol* **180:** 161–173. doi:10.1111/bph.15954

Zhang F, Yuan NJ, Lian D, Xie X, Ma WY. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, pp. 353–362. Association for Computing Machinery, New York.

Zhao Y, Cai H, Zhang Z, Tang J, Li Y. 2021. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat Commun* **12:** 5261. doi:10.1038/s41467-021-25534-2

Received February 15, 2024; accepted in revised form August 29, 2024.

A best-match approach for gene set analyses in embedding spaces

Lechuan Li, Ruth Dannenfelser, Charlie Cruz, et al.

Genome Res. 2024 34: 1421-1433 originally published online September 4, 2024
Access the most recent version at doi:10.1101/gr.279141.124